







## Modelling acute myeloid leukaemia in a continuum of differentiation states

H. Cho <sup>a</sup>, K. Ayers <sup>b</sup>, L. de Pillis<sup>c</sup>, Y.-H. Kuo <sup>d</sup>, J. Park <sup>c</sup>, A. Radunskaya <sup>b</sup> and R. C. Rockne <sup>e</sup>

<sup>a</sup>Department of Mathematics, University of Maryland, College Park, MD, USA; <sup>b</sup>Department of Mathematics, Pomona College, Claremont, CA, USA; <sup>c</sup>Department of Mathematics, Harvey Mudd College, Claremont, CA, USA; <sup>d</sup>Department of Hematological Malignancies Translational Science, Gehr Family Center for Leukemia Research, City of Hope, Duarte, CA, USA; <sup>e</sup>Division of Mathematical Oncology, City of Hope, Duarte, CA, USA

### ABSTRACT

Here we present a mathematical model of movement in an abstract space representing states of cellular differentiation. We motivate this work with recent examples that demonstrate a continuum of cellular differentiation using single-cell RNA-sequencing data to characterize cellular states in a high-dimensional space, which is then mapped into  $\mathbb{R}^2$  or  $\mathbb{R}^3$  with dimension reduction techniques. We represent trajectories in the differentiation space as a graph, and model directed and random movement on the graph with partial differential equations. We hypothesize that flow in this space can be used to model normal and abnormal differentiation processes. We present a mathematical model of haematopoiesis parameterized with publicly available single-cell RNA-Seq data and use it to simulate the pathogenesis of acute myeloid leukaemia (AML). The model predicts the emergence of cells in novel intermediate states of differentiation consistent with immunophenotypic characterizations of a mouse model of AML.

### ARTICLE HISTORY

Received 16 December 2017  
Accepted 20 April 2018

### KEYWORDS


Diffusion mapping; haematopoiesis; single-cell RNA-sequencing; developmental trajectories; nonlinear dimension reduction; cellular differentiation; acute myeloid leukaemia; differentiation continuum

## 1. Introduction

The recent advance of single-cell RNA-sequencing (scRNA-Seq) technologies has enabled a new, high-dimensional definition of cell states. In contrast to conventional gene expression analyses based on measuring the average levels in a tissue or given cell population, single-cell analysis captures the cellular heterogeneity and provides resolution at the level of individual cells within the tissue or cell population. This level of resolution coupled with genome-wide gene expression provides an unprecedented opportunity to quantitatively probe cellular behaviour, cellular variation and dynamics in a wide range of biological contexts.

There are on the order of 20,000 protein encoding genes that compose the transcriptome, which constitute a  $\mathbb{R}^{20,000}$  dimensional space. Therefore, the configuration of the

**CONTACT** R. Rockne  [rrockne@coh.org](mailto:rrockne@coh.org)

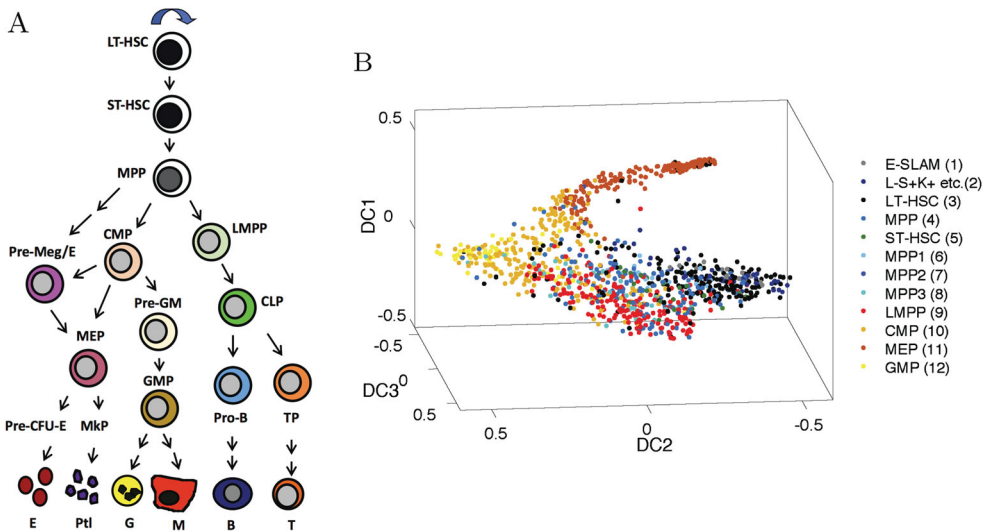
 Supplemental data for this article can be accessed here. <https://doi.org/10.1080/23737867.2018.1472532>

© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

transcriptome at a point in time can be represented as a coordinate vector in space. When a cell expresses genes, it ‘moves’ in this high-dimensional gene expression phenotype space. Over time, the sequence of locations in the space of a given cell creates a trajectory. Dimension reduction techniques are commonly used to map the larger space into a lower dimensional space, for instance,  $\mathbb{R}^2$  or  $\mathbb{R}^3$ , at which point the cells are clustered based on a similarity metric and recategorized. This process has revealed a continuum of cell phenotypes, with intermediate states connecting canonical cell states. The most prominent example of this process is in haematopoietic cell differentiation.

Normal haematopoiesis is long thought to occur through stepwise differentiation of haematopoietic stem cells following a tree-like hierarchy of discrete multipotent, oligopotent and then unipotent lineage-restricted progenitors (Figure 1A). The classical model of haematopoiesis considers differentiation as a stepwise process of binary branching decisions, famously represented as a potential landscape by Waddington (1957). However, this model is based on bulk characterization of prospectively purified immunophenotypic cell populations. Recent advances in scRNA-Seq technologies now allow resolution of single-cell heterogeneity and reconstruction of differentiation trajectories which have been applied to a number of different cellular systems, from haematopoiesis to breast endothelial cell differentiation (Bach et al., 2017; Hamey, Nestorowa, Wilson, & Göttgens, 2016; Nestorowa et al., 2016; Velten et al., 2017).



**Figure 1.** (A) Classic representation of a linear hierarchy of discrete cell states, from long-term haematopoietic stem cell (LT-HSC), short-term (ST)-HSC, multipotent progenitor (MPP) to committed common myeloid progenitor (CMP), pre-megakaryocyte/erythrocyte (Pre-Meg/E) and megakaryocyte–erythroid progenitor (MEP), pre-granulocyte/macrophage (Pre-GM), granulocyte–macrophage progenitor (GMP), and lymphoid-primed MPP (LMPP), common lymphoid progenitor (CLP) cells, on down to terminally differentiated cells such as erythrocytes (E) platelets (Plt), granulocytes (G) macrophages (M), B and T-cells. (B) The classical view is contrasted with a nonlinear continuum representation of haematopoietic cell differentiation states using diffusion map dimension reduction of scRNA-Seq data (figure recreated from data available in Nestorowa et al., 2016). Colours representing cell identities in (A) and (B) are coordinated. Cell types in (B) are a subset of cells represented in (A).

These efforts have led to the new view that haematopoietic lineage differentiation occurs as a continuous process, which can be mapped into a continuum of cellular and molecular phenotypes (Figure 1B). Haematopoietic malignancies such as acute myeloid leukaemia (AML) arise from dysregulated differentiation and proliferation of haematopoietic stem cells and progenitor cells upon accumulation of oncogenic genetic mutations and/or epigenetic alterations. Therefore, characterizing disordered haematopoiesis based on discretely defined phenotypic populations can be challenging. Moreover, ‘discrete’ phenotypic cell populations are in fact highly heterogeneous in terms of functional capacity and gene expression profiles. It is now possible to view pathologic haematopoiesis through a continuum of cellular and molecular phenotypes and capture the heterogeneity, differentiation plasticity and dysregulated gene expression associated with malignant transformation.

This new view of biology forces us to reconsider the mathematical approaches we use to model cell states and behaviours. Instead of building mathematical models which identify discrete cell populations and assign mathematical rules for their evolution and interactions, we may now consider a continuum of cellular states and model movement between these states in aggregate as a flow of mass on a structured graph. Modelling differentiation in this manner reduces the number of parameters and thus the complexity of the mathematical model by representing many cell populations and states in a single variable. At the same time, this increases biological resolution of the system by characterizing an infinite number of sub-states in a continuum representation. Here we consider a model of haematopoietic cell differentiation and associated disorders as a flow and transport process in a continuous differentiation space as a test system for a more general approach of modelling the temporal evolution of a continuum of cell states.

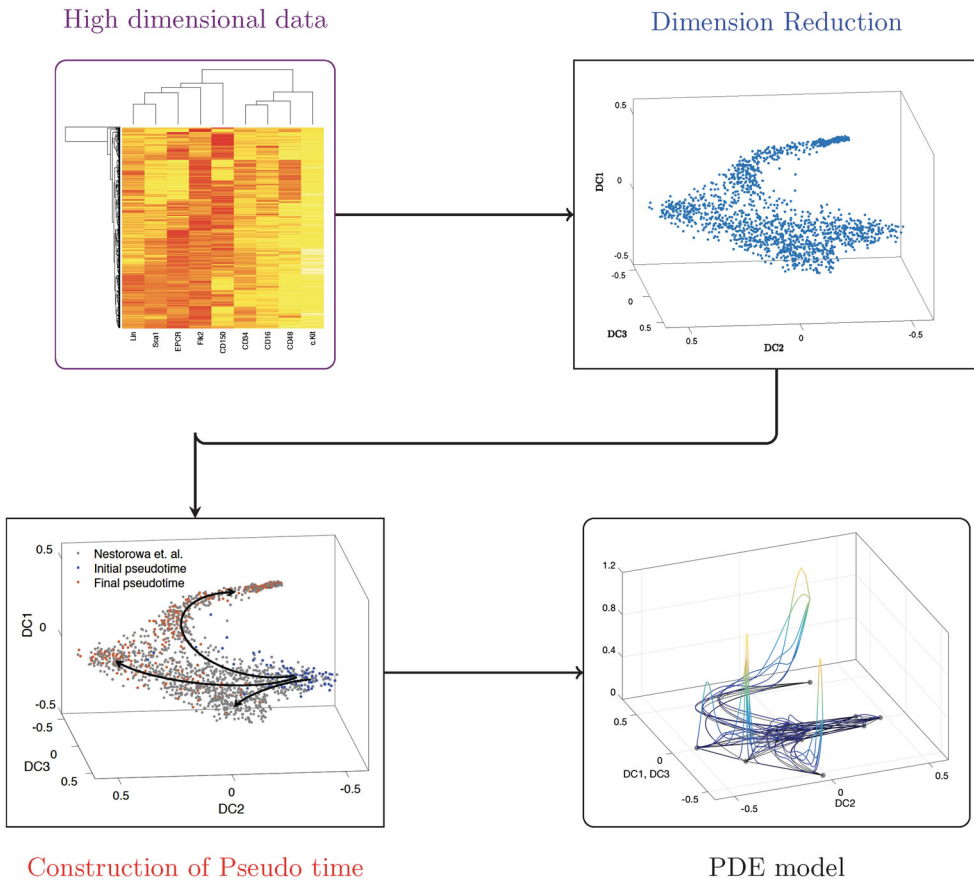
This article is structured as follows: first, we review the state of the art of dimension reduction methods that are used to construct and define haematopoietic differentiation spaces that can be represented as graphs, including a review of Schienbinger et al.’s method for modelling transport on a graph from reduced dimension gene expression data (2017). Then we introduce our partial differential equation (PDE) model of flow and transport on a graph, and illustrate the model on simple ‘Y’-shaped graph with symmetric and asymmetric differentiation. We then calibrate our model to a graph constructed from publicly available scRNA-Seq data of normal haematopoiesis. Finally, we use our model to simulate abnormal haematopoietic cell differentiation processes observed during the pathogenesis of AML, a form of aggressive hematologic malignancy. We conclude with a brief discussion of prior literature on modelling differentiation as a continuum, and the limitations and potential future applications of this modelling approach.

## 2. Construction of a differentiation continuum

In order to describe the entire modelling process, in this section, we briefly describe methods for reducing the dimension of high-dimensional scRNA-data, before reviewing pseudotime reconstruction techniques, and conclude this section by examining a technique from Schiebinger et al. (2017) for construction of a directed graph that represents haematopoietic differentiation space. While the focus of this paper is not dimension reduction techniques or pseudotime reconstruction, we summarize some of these techniques that are most relevant to our modelling approach, without advocating for one over another. We should emphasize that this is a review of already existing algorithms; the novel work

begins in Section 3. The relationship between time and pseudotime within a mathematical model of cell differentiation is analogous to the relationship between age-structured and stage-structured models in ecology. Cell differentiation data yield information about cells at various stages of differentiation, but generally do not provide time-specific data. A pseudotime model is one that considers the differentiation stage of a cell population instead of the time in which a cell is in a certain state.

In Figure 2, we lay out the steps required for going from high-dimensional data to construction of the PDE model. Section 2.1 will review various dimension reduction techniques, including a more thorough discussion of the technique used in our application, diffusion mappings. Section 2.2 summarizes techniques such as Wishbone and Wanderlust that are available for pseudotime reconstruction given dimension reduced data. Finally, Section 2.3 will give an overview of the technique presented in Schiebinger et al. (2017) for construction of a directed graph that indicates how cell populations evolve in pseudotime.



**Figure 2.** Flowchart of our modelling process: this chart organizes the steps taken toward constructing the PDE model. First, high-dimensional data such as single-cell RNA-sequencing (scRNA-Seq) are represented in two- or three-dimensional space through one of many dimension reduction techniques. Then, temporal events (pseudotime trajectories) are inferred from the dimension reduced data. We then use the reduced dimension representation and pseudotime trajectories to model flow and transport in the reduced space. Data is from Nestorowa et al. (2016).

