Inferring mechanisms of T cell expansion and contraction in response to acute infection from single-cell data

Jonas Mir

June 2021

TECHNISCHE UNIVERSITÄT MÜNCHEN

Fakultät für Medizin

Inferring mechanisms of T cell expansion and contraction in response to acute infection from single-cell data

Jonas Mir

Vollständiger Abdruck der von der Fakultät für Medizin der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. Percy A. Knolle

Prüfende/-r der Dissertation:

1. TUM Junior Fellow Dr. Michael Floßdorf
2. Prof. Dr. Dr. Fabian J. Theis

Die Dissertation wurde am 15.06.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Medizin am 09.11.2021 angenommen.

Acknowledgements

I deeply thank my supervisor Michael Floßdorf, who has always found the time to support me. His scientific curiosity and enthusiasm have motivated me greatly throughout the years. I particularly appreciated his simultaneous focus on small details as well as on the bigger picture. During the course of my PhD, I have learned a lot from him - also beyond science.

I would like to thank Fabian Theis as the second supervisor of this thesis. His valuable opinions were seminal for my work. Likewise, he has always made the entire expertise of his group and institute available to me. This exchange has had a strong influence on my research.

My thanks go to Dirk Busch and Veit Buchholz, who created the major experimental basis for this collaborative work. They have always placed their full trust in me so that I could develop my personal and professional skills further as a theoretician at a biological-medical institute.

I would like to thank my colleague Lena Toska for the fantastic teamwork during all these years. I cannot imagine more rewarding and effective interdisciplinary work than I have experienced with her. I have personally learned a lot from her direct style and pragmatism.

At this point, I would also like to thank some of my fellow PhD colleagues who have become good friends of mine. These include Lena Appel, Lena Toska, Philipp Angerer, Lisa Bast, Danica Bajic, Tilman Kurz, and Eliza Brett. They have made this stage of my life a lot of fun through many scientific/unscientific conversations, reviving swimming breaks in the Isar river, and countless after-work adventures.

Special thanks go to Johanna Dotterweich. Through her empathetic nature and interest in my situation, she was often able to give me the best advice without knowing my exact working environment.

Finally, I would like to thank my parents and my friends who have always listened to me and supported me even in difficult times. What sounds like a phrase is actually meant word by word.

Abstract

T cell responses to acute infection proceed along a course of distinct phases. Once antigen-inexperienced T cells are activated, they increase exponentially through proliferation, giving rise to large numbers of short-lived cells with distinct effector functions. After clearance of the pathogen, the T cell population substantially diminishes, leaving a small fraction of viable cells. These long-lived memory cells protect the host against re-infection and thus make a central contribution to adaptive immunity. The individual phases of the immune response are defined both by the cellular phenotypic composition and by the respective subset kinetics. Not only is a sound mechanistic understanding of these dynamics essential for improving vaccination strategies and immunotherapies, it also serves as a basis in the study of responses to chronic diseases and cancer. However, the fundamental mechanisms of the immune response at the T cell level are controversial in the literature. Moreover, there is a lack of studies that are quantitative, *in vivo* oriented, and comprehensive at the same time. With this dissertation, I intend to contribute to a better quantitative understanding of the mechanisms underlying T cell memory formation.

First, I address the problem of disentangling the influences of cell division and cell death on the growth kinetics of a proliferating cell population. To this end, I present a novel method that allows quantifying cell cycle speed of rapidly dividing cells *in vivo*. Moreover, the average length of each individual cell cycle phase can be estimated. This approach is based on labelling cells with a nucleoside analogue (NA) and subsequently measuring their NA abundance and DNA content. The method does not depend on the specific shape of the underlying phase length distributions, nor does it require knowledge of the unknown kinetics by which NA labelling efficiency decreases after administration. Using a stochastic simulation, I can validate the method and show that its assumptions are legitimate. Finally, I apply the method to dividing T cell populations subjected to acute infection, which yields results that I further use in the context of mathematical modelling.

In addition, I analyse time-resolved single-cell RNA sequencing (scRNA-seq) data from cytotoxic T cells during acute infection. Utilising a variety of trajectory inference methods, I derive developmental transitions between the identified phenotypes. The resulting differentiation scheme is consistent with the progressive model of T cell differentiation, in which long-lived memory-like cells become short-lived effector cells during infection-induced development. Furthermore, I find that the cessation of cell division, i.e. cells dropping out of cell cycle and entering a quiescent state, represents a major source of heterogeneity during the immune response. Based on the inferred model topologies, I fit stochastic compartment models to a variety of *in vivo* data at the single-cell level. These include single-cell fate mapping data, cell cycle dropout kinetics inferred from scRNA-seq data,

and phenotype-specific quantifications of division speed. The resultant *cell cycle dropout model*, which explicitly accounts for a division stop of the cells, can consistently explain the data, with all model parameters being identifiable. Even striking properties of the data, such as a high Gini coefficient of the clone size distribution, are correctly reproduced by stochastic simulations. Through model selection, I can show that the mechanistically simplest model, in which all subsets begin to turn quiescent at the same time and rate, best matches the data. Nevertheless, the model topology and the set of optimal parameters indirectly lead to the modelled subsets being in division to different extents. This model prediction has been confirmed in a subsequent experiment.

Lastly, I extend the cell cycle dropout model beyond the acute phase of infection. For this purpose, I employ a more complex model including cell death and division speed dynamics. Utilising data at later time points after infection, I show that the contraction of T cells is predominantly characterised by differences in proliferation and death rates between memory-like cells and effector cells. Interestingly, the time-dependencies of these rates seem to be very similar across the subsets.

In summary, I combine a variety of *in vivo* data at the single-cell level to infer mechanisms of T cell responses to acute infection, up to the memory phase. Additionally, I present a novel method to quantify cell cycle speed *in vivo*.

Zusammenfassung

T-Zell-Antworten auf akute Infektionen verlaufen in unterschiedlichen Phasen. Sobald Antigen-unerfahrene T-Zellen aktiviert werden, vermehren sich diese exponentiell durch Proliferation, wodurch eine große Anzahl kurzlebiger Zellen mit ausgeprägten Effektor-Funktionen entsteht. Nach der Eliminierung des Pathogens nimmt die Größe der T-Zell-Population stark ab, wobei ein geringer Anteil der Zellen überlebt. Diese langlebigen Gedächtniszellen schützen den Wirt vor erneuter Infektion und leisten somit einen zentralen Beitrag zur adaptiven Immunität. Die einzelnen Phasen der Immunantwort werden sowohl durch die zelluläre phänotypische Zusammensetzung als auch durch die jeweilige Kompartiment-Kinetik definiert. Ein fundiertes mechanistisches Verständnis dieser Dynamik ist nicht nur für die Verbesserung von Impfstrategien und Immuntherapien essentiell, sondern dient auch als Grundlage der Erforschung von Immunantworten auf chronische Krankheiten und Krebs. Die grundlegenden Mechanismen der Immunantwort auf T-Zell-Ebene werden in der Literatur jedoch kontrovers diskutiert. Zudem mangelt es an Studien, die gleichzeitig quantitativ, *in vivo*-orientiert und umfassend sind. Mit dieser Dissertation möchte ich zu einem tieferen quantitativen Verständnis derjenigen Mechanismen beitragen, die der T-Zell-Gedächtnisbildung zu Grunde liegen.

Zunächst gehe ich auf die Problematik ein, Einflüsse von Zellteilung und Zelltod auf die Wachstumskinetik einer proliferierenden Zellpopulation voneinander zu trennen. Hierzu stelle ich eine neue Methode vor, die es ermöglicht, Zellzyklusgeschwindigkeiten sich schnell teilender Zellen *in vivo* zu quantifizieren. Zudem erlaubt es die Methode, die durchschnittliche Dauer jeder einzelnen Zellzyklusphase abzuschätzen. Dieser Ansatz basiert auf der Markierung von Zellen mit einem Nukleosid-Analogon (NA) und der darauffolgenden Messung der NA-Menge und des DNA-Gehalts der Zellen. Die Methode hängt weder von der konkreten Form der zugrundeliegenden Verteilung der Phasenlängen ab noch erfordert sie Kenntnisse über die unbekannte Kinetik, mit der die NA-Markierungseffizienz nach Verabreichung abnimmt. Mithilfe einer stochastischen Simulation kann ich die Methode validieren und die getroffenen Annahmen rechtfertigen. Schließlich wende ich die Methode auf sich teilende T-Zell-Populationen während akuter Infektion an. Die so gewonnenen Ergebnisse berücksichtige ich später im Rahmen von mathematischer Modellierung weiter.

Außerdem analysiere ich zeitaufgelöste Einzelzell-RNA-Sequenzierungsdaten (scRNA-seq-Daten) zytotoxischer T-Zellen während akuter Infektion. Unter Verwendung diverser Methoden der Trajektorien-Inferenz leite ich Differenzierungsübergänge zwischen den identifizierten Phänotypen ab. Das resultierende Differenzierungsschema ist konsistent mit dem progressiven Modell der T-Zell-Differenzierung, in dem während infektionsinduzierter Entwicklung langlebige gedächtnisähnliche Zellen zu kurzlebigen Effektorzellen werden.

Darüber hinaus stelle ich fest, dass die Beendigung der Zellteilung, d.h. das Austreten der Zellen aus dem Zellzyklus und der Übergang in einen ruhenden Zustand, eine wichtige Quelle der Heterogenität während der Immunantwort darstellt. Basierend auf den abgeleiteten Modelltopologien passe ich stochastische Kompartmentmodelle an eine Vielzahl von *in vivo*-Daten auf Einzelzellebene an. Dazu gehören Einzelzell-Fate-Mapping-Daten, die aus scRNA-seq-Daten abgeleitete Zellzyklus-Austritts-Kinetik und phänotypspezifische Quantifizierungen der Teilungsgeschwindigkeit. Das resultierende *Zellzyklus-Austritts-Modell*, das den Teilungsstopp der Zellen explizit berücksichtigt, kann die vorliegenden Daten auf konsistente Weise erklären. Hierbei sind alle Modellparameter identifizierbar. Selbst auffällige Eigenschaften der Daten, wie z.B. ein hoher Gini-Koeffizient der Klon-Größenverteilung, werden durch stochastische Simulationen korrekt reproduziert. Durch Modellselektion kann ich zeigen, dass das mechanistisch einfachste Modell, in dem alle Kompartimente zur gleichen Zeit und mit der gleichen Rate beginnen, in den ruhenden Zustand einzutreten, die höchste Übereinstimmung mit den Daten erzielt. Dennoch führen die Kombination aus Modelltopologie und dem Satz optimaler Parameter indirekt dazu, dass sich die modellierten Kompartimente in unterschiedlichem Ausmaß in Teilung befinden. Diese Modellvorhersage konnte in einem nachfolgenden Experiment bestätigt werden.

Schließlich erweitere ich das Zellzyklus-Austritts-Modell über die akute Phase der Infektion hinaus. Zu diesem Zweck verwende ich ein komplexeres Modell, welches den Zelltod und dynamische Teilungsgeschwindigkeiten der Zellen berücksichtigt. Mithilfe von Daten zu späteren Zeitpunkten nach der Infektion zeige ich, dass die Kontraktion der T-Zellen vor allem durch Unterschiede in den Proliferations- und Sterberaten zwischen gedächtnisähnlichen Zellen und Effektorzellen geprägt wird. Interessanterweise scheinen die Zeitabhängigkeiten dieser Raten zwischen den betrachteten Phänotypen sehr ähnlich zu sein.

Insgesamt kombiniere ich eine Vielzahl von *in vivo*-Daten auf Einzelzell-Ebene, um die Mechanismen von T-Zell-Antworten während akuter Infektion bis hin zur Gedächtnisphase abzuleiten. Zusätzlich stelle ich eine neue Methode zur Quantifizierung der *in vivo*-Zellzyklusgeschwindigkeit vor.