## 2.1. Dimension reduction techniques

A broad range of techniques have been developed to provide insight into interpretation of high-dimensional biological data. These techniques provide a first step in our approach to modelling the evolution of cell states in a continuum and play a critical role in characterizing differentiation dynamics. We note that the application of different data reduction techniques, clustering methods and pseudotime ordering on the same data set will produce different differentiation spaces on which to build a dynamic model. We will use one particular dimension reduction approach as an example, but our framework allows one to select from a variety of approaches. In this section, we provide a brief description of a subset of such techniques to give the reader a sense of the field.

Several techniques have been developed to interpret the high-dimensional differentiation space, including principal component analysis (PCA), diffusion maps and t-distributed stochastic neighbour embedding (t-SNE). Each of these methods maps high-dimensional data into a lower dimensional space. As discussed in this section, different techniques produce different shapes and differentiation spaces, and so some techniques are better suited to certain data sets than others. For instance, one commonly used dimension reduction technique is PCA, a linear projection of the data. While PCA is computationally simple to implement, the limitation of this approach lies in its linearity – the data will always be projected onto a linear subspace of the original measurement space. If the data show a trend that does not lie in a linear subspace – for instance, if the data lie on an embedding of a lower dimensional manifold in Euclidean space that is not a linear subspace – then this trend will not be efficiently captured with PCA (Khalid, Khalil, & Nasreen, 2014).

In contrast, diffusion mapping and t-SNE, as well as a variant of t-SNE known as hierarchical stochastic neighbour embedding (HSNE), are nonlinear dimension reduction techniques. t-SNE, introduced by van der Maaten and Hinton (2008), is a machine learning dimension reduction technique that is particularly good at mapping high-dimensional data into a two- or three-dimensional space, allowing for the data to be visualized in a scatter plot.

Given a data set in  $\mathbb{R}^n : X = \{x_1, x_2, \dots, x_n\}$ , we can transform the Euclidean distances between two points into a probability distribution. Intuitively, this distribution gives the probability that data point  $x_j$  is a neighbour of point  $x_i$ , where the probability of being a neighbour of  $x_i$  has a Gaussian distribution (van der Maaten & Hinton, 2008):

$$p_{ji} = \frac{e^{-(\|x_i - x_j\|^2)/2\sigma^2}}{\sum_{k \neq i} e^{-(\|x_i - x_k\|^2)/2\sigma^2}}. \quad (1)$$

The t-SNE algorithm aims to find a map from the data set to two- or three-dimensional Euclidean space that minimizes the Kullback–Leibler divergence between the probability distributions in the original and reduced space. This optimization problem is often solved using gradient descent methods.

In van Unen et al. (2017), a new technique for examining high-dimensional mass cytometry data, known as HSNE is presented. Mass cytometry allows for the examination of several cellular markers on samples made up of vast quantities of cells. These data sets are truly ‘big’ in the sense that they are very large (a sample for each cell) as well as high-dimensional. Therefore, pre-existing dimension reduction techniques are not optimal for mass cytometry data. HSNE, as suggested by its name, is hierarchical by nature, allowing

for refinement in the level of detail. HSNE ultimately constructs a hierarchy of subsets of the dataset  $X$ :

$$X = \mathcal{L}^1 \supset \mathcal{L}^2 \supset \dots \supset \mathcal{L}^n.$$

The hierarchy begins with the data set itself ( $X = \mathcal{L}^1$ ). A weighted  $k$ -nearest neighbour (kNN) graph is constructed on the data set, and individual points, or ‘landmarks’, are selected from each node on the graph to represent the data set at the next, coarser, level,  $\mathcal{L}^2$ . This process is repeated as desired. These subsets can each be embedded in lower dimensional space. This hierarchical embedding scheme allows the user to view the data at different resolutions, from a broad overview (level  $\mathcal{L}^n$ ) to a more refined understanding of cell types associated with markers (intermediate levels). Starting with a certain subset  $\mathcal{O} \subset \mathcal{L}^s$ , the user is able to ‘drill in’ to the data by selecting a subset  $\mathcal{G} \subset \mathcal{L}^{s-1}$ . Thus, HSNE is an approach that is useful for data that require different levels of detail at different scales. An illuminating graphical representation of the HSNE process can be found in van Unen et al. (2017) (Figure 1).

Diffusion maps work by taking advantage of the relationship between heat diffusion and random walk Markov chains. Let  $X$  be a data set of size  $n$ . The diffusion map algorithm begins by considering a kernel function on pairs of data points; this function must be symmetric and nonnegative. The Gaussian kernel

$$k(x, y) = e^{-\frac{\|x-y\|^2}{\epsilon}}$$

is one popular choice. Similar to the conditional probability defined in Equation (1), the kernel  $k(x, y)$  is used to specify the probability of going from  $x$  to  $y$  in one step of a random walk on the data, found by normalizing the kernel to ensure the random walk probabilities integrate to 1:

$$p(x, y) = \frac{k(x, y)}{\sum_{y \neq x} k(x, y)}.$$

By letting the number of steps in this random walk go to infinity, we can consider the stationary distribution  $p_t$  of the Markov chain. This stationary distribution is used to formulate a new metric on the data space, known as the diffusion distance:

$$d(x_i, x_j) = \sum_{u \in X} |(p_t(x_i, u) - p_t(x_j, u))|^2.$$

Intuitively speaking, the diffusion distance between two points will be low if there are many paths in the random walk that connect them, and high if there are few. Because it is computationally expensive to repeatedly compute the diffusion distance between each pair of points, it is easier to map data points to a new Euclidean space using the function  $\phi : X \rightarrow \mathbb{R}^n$  defined as

$$\phi(x_i) = \begin{bmatrix} p_t(x_i, x_1) \\ p_t(x_i, x_2) \\ \vdots \\ p_t(x_i, x_n) \end{bmatrix}.$$

The Euclidean distance in this space, known as the diffusion space, is then equivalent to the diffusion distance in the data space. It can be demonstrated that the linearly independent



eigenvectors of the diffusion matrix (the transition matrix associated with the aforementioned Markov Chain) form a basis for the diffusion space. Therefore, by opting to keep the  $k$ -eigenvectors corresponding to the  $k$  largest eigenvalues, we obtain a map from the original data to a  $k$ -dimensional subspace of the diffusion space that most efficiently captures the structure of the data; this map is called the *diffusion* map. A more in-depth explanation can be found in Coifman et al. (2005).

Each of these dimension reduction methods has strengths and weaknesses depending on the question(s) being asked of the data. Moreover, each method will produce a distinctly different shape in the lower dimensional representation. Therefore, the choice of dimension reduction technique is a critical step in analysing any high-dimensional data set. For the purpose of analysing cell transition probabilities and inferring trajectories within the reduced space, Nestorowa et al. (2016) and others have chosen to use diffusion mapping to analyse cell differentiation.

## **2.2. Pseudotime ordering of differentiation states**

For data without temporal information, pseudotime methods are available to infer a sequence of biological states from single time point data. Diffusion mapping can be used to infer a 'diffusion pseudotime' (Haghverdi, Büttner, Wolf, Buettner, & Theis, 2016; Nestorowa et al., 2016). In particular, Haghverdi et al. (2016) develop an efficient diffusion pseudotime approach by rescaling the diffusion components by a weighted distance in terms of the eigenvalues, derived by considering a random walk according to a transition matrix that specifies the probability of transitioning from any single cell to another in an infinitesimal amount of time. Alternative pseudotime approaches include Wishbone (Setty et al., 2016) that uses shortest paths in a kNN graph constructed in diffusion component space to construct an initial ordering of cells, TASIC (Rashid, Kotton, & Bar-Joseph, 2017) that is able to incorporate time information and identify branches and incorporate time information in single-cell expression data by considering it as developmental processes emitting expression profiles from a finite number of states, and Monocle (Qiu, Hill, et al. 2017; Qiu, Mao, et al. 2017) that fits a principal graph (Mao, Wang, Goodison, & Sun, 2015) and uses a reversed graph embedding technique which simultaneously learns a low-dimensional embedding of the data and a graphical structure spanning the dataset.

Finally, when the data are collected at multiple time points, the transition rates between the nodes can be obtained after partitioning the cell data. For instance, Schiebinger et al. (2017) employ graph clustering (Levine et al., 2015; Shekhar et al., 2016) and optimal transport (OT) methods to understand the dynamics in the reduced space of cell data. We describe the OT method in an effort to provide a clear distinction between the OT algorithm and our modelling approach.

## **2.3. Optimal transport**

Schiebinger et al. (2017) have proposed a model and algorithm for constructing a directed graph oriented in pseudotime given temporal data. The OT algorithm itself is a classical problem studied in the mathematical area of Transportation Theory, which aims to optimally transport and allocate resources given certain cost functions. Schiebinger et al. (2017) apply this theory to a time series of reduced dimension single-cell gene expression profiles.

The time series is made up of a sequence of samples  $\{S_1, \dots, S_n\}$ , at different times  $t_i$  for  $i \in \{1, \dots, n\}$ . Suppose that each sample consists of points in  $\mathbb{R}^m$ . A distribution  $\hat{\mathbb{P}}_{t_i}$  is defined by each sample  $S_i$ . For each set  $A \subset \mathbb{R}^m$ :

$$\hat{\mathbb{P}}_{t_i}(A) = \frac{1}{|S_i|} \sum_{x \in S_i} \delta_x(A),$$

where  $\delta_x$  represents a Delta Distribution centred at  $x$ :

$$\delta_x(A) = \begin{cases} 1 & x \in A, \\ 0 & x \notin A. \end{cases}$$

Together, as a sequence, these inferred distributions  $\{\hat{\mathbb{P}}_{t_i}\}$  form what is known as an ‘empirical developmental process’. The goal is then to determine, as closely as possible, what the true underlying Markov developmental process  $\mathbb{P}_t$  is by examining what are known as transport maps between pairs  $\hat{\mathbb{P}}_{t_{i-1}}$  and  $\hat{\mathbb{P}}_{t_i}$ . A transport map for a pair  $(\hat{\mathbb{P}}_{t_{i-1}}, \hat{\mathbb{P}}_{t_i})$  is a distribution  $\pi$  defined on  $\mathbb{R}^m \times \mathbb{R}^m$  such that  $\hat{\mathbb{P}}_{t_{i-1}}$  and  $\hat{\mathbb{P}}_{t_i}$  are the two marginal distributions of  $\pi$ . Thus, given a function  $c(x, y)$  that represents the cost to transport some unit mass from  $x$  to  $y$ , the goal is to minimize

$$\iint_{\mathbb{R}^m \times \mathbb{R}^m} c(x, y) \pi(x, y) \, dx \, dy$$

subject to

$$\int_{\mathbb{R}^m} \pi(x, \cdot) \, dx = \hat{\mathbb{P}}_{t_{i-1}},$$

$$\int_{\mathbb{R}^m} \pi(\cdot, y) \, dy = \hat{\mathbb{P}}_{t_i}.$$

Schienbinger et al. further refine this algorithm by including a growth term in their transport plan to allow for cellular proliferation between time points. This differs from the classical OT algorithm in that the classical OT algorithm is formulated with conservation of mass in mind. OT can thus be used to estimate the ancestors and descendants of a set of cells. Cells are clustered using the Louvain-Jaccard community detection algorithm on the reduced dimension gene expression data in 20-dimensional space. Schienbinger et al. thus identified 33 cell nodes, which were then used as starting populations from which developmental trajectories could be analysed. These can be thought of as nodes on a graph visualized with force-directed layout embedding, and edges represent the motion in pseudotime.

In the following section, we assume that the first two steps in Figure 2 have been completed by one of the methods described above. In other words, we start with samples in high-dimensional space, we map the data to a lower dimensional space and then we produce pseudotime trajectories in this lower dimensional space. In the final step, we model the differentiation process in continuous (pseudo)-time and (reduced-dimensional) space using PDEs.



### 3. Modelling on the differentiation continuum

To illustrate our modelling technique, we assume that we have constructed a simple branched manifold or graph situated in the differentiation space. This graph is not a set of discrete nodes, rather, the graph and its edges represent a continuum of canonical states and intermediate states of differentiation. Assuming that the graph and the temporal evolution on the graph is obtained by any one of the various data analysis techniques summarized in Section 2 including OT (Schiebinger et al., 2017), diffusion pseudotime methods (Haghverdi et al., 2016; Nestorowa et al., 2016), Wishbone (Setty et al., 2016), TASIC (Rashid et al., 2017) and Monocle (Qiu, Hill, et al. 2017; Qiu, Mao, et al. 2017), we develop a PDE model that describes the dynamics in this differentiation continuum. Cell differentiation models in the continuous space have been developed in Gwiazda, Grzegorz, and Marciniak-czochra (2012) and Doumic, Marciniak-Czochra, Perthame, and Zubelli (2011) that extend the discrete multicompartment models (Lander, Gokoffski, Wan, Nie, & Calof, 2009; Lo et al., 2008; Marciniak-Czochra, Stiehl, Ho, Jäger, & Wagner, 2009; Stiehl & Marciniak-Czochra, 2011).

#### 3.1. PDE model on a graph

Let us define the graph  $G$  obtained in the differentiation continuum space. We comment that although we can consider a cell distribution on the actual reduced space, we further reduce our model on a graph that makes it convenient to employ the biological insights from the classical discrete models. The node set of  $G$  is denoted as  $\{v_k\}_{k=1}^{n_v}$ , where  $n_v$  is the total number of nodes, and the edge of the graph connecting in the direction from the  $i$ th to the  $j$ th node as  $e_{ij}$ . We also introduce an alternate description of the graph with respect to the edge that is more convenient for describing the PDE model. If the total number of nontrivial edges is  $n_e$ , we take  $\{e_k\}_{k=1}^{n_e}$  with the index mapping  $I: \mathcal{J} \rightarrow \{1, \dots, n_e\}$  on the set of nontrivial edges  $(i, j) \in \mathcal{J}$ , and the end points in the direction of cell transition as  $\{a_k\}_{k=1}^{n_e}$  and  $\{b_k\}_{k=1}^{n_e}$ , respectively. We remark that  $\bigcup_{k=1}^{n_e} \{a_k, b_k\} = \{v_k\}_{k=1}^{n_v}$ .