Contents

List of abbreviations

AIC_c	corrected Akaike information criterion
APC	professional antigen presenting cell
Att. *L.m.*-OVA	attenuated and *actA*-deficient *Listeria monocytogenes*-OVA
BrdU	5-bromodeoxyuridin
CI	confidence interval
CM	memory-associated cell
CTCMp	cycling central memory precursor T cell
CTEF	cycling effector T cell
CTEMp	cycling effector-memory precursor T cell
CTV	CellTrace^{TM} Violet
CV	coefficient of variation
EdU	5-ethynyl-2'-deoxyuridine
EFF	effector-associated T cell
FACS	fluorescence-activated cell sorting
TCM	central memory T cell
TCMp	central memory precursor T cell
TCR	T cell receptor
TEF	effector T cell
TEMp	effector-memory precursor T cell
LCMV Arm	lymphocytic choriomeningitis Armstrong
***L.m.*-OVA**	*Listeria monocytogenes*-OVA
MHC	major histocompatibility complex
N	naïve T cell
NA	nucleoside analogue
ODE	ordinary differential equation
OVA	ovalbumin
p.i.	post infection
phospho-Rb	phosphorylated retinoblastoma protein
scRNA-seq	single-cell RNA sequencing

Chapter 1

Introduction

In the last decade, immunotherapies with a direct or indirect focus on T cells were widely successful in the treatment of cancer and chronic diseases. Of particular note is the use of checkpoint inhibitors and adoptive T cell therapies. These sophisticated approaches all have their origin in the improved comprehension of the immune system and its interacting components. To better understand T cell responses in the context of chronic infection or cancer, comparisons with responses to acute infection can be very beneficial. This may allow phenotypic characterisation of early abnormalities throughout the response and reveal their mechanistic causes. Also, the improvement of vaccination strategies requires precise knowledge of the course of T cell responses to acute infection, which typically lead to the establishment of immunological memory. The progression of T cell responses occurs in distinct phases that differ both in kinetics and cellular phenotypic composition. This thesis quantitatively addresses the mechanisms of cytotoxic T cell responses following acute infection up to the memory phase.

The role of $CD8^+$ T cells in adaptive immunity

The immune system can be broadly divided into two components that interact with each other: the innate immune system and the adaptive immune system. The innate immune system typically provides the first barrier to pathogens. Once pathogens have penetrated the epithelial layers, phagocytic cells can recognise and engulf them. Depending on the infiltrated organ, macrophages, neutrophils, dendritic cells, or monocytes, for example, adopt this function; this is often accompanied by an inflammatory response in which cytokines and chemokines are secreted, see [Owen et al., 2013]. Such pathogen-unspecific responses are a prerequisite for triggering the adaptive immune response, which is initiated in secondary lymphoid organs such as lymph nodes and spleen.

Pathogenic antigens are processed by professional antigen-presenting cells (APCs) and presented to the adaptive immune system via the major histocompatibility complex (MHC) class I/II molecules. Dendritic cells in particular play a central role here. The B and T lymphocytes of the adaptive immune system have a highly specific receptor that recognises foreign epitopes. An epitope represents the part of the presented antigen that potentially

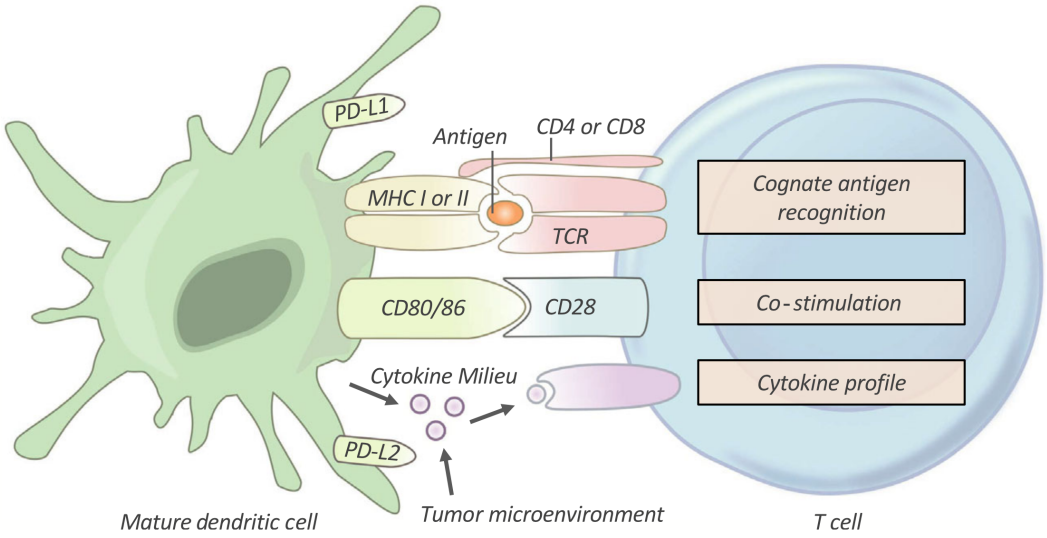

Figure 1.1: Signals necessary for the activation of naïve T cells. While co-stimulatory and cytokine-induced signals are non-specific, stimulation via the TCR is highly antigen-specific. Here, TCR and co-receptor are stimulated by professional antigen-presenting cells, e.g. dendritic cells. (Figure adapted from [Neeve et al., 2019], modified)

binds to the specific receptor of the immune cell. $CD4^+$ T helper cells are addressed via the MHC II presentation pathway. In the case of $CD8^+$ cytotoxic T cells, the presentation of epitopes occurs via MHC I. Analogously to pathogenic antigen presentation of APCs, the majority of cells in the body provide the immune system with a representation of their cell-internal peptides via MHC I. Therefore, the MHC I pathway is particularly relevant for controlling infections caused by viruses or intracellular bacteria. When a naïve, i.e. antigen-inexperienced, $CD8^+$ T cell encounters an antigen-presenting cell with cognate peptide-MHC I complex, an immune response can be induced. However, a potent response requires further signals, see Fig. 1.1 from [Neeve et al., 2019]. If the co-stimulatory signal via the CD28 co-receptor of the T cell is missing, this leads to apoptosis of the T cell, i.e. programmed cell death; if, on the other hand, the co-receptor of the T cell is stimulated without proper TCR stimulus, there is no response at all, see [Murphy et al., 2017]. For potent activation of the cell, a cytokine signal must be mediated in addition to the stimulation of TCR and co-receptor, see [Curtsinger et al., 1999; Neeve et al., 2019].

In mice, there are about 100 to 1000 naïve T cells specific for a unique epitope, see [Obar et al., 2008]. Generally, a single antigen has several epitopes. Thus, an infection leads to a polyclonal T cell response. According to the current opinion in the field, almost all specific $CD8^+$ T cells are recruited into the immune response, cf. [Van Heijst et al., 2009].

A typical acute infection consists of three phases, see [Khan et al., 2015]. During the acute phase, T cells divide exponentially and eliminate the pathogen. The peak of the $CD8^+$ T cell immune response is infection-dependent, but usually lies around 6 d to 9 d after infection, see [De Boer et al., 2003; Porter et al., 2006; Schlub et al., 2009]. However, it is not clear to what extent the immune response of $CD8^+$ T cells is programmed (by cell-intrinsic or -extrinsic factors) after initiation, or to what extent the response is dynamically regulated by environmental signals, such as antigen concentration, see [Badovinac et al., 2002; Buchholz et al., 2016; Khan et al., 2015]. Next, the contraction phase of the response is initiated. At this point, approximately 90 % to 95 % of all cells die, cf. [De Boer et al., 2001; Harty et al., 2008]. The surviving minority of the cells consists of memory cells, whose numbers are kept stable in the memory phase by homeostatic proliferation. These cells form an important part of the immunological memory and can initiate a rapid and efficient immune response upon re-challenge with the same pathogen.

These three different phases of the immune response are characterised by different kinetics in terms of proliferation and cell death. They are also distinguished by different cellular compositions. For instance, the peak of the immune response is mainly dominated by short-lived effector T cells (TEFs), which have an increased killing capacity of infected cells. During the memory phase, long-lived central memory T cells (TCMs), effector memory T cells (TEMs), and tissue-resident memory T cells (TRMs) are mainly present. The TCM cells in particular are characterised by the fact that they circulate between blood and lymphatic system during the memory phase, see [Khan et al., 2015]. This circulation is regulated, among other factors, by the surface protein CD62L, which is at the same time an important memory marker. Overall, it is unclear when exactly the individual phenotypes arise and what their lineage relations are. Cells that are already memory-like during the acute phase could thus be precursors of the later TCM and TEM cells and are notationally referred to as TCMp and TEMp cells, respectively.

Mechanistic inference of T cell kinetics and diversification

From the above considerations, one fundamental question arises: what mechanisms drive response heterogeneity at the single-cell level and how exactly do antigen-inexperienced cells differentiate into the observed (highly specialised) phenotypes? In answering this question, one is dealing with highly complex, dynamic *in vivo* processes, so data has to be retrieved in the form of snapshot data, depicting the phenotypic T cell composition in selected organs at different time points. In order to interpret these snapshot data jointly under a particular set of biological hypotheses, data-driven mathematical modelling has proven to be a valuable tool in immunology, see [Andrew et al., 2007]. In particular,

deterministic compartment models are by far the most popular ones to date, as they are based on simple ordinary differential equations (ODEs), on the one hand, and establish a direct link to already existing biological concepts, such as immunological phenotypes, on the other hand, see [Höfer et al., 2019]. In this sense, the models benefit from experimental advances in immunostaining and fluorescence-activated cell sorting (FACS). In the case of cell type compartments, the kinetics of the system can be captured in terms of cellular programs such as proliferation, differentiation, and cell death. In the simplest case, these processes are described in terms of constant rate processes, leading to linear equations.

Models of this type have been used, for example, by R. J. de Boer and colleagues to describe population data of $CD8^+$ T cells during acute infection up to the memory phase, see [De Boer et al., 2003, 2001]. Here, the authors assumed that naïve cells initially enter a highly proliferative activated state upon recruitment. From this state, they can then differentiate into the memory compartment. After the acute phase has ended, these memory cells can then be reactivated by transitioning into the activated compartment, similar to the naïve cells before. Interestingly, it could be shown at the single-cell level in [Graef et al., 2014] that single TCM cells indeed showed a similar response pattern as naïve cells did (see below). However, R. J. de Boer and colleagues make no further statements about the phenotypic composition of the activated compartment, while identifying important kinetic parameters of the entire $CD8^+$ T cell immune response.

P. D. Hodgkin and colleagues introduced another class of models by means of the *cyton model*, see [Hawkins et al., 2007; Subramanian et al., 2008]. The primary aim was to explain the kinetics of dividing T and B lymphocyte populations and the associated heterogeneity. Formally, the cyton model can be expressed via coupled generation-structured integral equations. Alternatively, the equations can be reformulated as partial differential equations (PDEs), see [De Boer et al., 2013]. Often, skewed distributions are chosen for division- and survival-time, e.g. log-normal distributions. The first cell division is considered separately here, as there is *in vitro* evidence for the first cell division taking longer than all subsequent divisions, see [Deenick et al., 2003; Dowling et al., 2014; Hawkins et al., 2009]. The equations formulated in this way can then be solved numerically. Additionally, in the cyton model, each cell is assigned a time until its next division and a time until death. These times are assigned independently and stored in the cell's *cyton*. This unit thus decides the fate of the cell, depending on which of the two times is smaller, and at the same time cancels the occurrence of the other process, i.e. censors it. The authors fit their models mainly to a combination of population data and kinetic data concerning cell division and cell death. The cyton model is considered in combination with the concept of *division destiny*, also developed by P. D. Hodgkin, see [Hawkins et al., 2007; Turner et al.,

2008]. In its original formulation, the division destiny represents the maximum number of divisions a cell will undergo until it stops dividing. This "destiny" is programmed early after activation and follows a distribution shaped by the cells under consideration and their particular environmental conditions. Here, the sum of all stimuli a cell receives seems to shape this distribution, see [Marchingo et al., 2014]. More recent studies by the same group suggest that division destiny is likely to be a distributed time that a cell can be in division rather than a sheer number of divisions, see [Heinzel et al., 2017]. In the same study, the authors confirm a similar mechanism for regulating cell death and further show that this mechanism is independent of the timer programming division destiny.

The exact differentiation hierarchies during primary infection are discussed controversially in the literature. The most widely accepted model in the past has been the *linear model* of differentiation, see [Opferman et al., 1999; Wherry et al., 2003]. Here, naïve cells directly develop effector properties upon recruitment. During contraction, most of these cells die, with a few surviving to mature into central memory cells. More recent modelling studies inferring such a differentiation pathway based on population data and deterministic compartment models can be found, for example, in [Crauste et al., 2017; Terry et al., 2012]. Likewise, the *decreasing potential model* should be mentioned at this point, suggested by R. Ahmed and colleagues, see [Crotty et al., 2004], where recruited naïve cells also immediately acquire effector properties after activation. These then continue to mature under constant exposure to environmental stimuli. During this process, effector cells can differentiate into memory-like cells. The potency of the resulting memory cells in terms of long-lasting protective capacity strongly depends on the developmental stage of the corresponding effector cell, cf. [Kalia et al., 2010]. The effector cells thus lose their potential to generate potent memory cells in the course of their maturation. Terminal effectors have finally lost this potential completely, while the authors claim that this very last state of effector differentiation is related to higher antigen stimuli.