We denote  $u(x, t)$  as the cell distribution on the graph  $G$  at the differentiation continuum space location  $x \in G$  and time (or pseudotime)  $t$ . Thus, we follow the dynamics of the cell density at  $x \in G$ . 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–diffusion–reaction equation (Evans, 2010) as

$$\frac{\partial u_k}{\partial t} = -\frac{\partial}{\partial x}(V_k(x)u_k) + R_k(x)u_k + \frac{D_k(x)}{2W_k(x)} \frac{\partial}{\partial x} \left( w_k(x) \frac{\partial u_k}{\partial x} \right), \quad x \in e_k = \overline{a_k b_k}, \quad (2)$$

where  $x$  is a one-dimensional variable parameterized on each edge  $e_k$  from  $a_k$  to  $b_k$ . The advection coefficient  $V_k(x)$  models the cell differentiation and the transition between the different cell types, that is, the nodes. The transition rate per unit time (e.g. day<sup>-1</sup>) or pseudotime can be taken as  $V_k(x)$  computed using the periods of cell differentiation. For instance,  $V_k(x)$  can be computed by smoothly interpolating the speed of cell differentiation from the multicompartment discrete models as  $V_k(x) = V_{I(i,j)}(x) = \phi(c_i, c_j)$ , where  $c_n$  is the differentiation rate of cell type  $v_n$  and  $\phi$  is an interpolation function.<sup>1</sup>

Cell proliferation and apoptosis can be modelled by the reaction coefficient  $R_k(x)$ . Similar to  $V_k$ , if only the proliferation at the discrete cell types are available, we interpolate as

$R_k(x) = R_{I(i,j)}(x) = \phi(r_i, r_j)$ , where  $r_n$  is the growth rate at node  $v_n$ . In addition to natural proliferation and apoptosis, this term can also model abnormal tumorous cell growth or the effect of targeted therapy by localized Gaussian or Dirac-delta functions centred at the location of the corresponding cell type on the graph.

The diffusion term represents the instability on the phenotypic landscape of the cells that should be taken account into when modelling the macroscopic cell density. In particular, we consider the diffusion term in Equation (2) in such form that is appropriate to model the dynamics on a graph that is reduced from a higher dimensional narrow domain. It involves two parameters  $D_k(x)$  and  $w_k(x)$  describing the magnitude of cell fluctuation and the width of the narrow domain around the edge, respectively. Considering the phenotypic fluctuation of the cell density as a random process subject to Brownian motion with magnitude  $\sigma$ , the diffusion term becomes  $D_k = \sigma^2$  and  $w_k = 1$  (Evans, 2010). In addition, the width or the area of the cross section of the narrow domain that is vertical to the projecting edge can be taken as  $w_k(x)$ , which is called Fick-Jacobs equation for deterministic PDEs (Valdes & Guzman, 2014; Zwanzig, 1992) and can be similarly derived for stochastic PDEs (Cerrai & Freidlin, 2017; Freidlin & Hu, 2013).  $w_k(x)$  can be measured as the length of maximal fluctuation in the vertical direction along the graph.

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  $N_I \doteq \{v_k \notin \bigcup_{j=1}^{n_e} \{b_j\}, k=1, \dots, n_v\}$ , e.g. stem cells, the final nodes without outflow  $N_F \doteq \{v_k \notin \bigcup_{j=1}^{n_e} \{a_j\}, k=1, \dots, n_v\}$ , e.g. the most differentiated cells, and the intermediate nodes,

$$N_T \doteq \left[ \bigcup_{j=1}^{n_e} \{a_j\} \right] \cap \left[ \bigcup_{j=1}^{n_e} \{b_j\} \right].$$

On the intermediate nodes  $v_n \in N_T$ , mixed boundary conditions can be imposed to balance the cell inflow and outflow as

$$\sum_{(i,n) \in \mathcal{J}} \mathcal{B}_{I[i,n]}(u, b_{I[i,n]}) = \sum_{(n,j) \in \mathcal{J}} \mathcal{B}_{I[n,j]}(u, a_{I[n,j]}), \tag{3}$$

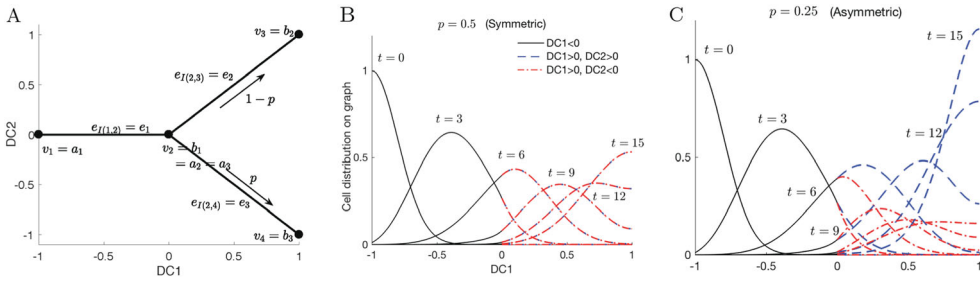
where  $\mathcal{B}_{I[i,j]}(u, x) \doteq V_{I[i,j]}(x)u(x) - D_{I[i,j]}(x)(\partial/\partial x)w_{I[i,j]}u(x)|_{x_{I[i,j]}}$ , and  $b_{I[i,n]}$  is the right end point of the edge between nodes  $i$  and  $n$ . Similarly,  $a_{I[n,j]}$  is the left end point of the edge between nodes  $n$  and  $j$ . In addition, continuity conditions are taken as Dirichlet boundary conditions as follows:

$$u(b_{I[i,n]}) = u(a_{I[n,j]}), \quad \text{for all } (i, n) \in \mathcal{J}, (n, j) \in \mathcal{J},$$

for a fixed  $n$ . The cell outflow boundary conditions on the final nodes  $v_n \in N_F$  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  $\mathcal{J}$ . Similarly this can be imposed on the initial nodes  $v_n \in N_I$  as  $(\partial/\partial x)u(a_{I[n,j]}) = \alpha_n, (n, j) \in \mathcal{J}$  or  $u(a_{I[n,j]}) = \alpha_n, (n, j) \in \mathcal{J}$ ,



**Figure 3.** We use a simple ‘Y’-shaped graph to illustrate our model. (A) The graph is defined by four nodes  $\{v_k\}_{k=1}^4$  and three edges  $e_1 = e_{I(1,2)}$ ,  $e_2 = e_{I(2,3)}$  and  $e_3 = e_{I(2,4)}$  within two components of a diffusion map (DC1, DC2). The transfer rate from  $v_2$  to  $v_3$  and  $v_4$  is taken to be proportional to  $1-p$  and  $p$ , respectively. (B) The evolution of the cell density solution from the initial condition ( $t = 0$ ) concentrated at the left end,  $DC1 = -1$ , to a density concentrated at the right end,  $DC1 = 1$ , at  $t = 15$ . In the symmetric case,  $p = 0.5$ , the two branches evolve in the same way; (C) in the asymmetric case,  $p = 0.25$ , the cell density is larger at  $t = 15$  on the upper branch, shown in blue dashed line, compared to the lower branch, shown in red dash-dotted line. (Color online).

depending on whether the model describes the cell inflow flux or a prescribed density. In particular, the prescribed value when  $u(a_{I[n,j]})$  represents the density of stem cells, one can model the discrete stem cell state as a separate ordinary differential equation (ODE) and impose its solution as the boundary condition at  $a_{I[n,j]}$  (Doumic et al., 2011; Gwiazda et al., 2012). This approach makes it possible to distinguish stem cell proliferation into the division that remains as stem cell and the one that differentiates to a matured cell.

### 3.1.1. Example on a Y-shaped graph

To illustrate our approach, we apply the PDE model given in Equation (2) to a simple Y-shaped graph. This example is motivated by cell differentiation data that reveal multiple branching procedures in the continuous space (Haghverdi, Buettner, & Theis, 2015; Moris, Pina, & Arias, 2016; Rizvi et al., 2017; Velten et al., 2017), therefore we assume the simplest case that the differentiated cells have two different cell fates with one branching. For instance, assume that the cell data projected onto the first two diffusion components, DC1 and DC2, are as in Figure 3A and the temporal direction of cell differentiation is from left to right, as indicated by the arrows in the Figure. We define the Y-shaped graph with four nodes  $v_1 = (-1, 0)$ ,  $v_2 = (0, 0)$ ,  $v_3 = (1, 1)$  and  $v_4 = (1, -1)$ , and three edges  $e_1 = e_{I(1,2)} = \overline{v_1 v_2}$ ,  $e_2 = e_{I(2,3)} = \overline{v_2 v_3}$  and  $e_3 = e_{I(2,4)} = \overline{v_2 v_4}$ . This corresponds to the set of nontrivial edges  $\mathcal{J} = \{(1, 2), (2, 3), (2, 4)\}$  and index mapping  $I$  on  $\mathcal{J}$  as  $I(1, 2) = 1$ ,  $I(2, 3) = 2$  and  $I(2, 4) = 3$  that yields the end points of the edges  $a_k$  and  $b_k$  as  $v_1 = a_1$ ,  $v_2 = b_1 = a_2 = a_3$ ,  $v_3 = b_2$  and  $v_4 = b_3$ . For simplicity, we assume that the edges are straight lines and parameterize the variables on each edge as  $e_1(x) = (x - 1, 0)$ ,  $e_2(x) = (x, x)$  and  $e_3(x) = (x, -x)$ , so that  $x \in [0, 1]$ . When there is possibility for confusion, we use subscripts on the  $x$ -variables to specify which edge is parameterized. So, for example,  $x_2$  parameterizes  $e_2$ . Then, the PDE model on each parameterized edge  $e_k$  can be written as

$$\frac{\partial u_k(x)}{\partial t} = -V_k(x) \frac{\partial u_k(x)}{\partial x} + \frac{D_k}{2W_k} \frac{\partial}{\partial x} \left( W_k \frac{\partial u_k(x)}{\partial x} \right) \quad x \in e_k, \quad k=1,2,3. \quad (4)$$

We consider the case that the cells transfer from  $v_1$  to  $v_2$  in  $n_T = 5$  unit time, differentiate into each cell type with proportion  $p$  and  $1-p$ , and accumulate at  $DC1 = 1$ , where the cells are fully differentiated.<sup>2</sup> Here, we simplify the differentiation rate to be constants assuming that the single branching Y graph lies locally and close enough in the differentiation space that the differentiation rate does not change. Then,

$$V_1(x) = \frac{1}{n_T}, \quad V_2(x) = \frac{1-p}{n_T}(1-x^2), \quad V_3(x) = \frac{p}{n_T}(1-x^2), \quad (5)$$

where  $V_2$  and  $V_3$  reflect the accumulation at cell types  $v_3$  and  $v_4$  ( $x = 1$ ). Also, we assume that the differentiation process is subject to fluctuations such as trans-differentiation (cross-lineage) and de-differentiation (stem state reversion) that is modelled as Brownian motion with a constant variance  $\sigma$  so that  $D_k = \sigma^2 = 0.02$ . Also, the maximal fluctuation in the vertical direction of the edge is assumed to be a constant that is independent of  $x$  and  $w_1 = 2w_2 = 2w_3$  so that the fluctuation in the vertical direction reduces by half in  $e_2$  and  $e_3$ .  $w_k$  cancels out in the diffusion term in Equation (4). Figure 3 plots the two examples of symmetric differentiation  $p = 0.5$  and asymmetric differentiation  $p = 0.25$ .

In this example, to demonstrate our model focusing on the cell differentiation and branching, we assume that the proliferation is zero as  $R_k = 0$  (see Appendix for the detail of modelling  $R_k$ ). The boundary type of the nodes are classified, according to our description above, as  $N_I = \{v_1\}$ ,  $N_T = \{v_2\}$  and  $N_F = \{v_3, v_4\}$ . Thus, we impose the *gluing* boundary condition, as in Equation (3) at  $v_2$ , as

$$-V_1(b_1)u(b_1) + Dw_1 \frac{\partial}{\partial x} u_1(b_1) = \sum_{k=1}^2 \left( -V_k(a_k)u_k(a_k) + Dw_k \frac{\partial}{\partial x} u_k(a_k) \right),$$

with continuity conditions  $u_1(b_1) = u_2(a_2) = u_3(a_3)$ . In addition, an inflow boundary condition is imposed at  $v_1$ , and reflecting boundary conditions at the end nodes  $v_3$  and  $v_4$  as  $u_1(a_1, t) = (1/\sqrt{0.08\pi}) \exp[-(-(1/n_T)t)^2/0.08]$ ,  $\partial u_2(b_2)/\partial x_2 = 0$  and  $\partial u_3(b_3)/\partial x_3 = 0$ . The Dirichlet condition of  $u_1(a_1, t)$  is given to resemble the transition of a certain cell state to fully differentiated cells from the initial distribution

$$u_1(x, t = 0) = \frac{1}{\sqrt{0.08\pi}} \exp \left[ -\frac{x^2}{0.08} \right],$$

$$u_i(x, t = 0) = \frac{1}{\sqrt{0.08\pi}} \exp \left[ -\frac{(x+1)^2}{0.08} \right], \quad i = 2, 3.$$

Simulations of this simple model are shown in Figure 4, where densities on edges  $e_2$  and  $e_3$  are plotted in different colours. We see that an initial cell distribution concentrated near the cell state  $v_1$  moves to the right as the cells differentiate, branches at  $v_2$  and becomes absorbed at the fully differentiated cell states  $v_3$  and  $v_4$ . In the symmetric case, when  $p = 0.5$ , the density is the same on each of the two branches to the right of  $v_2$ , so that the two curves are plotted on top of each other. When  $p = 0.25$ , the density profile is not symmetric: more cells move along the upper branch than on the lower branch. This provides a simple illustration of the mathematical details of our modelling framework, which we apply to more complicated graph structure derived from data as follows.

## 4. Simulation results

In this section, we employ the framework developed in Section 3.1 to mouse haematopoietic stem and progenitor cell (HSPC) data in Nestorowa et al. (2016). See Appendix for details, including the model parameters and simulation codes.

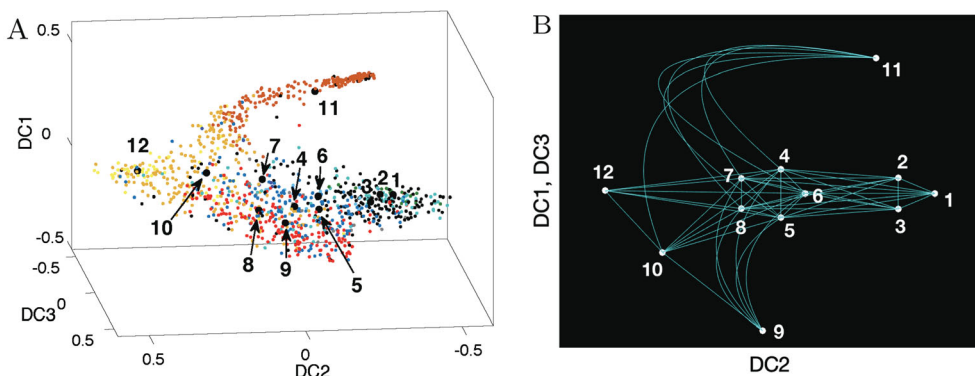
### 4.1. Model of normal adult haematopoiesis

To calibrate our model, we first apply it to normal haematopoietic cell differentiation trajectories identified in Nestorowa et al. (2016). Nestorowa et al. characterize early stages in haematopoiesis with 12 cell types, shown in Table 1 and Figure 4, including E-SLAM (CD48–CD150 +CD45+EPCR+), long-term HSCs (LT-HSCs), short-term HSCs (ST-HSCs), lymphoid-primed multipotent progenitors (LMPPs), multipotent progenitors (MPPs), megakaryocyte–erythroid progenitors (MEPs), common myeloid progenitors (CMPs), and granulocyte–macrophage progenitors (GMPs). We consider these 12 cell types as the 12 nodes,  $v_k$ , in our graph, and add 51 edges to model the haematopoietic

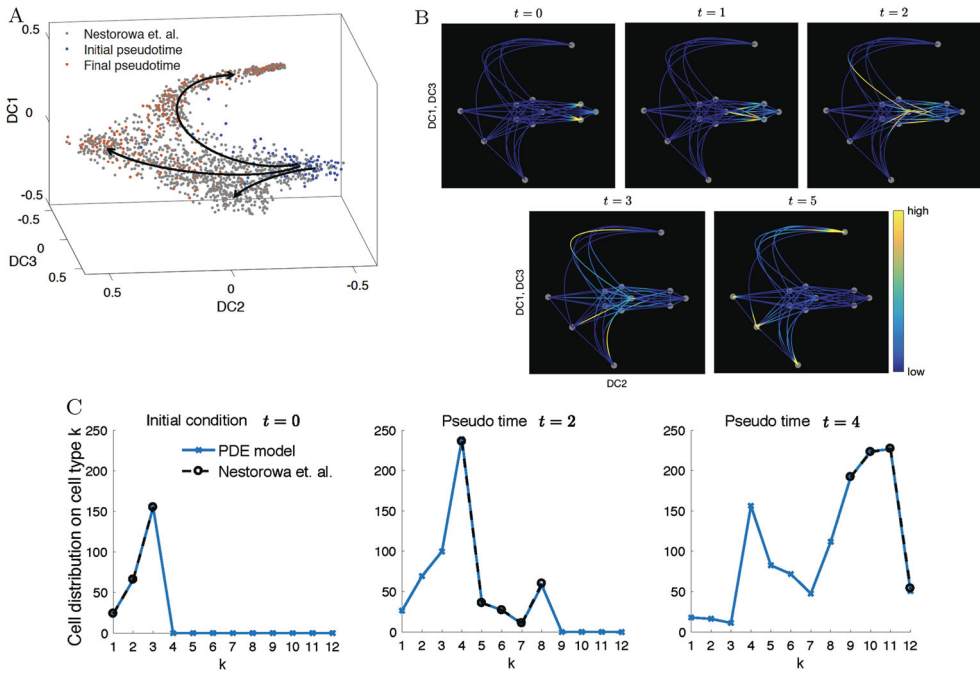
**Table 1.** Index of cell identities and labels.

Cell identities and labels			
ID	Cell type	ID	Cell type
1	E-SLAM	7	MPP2
2	L-S+K+ CD34– Flk2+ CD48– CD150+	8	MPP3
3	LT-HSC	9	LMPP
4	MPP	10	CMP
5	ST-HSC	11	MEP
6	MPP1	12	GMP

Notes: LT-HSC, ST-HSC: long- and short-term haematopoietic stem cells; MPP: multipotent progenitors, lymphoid-primed multipotent progenitors; CMP: common myeloid progenitors; MEP: megakaryocyte–erythroid progenitors; GMP: granulocyte–macrophage progenitors.



**Figure 4.** (A) For the 12 cell types identified in Nestorowa et al. (2016), the centre of mass of each cell type is used to define a node on an abstracted graph. (B) Edges between nodes are constructed based on inferred trajectories on the graph based on diffusion pseudotimes starting from nodes 1–3 to nodes 4–8, then to the progenitor nodes 9–12. The graph represents a continuum of cell states (edges) that includes identification of canonical cell states along the continuum (nodes 1–12) (Table 1).



**Figure 5.** (A) The cell data coloured by pseudotime analysis produced by the Wanderlust algorithm applied to data mapped to diffusion space in Nestorowa et al. (2016). The initial point in pseudotime is taken from the HSC cells and the final pseudotime from the progenitor cells. (B) Cell distribution computed by the PDE model on the graph from  $t = 0$  to  $t = 5$ . The cells flow from E-SLAM and LT-HSC nodes on the right to the LMPP, CMP, MEP and GMP nodes on the other three ends (bottom, top and left), following the pseudotime trajectories identified in (A). (C) Comparison of the cell type distribution computed by the PDE model described in Equation (2) and the reference data from Nestorowa et al. (2016). The reference distribution (Nestorowa et al.) is computed by clustering the initial, middle and final pseudotime cells from (A) into 12 cell nodes. By considering  $t = 4$  as the final pseudotime in the PDE model, the values of the solution at the nodes agree well with the reference data.

cell hierarchy (see Figure 1A) and pseudotime computed in Nestorowa et al. (2016) (see Figure 5A). This graph represents a continuum of canonical and intermediate states of haematopoietic differentiation with nodes and edges, respectively. The spatial variable in our PDE model represents the differentiation state of the cell.

The coloured and labelled clustered cell data and the corresponding graph are shown in Figure 2. The location of the nodes on the graph is not chosen to be identical to the data, but for an illustrative purpose to represent DC2 and DC1/DC3. The edges are chosen according to the pseudotime progression from the E-SLAM and HSCs (nodes 1–3) to the progenitor cells (nodes 9–12).

The parameters of the PDE model of cell differentiation under normal conditions are chosen to reproduce the distribution of cell types from Nestorowa et al. (2016) at the initial and final pseudotime (Figure 5C). Considering the data in Nestorowa et al. (2016) grouped by sorting gate of LT-HSC, HSPC, and progenitor cells, we denote the subsets of nodes that are representative of each group as  $\mathcal{I}_1 = \{1, 2, 3\}$  for HSC,  $\mathcal{I}_2 = \{4, \dots, 8\}$  for HSPC and  $\mathcal{I}_3 = \{9, \dots, 12\}$  for progenitor cells, where we also take  $N_I = \mathcal{I}_1$ ,  $N_T = \mathcal{I}_2$  and



$N_F = \mathcal{I}_3$ . The reference distribution is computed by counting the relative number of cells in each cluster at the initial and final pseudotime. The initial and final cell distribution is concentrated on nodes 1–3 of  $\mathcal{I}_1$  and 9–12 of  $\mathcal{I}_3$ , respectively.

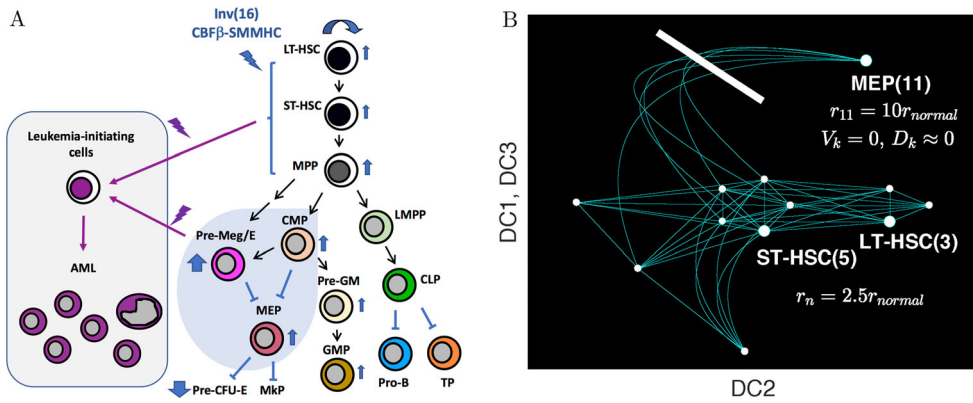
The distribution of cells in the remaining states, represented by nodes 4–8 of  $\mathcal{I}_2$ , goes from 0 at time  $t = 0$  to positive at time  $t = 2$ , and reduces at  $t = 4$ . We remark that the ratios of the number of cells in each node are used to compute the advection coefficients  $V_k$  in Equation (A3), where we take the drift  $V_{I[i,j]}$  from cell type  $i$  to another cell type  $j$  to be proportional to the ratio plotted in Figure 5C. For instance, the outflow from  $v_5$  to nodes 9–12 is taken to be proportional to the reference distribution at pseudotime  $t = 4$ . With the ratios fixed, we assume a constant parameter that represents the differentiation rate on each node, and find the values that reproduce the given cell data by simple root-finding algorithms such as the secant method. The range of the values are  $0 \leq V_k \leq 3$ . The detailed procedure is explained in Appendix.

The diffusion coefficient is taken as  $D_k = D_{I(i,j)} = 10^{-2}$  within either subsets of nodes  $i, j \in \mathcal{I}_1$  or  $i, j \in \mathcal{I}_3$ , and  $D_k = 10^{-3}$  on the other edges. The magnitude  $D_k = 10^{-2}$  corresponds to the phenotypic fluctuation of  $2.5456 \times 10^{-2}$  in the diffusion space and  $D_k = 10^{-3}$  takes into account of the increased average distance between the nodes that yields smaller diffusion coefficient due to relatively smaller fluctuation. We assume that the proliferation of the progenitor cell nodes are a constant as  $r_n = 1.3648$  at  $t \leq 2$  and  $r_n = 0.4$  at  $t > 2$  for  $n \in \mathcal{I}_2 \cup \mathcal{I}_3$ , where the proliferation rate reflects the increment of cell number from HSCs to progenitor cells in the data. Also, the proliferation at the HSC nodes is assumed to be negligible compared to progenitor cells as  $r_n = r_{n \in \mathcal{I}_2 \cup \mathcal{I}_3} \times 10^{-2}$  for  $n \in \mathcal{I}_1$  (Passequé, Wagers, Giuriato, Anderson, & Weissman, 2005). See Appendix for the model parameters and detailed discussion.

For the implementation, we discretize the system using a fourth-order finite difference method and 100 grid points on each edge, and a third-order Runge–Kutta method in time with time step  $10^{-4}$ . Figure 5C compares the solution to the PDE in the normal condition to the reference distribution. The initial condition of the PDE is taken as the initial reference distribution, and we compute the solution up to time  $t = 5$ . The solution at  $t = 4$  is similar to the reference distribution at final pseudotime. Also, the solution at  $t = 2$  is close to the distribution of the remaining cells excluding the initial and final cells. Figure 5B shows the cell distribution on the graph from time  $t = 0$  to  $t = 5$ . We observe that the cell density shifts from the initial nodes 1–3 representing HSCs, to nodes 9–12 representing progenitor cells.

## 4.2. Acute myeloid leukaemia

AML results from aberrant differentiation and proliferation of transformed leukaemia-initiating cells and abnormal progenitor cells. Parallel to the hierarchy of normal haematopoiesis, malignant haematopoiesis has also been considered to follow a hierarchy of cells at various differentiation states although with certain levels of plasticity (see Figure 6). Given the aberrant differentiation and plasticity associated with the pathology of AML, modelling in a continuous differentiation space offers the advantage over discrete models that all pathological and intermediate cell states can be captured. With our model calibrated to data obtained from normal haematopoietic differentiation trajectories, we now model the progression of AML using a genetic knock-in mouse model that recapitulates somatic acquisition of a chromosomal rearrangement, *inv(16)(p13q22)*



**Figure 6.** (A) AML is a cancer of aberrant differentiation and proliferation of haematopoietic progenitor cells. Previous studies demonstrated that expression of *inv(16)* leukaemogenic fusion protein CBF $\beta$ -SMMHC results differentiation block at multiple haematopoietic stages along with the expansion of preleukaemic stem/progenitor cells and abnormal myeloid progenitors, including CMP, Pre-Meg/E and MEP. These preleukaemic stem/progenitor cells and abnormal myeloid progenitors are susceptible to malignant transformation into leukaemia-initiating cells that drive and sustain AML pathogenesis. (B) Schematic illustration of AML pathogenesis in the differentiation continuum. To simulate *inv(16)*-driven AML, the proliferation  $R_k(x)$  connecting the nodes 3, 5 and 11 is increased and the flow toward the node 11,  $V_k(x)$  and  $D_k(x)$  for  $k = I(i, 11)$  is blocked.

(Liu et al., 1993,9), commonly found in approximately 12 % of AML cases. *Inv(16)* rearrangement results in expression of a leukaemogenic fusion protein CBF $\beta$ -SMMHC, which impairs differentiation of multiple haematopoietic lineages at various stages (Castilla et al., 1999; Kuo, Gerstein, & Castilla, 2008; Kuo et al., 2006).