An important advance in the analysis of T cell differentiation is the experimental transition to the single-cell level. By labelling cells with congenic markers or barcodes, it was possible to trace the fate of individual naïve cells after infection *in vivo*, see [Buchholz et al., 2013; Gerlach et al., 2013; Stemberger et al., 2007]. Here, *single-cell fate mapping* data showed that despite overall T cell responses being robust and reproducible, individual responses can be very heterogeneous in terms of clone size and phenotypic composition, cf. [Buchholz et al., 2016]. These data thus directly rule out extreme formulations of a *bifurcation model* in which whole clones are assumed to consist exclusively of one phenotype. Variants of the bifurcation model, especially with a first asymmetric cell division, see e.g. [Chang et al., 2007; King et al., 2012], are in principle compatible with these data. In [Gerlach et al.,

2013], however, the authors exclude asymmetric cell division as the most important source of heterogeneity in their data.

Contrary to the linear model, there is evidence for the reverse differentiation order of the cells. In such a *progressive model*, TCMp cells downregulate CD62L early to enter the TEMp or TEF compartment, cf. [Gattinoni et al., 2017]. In [Schlub et al., 2010, 2009], a division-linked differentiation concept was proposed, in which these diversifications are induced by greater proliferation. Using single-cell fate mapping data together with stochastic compartment modelling, the authors in [Buchholz et al., 2013] were able to show through model comparisons that only a progressive model could be reconciled with the stochasticity of their data. Here, cells differentiate from naïve $\rightarrow$ TCMp $\rightarrow$ TEMp $\rightarrow$ TEF. Along the developmental pathway, division speed increases: according to the authors, a higher proliferation rate results precisely from differentiation, which is conceptually very different from the concept put forward in [Schlub et al., 2009]. The strength of the approach taken in [Buchholz et al., 2013] is that the high informative single-cell progeny data leads to a reliable identification of the model parameters by stochastic compartment modelling and allows to confidently reject the vast majority of the tested topologies. First, it needs to be highlighted that such data, even if only available for one time point, always contain information about the underlying dynamics. This is because the initial condition is known exactly. Furthermore, higher moments of the cell number distribution, such as covariances, can be added to inform the model.

In a meta-study from [Miles et al., 2019], the authors examined several published data sets, including data from [Buchholz et al., 2013], to see if they were more likely to be described by a progressive or a linear model. They used deterministic compartment models and found that progressive models systematically outperformed linear models. Nevertheless, some simplifying assumptions were made in the original study from [Buchholz et al., 2013] that should be viewed critically. First, exponential waiting times were implicitly assigned to all processes. In contrast, it could be shown in a large number of *in vitro* studies that cell divisions can be well described by log-normal distributions. As pointed out above, the first cell division takes much longer than the subsequent ones, cf. [De Boer et al., 2005; Deenick et al., 2003; Lee et al., 2008]. Especially for the first division, the variability, described by a Markov process, could be strongly overestimated. Furthermore, rates for differentiation and proliferation were assumed to be time-independent, i.e. the rates were averaged over time. Cell death was disregarded for simplicity. Also, the exit of cells from the cell cycle was not taken into account, i.e. the division rates were averaged over dividing and quiescent cells. Furthermore, all considerations were restricted to the expansion phase of the acute infection.

A completely different way of looking at T cell diversification with single-cell resolution is given by single-cell RNA sequencing (scRNA-seq) technologies. Not only because of the experimental advances, but also because of the simultaneously evolving analysis tools, this approach has proven to be enormously important in answering questions of cellular fate decisions, cf. [Angerer et al., 2017; Griffiths et al., 2018; Luecken et al., 2019]. Here, for example, developmental pathways can be revealed *in vivo* by trajectory inference methods. The continuous nature of cell states is highlighted and phenotypic marker combinations can be challenged, confirmed, or refined. For instance, in [Grassmann et al., 2020] we were able to identify the progressive differentiation of $CD8^+$ T cells during chronic infection. Other examples of application in the T cell context can be found, for example, in [Lönnberg et al., 2018; Pace et al., 2018]. Furthermore, some approaches jointly evaluate transcriptome and population data: in [Fischer et al., 2019], the authors developed a probabilistic framework that, starting from discrete compartments, allows an estimation of the underlying parameter dynamics.

The need to disentangle cell division and death

In the compartment models described above, mere absolute population data are not sufficient to reliably allow simultaneous estimation of proliferation and death rates. At best, only the difference of these rates for each compartment is determined by the cell number data in terms of an *effective* growth rate. Although cell death does not appear to play a major role in the onset of the immune response, there is a time interval in which cell division and cell death possibly overlap, see [Garrod et al., 2012]. To mechanistically describe the course of the immune response beyond the exponential phase, it is necessary to disentangle these two mechanisms from each other using appropriate data. As it seems sufficient to restrict the focus to only one of these processes, we have, in this work, chosen to focus on cell division. For data-driven quantifications of cell death, we refer to [Ganusov et al., 2013; Kranich et al., 2020; Owen et al., 2013]. During cell division, mother cells progress through the phases $G_1 \to S \to G_2 \to M$. During S-phase, doubling of DNA content from diploid ($2N$) to tetraploid ($4N$) takes place. G_1 and G_2 represent *gap*-phases, while mitosis occurs during M-phase. G_2- and M-phase are often summarised as G_2M-phase. After having completed the cell cycle, daughter cells can either divide further, i.e. re-enter G_1-phase, or exit the cell cycle and become quiescent (G_0-phase).

An important experimental approach to measuring cell cycle speed is provided by labelling cells with a nucleoside analogue, e.g. 5-bromodeoxyuridine (BrdU). BrdU is a thymidine analogue that is incorporated into the DNA of dividing cells during their S-phase. The proportion of cells labelled with BrdU can be determined by subsequent FACS measure-

ment. For *in vivo* studies, BrdU is either injected as a pulse or administered orally. The exact degree of its toxicity and the possible influence on the analysis is controversial, see [Ganusov et al., 2013] for a review of the literature on this topic. Within the framework of a simple compartment model considering labelled and unlabelled cells, the labelling phase dynamics of the proportion of $BrdU^+$ cells only depends on labelling time and proliferation rate, but not on the death rate. Thus, division speed can be inferred directly, cf. [Ganusov et al., 2013]. In our case of rapidly dividing cells during acute infection, BrdU is not sufficient for quantitative analyses: Due to the possible time-dependence of division speed, we cannot afford labelling over long periods, which is why BrdU must necessarily be injected. After injection, the temporal-spatial labelling efficiency will not constantly be given by 100 %. In case of short labelling times, BrdU mainly labels those cells that are in S-phase at the time of administration. However, this measure is by no means informative about division speed of the cells. Here, it is mostly those cells that enter S-phase during the availability of BrdU that matter. Since the above-mentioned labelling phase model implicitly neglects the temporal extension of the S-phase, one would systematically overestimate division speed. Dual-pulse chase experiments, in which another label, e.g. 5-Ethynyl-2'-deoxyuridine (EdU), is administered shortly before the time of analysis, attempt to capture cells that newly enter S-phase, see [Gitlin et al., 2015]. We conclude that mere BrdU measurements are insufficient for our particular objectives. However, for long labelling periods and when looking at steady state systems (e.g. memory phase of the immune response or hematopoiesis), BrdU is certainly a good approach to quantify turnover rates.

Another common method to measure cell cycle speed *in vivo* is provided by CellTrace^{TM} Violet (CTV), or analogous intracellular fluorescent dyes. These are added to the cells *in vitro* before transfer into the organism and are based on the amount of the substance being approximately halved upon cell division. This results in typical profiles after subsequent FACS analysis, from which the generation-dependence of the cells can be inferred. Using these data to fit simple generation-structured compartment models, one obtains the proliferation rate of the cells, see [De Boer et al., 2005; Heinzel et al., 2017]. However, such an approach is unsuitable for our purpose: After a maximum of seven to eight divisions the dye gets diluted and thus can no longer be distinguished from the background. During acute infection, however, cells can divide more than ~ 15 times until day 8 post infection (p.i.), see chapter 5. For a comprehensive overview of other experimental methods for measuring division speed and their corresponding mathematical analysis, we refer to [De Boer et al., 2013].

In addition to informative data, the quantification of division speed also requires several assumptions about the cell cycle. Such a model of the cell cycle can be defined, for example, by assumptions regarding the distribution of the individual phase lengths and their statistical dependencies. Probably the oldest and best-known model is the Smith-Martin model, cf. [Smith et al., 1973]. This model divides the cell cycle into two phases A and B, with the A-phase mainly being associated with the G_1-phase and the B-phase predominantly with the S- and G_2M-phase. J. A. Smith and L. Martin found that a model in which the A-phase carries the full stochasticity in the form of an exponential waiting time distribution with the B-phase being constant best described their data. In [De Boer et al., 2005], the authors reformulated this model in terms of a system of delay differential equations. However, recent *in vitro* measurements show that the inter-division time distribution (as well as the time-to-death distribution) of proliferating lymphocytes can be well described by log-normal distributions, see [Deenick et al., 2003; Dowling et al., 2014; Hawkins et al., 2009]. Therefore, there are several models with more realistic assumptions, which are mostly limited to the *in vitro* case, see e.g. [Dowling et al., 2014; Weber et al., 2014]. Complicating factors in building consistent (*in vivo*) models include consideration of quiescence, knowledge of the exact distribution of phase lengths and their statistical dependence, cell death, the unknown pharmacokinetics of nucleoside analogues used, and variability in the underlying phase distributions.

Added value and overview of this thesis

Most of the experimental foundation for this thesis was established by my colleague Albulena Toska and the laboratory of Veit Buchholz. The close collaboration of theory and experiment was central to the findings presented in chapters 3, 4, and 5. Fig. 1.2 shows a schematic order of the results presented in this thesis. In chapter 6, all results will be discussed in detail. For further explanation of the utilised methods, we refer to chapter 2.

In chapter 3 of this thesis, we begin by presenting a newly developed method that allows quantifying cell cycle speed of rapidly dividing cells *in vivo*. We apply the method to dividing $CD8^+$ T cell populations during various acute infections. The obtained results are then further used during mathematical modelling in chapters 4 and 5 to ensure identifiability of all kinetic model parameters. The formalism offers an important contribution to quantitative immunological research and has already been published in [Kretschmer et al., 2020].

Chapter 4 first focuses on the analysis of time-resolved scRNA-seq data from $CD8^+$ T cells during acute infection. Using a variety of trajectory inference methods, it can be

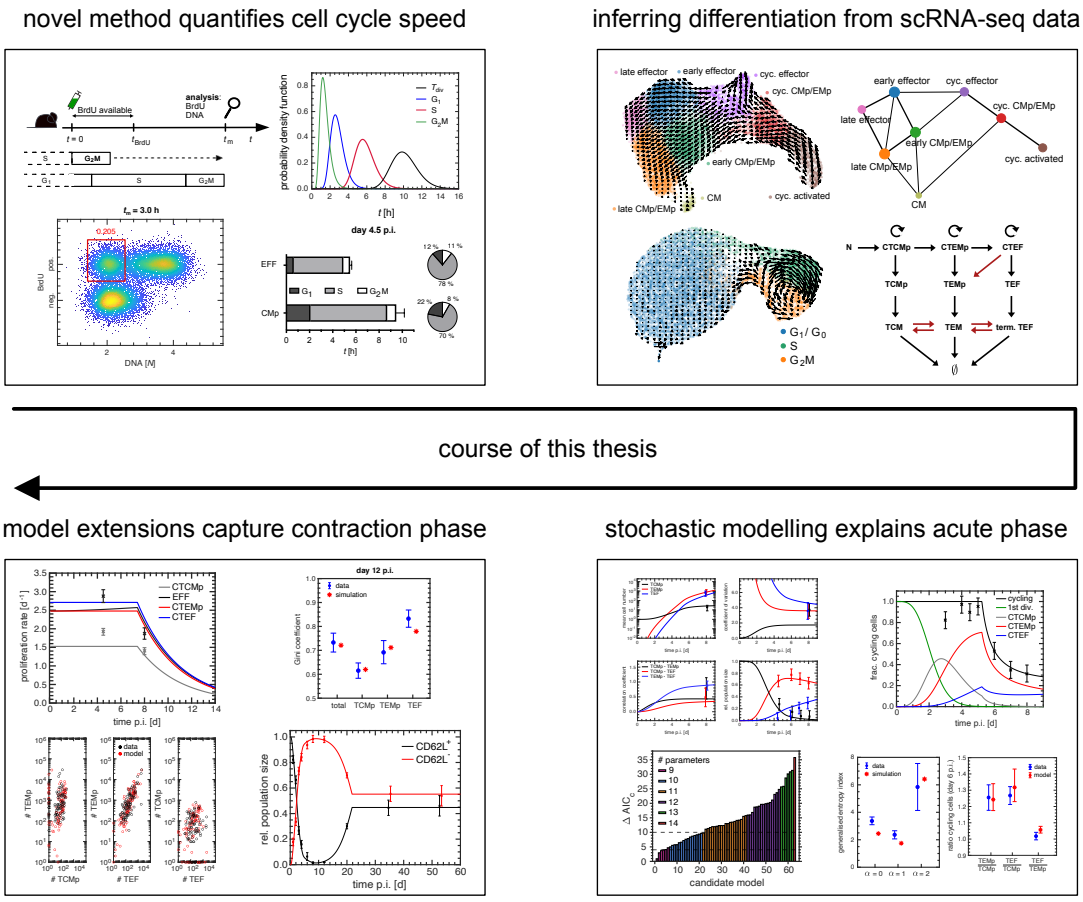

Figure 1.2: Overview of this thesis. First, a novel cell cycle quantification method is presented (chapter 3) which results in cell cycle estimates that are used further on. The analysis of scRNA-seq data provides possible model topologies of $CD8^+$ T cell differentiation during acute infection (chapter 4, first part). Mathematical models based on the previously derived topology are fitted to a variety of single-cell data to mechanistically explain the acute phase of infection (chapter 4, second part). The best candidate model gets extended to also capture the contraction phase of the immune response (chapter 5).

shown that cells differentiate mainly progressively during division. Moreover, the exit of cells from the cell cycle can be identified as a major source of heterogeneity. Second, stochastic compartment models with topologies consistent with the scRNA-seq data are fitted to a variety of population data and kinetic data from the expansion phase of the response. We provide a mechanistic understanding of how the simplest variant of the *cell cycle dropout model* explains the observed heterogeneity. Furthermore, we make a model prediction about the phenotypic cell cycle dropout dynamics, which can be confirmed in a subsequent experiment.

In chapter 5, we extend the cell cycle dropout model to the contraction phase of the response by considering cell death and division speed dynamics. Using additional data at later time points, T cell contraction shows to be mainly regulated by kinetic differences across the subsets. Overall, by combining informative *in vivo* data at the single-cell level, we can infer mechanisms underlying $CD8^+$ T cell responses to acute infection, up to the memory phase.

List of contributed articles

The author of this dissertation has participated in the preparation and publication of several articles, resulting in the following co-authorships:

- Kretschmer*, L., Flossdorf*, M., **Mir, J.**, Cho, Y., Plambeck, M., Treise, I., Toska, A., Heinzel, S., Schiemann, M., Busch, D. H., Buchholz, V. R. (2020). "Differential expansion of T central memory precursor and effector subsets is regulated by division speed". *Nature Communications* **11**.1.

- Schober, K., Fuchs, P., **Mir, J.**, Hammel, M., Fanchi, L., Flossdorf, M., Busch, D. H. (2020). "The CMV-specific CD8+ T cell response is dominated by supra-public clonotypes with high generation probabilities". *Pathogens* **9**.8.

- Grassmann*, S., Mihatsch*, L., **Mir, J.**, Kazeroonian, A., Rahimi, R., Flommersfeld, S., Schober, K., Hensel, I., Leube, J., Pachmayr, L. O., Kretschmer, L., Zhang, Q., Jolly, A., Chaudhry, M. Z., Schiemann, M., Cicin-Sain, L., Höfer, T., Busch, D. H., Flossdorf, M., Buchholz, V. R. (2020). "Early emergence of T central memory precursors programs clonal dominance during chronic viral infection". *Nature Immunology* **21**.12, pp. 1563–1573.

- Bortoluzzi, S., Dashtsoodol, N., Engleitner, T., Drees, C., Helmrath, S., **Mir, J.**, Toska, A., Flossdorf, M., Öllinger, R., Solovey, M., Colomé-Tatché, M., Kalfaoglu, B., Ono, M., Buch, T., Ammon, T., Rad, R., Schmidt-Supprian, M. (2021). "Brief homogeneous TCR signals instruct common iNKT progenitors whose effector diversification is characterized by subsequent cytokine signaling". *Immunity* **54**.11, pp. 2497-2513.e9.

- Falcomatà*, C., Bärthel*, S., Widholz, S. A., Schneeweis, C., Montero, J. J., Toska, A., **Mir, J.**, Kaltenbacher, T., Heetmeyer, J., Swietlik, J. J., Cheng, J. Y., Teodorescu, B., Reichert, O., Schmitt, C., Grabichler, K., Coluccio, A., Boniolo, F., Veltkamp, C., Zukowska, M., Vargas, A. A., Paik, W. H., Jesinghaus, M., Steiger, K., Maresch, R., Öllinger, R., Ammon, T., Baranov, O., Robles, M. S., Rechenberger, J., Kuster, B., Meissner, F., Reichert, M., Flossdorf, M., Rad, R., Schmidt-Supprian, M., Schneider, G., Saur, D. (2022). "Selective multi-kinase inhibition sensitizes mesenchymal pancreatic cancer to immune checkpoint blockade by remodeling the tumor microenvironment". *Nature Cancer*, DOI: 10.1038/s43018-021-00326-1.

(* indicates equal contribution)

Chapter 2

Methods

This chapter provides the basic methods which will be applied in chapters 4 and 5. These methods can be broadly divided into two parts: first, methods for data-driven mathematical modelling, and second, methods for the analysis of scRNA-seq data. We begin by outlining the framework of the mathematical modelling approach adopted in this work.

2.1 Data-driven mathematical modelling

We explain the mathematical formalism used in this thesis to capture the dynamics of stochastic moments of (cell number) distributions, and how model parameters can be estimated with the inclusion of informative data. The identifiability of the resulting parameters is addressed, as well as the comparison between different candidate models.

2.1.1 Stochastic modelling framework

In order to summarise an important part of our population data in chapters 4 and 5 in terms of their stochastic moments, we need equations that capture their time evolution. In the following, we will outline the derivation of these equations, based on the approach and notation that have been put forward in [Lestas et al., 2008]. The central starting point for this approach is the master equation, which is given by

$$\frac{\mathrm{d}P(k,t)}{\mathrm{d}t} = \sum_i \left[W_i(k-r_i)P(k-r_i,t) - W_i(k)P(k,t)\right] \tag{2.1}$$

where $P(k,t)$ denotes the probability of the n-dimensional state vector $x(t)$ being realised by the concrete state k at time t, subject to the initial condition $x(t_0) =: k_0$. In our case, the state vector is represented by a collection of absolute cell numbers at time t, associated to the respective cell type compartments under consideration. $W_i(x(t))$ represents the rate by which the elementary reaction i occurs, as one of the m possible independent reactions. The occurred reaction i causes a discrete change of the state according to $x(t) \longrightarrow x(t)+r_i$, where r_i characterises the reaction-specific replacement of the state. The phenomenological

equation 2.1 provides the time evolution of the probability of the system state $x(t) = k$ at time t by a simple balance consideration. In this thesis we are looking at reaction processes in the context of cells, and therefore assume both $W_i(x(t)) \geq 0$ and $x(t) \geq 0$, for all times t. Using eq. 2.1 and the definition of the expectation value, it can be concluded that

$$\begin{aligned}
\frac{\mathrm{d}\langle x(t)\rangle}{\mathrm{d}t} &= \sum_k k \frac{\mathrm{d}P(k,t)}{\mathrm{d}t} \\
&= \sum_{k,i} \left[kW_i(k-r_i)P(k-r_i,t) - kW_i(k)P(k,t) \right] \\
&= \sum_{k,i} (k+r_i)W_i(k)P(k,t) - \sum_{k,i} kW_i(k)P(k,t) \\
&= \sum_{i,k} r_i W_i(k)P(k,t) =: \langle f(x(t))\rangle, \qquad (2.2)
\end{aligned}$$

where from the second to the third line $k = k' + r_i$ was substituted and subsequently relabelled back to $k' \longrightarrow k$, leaving the sum unaffected. Eq. 2.2 shows that the dynamics of the mean state vector can be written as the mean of some function $f(x(t))$. If $W_i(x_j)$ is linear in all x_j, f can be expressed in the form

$$f(x(t)) = Ax(t) + f_0, \qquad (2.3)$$

where A is a $n \times n$ matrix, containing all the reaction parameters and f_0 is a vector with constant entries. Since in our case we will only consider simple rate processes such as cell division, cell death, and differentiation, the assumption of linearity is appropriate. Then, eq. 2.2 can be rewritten as

$$\frac{\mathrm{d}\langle x(t)\rangle}{\mathrm{d}t} = A\langle x(t)\rangle + f_0, \qquad (2.4)$$

thereby providing n coupled equations for the first moment of the individual cell types $x_i(t)$. In the same way, equations for the second moment of the cell number distribution can be derived as

$$\begin{aligned}
\frac{\mathrm{d}\left\langle x(t)x(t)^{\mathrm{T}}\right\rangle}{\mathrm{d}t} &= \sum_k kk^{\mathrm{T}}\frac{\mathrm{d}P(k,t)}{\mathrm{d}t} \\
&= \sum_{k,i}(k+r_i)(k+r_i)^{\mathrm{T}}W(k)P(k,t) - \sum_{k,i}kk^{\mathrm{T}}W_i(k)P(k,t) \\
&= \sum_{k,i}\left(kk^{\mathrm{T}} + r_ir_i^{\mathrm{T}} + kr_i^{\mathrm{T}} + r_ik^{\mathrm{T}}\right)W_i(k)P(k,t) - \sum_{k,i}kk^{\mathrm{T}}W_i(k)P(k,t) \\
&= \sum_{k,i}\left(kr_i^{\mathrm{T}} + r_ik^{\mathrm{T}} + r_ir_i^{\mathrm{T}}\right)W_i(k)P(k,t) \\
&= \sum_k kf^{\mathrm{T}}(k)P(k,t) + \sum_k f(k)k^{\mathrm{T}}P(k,t) + \sum_{k,i}r_ir_i^{\mathrm{T}}W_i(k)P(k,t) \\
&= \left\langle x(t)f(x(t))^{\mathrm{T}}\right\rangle + \left\langle f(x(t))x(t)^{\mathrm{T}}\right\rangle + \left\langle \sum_i r_ir_i^{\mathrm{T}}W_i(x(t))\right\rangle . \qquad (2.5)
\end{aligned}$$

Considering that the covariance Σ is defined by

$$\Sigma(t) = \left\langle x(t)x(t)^{\mathrm{T}}\right\rangle - \langle x(t)\rangle\left\langle x(t)^{\mathrm{T}}\right\rangle , \qquad (2.6)$$

and the fact that

$$\left\langle f(x(t))x(t)^{\mathrm{T}}\right\rangle = \left\langle \left(f(x(t)) - \langle f(x(t))\rangle\right)\left(x(t) - \langle x(t)\rangle\right)^{\mathrm{T}}\right\rangle + \langle f(x(t))\rangle\left\langle x(t)^{\mathrm{T}}\right\rangle , \qquad (2.7)$$

eq. 2.5 can be rewritten in terms of the covariance as

$$\begin{aligned}
\frac{\mathrm{d}\Sigma(t)}{\mathrm{d}t} = &\left\langle \left(f(x(t)) - \langle f(x(t))\rangle\right)\left(x(t) - \langle x(t)\rangle\right)^{\mathrm{T}}\right\rangle + \\
&+ \left\langle \left(x(t) - \langle x(t)\rangle\right)\left(f(x(t)) - \langle f(x(t))\rangle\right)^{\mathrm{T}}\right\rangle + \\
&+ \left\langle \sum_i r_iW_i(x(t))r_i^{\mathrm{T}}\right\rangle . \qquad (2.8)
\end{aligned}$$

Again, assuming linearity, there holds $f(x(t)) - \langle f(x(t))\rangle = A\left(x(t) - \langle x(t)\rangle\right)$ and therefore,

$$\frac{\mathrm{d}\Sigma(t)}{\mathrm{d}t} = A\Sigma(t) + \Sigma(t)A^{\mathrm{T}} + \sum_i r_iW_i\left(\langle x(t)\rangle\right)r_i^{\mathrm{T}}. \qquad (2.9)$$

Together, equations 2.4 and 2.9 form a closed system of exact equations due to the linearity assumption 2.3, while for non-linear f, truncation schemes must be employed. Making use

of the symmetry property of the covariance matrix, i.e. $\Sigma_{ij} = \Sigma_{ji}$, this leads to $(3n+n^2)/2$ equations. Equations of higher moments, such as skewness and kurtosis, can be derived analogously. In this thesis, however, we only consider the dynamics up to the covariance, as it turns out that higher moments are no longer informative due to a strong bias.