Our previous studies using the *inv(16)* AML mouse model demonstrate that expression of CBF $\beta$ -SMMHC leukaemogenic fusion protein results in expansion of preleukaemic haematopoietic stem and progenitor populations susceptible to transformation into leukaemia-initiating cells which can initiate and propagate AML. Most notable was the increase in abnormal myeloid progenitors, which had an MEP-like immunophenotype and a CMP-like differentiation potential (Kuo et al., 2006). Further separation of MEPs with additional phenotypic markers (Pronk et al., 2007) show a predominant increase in pre-megakaryocyte/erythroid (Pre-Meg/E) population (ranging from 5 to 12 fold) accompanied by impaired erythroid lineage differentiation (Figure 6A) (Cai et al., 2016). This refined phenotypic Pre-Meg/E population consists partly of the CMP and MEP populations using conventional markers (Akashi, Traver, Miyamoto, & Weissman, 2000; Nestorowa et al., 2016).

The simulation of *inv(16)*-initiated AML pathogenesis is motivated by our previous observations that AML is preceded by expansion of preleukaemic myeloid progenitor cells, particularly the Pre-Meg/E and MEP-like populations with impaired differentiation. These abnormal progenitors are predisposed to subsequent cooperating events necessary to transform to overt AML (Cai et al., 2016; Castilla et al., 1999; Kuo et al., 2006). To simulate AML pathogenesis, we increase the proliferation rate of MEP (node 11) by 10 times, that is,  $r_{I[i,11]} = 10r_{normal}$ , to reflect the abnormal expansion of MEP-like cells (ranging from 5 to 12 fold based on previous data) (Cai et al., 2016; Kuo et al., 2006). Here,

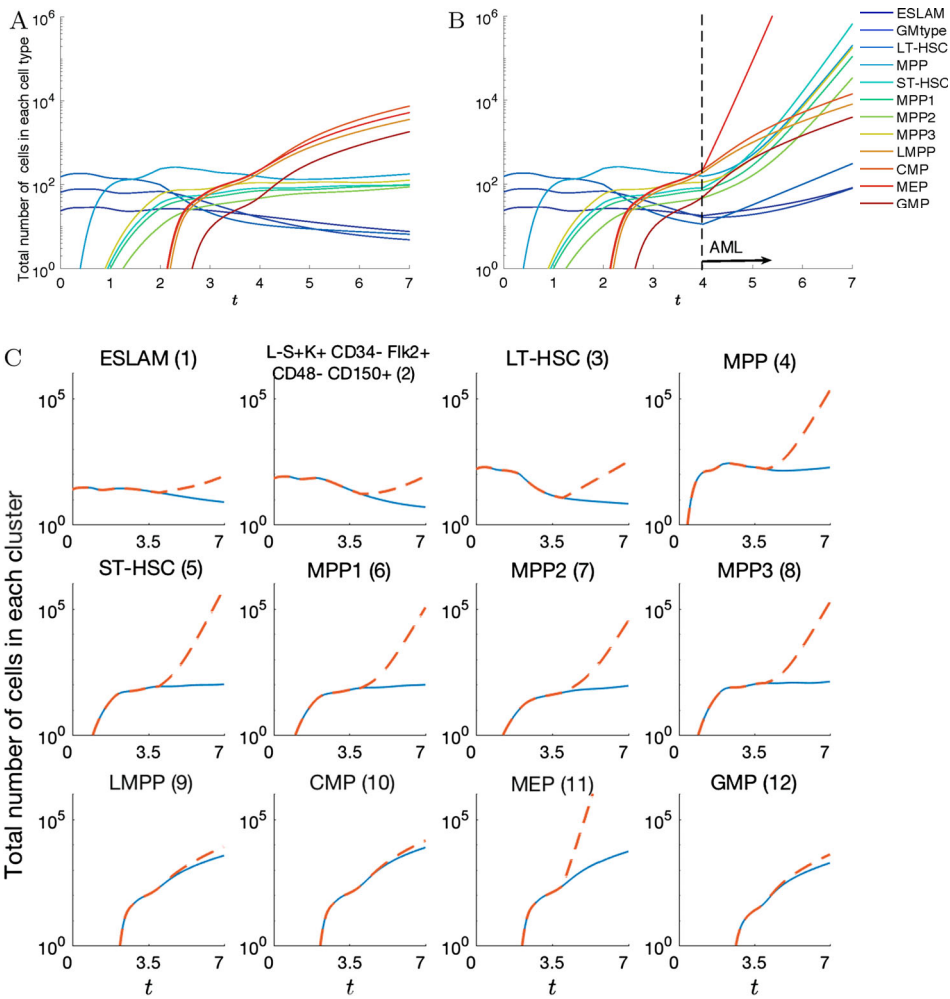
$r_{\text{normal}}$  is the value that is used in the normal condition in Section 3.1. Thus, the proliferation is assumed to be maximal at the MEP node,  $R_k(v_{11})$  and the proliferation of intermediate cells that are phenotypically close to MEP, that is,  $R_{I[i,11]}(x)$  near  $x = v_{11}$ , also increase. Also, the flow to the MEP is blocked by taking zero advection coefficient on the edge that is connected to  $v_{11}$ , i.e.  $V_{I(i,11)}(x) = 0$ . We also lower diffusion by 10 as  $D_{I(i,11)}(x) = D_{\text{normal}}/10$  to model the phenotypic fluctuations and imperfect differentiation block involved in AML pathogenesis. The differentiation block is imperfect because there is a continuum of leukaemic cell phenotypes (states).

In addition, the proliferation rate of LT-HSC and ST-HSC (nodes 3 and 5), that is,  $r_3$  and  $r_5$ , is increased by 2.5 times as  $2.5r_{\text{normal}}$  (Figure 6B). We model the induction of the leukaemogenic fusion protein CBF $\beta$ -SMMHC resulting from the chromosome inversion *inv(16)* (p13q22) as the ‘start of AML’. In this murine model of AML, *inv(16)* is the initial founder event that results in differentiation block and expansion of abnormal progenitors, which are predisposed to subsequent cooperating events necessary to transform to overt AML (Cai et al., 2016; Castilla et al., 1999; Kuo et al., 2006). The approach used here directly models the sequence of events observed during leukaemia initiation. Finally, we denote  $t_{\text{AML}}$  as the effective time that the aforementioned 10-fold proliferation change in MEP and other abnormal differentiation and proliferations due to AML are observed. The other parameters except the ones described in this section follow the ones from Section 4.1.

Figure 7 shows the total number of cells in each cell type in the normal and AML conditions starting at  $t = 4$ . In the normal condition, the CMP, MEP and LMPP cells dominate the population after  $t \geq 4$ . However, in the AML case, the MEP cells increase up to 100 times of the normal condition after a single pseudotime and dominate the population. Figure 7C plots the number of cells in each cell type separately, where we can observe the increasing number of cells not only in MEP, but also in the intermediate cell types, 4–8. Figure 8 compares the cell distribution on the graph between the normal and the AML case. In the AML case, the peak is shown on the edges near MEP cells.

The continuum of intermediate cell types, represented as numbers of cells along the edges of the graph are plotted in Figure 9. The cell distribution in the normal case at  $t = 1$  and  $t = 3$  shows the cell population moving on the edges from HSCs to progenitor states. Under normal haematopoiesis, we observe the flow of cells along the continuum from a stem cell like state to a progenitor state, with an even distribution of all types of progenitor cells. However in the AML case, we predict the emergence of novel intermediate cell types, including a mixed CMP-MPP3 and CMP-MEP cell type. These indeterminate cells may exhibit phenotypic and/or functional properties of both cell types on either side of the edge (node  $i$  and/or node  $j$ ). This cell state may be unstable, phenotypically plastic, may be in an abnormal state or process of differentiation, or perhaps even undergoing a selection pressure to induce transformation. Of note, this prediction of a mixed CMP-MEP cell type echoes the biological observation that abnormal myeloid progenitors seen during AML progression exhibit an MEP-like immunophenotype with a CMP-like functional readout (Kuo et al., 2006). This mixed identity/functionality coincides with a strong differentiation block toward erythrocyte and megakaryocytes (Cai et al., 2016).

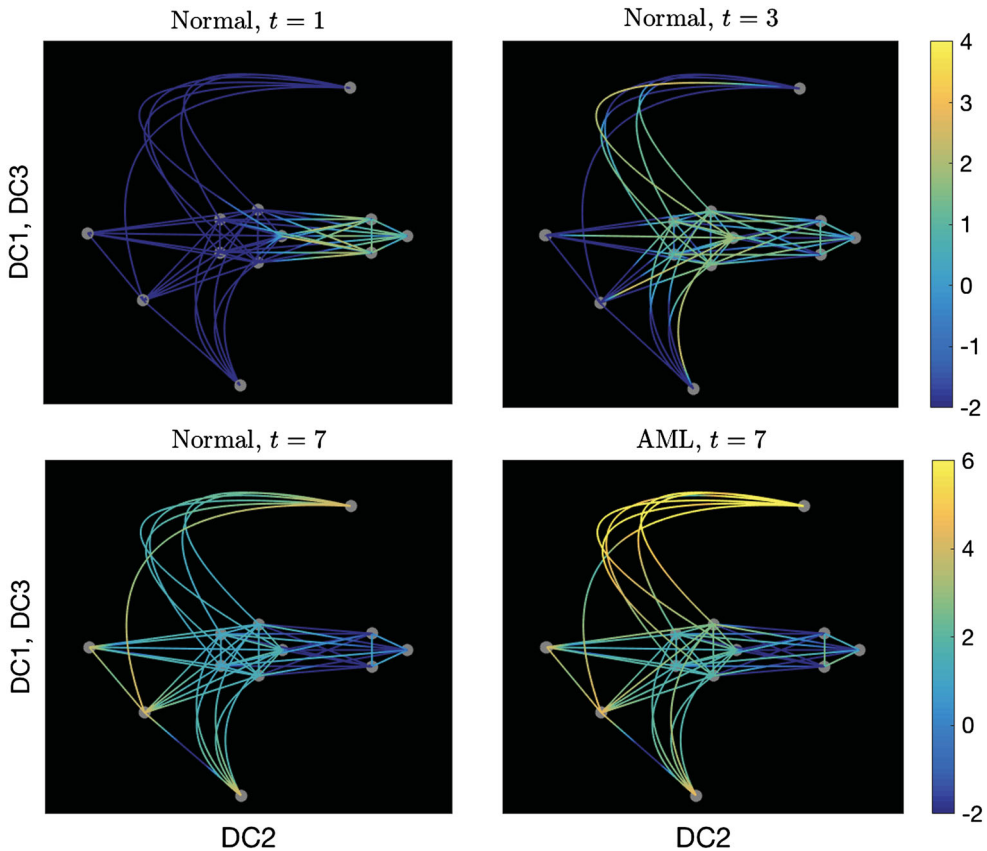
This highlights the advantage of modelling pathologic conditions in a continuum of cell states as the phenotypic properties and the differentiation processes are often abnormal during pathogenesis. This approach also circumvents the limitations of varying phenotypic definitions used in different studies in the literature (e.g. MEP vs. Pre-Meg/E) and the



**Figure 7.** Total number of cells in each node up to  $t = 7$  in (A) normal condition and (B) AML pathogenesis. The AML simulation started at  $t = 4$ . Compared to the normal case, cells in MEP, LT-HSC and ST-HSC increase as well as other cell types. (C) compares the number of cells between the normal and the AML case for each cell type individually.

varying degree of heterogeneity within phenotypically defined cell populations in health and in disease.

We also simulated AML starting at different time points from  $t = 1$  to  $t = 6$ . Since our initial condition assumes that the cells have not yet developed to MEP, the total number of cells is maximized when the AML occurs after a critical amount of cells have differentiated into an MEP state. Figure 10 shows the results of model simulations, where we observe that the number of cells is maximal at later times when AML is started at  $t = 3$ . From these simulations, we infer that the short- and long-term evolution of AML may depend on the state and composition of the haematopoietic landscape at the time of AML initiation.



**Figure 8.** The cell distribution on the graph in a  $\log_{10}$  scale, comparing the normal and AML conditions at  $t = 7$ . The AML condition shows increased density on the edges near the MEP state (node 11) at  $t = 7$ .

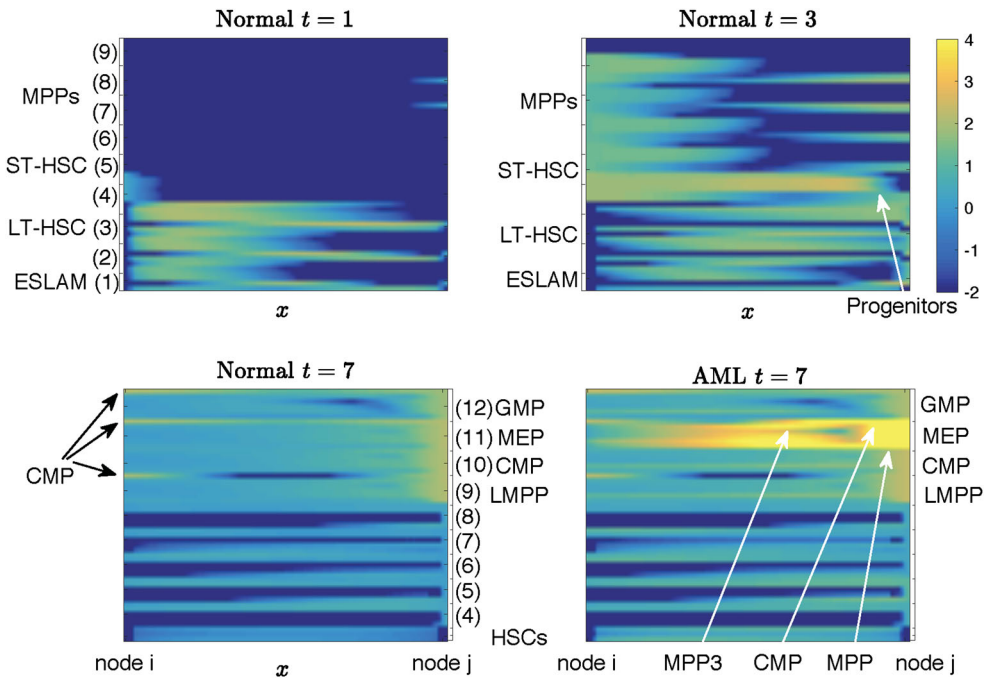
## 5. Discussion

We present a mathematical model of movement in an abstract space representing states of cellular differentiation. We represent trajectories in the differentiation space as a graph and model directed and random movement on the graph with PDEs. We demonstrate our modelling approach on a simple graph and then apply our model to haematopoiesis with publicly available scRNA-Seq data. We calibrate the PDE model to pseudotime trajectories in the diffusion map space and use the model to predict the early stages of pathogenesis of AML.