In chapters 4 and 5 we discuss the *cell cycle dropout model* in great detail, see Fig. 4.9B for a scheme of the model topology. The corresponding state vector of this model is given by

$$x(t) = (\mathrm{N}_1, ..., \mathrm{N}_5, \mathrm{CTCMp}, \mathrm{CTEMp}, \mathrm{CTEF}, \mathrm{TCMp}, \mathrm{TEMp}, \mathrm{TEF})^{\mathrm{T}},$$

where the five naïve states N_1 to N_5 represent auxiliary compartments that allow for non-exponential activation mechanisms. The waiting time associated with the entire process $\mathrm{N}_1 \to ... \to \mathrm{N}_5$ is distributed according to the convolution of the individual exponential distributions. In this case, the convolution is given by a generalised Erlang distribution. In our case, the explicit inclusion of naïve cells into the state vector implies $f_0 = 0$. The remaining compartments represent the actual memory, effector-memory, and effector-phenotypes, each in the configurations *cycling* (marked with a "C") and *non-cycling*. By definition, only the cycling compartments participate in cell division, while only the non-cycling ones are assigned a death rate. For notational simplicity, we will abbreviate the ith component of the state vector above by x_i. Let p_i denote the proliferation rates of x_i, $i \in \{6, 7, 8\}$, and d_j the corresponding death rates, $j \in \{9, 10, 11\}$, respectively. Differentiation rates associated with transitions $x_i \longrightarrow x_j$ are defined as $r_{i\to j}$. After applying eq. 2.4 on this particular choice of elementary reactions, we obtain the following equations for the means of the cell numbers:

$$\begin{aligned}
\dot{\langle x_1 \rangle} &= -r_{1\to 2}\langle x_1\rangle \\
\dot{\langle x_2 \rangle} &= r_{1\to 2}\langle x_1\rangle - r_{2\to 3}\langle x_2\rangle \\
\dot{\langle x_3 \rangle} &= r_{2\to 3}\langle x_2\rangle - r_{3\to 4}\langle x_3\rangle \\
\dot{\langle x_4 \rangle} &= r_{3\to 4}\langle x_3\rangle - r_{4\to 5}\langle x_4\rangle \\
\dot{\langle x_5 \rangle} &= r_{4\to 5}\langle x_4\rangle - r_{5\to 6}\langle x_5\rangle \\
\dot{\langle x_6 \rangle} &= 2r_{5\to 6}\langle x_5\rangle - \langle x_6\rangle(r_{6\to 7} - p_6 + r_{6\to 9}) \\
\dot{\langle x_7 \rangle} &= r_{6\to 7}\langle x_6\rangle - \langle x_7\rangle(r_{7\to 8} - p_7 + r_{7\to 10}) \\
\dot{\langle x_8 \rangle} &= r_{7\to 8}\langle x_7\rangle + \langle x_8\rangle(p_8 - r_{8\to 11}) \\
\dot{\langle x_9 \rangle} &= r_{6\to 9}\langle x_6\rangle - \langle x_9\rangle(d_9 + r_{9\to 10}) \\
\dot{\langle x_{10} \rangle} &= r_{7\to 10}\langle x_7\rangle - \langle x_{10}\rangle(d_{10} + r_{10\to 11}) + r_{9\to 10}\langle x_9\rangle \\
\dot{\langle x_{11} \rangle} &= r_{8\to 11}\langle x_8\rangle - d_{11}\langle x_{11}\rangle + r_{10\to 11}\langle x_{10}\rangle.
\end{aligned} \tag{2.10}$$

In practice, we set $r_{1\to2} = r_{2\to3} = r_{3\to4} = r_{4\to5} = r$, in order to enforce an Erlang-distributed time for recruitment and first cell division with shape parameter $k = 4$ (number of transitions) and rate parameter r. The coefficient of variation (CV) of the initial net waiting time is given by $1/\sqrt{k}$. Hence, the CV is effectively reduced by 50 % compared to an exponential distribution with $k = 1$. The need to remove variability will be discussed in chapter 4. Applying eq. 2.9 on this same system of reactions gives the equations for the covariance matrix. Here we refer to appendix B for the resulting set of equations. Note that moment equations can also be derived by applying the equivalent formalism based on the probability generating function, see for example [Thattai et al., 2001].

Instead of working with the entries of the covariance matrix directly, we have chosen to work with the CVs and the pairwise (Pearson) correlations ρ_{ij}. These measures lead to more independent summary statistics and are thus better suited for model fitting. They can be computed from the covariance matrix elements as

$$\mathrm{CV}_i = \frac{\sqrt{\Sigma_{ii}}}{\langle x_i \rangle} = \frac{\sigma_i}{\langle x_i \rangle}, \tag{2.11}$$

$$\rho_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}}\sqrt{\Sigma_{jj}}} = \frac{\Sigma_{ij}}{\sigma_i \sigma_j} \qquad (i \neq j). \tag{2.12}$$

Whenever we are interested in moments of sums of individual state vector components, the following rules apply:

$$\langle x_i + x_j \rangle = \langle x_i \rangle + \langle x_j \rangle, \tag{2.13}$$

$$\mathrm{var}(x_i + x_j) = \mathrm{var}(x_i) + \mathrm{var}(x_j) + 2\,\mathrm{cov}(x_i, x_j), \tag{2.14}$$

$$\mathrm{cov}(x_i + x_j, x_k) = \mathrm{cov}(x_i, x_k) + \mathrm{cov}(x_j, x_k). \tag{2.15}$$

This is relevant for our considerations insofar as we have to calculate the moments of the experimentally observed phenotypes. In the model, these phenotypes are composed of several sub-compartments x_i. For our single-cell progeny data, the initial condition is given by $x_0 = (1, 0, ..., 0)^{\mathrm{T}}$. The exact knowledge of the initial state provides a central added value here. At this point, it should be emphasised that the choice of this methodology to quantify the stochasticity of the corresponding data is entirely based on the findings in [Buchholz et al., 2013]. In the course of this work, we will consider dynamic rates, such as $p_k = p_k(t)$ and $r_{l\to m} = r_{l\to m}(t)$. This corresponds to W_i being explicitly time-dependent, i.e. $W_i = W_i(x(t), t)$. Note that this modification has no effects on the derivation of the moment equations above and that the Markovian property of the system is preserved at every point in time.

The above considerations were all coded in MATLAB (version R2020a). First, code that derives analytical expressions of the moment equations based on the particular choice of elementary reactions was written. For this purpose, the *Symbolic Math Toolbox*$^{\text{TM}}$ was used. Equations were passed to MATLAB's ODE solver, which then calculated a numerical solution for a fixed set of parameters on the time interval $[0; t_{\text{end}}]$ and the above initial condition. Depending on the particular case, t_{end} varied between 8 d and 60 d. MATLAB's standard ODE solver for non-stiff equations, *ode45*, was called with default parameters. This solver is based on a pair of explicit Runge-Kutta formulas of orders four and five, more precisely the Dormand-Prince pair, see [Dormand et al., 1986].

2.1.2 Parameter estimation

Let $D = \{(t_1, y_1), ..., (t_N, y_N)\}$ be the data set under consideration, including N data points. Here, y_i represent the measured observables at corresponding time points t_i. It is assumed that the time points are not affected by experimental noise. In our particular case, the y_i are given by moments of the TCMp, TEMp, and TEF cell number distributions and their relative population sizes. Also, kinetic properties such as cell cycle speed and fractions of cycling cells contribute here. We can compute all these observables for a given choice of parameter set θ_j, $j \in \{1, ..., M\}$, in terms of the considerations above, i.e. by solving the moment equations for the mean and the covariance matrix. Let $f_i(t_i, \theta)$, $i \in \{1, ..., N\}$, denote these computed observables associated with the corresponding measurements y_i. Assuming additive and normally distributed measurement noise ϵ with zero mean and parameter-dependent variance $\sigma^2(\theta)$, i.e. $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2(\theta))$, the relation between data and model can be expressed as $y_i = f_i(t_i, \theta) + \epsilon_i$. In the following, we outline how the optimal parameter set θ_{opt} can be obtained under these assumptions.

Least-squares fitting

For the above considerations, the likelihood function $L_D(\theta)$ reads

$$L_D(\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2(\theta)}} \exp\left(-\frac{(y_i - f_i(t_i, \theta))^2}{2\sigma_i^2(\theta)}\right). \tag{2.16}$$

The optimal parameter set θ_{opt} is thus obtained by maximising the likelihood function. Defining the log-likelihood function, $l_D(\theta) := \ln(L_D(\theta))$, the same strategy holds, as the logarithm is a strictly monotone function. Hence we get,

$$\begin{aligned}\theta_{\text{opt}} &= \text{argmax}_\theta \; l_D(\theta) \\ &= \text{argmax}_\theta \; \ln\left(\left(\frac{1}{\sqrt{2\pi}}\right)^N \left(\prod_{i=1}^{N} \frac{1}{\sigma_i(\theta)}\right) \exp\left(-\sum_{i=1}^{N} \frac{(y_i - f_i(t_i,\theta))^2}{2\sigma_i^2(\theta)}\right)\right) \\ &= \text{argmax}_\theta \left[-N\ln\left(\sqrt{2\pi}\right) - \sum_{i=1}^{N} \ln(\sigma_i(\theta)) - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{y_i - f_i(t_i,\theta)}{\sigma_i(\theta)}\right)^2\right]. \end{aligned} \tag{2.17}$$

The first term in eq. 2.17 is θ-independent and consequently does not affect the optimisation procedure. Furthermore, if we assume constant errors σ_i, minimising the negative log-likelihood yields

$$\theta_{\text{opt}} = \text{argmin}_\theta \sum_{i=1}^{N}\left(\frac{y_i - f_i(t_i,\theta)}{\sigma_i}\right)^2. \tag{2.18}$$

Thus, the sum in eq. 2.18 represents the maximum likelihood estimator for Gaussian distributed errors. Moreover, if the model only depends linearly on the parameters, this sum of squares follows a χ^2-distribution of $N - M$ degrees of freedom, cf. [Hastie et al., 2009].

All errors related to the fate mapping data were calculated using the bootstrap method, see [Efron et al., 1993], utilising $N = 10^5$ bootstrapped samples. Uncertainties regarding fractions of cycling cells and proliferation rates were estimated in terms of the standard error of the mean (SEM). The error of the cell cycle dropout fractions in chapter 4 was estimated during least-squares fitting since too few data points were available. The inclusion of uncertainty estimations into the fitting problem is explained below. Least-squares fitting was performed with MATLAB's function *lsqnonlin* (version R2020a), MATLAB's standard function for non-linear least-squares fitting problems. Here, we used the trust-region-based variant *Trust-Region-Reflective Least-Squares Algorithm*. During parameter estimation, parameters were constrained to the interval $[0; \infty]$. Furthermore, central finite differences were used. The tolerances of the objective function, as well as the step size were chosen to be 10^{-4}, whereby the maximum number of function evaluations was 10^4. The maximum number of iterations per fit was set to 10^2. To find the global minimum of the objective function, starting values were chosen randomly. For this purpose, the Latin hypercube method was used, see [McKay et al., 2000], to enable efficient screening of the multidimensional parameter space. When compiling the random sample of $N = 10^3$ pa-

rameter sets using MATLAB's function *lhsdesign*, hypercubes of edge length $a = 10$ were generated, thereby covering the subspace of all biologically relevant parameter values.

Estimation of uncertainties

As mentioned before, we had to estimate the error associated with fractions of cycling cells in chapter 4. In this case we have $\sigma_i = \sigma_i(\theta) = \theta_j$, i.e. the error is considered as an additional parameter. The second term in eq. 2.17 must therefore be included in the optimisation procedure. The extended least-squares sum is thus given by

$$\theta_{\text{opt}} = \text{argmin}_\theta \sum_{i=1}^{N} \left[2\ln(\sigma_i(\theta)) + \left(\frac{y_i - f_i(t_i, \theta)}{\sigma_i} \right)^2 \right], \tag{2.19}$$

where constant σ_i can actually be omitted in the sum. Optimisations of this form can also be handled with MATLAB's standard function *lsqnonlin*. However, since this function is passed unsquared summands of the form $\sqrt{2\ln(\sigma_i)}$, problems arise for error values smaller than one. To avoid this, we rewrite eq. 2.19 as

$$\begin{aligned} \theta_{\text{opt}} &= \text{argmin}_\theta \sum_{i=1}^{N} \left[2\ln(\sigma_i(\theta)) + \left(\frac{y_i - f_i(t_i, \theta)}{\sigma_i(\theta)} \right)^2 \right] \\ &= \text{argmin}_\theta \sum_{i=1}^{N} \left[2\ln(C\sigma_i(\theta)) - 2\ln(C) + \left(\frac{y_i - f_i(t_i, \theta)}{\sigma_i(\theta)} \right)^2 \right] \\ &= \text{argmin}_\theta \sum_{i=1}^{N} \left[\sqrt{2\ln(C\sigma_i(\theta))}^2 + \left(\frac{y_i - f_i(t_i, \theta)}{\sigma_i(\theta)} \right)^2 \right], \end{aligned} \tag{2.20}$$

introducing a constant and large scaling factor $C \gg 1$. From the second to the third line we excluded the constant offset of $N\ln(C)$ from the negative log-likelihood. In our case eq. 2.20 together with the choice of $C = 10^4$, has proven to be robust for all considerations made in this thesis. Notably, the least-squares sum now depends on the explicit choice of C.

Identifiability of model parameters and predictions

It shall be assumed that the sum in 2.18 or 2.19 gives a value of χ^2_{min} at the global minimum, i.e. at $\theta = \theta_{\text{opt}} \in \mathbb{R}^M$. Note that with this choice of notation we do not want to imply that the sum obeys a χ^2-distribution. Since we consider a non-linear fitting problem, we use the *profile likelihood* method, see [Venzon et al., 1988], to estimate confidence intervals

of the single components of θ_{opt}. Here, projections of M-dimensional regions around the global minimum that are bounded by fixed $\Delta\chi^2 = \chi^2(\theta) - \chi^2_{\text{min}}$ are considered. Given one degree of freedom, the 68 % quantile of the χ^2-distribution corresponds to $\Delta\chi^2_{0.68} \approx 1$, while the 95 % quantile is associated with $\Delta\chi^2_{0.95} \approx 4$. In practice, for the 95 % confidence bounds of parameter θ_j, we fix this parameter in the vicinity of its optimal value $\theta_{j,\text{opt}}$ and perform least-squares fitting regarding the remaining $M-1$ parameters. This will result in $\chi^2 > \chi^2_{\text{min}}$, by definition of the global minimum. By systematically increasing or decreasing θ_j, upper and lower bounds are gained whenever the critical value of $\chi^2_{\text{min}} + 4$ is reached. Plotting χ^2 over θ_j gives the profile likelihood of parameter θ_j.

Confidence intervals of model predictions were similarly addressed in terms of prediction profile likelihoods, also known as *predictive likelihoods*. This concept was discussed thoroughly in [Hinkley, 1979] and was applied to ODE-based dynamic models in [Kreutz et al., 2012]. Here, an artificial data point is added to the fitting problem with a sufficiently large weight to be considered in the minimisation procedure, while excluding its contribution from the likelihood computation. This additional data point represents a possible realisation of our observable near the global minimum, whose confidence interval we intend to calculate. Now the value of this variable is successively increased or decreased around its best fit value for fixed time t. In this way a profile likelihood is obtained which leads to confidence intervals of desired confidence levels.

The calculation of profile likelihoods was coded in MATLAB (version R2020a). As before, *lsqnonlin* was used for non-linear least-square fitting with the same choice of parameters as above. The fixed parameter was incremented or decremented with a relative step size of 0.5 % of its value at the global minimum. The initial condition of the fitting procedure was set to the parameter solution of the previous minimum. Also, the (reduced) parameter set at the global minimum was checked for each value of the fixed parameter of interest. Predictive likelihoods were computed in a similar fashion, with additionally setting errors of "artificial" data points to $\sigma = 10^{-2}$. 68 % confidence bounds for the model predictions were obtained this way, see section 4.7.

Model selection

All model comparisons in this work are based on the corrected Akaike information criterion (AIC_c), cf. [Akaike, 1974]. Here, the term "correction" refers to the consideration of

finite sample sizes N. Let M denote the number of model parameters and $L_D(\theta_{\text{opt}})$ the maximum likelihood of this same model. Then, the AIC_c reads

$$\text{AIC}_\text{c} = 2M - 2\ln(L_D(\theta_{\text{opt}})) + \frac{2M^2 + 2M}{N - M - 1}. \tag{2.21}$$

When performing least-squares fitting, we do not calculate the likelihood explicitly. Instead, we calculate θ-dependent parts of the negative log-likelihood that we denoted with χ^2_{min} so far. Considering the exact relation between χ^2_{min} and the likelihood, we get

$$\overline{\text{AIC}_\text{c}} = \chi^2_{\text{min}} + 2M + \frac{2M^2 + 2M}{N - M - 1}, \tag{2.22}$$

where χ^2_{min} arises either from eq. 2.18 or eq. 2.19, depending on whether uncertainties have been included in the fitting procedure or not. The last two terms in eq. 2.22 penalise increased model complexity and counteract the decision for overfitted models. The over-line emphasises that the given measure misses some of the constant terms irrelevant to the optimisation. However, in practice, we look at differences between AIC_c values, and hence these constant contributions cancel out. For simplicity, we omit the over-line from now on when referring to the AIC_c measure. When using the AIC_c for model selection, the best model is associated to a minimal AIC_c of $\text{AIC}_\text{c}^{(\text{min})}$. Differences with respect to this best model and candidate model j are considered as $\Delta\text{AIC}_\text{c}{}^{(j)} = \text{AIC}_\text{c}^{(\text{min})} - \text{AIC}_\text{c}{}^{(j)}$. The exponential $\exp(\Delta\text{AIC}_\text{c}{}^{(j)}/2)$ is referred to as the relative likelihood of model j with respect to the best model. In the field of systems biology, the consensus is that a model j can be confidently rejected during selection, if $|\Delta\text{AIC}_\text{c}{}^{(j)}| \gtrsim 10$, while differences of $|\Delta\text{AIC}_\text{c}{}^{(j)}| \lesssim 4$ are considered not to be relevant for decision; for differences in between, no reliable decisions can be made, see [Burnham et al., 2002].

In the case of a very large number of nested candidate models, we used the *forward selection* method for model selection, see [Fröhlich et al., 2019; Hastie et al., 2009]. Here, one starts from the simplest model and systematically expands it to more complex models until the selection criterion (here AIC_c) cannot be improved any further. However, this is a greedy method that does not guarantee to find the model with the smallest AIC_c among all candidate models.

2.2 Single-cell RNA-sequencing analysis

This section presents the methods that were used to analyse scRNA-seq data. The data is represented by a time series of OT-I T cell transcripts during acute *Listeria mono-*

cytogenes–OVA (*L.m.*-OVA) infection at day 4.5, 8, and 12 p.i. Cells were harvested primarily from the spleen but at day 12 p.i. could also be obtained from the lymph nodes. Data preprocessing and analysis were mainly performed in Python (3.8) using the package *SCANPY* (version 1.7.1), see [Wolf et al., 2018]. SCANPY offers a variety of implementations of state-of-the-art methods and is based on the *AnnData* class (version 0.7.5), see [Wolf et al., 2018]. AnnData defines a data type that is used to store a variety of data efficiently and in a scalable manner. In the following, details on the concrete analysis steps and used methods are provided.

2.2.1 Data preprocessing and quality control

We briefly comment on how the underlying count matrices were gained and outline our choice of cells for the subsequent analysis procedure. We mainly followed the preprocessing guidelines presented in [Luecken et al., 2019].

Generation of count matrices

Library preparation was achieved using a droplet-based approach by *10X Genomics*. For this purpose, the *Chromium Single Cell 3' & Gel Bead Kit v3* was utilised. In addition, libraries of spleen and lymph nodes were pooled at each time point using Hashtag oligonucleotides (HTOs), thereby attaching an organ-specific label to the surface of the cells. Following the same principle, antibody-derived tags (ADTs) were also used to label important surface proteins such as CD127 and CX3CR1. Both the HTOs and the ADTs were obtained from *Biolegend*. The pooled libraries were then paired-end-sequenced on an Illumina Novaseq 6000 S2 sequencer. All experimental steps were carried out by my colleague Albulena Toska. Alignment to the mouse reference genome *GRCm38* (release 8.4) was done by the 10X Genomics software *Cell Ranger* (version 3.1.0) using default parameters. In addition, Cell Ranger performed counting of unique molecular identifiers (UMIs), basic filtering of cells, and demultiplexing with respect to gene expression, HTO, and ADT counts. The resulting *.bam*-files were then processed by the software tool *velocyto* (version 0.17.17), called with default parameters to provide count matrices for spliced and unspliced mRNA molecules, respectively, see [La Manno et al., 2018]. The obtained *.loom*-files were the basis of all subsequent considerations, while the matrices were concatenated across time points. Of course, the HTO and ADT information were not considered as gene dimensions here.

Basic filtering of cells and genes

Starting with 17502 cells, we first removed 525 cells that had signatures different to T cells. For example, erythrocytes and B cells, expressing high levels of *Hba-a1* or *Cd74* and *Ighm*, were excluded. Moreover, cells were filtered out that expressed less than 300 genes. Dying cells were classified as cells having mitochondrial gene counts above 7.5 % of their total counts and were excluded from the analysis. The organ label was assigned to each cell based on its respective HTO counts. Thresholds were defined individually for each time point and each HTO label, ranging between 2 to 120. Cells that could not be assigned any label were removed. In contrast, cells that were positive for more than one HTO label were not initially removed, but were assigned the label "multiplet". This way, 1424 cells were classified as *inter-HTO doublets*. Below we explain the subsequent considerations regarding these cells. Overall, 3000 cells were excluded by this choice of filters, while 14502 cells remained. In addition, gene filters were set to 20 counts for spliced and to 10 counts for unspliced mRNA molecules giving rise to 10766 remaining genes in the expression matrix.

Removal of doublets

In order to remove doublets, we used the Python package *Solo* (version 0.6), see [Bernstein et al., 2020]. Solo partially consists of a variational autoencoder that both embeds the cells under consideration and also serves as a first pre-trained classifier. A neural network, which learns to distinguish doublet states from measured cell states then acts as an additional classifier. The doublet states are obtained from the measured cells by simulations, assuming that the fraction of doublets in the data set is small. Therefore, no training data needs to be specified within this semi-supervised approach. Nevertheless, the accuracy improves by providing labelled data. While the previous considerations allowed us to reliably identify cells positive for two (or more) HTOs as inter-HTO doublets, they did not provide information on doublets within the same HTO class. To identify these *intra-HTO doublets*, we applied Solo on each sample individually, thereby providing a list of known doublets. As recommended by the authors, we passed cell- and gene-filtered count matrices to the tool for this purpose. In this way, Solo identified 2213 intra-HTO doublets, in addition to the 1424 previously classified inter-HTO doublets. Notably, the intra-HTO doublets showed significant overlap with the inter-HTO doublets in dimensionally reduced embeddings (not shown). On day 4.5 p.i. only a few doublets could be classified this way. We suspect that the strong cell cycle activity at day 4.5 compared to the other time points might be a possible confounding factor in the analysis. This is because cells in

their G_2M-phase are naturally hard to distinguish from real doublets. However, this is no severe problem as the day 4.5 time point had the smallest cell number and therefore the smallest fraction of doublets to be expected. Removing all doublets (inter- and intra-HTO doublets) gave rise to 10865 cells being available for downstream analysis. Finally, gene counts were per-cell normalised and (log+1)-transformed.

Cell cycle classification

Cell cycle classification was performed according to SCANPY's working flow. To this end, we used the cell filtered matrix from above but considered it first with raw count data. Genes were filtered out that were expressed in less than three cells. Counts were (log+1)-transformed and scaled, i.e. gene counts were set to zero mean and unit variance. Then, based on Seurat's standard vignette, see [Satija et al., 2015], a continuous S and G_2M score was calculated for each cell. For this purpose, a list of S- and G_2M-phase genes was taken from [Tirosh et al., 2016]. The classification was done using both S and G_2M scores and a linear decision boundary in this two-dimensional space by employing the identity function. Cells for which both scores had been negative were assigned a G_0/G_1 label. All labels were saved in the previously manipulated AnnData object.