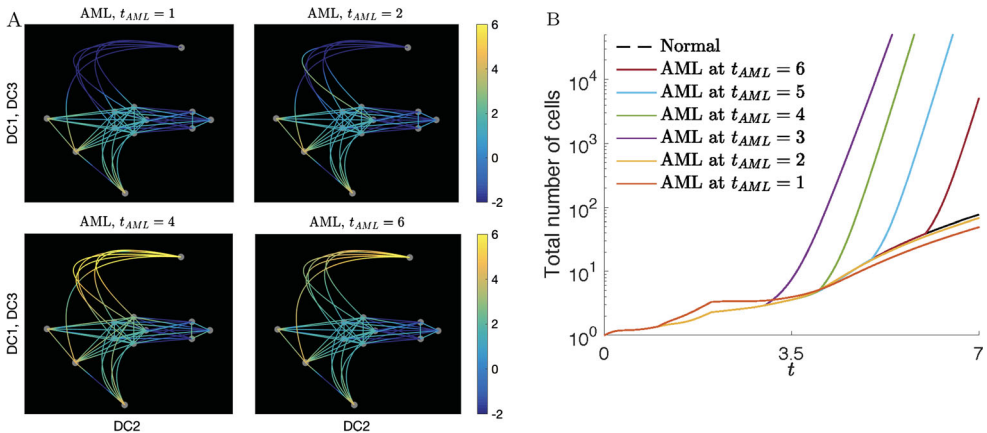
A more traditional approach for modelling the process of cell differentiation is to use a discrete collection of ODEs that describe dynamics of cells at  $n$  different maturation stages and the transition between those stages (cf. Lander et al., 2009; Lo et al., 2008; Marciniak-Czochra et al., 2009; Stiehl & Marciniak-Czochra, 2011). These discrete models are also referred to as ‘multicompartmental models’, and are based on the biological assumption that in each lineage of cell precursors there are discrete steps in the maturation process that are followed sequentially (cf. Lord, 1997; Uchida, Fleming, Alpern, & Weissman, 1993).

This view of the differentiation process being discrete does not capture biological observations that indicate that cell differentiation is more likely a continuous process, and that





**Figure 9.** The continuum of cell states can be visualized as the density of cells along the 51 edges of the graph (rows) connecting node  $i$  (left) to node  $j$  (right) for all nodes  $i, j$ . Cell distribution ( $\log_{10}$  scale) on the edge comparing the normal condition and AML. In addition to an accumulation of MEP cells, novel intermediate cell states emerge resulting from the differentiation block and increased proliferation rate resulting from AML. These novel cell states are indicated with white arrows and generally fall between the CMP, MPP and MEP canonical cell states. The presented edges in the first row ( $t < 4$ ) are lexicographically ordered with respect to the left end ( $a_n$ ) to visualize which nodes are the differentiating cells departing from and with respect to the right end ( $b_n$ ) in the second row ( $t > 4$ ) to visualize which nodes are the arriving cells differentiated into.



**Figure 10.** (A) Cell distribution on the graph at  $t = 7$  for AML occurring at different times,  $t_{AML} = 1, 2, 4$  and  $6$ . MEP (11) blows up when AML occurs after  $t \geq 2$ . The dominating intermediate cells are also distinct. (B) Relative total number of cells when AML occurs at  $t_{AML} = 1$  to  $t_{AML} = 6$  compared to the normal case (dashed line) up to time  $t = 7$ . The total number of cells is maximized when AML occurs at  $t_{AML} = 3$ .



maturation may, in fact, even be decoupled from cell division (cf. Dontu, Al-Hajj, Abdallah, Clarke, & Wicha, 2003; Doumic et al., 2011). A number of mathematical models have been created that aim to capture the continuous process of cell differentiation (Adimy, Crauste, & Ruan, 2005; Alarcon, Getto, Marciniak-Czochra, & Vivanco, 2011; Bélair, Mackey, & Mahaffy, 1995; Colijn & Mackey, 2005; Doumic et al., 2011; Doumic-Jauffret, Kim, & Perthame, 2010; Gwiazda et al., 2012; Pujo-Menjouet, Crauste, & Adimy, 2004).

For example, in Doumic et al. (2011), the authors present a model of cell differentiation that assumes that the dynamics of differentiated precursors can be approximated by a continuous maturation model. The model is created by extending the multicompartment discrete system of Marciniak-Czochra et al. (2009). The authors provide a careful comparison that shows that the continuous structured population model is not a mathematical limit of the discrete multicompartment model. In particular, the dynamics of the continuous model allow commitment and maturation of cell progenitors to be a continuous process that can take place between cell divisions. They do show, however, that there is overlap in model dynamics with a particular choice of maturation rate. In Gwiazda et al. (2012), the authors subsequently developed a continuous model that can be viewed as a generalization that admits both the continuous model of Doumic et al. (2011) and the discrete model of Marciniak-Czochra et al. (2009) as special cases. In Prokharau, Vermolen, and García-Aznar (2014), the authors develop a PDE-based continuous model of cell differentiation that allows cells to differentiate into an arbitrary number of cell types. A particular differentiation trajectory can be determined by any number of parameters, such as biochemical factors, the current differentiation state or just by a random variable, so their approach allows differentiation to be either a deterministic or a stochastic process.

The modelling approach we present differs from previous cell differentiation models in that it is centred on capturing cell differentiation dynamics that take place within a space that has been created via a dimension reduction transformation of high-dimensional data. Within that reduced space, our model assumes that maturation and differentiation take place along a continuous trajectory. (The dimension reduction outcomes on the data sets we tested indicate that the trajectory will, in fact, be continuous.) Cells can differentiate along an arbitrary number of paths with an arbitrary number of end states, all of which are determined by the data set and dimension reduction technique employed. Thus, the reduced differentiation space is not predetermined, but is generated as a function of the dimension reduction technique and the data set of interest.

Although methods exist to characterize differentiation trajectories, such as OT (Schiebinger et al., 2017) and diffusion pseudotime methods (Haghverdi et al., 2016), an advantage of our approach is the ability to use a mathematical model to predict the outcomes of abnormal trajectories and to *perturb* the system mathematically with the model. We use this advantage of the mathematical model to simulate and explore AML pathogenesis based on immunophenotypic characterization of a mouse model for *inv(16)* AML. Our simulation results are consistent with the evolution of *inv(16)*-driven AML and predict dynamics in canonical cell populations as well as cells in novel, intermediate states of differentiation. The intermediate cell states such as CMP-MEP seen in our simulation is consistent with previous observations that CBF $\beta$ -SMMHC expressing phenotypic MEP cells confer CMP-like progenitor cell activity (Kuo et al., 2006). Given the phenotypic plasticity and aberrant differentiation occurring during leukaemia evolution, it is particularly informative to model cell dynamics in a continuum of differentiation space.

The novelty and power of this modelling approach is the ability to capture and predict dynamics of many interconnected cell types. We now consider a continuum of cellular states, and model movement between these states in aggregate by representing many cell populations and states in a single variable. This approach increases biological resolution of the system by characterizing an infinite number of sub-states in a continuum representation and allows us to make predictions with one equation and very few model parameters, which can be directly calibrated to experimental data, for example, with time-series cell differentiation experiments. These data could be used in place of the inferred pseudo-time methods to construct more realistic differentiation trajectories, as well as estimate parameters such as the transport rates between locations in the differentiation space. We note that this is not equivalent to rates of cellular differentiation, since this allows inference of transition between intermediate states of differentiation which may not be directly calculated from differentiation assays which rely on specific lineage markers.

A limitation of our approach is that it does not include physical properties of the living biological system, such as the cellular microenvironment, which is known to play a critical role in the transformation of cell state and function. Furthermore, we recognize and acknowledge that cellular state transition dynamics as represented as a projection in a low-dimensional space is an approximation of the dynamics in the original high-dimensional space. Moreover, the dynamics observed and predicted in the lower dimensional space critically depend on the method of dimension reduction. This logic motivates our use of diffusion maps as the method to construct the differentiation space.

In addition, our current model assumes that the cell properties of the intermediate cell types change linearly between the node cell types. Although it is reasonable to assume that the overall cell properties in the macro scale changes linearly depending on the distance in the phenotypic space when no other information in between is given, our future work involves using the expression levels of the intermediate cells that are related to cell dynamics, e.g. cell cycle, differentiation and proliferation, to develop more appropriate models for the intermediate cells. A limitation of the Nestorowa et al. (2016) data set is that it includes only stem and committed progenitor cells, and lacks a population of fully differentiated cells (e.g. erythrocytes, platelets, B-cells, T-cells, etc.), which yields an incomplete differentiation trajectory. Although we note that the stem and progenitor cell populations are the leukaemia-initiating cell populations most immediately relevant to the pathogenesis of *inv(16)*-driven AML (Cai et al., 2016). Data sets covering the full spectrum of differentiation trajectory during normal and abnormal (AML) haematopoiesis will enable modelling of differentiation blocks occurring at later stages of differentiation.

However, despite these limitations, we contend that this kind of analysis is a critical and valuable first step toward understanding the evolution of the higher dimensional system, and that low-dimensional approximations have value, particularly when predictions in the lower dimensional space can be experimentally validated. We postulate that when dynamics in low-dimensional representations are sufficiently characterized, they may eventually be used as a surrogate for high-dimensional data, thus reverting the trend of 'big data' back down to more informative 'small data'.

We note that our modelling approach can be applied to any data set or manifold shape. As more normal and abnormal cellular state transitions are characterized at single-cell resolution, we may apply similar computational and modelling methods to those systems. We emphasize our modelling approach is general and is not tailored or adapted to

haematopoiesis in particular. Future applications of this approach may be useful to model the effects of therapies which target specific states of differentiation or the differentiation process itself, including other hematologic malignancies.

## Notes

1. The interpolation function can be taken, for instance, as a linear function  $\phi(c_i, c_j) = (c_j - c_i)(x - a_k)/(b_k - a_k) + c_i$ , where  $k = I(i, j)$ . This assumes that the cell property changes linearly in terms of the distance in the diffusion component space (Doumic et al., 2011; Gwiazda et al., 2012). In addition, the values of  $V_{I(n,j)}(x)$  near  $x = v_n$  will take into account of the ratio of cells that branch out to different cell types  $v_j$ , while the values of  $V_{I(i,n)}(x)$  consider the ratio of cells that are flowing in from different cell types  $v_i$ .
2. Using the notation in Appendix,  $\gamma_3 = p$  and  $\gamma_4 = 1 - p$ .

## Acknowledgments

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

## Disclosure statement

No potential conflicts of interest are disclosed by the authors.

## Funding

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under award number P30CA033572 and R01CA178387.

## ORCID

H. Cho  <https://orcid.org/0000-0003-3005-8943>

K. Ayers  <https://orcid.org/0000-0003-1489-5872>

Y.-H Kuo  <https://orcid.org/0000-0003-2595-0419>

J. Park  <https://orcid.org/0000-0002-4470-462X>

A. Radunskaya  <https://orcid.org/0000-0002-2353-5046>

R. Rockne  <http://orcid.org/0000-0002-1557-159X>

## References

- Adimy, M., Crauste, F., & Ruan, S. (2005). A mathematical study of the hematopoiesis process with applications to chronic myelogenous leukemia. *SIAM Journal on Applied Mathematics*, 65(4), 1328–1352.
- Akashi, K., Traver, D., Miyamoto, T., & Weissman, I. L. (2000). A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature*, 404(6774), 193–197.
- Alarcon, T., Getto, P., Marciniak-Czochra, A., & Vivanco, D. (2011). A model for stem cell population dynamics with regulated maturation delay. *Discrete and Continuous Dynamical Systems*, 2011(special), 32–43.
- Bach, K., Pensa, S., Grzelak, M., Hadfield, J., Adams, D. J., Marioni, J. C., . . . Khaled, W. T. (2017). Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nature Communications*, 8(1), 2128.
- Bélair, J., Mackey, M. C., & Mahaffy, J. M. (1995). Age-structured and two-delay models for erythropoiesis. *Mathematical biosciences*, 128(1–2), 317–346.