Highly variable genes

Most of the downstream analyses were not based on the full set of genes, but only on highly variably expressed ones. We classified genes as being highly variable or not, using the R (version 3.6.1) package *scran* (version 1.12.1), see [Lun et al., 2016]. Here, the functional relation between variance and mean is estimated based on the normalised and log-transformed expressions. The functional relation obtained in this way is interpreted as the technical variance. The biological component of the variance is defined as the difference between the observed and the estimated variance. Genes with the highest biological variance and a false discovery rate below 0.05 are classified as highly variable. Applying scran to our data resulted in 3496 genes being classified as highly variable.

2.2.2 Dimensionality reduction, clustering, and visualisation

To visualise and interpret the high-dimensional data set at hand, its dimensions need to be reduced. In addition, cells must be identified with known immunological phenotypes to infer possible developmental transitions between them. Therefore, we briefly outline all essential methods that were used to achieve this.

Uniform manifold approximation and projection

The *Uniform Manifold Approximation and Projection* (UMAP) method was used to visualise our data, see [McInnes et al., 2018]. In contrast to principal component analysis (PCA), cf. [Pearson, 1901], UMAP is a non-linear method based on manifold learning. The authors claim that their method is particularly robust and preserves global properties of the data extremely well. Like many other methods, UMAP utilises the k-nearest neighbour (kNN) graph representation of the data. kNN graphs are directed graphs that store information about neighbourhood relationships of cells from high dimensional space. Here, k refers to the number of neighbours considered per cell and thus defines a scale of locality. We computed kNN graphs utilising SCANPY's implementation with $k = 30$ neighbours and $n = 50$ principal components (PCs). Note that computing this graph based on $n = 50$ PCs can ideally be considered as noise removal since relevant biological effects can possibly be captured in the first few PCs. In our analyses, UMAPs were computed with default parameters (version 0.5.1).

Clustering

We used the Leiden algorithm, see [Traag et al., 2019], to cluster cells in an unsupervised fashion. It addresses the community detection problem by directly maximising the modularity of the kNN graph hierarchically. The modularity of a graph is a scalar between zero and one that quantifies the connectivity of nodes within a cluster (number of intra-edges) compared to the connectivity of nodes across different clusters (number of inter-edges), taking into account the actual size of the clusters. While at the start each cell is assigned a different cluster label, labels are changed iteratively in a greedy fashion until there is no further modularity gain. Cells belonging to the same cluster are summarised as supercells and the procedure is repeated. Depending on the specified resolution parameter r, a final number of cluster labels is obtained. Hence the number of resulting clusters has no biological meaning per se and is controlled by the resolution. The Leiden algorithm represents a conceptual refinement of the Louvain algorithm, see [Blondel et al., 2008]. This is because an additional step in the algorithm guarantees that all sub-parts of a cluster will be connected. In this work, we used SCANPY's implementation of the Leiden algorithm (version 0.8.1) based on the previous kNN graph and a resolution parameter of $r = 0.8$.

Cluster annotation

Gene expression was visualised by plotting (log+1)-transformed raw counts. Expression profiles of known marker genes and cell cycle classification labels were compared with the obtained Leiden clusters. In addition, differential expression testing was performed across the Leiden clusters, assuming the clusters have biological significance. Here, SCANPY's standard implementation was used with default parameters, utilising unpaired t-tests while p-values were corrected in terms of the Benjamini-Hochberg procedure, see [Benjamini et al., 1995]. Altogether, this allowed for accurate annotation of the available clusters.

2.2.3 Trajectory inference

The central focus of our analyses lies in the inference of developmental pathways. Of particular relevance are transitions between whole cell clusters, as we would like to utilise these findings in the context of compartment modelling. Since it is important to approach this matter in an unbiased fashion, we only considered methods that did not require prior knowledge of directionality, e.g. in terms of a "root cell". Below we briefly present these methods.

RNA velocities

To derive differentiation pathways during acute infection, we used the method of *RNA velocities*, see [La Manno et al., 2018]. Fitting a simplified transcription model to spliced and unspliced counts of a fixed gene Y, the degradation rate of Y can be inferred by considering cells that are in steady state with respect to Y. Under certain assumptions, an RNA velocity can be calculated for each cell concerning gene Y. This velocity is positive in case of upregulation and negative for downregulation of gene Y. Such consideration for all genes provides an extrapolation of a cell's state into its near future, which can be projected into a (non-linear) embedding of choice. This model will be referred to as the *deterministic model* in the following. The underlying transcription model was modified in [Bergen et al., 2020] such that the steady state assumption mentioned above is no longer required. This was achieved by taking into account full transcription dynamics. In chapter 4, we will exclusively present results from this *dynamic model*. This is because the dynamic model seems to be slightly more robust compared to the deterministic model, especially at the boundaries of the embedding. Overall, the results of both models were very similar, suggesting that the steady state assumption seems to hold approximately (at least for the driver genes). RNA velocities were computed with *scVelo* (version 0.2.3), i.e.

the SCANPY-compatible Python package from [Bergen et al., 2020]. To this end, first and second moments of gene expression for spliced and unspliced counts were computed based on the previous kNN graph ($k = 30$). Note that whenever considering embeddings that were based on ~ 100 cell cycle genes only, we instead considered kNN graphs that resulted from a reduced number of PCs ($n = 3$). The velocity graph was obtained from the RNA velocities by taking into account cosine similarities between the velocity vector of one cell and its distance vector with respect to another cell. Therefore, this graph contains information about transition probabilities between cells. Directions of developmental pathways were projected into UMAP embeddings, cf. [La Manno et al., 2018], and averaged on a grid.

Partition-based graph abstraction

Partition-based graph abstraction (PAGA) directly addresses our problem of inferring transitions between clusters, see [Wolf et al., 2019]. To this end, PAGA assumes (Leiden) clustered kNN graphs. The method compares the number of inter-edges between two different partitions with the expected number that is obtained by randomly connecting half-edges (thereby forming full edges). The latter procedure is considered the null model of the method, allowing to test for disconnectedness between partitions. Based on further heuristic considerations, a measure of connectivity is defined, which is represented graphically as the thickness of connections between two partitions. Utilising RNA velocities, PAGA can be used to map transitions between clusters. Similarly, as in the RNA velocity method above, a transition matrix is concluded. An overall transition measure from cluster i to j is obtained by summing up cellular transitions. Considering the inverse transition j to i results in a difference between these transition measures. t-tests are used to decide whether differences are significant, with the sign of the differences determining the respective transition direction. Although PAGA tends to overestimate the number of inter-edge connections within its null model, it has been shown to provide highly robust results, as demonstrated in the meta-study by [Saelens et al., 2019]. With PAGA being part of SCANPY's standard functions, we used PAGA to map connections between clusters with a connectedness threshold of $\alpha = 0.05$. When mapping transitions using precomputed velocity transitions from the dynamic model, we applied the same threshold.

CellRank

Based on the inferred directionality from RNA velocities, the SCANPY-compatible Python package *CellRank* (version 1.2.0) offers further applications, see [Lange et al., 2020]. Importantly, uncertainties of the RNA velocities are included and stochasticity of fate decisions is taken into account. The latter is achieved via Markov state modelling, whereby Markov chains are broken down into individual macrostates. Mathematically this is realised by using a generalised version of the Perron-cluster cluster analysis (G-PCCA), cf. [Reuter et al., 2018]. These macrostates can represent either initial, final, or transient states. For cells associated with transient states, *absorption probabilities* can be calculated, i.e. probabilities to translate into the respective terminal states. Velocity-kernel-based transition matrices were computed using the *deterministic* mode. Three final states were chosen manually. Absorption probabilities with respect to these final states were computed using default parameters, which gave rise to *fate maps*. Additionally, genes were inferred whose expression correlated with transitioning into the respective terminal lineages. Genes that showed high correlations and/or were known to have significant biological relevance were selected and their expression was plotted along pseudo-time. Here, RNA velocity pseudo-time was chosen which is based on the previously mentioned velocity graph.

CytoTRACE

As an alternative method to infer developmental trajectories in an unbiased way, we used the R package *CytoTRACE* (version 0.3.3), see [Gulati et al., 2020]. Analysing a large number of scRNA-seq data sets with annotated differentiation pathways, the authors have identified the *number of expressed genes per cell* as a robust correlate for stemness. The authors take advantage of this result in the following way: First, those genes are identified that show a high correlation with the number of expressed genes per cell. Based on these genes, a score is calculated for each cell. This score is normalised to take values between zero (differentiated) and one (undifferentiated). The CytoTRACE score obtained in this way thus captures the relative degree of differentiation of the cells. While this approach is entirely heuristic, the underlying reasons for the number of expressed genes per cell correlating so well with stemness remain unclear. In our analysis, we used the CytoTRACE package with default parameters while treating the time series data as one batch. To this end we provided a preprocessed but unnormalised count matrix as the recommended working flow suggests.

Chapter 3

Quantifying cell cycle speed

In this chapter, we present a method that allows the calculation of mean cell cycle lengths of rapidly dividing cells *in vivo*. We begin by explaining the mathematical framework of this formalism. The results and assumptions of the method are validated using stochastic simulations. We then apply the new method to quantify the cell cycle of dividing T cell subpopulations during acute infection. Those results will be considered in the next chapters during mathematical modelling of single T cell responses. This method was incorporated into the previously published manuscript [Kretschmer et al., 2020], where it was used to show that cell cycle speed of memory precursor cells is more strongly modulated by antigen availability than cell cycle speed of effector cells.

3.1 Quantification of cell cycle lengths

Here we look at a population of cells that grows due to cell division. The cell cycle length T_{div} of a single cell should obey the general and unknown distribution $D(T_{\mathrm{div}})$, which results from the convolution of the individual phase distributions. We assume that the individual cell cycle phases are totally uncoupled. This assumption will be discussed later in chapter 6. On the level of individual cells, we can assign each cell an age a and a time-until-next division δ, such that $a + \delta = T_{\mathrm{div}}$. After the transient phase, the system is in a state of equilibrium, which we would like to study in more detail below. In this section, we provide a simple experimental setup and an associated mathematical formalism to it. First, we consider the time-until-next division representation, then we progress to the age representation. Both approaches will be used in the course of this work. Assumptions and approximations made in this section are discussed in the following sections with the aid of a stochastic simulation.

3.1.1 Experimental framework

Let us consider the experimental design that is depicted in Fig. 3.1A: At $t_0 = 0$, mice are injected a nucleoside analogue (NA) that gets incorporated into the cells' genomes