- Cai, Q., Jeannot, R., Hua, W-K. K., Cook, G. J., Zhang, B., Qi, J., . . . Kuo, Y. H. (2016). CBF $\beta$ -SMMHC creates aberrant megakaryocyte-erythroid progenitors prone to leukemia initiation in mice. *Blood*, 128(11), 1503–1515.
- Castilla, L. H., Garrett, L., Adya, N., Orlic, D., Dutra, A., Anderson, S., . . . Liu, P. P. (1999). The fusion gene Cbfb-MYH11 blocks myeloid differentiation and predisposes mice to acute myelomonocytic leukaemia. *Nature Genetics*, 23(2), 144–146.
- Cerrai, S., & Freidlin, M. (2017). SPDEs on narrow domains and on graphs: An asymptotic approach. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 53, 865–899.
- Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., . . . Zucker, S. W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences*, 102(21), 7426–7431.
- Colijn, C., & Mackey, M. C. (2005). A mathematical model of hematopoiesis–I. Periodic chronic myelogenous leukemia. *Journal of Theoretical Biology*, 237(2), 117–132.
- Dontu, G., Al-Hajj, M., Abdallah, W. M., Clarke, M. F., & Wicha, M. S. (2003). Stem cells in normal breast development and breast cancer. *Cell Proliferation*, 36, 59–72.
- Doumic, M., Marciniak-Czochra, A., Perthame, B., & Zubelli, J. P. (2011). A structured population model of cell differentiation. *SIAM Journal on Applied Mathematics*, 71(6), 1918–1940.
- Doumic-Jauffret, M., Kim, P. S., & Perthame, B. (2010). Stability analysis of a simplified yet complete model for chronic myelogenous leukemia. *Bulletin of Mathematical Biology*, 72(7), 1732–1759.
- Evans, L. C. (2010). *Partial differential equations* (2nd ed.). Providence: American Mathematical Society.
- Freidlin, M., & Hu, W. (2013). On diffusion in narrow random channels. *Journal of Statistical Physics*, 152, 136–158.
- Gwiazda, P., Grzegorz, J., & Marciniak-czochra, A. (2012). Models of discrete and continuous cell differentiation in the framework of transport equation. *SIAM Journal on Mathematical Analysis*, 44(2), 1103–1133.
- Haghverdi, L., Buettner, F., & Theis, F. J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, 31(18), 2989–2998.
- Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., & Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13(10), 845–848.
- Hamey, F. K., Nestorowa, S., Wilson, N. K., & Göttgens, B. (2016). Advancing haematopoietic stem and progenitor cell biology through single-cell profiling. *FEBS Letters*, 590(22), 4052–4067.
- Hao, S., Chen, C., & Cheng, T. (2016). Cell cycle regulation of hematopoietic stem or progenitor cells. *International Journal of Hematology*, 103(5), 487–497.
- Khalid, S., Khalil, T., & Nasreen, S. (2014, August). A survey of feature selection and feature extraction techniques in machine learning. In *Science and information conference (SAI) 2014*. London: IEEE.
- Kuo, Y-H. H., Gerstein, R. M., & Castilla, L. H. (2008). Cbfb-SMMHC impairs differentiation of common lymphoid progenitors and reveals an essential role for RUNX in early B-cell development. *Blood*, 111(3), 1543–1551.
- Kuo, Y-H. H., Landrette, S. F., Heilman, S. A., Perrat, P. N., Garrett, L., Liu, P. P., . . . Castilla, L. H. (2006). Cbfb-SMMHC induces distinct abnormal myeloid progenitors able to develop acute myeloid leukemia. *Cancer Cell*, 9(1), 57–68.
- Lander, A. D., Gokoffski, K. K., Wan, F. Y. M., Nie, Q., & Calof, A. L. (2009). Cell lineages and the logic of proliferative control. *PLOS Biology*, 7(1), e1000015.
- Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., Amir, E. D., Tadmor, M., . . . Nolan, G. P. (2015). Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1), 184–197.
- Liu, P., Tarlé, S. A., Hajra, A., Claxton, D. F., Marlton, P., Freedman, M., . . . Collins, F. S. (1993). Fusion between transcription factor CBF beta/PEBP2 beta and a myosin heavy chain in acute myeloid leukemia. *Science*, 261(5124), 1041–1044.

- Liu, P. P., Wijmenga, C., Hajra, A., Blake, T. B., Kelley, C. A., Adelstein, R. S., . . . Collins, F. S. (1996). Identification of the chimeric protein product of the CBF $\beta$ -MYH11 fusion gene in inv(16) leukemia cells. *Genes, Chromosomes & Cancer*, 16(2), 77–87.
- Lo, W.-C., Chou, C.-S., Gokoffski, K., Wan, F., Lander, A., Calof, A., . . . Nie, Q. (2009). Feedback regulation in multistage cell lineages. *Mathematical Biosciences and Engineering*, 6(1), 59–82.
- Lord, B. I. (1997). Growth factors and the regulation of haemopoietic stem cells. In C. Potten (Ed.), *Stem cells* (pp. 401–422). Cambridge, UK: Academic Press.
- van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.
- Mao, Q., Wang, L., Goodison, S., & Sun, Y. (2015). Dimensionality reduction via graph structure learning. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. KDD '15* (pp. 765–774). New York, NY: ACM.
- Marciniak-Czochra, A., Stiehl, T., Ho, A. D., Jäger, W., & Wagner, W. (2009). Modeling of asymmetric cell division in hematopoietic stem cells—regulation of self-renewal is essential for efficient repopulation. *Stem Cells and Development*, 18(3), 377–386.
- Moris, N., Pina, C., & Arias, A. M. (2016). Transition states and cell fate decisions in epigenetic landscapes. *Nature Reviews Genetics*, 17(11), 693–703.
- Nestorowa, S., Hamey, F. K., Sala, B. P., Diamanti, E., Shepherd, M., Laurenti, E., . . . Göttgens, B. (2016). A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood*, 128(8), e20–e31.
- Passegué, E., Wagers, A. J., Giuriato, S., Anderson, W. C., & Weissman, I. L. (2005). Global analysis of proliferation and cell cycle gene expression in the regulation of hematopoietic stem and progenitor cell fates. *The Journal of Experimental Medicine*, 202(11), 1599–1611.
- Pietras, E. M., Warr, M. R., & Passegué, E. (2011). Cell cycle regulation in hematopoietic stem cells. *The Journal of Cell Biology*, 195(5), 709–720.
- Prokharau, P. A., Vermolen, F. J., & Garcia-Aznar, J. M. (2014). A mathematical model for cell differentiation, as an evolutionary and regulated process. *Computer Methods in Biomechanics and Biomedical Engineering*, 17(10), 1051–1070.
- Pronk, C. J. H., Rossi, D. J., Mansson, R., Attema, J. L., Norddahl, G. L., C. K. F. Chan, . . . Bryder, D. (2007). Elucidation of the phenotypic, functional, and molecular topography of a myeloerythroid progenitor cell hierarchy. *Cell Stem Cell*, 1(4), 428–442.
- Pujo-Menjouet, L., Crauste, F., & Adimy, M. (2005). On the stability of a nonlinear maturity structured model of cellular proliferation. *Discrete and Continuous Dynamical Systems*, 12(3), 501–522.
- Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A., Trapnell, C., & Cellular Biology Program (2017b). Single-cell mRNA quantification and differential analysis with census. *Nature Methods*, 14(3), 309–315.
- Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., . . . Trapnell, C. (2017a). Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*, 14(10), 979–982.
- Rashid, S., Kotton, D. N., & Bar-Joseph, Z. (2017). TASIC: Determining branching models from time series single cell data. *Bioinformatics*, 33(16), 2504–2512.
- Rizvi, A. H., Camara, P. G., Kandror, E. K., Roberts, T. J., Schieren, I., Maniatis, T., . . . Rabadan, R. (2017). Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nature Biotechnology*, 35(6), 551–560.
- Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., . . . Lander, E. S. (2017). Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming. *BioRxiv*. doi:10.1101/191056
- Setty, M., Tadmor, M. D., Reich-Zeliger, S., Angel, O., Salame, T. M., Kathail, P., . . . Pe'er, D. (2016). Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature Biotechnology*, 34, 637–645.
- Shekhar, K., Lapan, S. W., Whitney, I. E., Tran, N. M., Evan, Z., Kowalczyk, M., . . . Sanes, J. R. (2016). Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell*, 166(5), 1308–1323.e30.



Stiehl, T., & Marciniak-Czochra, A. (2011). Characterization of stem cells using mathematical models of multistage cell lineages. *Mathematical and Computer Modelling*, 53(7–8), 1505–1517.

Uchida, N., Fleming, W. H., Alpern, E. J., & Weissman, I. L. (1993). Heterogeneity of hematopoietic stem cells. *Current Opinion in Immunology*, 5(2), 177–184.

van Unen, V., Höllt, T., Pezzotti, N., Li, N., Reinders, M. J. T., Eisemann, E., . . . Lelieveldt, B. P. F. (2017). Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types. *Nature Communications*, 8(1), 1740.

Valdes, C. V., & Guzman, R. H. (2014). Fick-Jacobs equation for channels over three-dimensional curves. *Physical Review E*, 90(5), 1–11.

Velten, L., Haas, S. F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B. P., . . . Steinmetz, L. M. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. *Nature Cell Biology*, 19(4), 271–281.

Waddington, C. H. (1957). *The strategy of the genes. A discussion of some aspects of theoretical biology*. London: George Allen & Unwin.

Zwanzig, R. (1992). Diffusion past an entropy barrier. *The Journal of Physical Chemistry*, 96(10), 3926–3930.

### Appendix. Model and parameters

Here we present the PDE model and parameter calculations used to produce the results presented in Section 4. MATLAB files used to generate the figures and results can be downloaded from <https://github.com/heyrim/Modeling-acute-myeloid-leukemia-in-a-continuum-of-differentiation-states>. The cell distribution  $u(x, t)$  is computed on the graph  $G$  as

$$\frac{\partial u_k}{\partial t} = -\frac{\partial}{\partial x}(V_k(x)u_k) + R_k(x)u_k + \frac{D_k(x)}{2} \frac{\partial^2 u_k}{\partial x^2}, \quad x \in e_k, \quad k=1, \dots, 51, \tag{A1}$$

where  $u_k$  is the solution projected on the edge  $e_k$  as  $u_k(x, t) \doteq u(x, t)|_{x \in e_k}$  and  $\{e_k\}_{k=1}^{51}$  are the 51 edges connecting the 12 nodes  $\{v_n\}_{n=1}^{12}$  as in Figure 2B. We assume that the edges are unit length as  $e_k = [a_k, b_k] = [0, 1]$  and find the coefficients in Equation (A1) that are scaled to the unit length edge.

The total number of cells can be computed as  $\rho(t) \doteq \sum_{k=1}^{51} \int_{e_k} u_k(x, t) dx$ , and we compute the number of cells in the  $n$ th cluster as

$$\rho_n(t) \doteq \sum_{k=I(n,j)} \int_{a_k}^{(a_k+b_k)/2} u_k(x, t) dx + \sum_{k=I(i,n)} \int_{(a_k+b_k)/2}^{b_k} u_k(x) dx. \tag{A2}$$

Alternatively, since the boundary of the cell types are not distinctive, one can compute it as a weighted sum along the edges adjacent to the node  $n$  with linear weight functions such as  $\omega(x) = -x + 1$  and  $1 - \omega(x)$  along the entire edge.

To obtain the transfer rate between the cell nodes, we assume three discrete pseudotimes at those three sorted groups starting from LT-HSC to HSPC, and finally to progenitor cells. As remarked in Section 4.1, we consider subsets of nodes  $\mathcal{I}_1 = \{1, 2, 3\}$  as HSC,  $\mathcal{I}_2 = \{4, \dots, 8\}$  as HSPC and  $\mathcal{I}_3 = \{9, \dots, 12\}$  as progenitor cell group. This follows the cell data in Nestorowa et al. (2016) that is classified with ComBat from the SVA package using the sorting gate of LT-HSC, HSPC and progenitor, and then processed with diffusion mapping initialized from a subpopulation of LT-HSC to the progenitor cells of different lineage of erythroid, granulocyte–macrophage and lymphoid. Accordingly, we consider three discrete pseudotimes considering LT-HSC ( $t_0$ ), HSPC ( $t_1$ ) and progenitor ( $t_2$ ) and compute the number of cells in each node that is summarized in Table A1. We comment that diffusion pseudotime is not a physical time unit (i.e. days) and that the differentiation process is modelled based on the inferred pseudotime trajectories with the following mapping of pseudotimes  $t_0 = 0$ ,  $t_1 = 2$  and  $t_2 = 4$ . The time mapping procedure can be refined with time-series differentiation assay data. The transfer rates between the nodes are taken from the ratios at each pseudotime.



We compute the ratio as time independent within the subsets as follows:

$$\gamma_n \doteq \bar{\rho}_n / \sum_{n \in \mathcal{I}_l} \bar{\rho}_n, \quad n \in \mathcal{I}_l,$$

that is,  $\gamma_1 = 24/245, \gamma_2 = 66/245, \gamma_3 = 155/245$  for  $\mathcal{I}_1, \gamma_4 = 236/370, \gamma_5 = 36/370, \gamma_6 = 27/370, \gamma_7 = 11/370, \gamma_8 = 60/370$  for  $\mathcal{I}_2$ , and  $\gamma_9 = 192/696, \gamma_{10} = 223/696, \gamma_{11} = 227/696, \gamma_{12} = 54/696$  for  $\mathcal{I}_3$ . We remark that the transfer rates can be time dependent as  $\gamma_n(t)$  if the data are collected at sequential timepoints, which is one way that the model could be parameterized.

We take these values as the in and out transfer rate imposed in the advection coefficient. For each node, we assume a constant parameter  $c_n \neq 0$  that determines the magnitude of the advection coefficient, that is, the speed of the cell differentiation. We take the transfer in rate at the node  $v_n, n \in \mathcal{I}_l$ , as  $V_{I(i,n)}(b_{I(i,n)}) = \gamma_i c_n$ , for  $i \in \mathcal{I}_{l-1}$ , and transfer out rate as  $V_{I(n,j)}(a_{I(n,j)}) = \gamma_j c_n$ , for  $j \in \mathcal{I}_{l+1}$ . Using the fixed transfer rates at the nodes, the advection coefficient is linearly interpolated as  $V_{I(i,j)}(x) = V_{I(i,j)}(a_{I(i,j)}) + (V_{I(i,j)}(b_{I(i,j)}) - V_{I(i,j)}(a_{I(i,j)}))x$ , that is,

$$V_{I(i,j)}(x) = \gamma_j c_i + (\gamma_i c_j - \gamma_j c_i)x, \quad i \in \mathcal{I}_1, j \in \mathcal{I}_2. \tag{A3}$$

In addition, we apply the weight  $(1 - x^2)$  to model the accumulation of cells at the progenitor nodes  $j \in \mathcal{I}_3$ ,

$$V_{I(i,j)}(x) = (\gamma_j c_i + (\gamma_i c_j - \gamma_j c_i)x)(1 - x^2), \quad i \in \mathcal{I}_2, j \in \mathcal{I}_3,$$

and take  $V_{I(i,j)}(x) = 0$ , for other pairs of nodes. For instance,  $V_{I(i,j)}(x) = 0$ , for  $i, j \in \mathcal{I}_1$  within the same hierarchy of cells, and the transition between these nodes are only governed by diffusion. The constant parameter at each node  $c_n$  is taken to reproduce the cell distribution as in Figure 5 as follows:

$$\begin{aligned} c_1 = c_2 = c_3 = 1.0, \quad c_4 = 1.2898, \quad c_5 = 0.9535, \quad c_6 = 0.9488, \\ c_7 = 0.8060, \quad c_8 = 0.8263, \quad c_9 = c_{10} = c_{11} = c_{12} = 1.0, \quad \text{for } t < t_1, \\ c_1 = c_2 = c_3 = 1.0, \quad c_4 = 1.7898, \quad c_5 = 1.4535, \quad c_6 = 1.4488, \quad c_7 = 1.3060, \\ c_8 = 1.3263, \quad c_9 = 1.7992, \quad c_{10} = 1.4380, \quad c_{11} = 1.5070, \quad c_{12} = 2.6347, \quad \text{for } t \geq t_1. \end{aligned}$$

The values are computed by a simple root-finding algorithm such as the secant method.

The diffusion coefficients on the edges are taken as  $D_k(x) = D_{I(i,j)}(x) = 10^{-2}$  between the nodes that are within  $i, j \in \mathcal{I}_1$  and  $i, j \in \mathcal{I}_2$ , assuming that the perturbation of the cells that are in unit psuedotime in the rescaled edges is in the order of  $\sqrt{2L^2} \times 10^{-2} \approx 2.5456 \times 10^{-2}$ , where  $L$  is the average length of the edges within  $i, j \in \mathcal{I}_1$  and  $i, j \in \mathcal{I}_2$ . Considering that the average length of the other combinations of  $(i, j)$  is increased by threefold, we take  $D_{I(i,j)}(x) = 10^{-3}$ .

The proliferation rate is also obtained by the secant method to match the given data in Table A1 at  $t_1$  and  $t_2$ . The computed values are  $r_n = 1.3648, t < t_1$  and  $r_n = 0.4, t \geq t_1$  for the HSPC and progenitor cells  $n \in \mathcal{I}_2 \cup \mathcal{I}_3$ . In addition, the fact that LT-HSC cells proliferate relatively less than the progenitor cells (Passequ et al., 2005) is imposed as  $r_n = r_{n \in \mathcal{I}_2 \cup \mathcal{I}_3} \times 10^{-2}$  for  $n \in \mathcal{I}_1$ . The intermediate level of proliferation is linearly interpolated as

$$R_k(x) = R_{I(i,j)}(x) = r_i + (r_j - r_i)x, \tag{A4}$$

assuming that the overall proliferation of intermediate cell states changes gradually. If the time variable is taken as the actual time, the rate in each node can be computed considering the proportion of proliferating stem cells (5–10%) and cell cycle (36–145 days) (Hao, Chen, & Cheng, 2016;

**Table A1.** Number of cells in each node (ID) at three distinct psuedotimes  $t_0, t_1$  and  $t_2$ .

ID	1	2	3	4	5	6	7	8	9	10	11	12
$t_0$	24	66	155	0	0	0	0	0	0	0	0	0
$t_1$	0	0	0	236	36	27	11	60	0	0	0	0
$t_2$	0	0	0	0	0	0	0	0	192	223	227	54

Notes: We notate it as  $\bar{\rho}_{ID}(t)$ . The cell numbers are plotted in Figure 5C comparing to our simulation.

Pietras, Warr, & Passequé, 2011). Moreover, the abnormal proliferation of cancerous cells with cell cycle  $\lambda$  and apoptosis of the differentiated cells with rate  $d$  at expression level  $x^*$  can be modelled with a localized Gaussian function with variance  $\epsilon$  as  $R_k(x) = (\ln(2)/\lambda) \exp[-(x - x^*)^2/\epsilon]$  and  $R_k(x) = -d \exp[-(x - x^*)^2/\epsilon]$ , respectively. The choice of localized Gaussian function assumes that the centre  $x^*$  is the location in the diffusion space that most closely resembles the ‘prototypical’, or ‘ideal’ cell type identity.

The described parameters are summarized in Table A2.

The initial condition is taken by considering the cell data at pseudotime  $t_0$  with ratios  $\gamma_1^0 = 25/296$ ,  $\gamma_2^0 = 78/296$ ,  $\gamma_3^0 = 193/296$ ,  $\gamma_k^0 = 0$  for  $k = 4, \dots, 12$ . We remark that this is shown in Figure 5C. Accordingly, the initial distribution is taken as

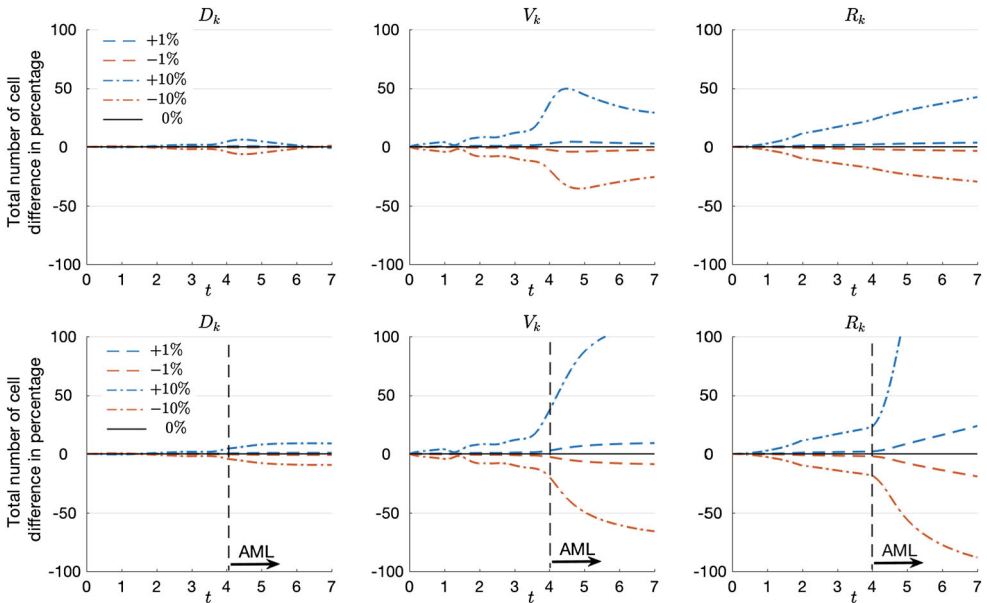
$$u_k(x, t_0) = u_{I(i,j)}(x, t_0) = \gamma_i^0 e^{-x^2/0.08} + \gamma_j^0 e^{-(x-1)^2/0.08}, \quad x \in e_k.$$

With this choice, the total number of cells in each node  $\rho_n(t_0)$  computed as in Equation (A2) is similar to the given ratios  $\gamma_n^0$ . The boundary condition defined as in Equation (3) around the node

**Table A2.** Summary of the required data and corresponding parameters.

Biological meaning and parameters	
$V_k(x)$	Cell differentiation rate $c_k$ , branching ratio $\gamma_k$
$R_k(x)$	Growth rate $r_k$
$D_k(x)$	Phenotypic fluctuation $\sigma_k, W_k$

Note: In our simulation,  $V_k$  and  $R_k$  are estimated from  $\bar{\rho}_k$  in Table A1.



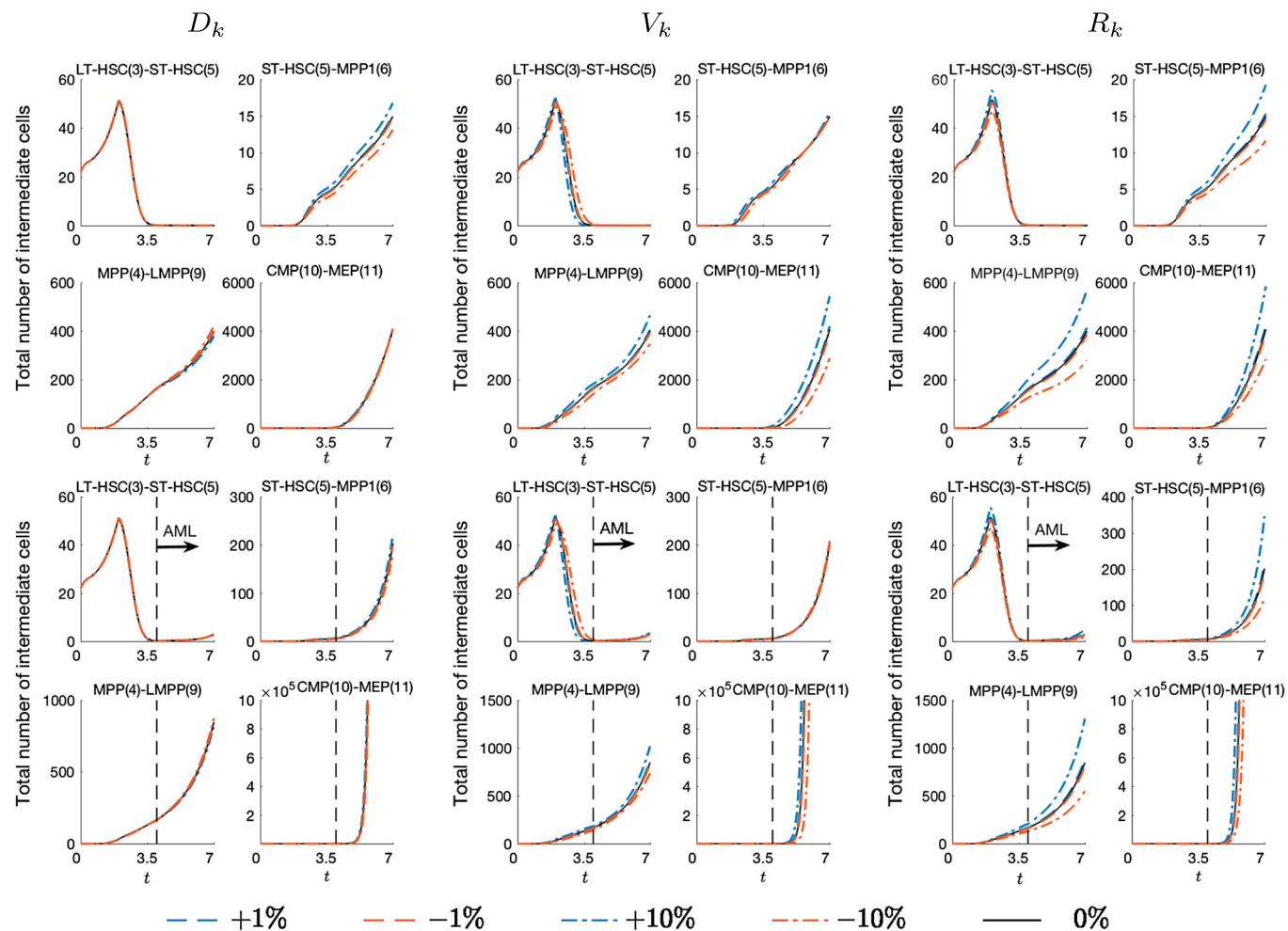
**Figure A11.** Change in the total number of cells  $\rho(t)$  in percentage with respect to the model parameters, diffusion  $D_k$ , advection  $V_k$  and reaction  $R_k$ . We test the cases where the coefficients change their values by  $-10\%$ ,  $-1\%$ ,  $1\%$  and  $10\%$ . The results are sensitive to the reaction and advection coefficients particularly in the AML condition. On the other hand, the results are less dependent on the diffusion coefficient.

$v_n$ , that is, at  $x = b_{I(i,n)}$  and  $x = a_{I(n,j)}$ , becomes

$$\sum_{(i,n) \in \mathcal{J}} \left[ \gamma_i c_n u_{I[i,n]} - D_{I[i,n]} \frac{\partial}{\partial x} u_{I[i,n]} \right] = \sum_{(n,j) \in \mathcal{J}} \left[ \gamma_j c_n u_{I[n,j]} - D_{I[n,j]} \frac{\partial}{\partial x} u_{I[n,j]} \right], \quad (\text{A5})$$

with continuity boundary conditions  $u_{I(n,j)}(a_{I(n,j)}) = u_{I(i,n)}(b_{I(i,n)})$  for fixed  $n$ . Condition (A5) reduces to  $\sum_{(i,n) \in \mathcal{J}} D_{I[i,n]} \frac{\partial}{\partial x} u_{I[i,n]}(x) = \sum_{(n,j) \in \mathcal{J}} D_{I[n,j]} \frac{\partial}{\partial x} u_{I[n,j]}(x)$  in our model since  $\sum_{(i,n) \in \mathcal{J}} \gamma_i c_n = \sum_{(n,j) \in \mathcal{J}} \gamma_j c_n = c_n$ .

*Sensitivity of model parameters.* We test the sensitivity of the results with respect to the parameters in the diffusion, advection and reaction coefficient. The values of  $D_k$ ,  $V_k$  and  $R_k$  are varied by  $-10\%$ ,  $-1\%$ ,  $1\%$  and  $10\%$ , and Figure A11 presents the difference in the total number of cells  $\rho(t)$  in percentage. While it is expected that the total number of cells are sensitive to the reaction coefficient, since it governs the proliferation rate, it also strongly depends on the advection coefficient as well, particularly in the AML condition. On the other hand, the results are less dependent on the diffusion coefficient. The number of intermediate cells while varying the coefficients are plotted in Figure A12. In particular, we present the dynamics of LT-HSC(3)-ST-HSC(5), ST-HSC(5)-MPP1(6), MPP(4)-LMPP(9) and CMP(10)-MEP(11) cells in the normal and AML conditions. We observe similar results as in the total number of cells; however, the overall trend of the dynamics is independent of the variation in the coefficients.



**Figure A12.** Number of intermediate cells with respect to the model parameters, diffusion  $D_k$ , advection  $V_k$  and reaction  $R_k$ . The results are computed by varying the coefficients by  $-10\%$ ,  $-1\%$ ,  $1\%$  and  $10\%$ . Although the result varies from the reference case ( $0\%$ ), the overall trend of the cell dynamics is observed to be similar.