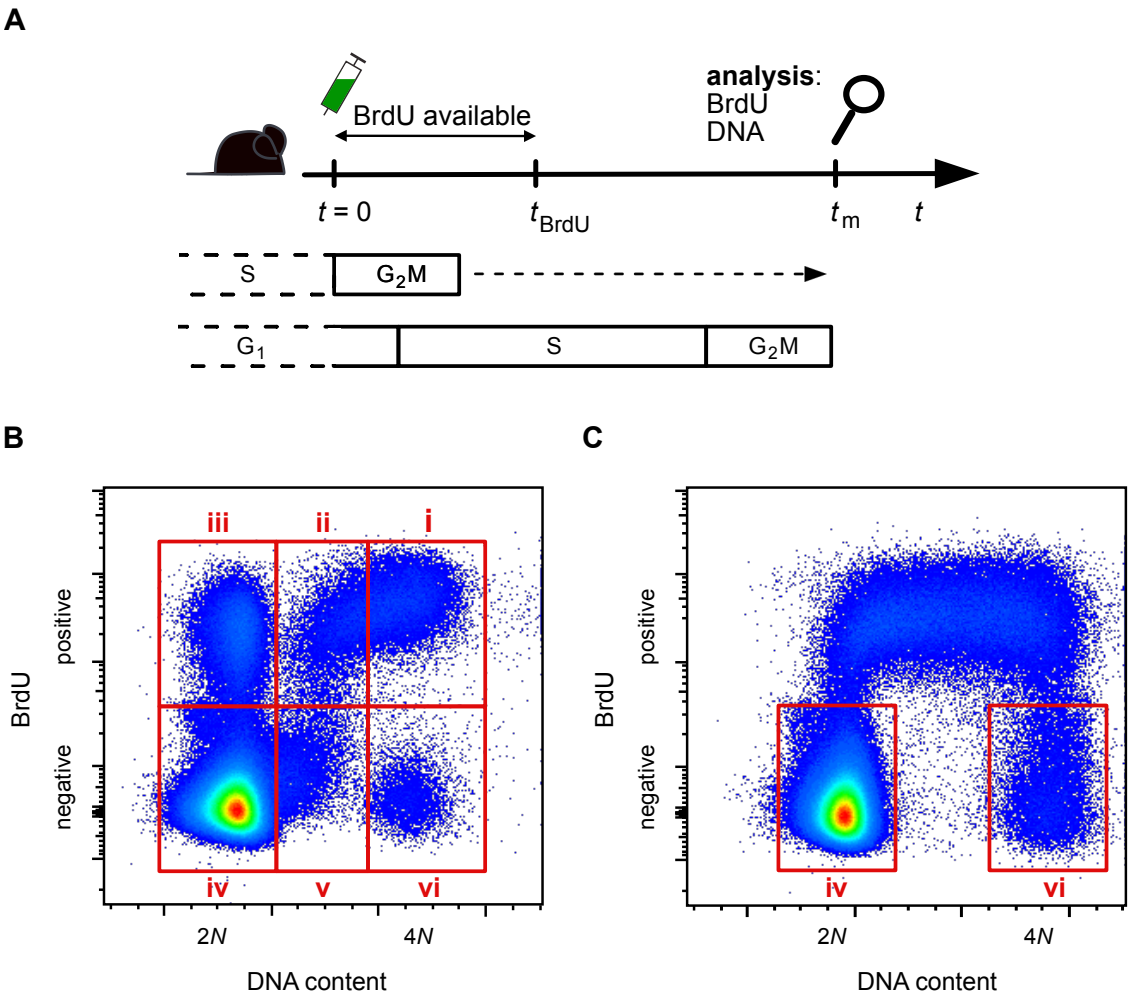

Figure 3.1: Sketch of the experimental setup and an example of the resulting data used for quantifying cell cycle speed. **(A)** Experimental setting that consists of BrdU labelling and subsequent measurement of the DNA content of the cells. Shown are the extreme cell configurations that give contributions to gate iii (see (B)), whose cells carry information about the average cell cycle speed. Here, t_{m} denotes the time point of measurement, while t_{BrdU} represents the time at which the BrdU concentration is no longer sufficient to label cells so that they are considered positive in the subsequent FACS analysis. **(B)** Gate notation (red) and BrdU/DNA profile for double-positive thymocytes with an analysis time of $t_{\mathrm{m}} = 4.0$ h. **(C)** Same conditions as in (B), with measurement time point of $t_{\mathrm{m}} = 0.5$ h. (Experiments conducted by Sabrina Bortoluzzi and Marc-Schmidt-Supprian)

during DNA synthesis in S-phase. For the purpose of illustration, we will choose the thymidine analogue bromdesoxyuridin (BrdU) as representative example[1]. After injection, degradation of BrdU begins, where labelling efficiency is an unknown function of time. By

[1]Another experimentally very common nucleoside analogue is 5-ethynyl-2'-deoxyuridine (EdU).

t_{BrdU} we mean the time at which BrdU is no longer present in the required concentration necessary to label cells such that they will be classified as BrdU^+ during measurement. At measurement time t_{m}, the analysis takes place, measuring BrdU and DNA content of the cells. This is done within the framework of FACS staining and subsequent flow cytometric analysis. The DNA content characterises the state of the cell within its cell cycle: diploid DNA content ($2N$) corresponds to the G_1-phase, while tetraploid content ($4N$) indicates the G_2- or M-phase. DNA content in between corresponds to the S-phase. As mentioned before, the idea of NA labelling and subsequent DNA content measurement is by no means novel, cf. [Weber et al., 2014]. The decisive contribution on our part is the quantitative analysis of these data in the context of *in vivo* experiments of rapidly dividing cells. Here we circumvent the unknown degradation kinetics of BrdU. Moreover, we are able to estimate the means of all phase length distributions without making strong assumptions about the exact shape of these distributions, see the following sections.

Fig. 3.1B shows the typical BrdU/DNA profile of such an experiment for the measurement time $t_{\mathrm{m}} = 4.0$ h. This time point was adapted to the kinetic properties of the dividing cell population. In this case, $\mathrm{CD4}^+$ and $\mathrm{CD8}^+$, i.e. double-positive, thymocytes were considered. The quadrants of the plot are numbered counter-clockwise. During progression through the S-phase, cells synthesise DNA and consequently increase their DNA content from $2N$ to $4N$. After division, a discrete jump in DNA content from $4N$ to $2N$ takes place. As soon as the S-phases of the cells overlap with the BrdU availability window, cells become positive for BrdU. Since the BrdU label is inherited, cells remain in one of the upper quadrants also after division. BrdU dilution can be neglected, as we are only following one cell division at most. It should be noted that due to the limited BrdU availability, at an unknown (point in) time, cells stop incorporating BrdU. This effect leads to the convenient "gap" along the labelled DNA axis and allows for precise identification of the cells in gate iii. At the same time these cells are of particular interest, since they represent exactly those daughter cells that originate from mother cells whose S-phase overlapped with the BrdU availability window and which then divided in the course of the experiment. Hence, this gate captures information about the average speed at which cells divide. After having formalised this intuition, we get back to more practical considerations in section 3.2.

3.1.2 Mathematical description

Let $p^{\mathrm{ss}}_{\mathrm{G_2M}}$ be the distribution of $\mathrm{G_2M}$ lengths $T_{\mathrm{G_2M}}$ in steady state. Throughout this thesis we consider the $\mathrm{G_2M}$-phase as a whole and do not break it down into its two sub-phases G_2 and M. Furthermore, let $p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta)$ be the steady state distribution of cells having a time-

to-next division (tnd) of δ left. Assuming, cells become positive for BrdU immediately upon incorporation, the probability of a cell being in gate iii at measurement time (see Fig. 3.1B for gate notation) can be written as

$$\Pr(\text{gate iii}) = \int_0^\infty \mathrm{d}T_{\mathrm{G_2M}} \int_{T_{\mathrm{G_2M}}}^{\max\left(t_\mathrm{m}, T_{\mathrm{G_2M}}\right)} \mathrm{d}\delta\, p^{\mathrm{ss}}_{\mathrm{G_2M}}(T_{\mathrm{G_2M}}) p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta), \tag{3.1}$$

where this exact choice of integral bounds corresponds to all possible cell configurations required to contribute to that gate, see Fig. 3.1A. This is also the reason why the length of the G_2M-phase plays a role here at all. Assuming that the distribution of G_2M lengths is sufficiently well described by a Dirac-function, then this allows us to simply account for the outer integral by replacing $T_{\mathrm{G_2M}}$ with its mean value in the boundaries of the inner integral. Thus we get

$$\Pr(\text{gate iii}) \approx \int_{\langle T_{\mathrm{G_2M}}\rangle}^{\max\left(t_\mathrm{m}, \langle T_{\mathrm{G_2M}}\rangle\right)} \mathrm{d}\delta\, p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta). \tag{3.2}$$

In order to proceed, the so far unknown function $p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta)$ must be known. The concrete functional form can be derived from a variant of the von Foerster equation, c.f. [Von Foerster, 1959], with the steady state constraint added. Based on [Dowling et al., 2005], we provide a detailed derivation in supplementary information A. The analytical solution reads

$$p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta) = c\mathrm{e}^{c\delta}\left(1 - 2\int_0^\delta \mathrm{d}\delta'\, D(T_{\mathrm{div}}{=}\delta')\mathrm{e}^{-c\delta'}\right), \tag{3.3}$$

where c is interpreted as an average growth factor of the population. The relation between c and the division-time distribution $D(T_{\mathrm{div}})$ follows from eq. 3.3 and the fact that the distribution $p^{\mathrm{ss}}_{\mathrm{tnd}}$ must vanish at infinity, i.e. $\lim_{\delta\to\infty} p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta) = 0$, since we implicitly assume that all cells are dividing[2]. Hence we get the relation

$$\int_0^\infty \mathrm{d}\delta'\, D(T_{\mathrm{div}} = \delta')\mathrm{e}^{-c\delta'} = \frac{1}{2}. \tag{3.4}$$

[2]This assumption can be easily released by defining $\widetilde{D}(T_{\mathrm{div}})$ as the division-time distribution of dividing cells only, thereby excluding finite contributions from $T_{\mathrm{div}} = \infty$, c.f. [Powell, 1956].

For synchronous cell division of inter-division time T_0, that is $D(T_{\mathrm{div}}) = \delta_{\mathrm{D}}(T_{\mathrm{div}} - T_0)$, where δ_{D} denotes the Dirac-function, it follows directly that

$$c = \frac{\ln(2)}{\langle D(T_{\mathrm{div}})\rangle}. \tag{3.5}$$

For distributions with finite variability, eq. 3.5 holds only approximately. Interestingly, this approximation also applies to realistic distributions with moderate to large variability, which we will show in section 3.3. In the following, we will abbreviate $\langle D(T_{\mathrm{div}})\rangle =: \langle T_{\mathrm{div}}\rangle$.

Let us further assume that the division-time distribution has the property of being zero for small waiting times. Put differently, there exists a minimum division-time $T_{\mathrm{div}}^{\mathrm{min}}$ with $D(T_{\mathrm{div}} < T_{\mathrm{div}}^{\mathrm{min}}) = 0$. Conceptually this is not a strong assumption, since, as we discussed in chapter 1, observed *in vitro* division-time distributions suppress small division-times. Considering times-to-next division δ that are smaller than $T_{\mathrm{div}}^{\mathrm{min}}$, eq. 3.3 simplifies to $p_{\mathrm{tnd}}(\delta|\delta < T_{\mathrm{div}}^{\mathrm{min}}) = c\mathrm{e}^{c\delta}$. Thus we can express the probability of cells having a time-to-next division in between δ_1 and δ_2 as

$$p_{\mathrm{tnd}}^{\mathrm{ss}}(\delta_1 < \delta < \delta_2) = \int_{\delta_1}^{\delta_2} \mathrm{d}\delta\, p_{\mathrm{tnd}}^{\mathrm{ss}}(\delta|\delta < T_{\mathrm{div}}^{\mathrm{min}}) = \mathrm{e}^{c\delta_2} - \mathrm{e}^{c\delta_1}, \tag{3.6}$$

while assuming $\delta_1 < \delta_2 < T_{\mathrm{div}}^{\mathrm{min}}$. This expression is also obtained, if we try to use a distribution that manifestly incorporates the approximation property. For example, if we again consider synchronous cell division from before, eq. 3.6 follows directly, where the minimum division-time is given by $T_{\mathrm{div}}^{\mathrm{min}} = T_0$. Eq. 3.6 defines a very handy expression since the full distribution of division-times has cancelled out and the only link to it is given by relation 3.5.

Let us return to our experimental setting. Our next goal is to use these considerations to write down the probability of a cell being in the informative gate iii during time of measurement. Eq. 3.6 allows calculating the number of mother cells at $t_0 = 0$ that will give rise to two daughter cells in gate iii at time t_{m}. Defining the initial cell number $N(t{=}0) =: N_0$, this number of mother cells reads $N_0 p_{\mathrm{tnd}}^{\mathrm{ss}}\left(\langle T_{\mathrm{G_2M}}\rangle < \delta < t_{\mathrm{m}}\right)$. Since the population further grows between t_0 and t_{m}, the probability to find a cell in gate iii at t_{m} is given by

$$\Pr(\text{gate iii}) = \frac{2N_0 p_{\mathrm{tnd}}^{\mathrm{ss}}\left(\langle T_{\mathrm{G_2M}}\rangle < \delta < t_{\mathrm{m}}\right)}{N(t{=}t_{\mathrm{m}})}. \tag{3.7}$$

Here, the correct normalisation $N(t{=}t_\mathrm{m})$ directly follows from the definition of c as the mean growth factor of the population. Hence, we get

$$N(t{=}t_\mathrm{m}) = N_0\,\mathrm{e}^{ct_\mathrm{m}}. \tag{3.8}$$

Inserting eq. 3.8 into eq. 3.7 finally gives the primary equation

$$\Pr(\text{gate iii}) = 2\left(1 - \mathrm{e}^{c\left(\langle T_{\mathrm{G_2M}}\rangle - t_\mathrm{m}\right)}\right), \tag{3.9}$$

meaning that if the left-hand side of eq. 3.9 is measured, the parameter of interest c can be computed. However, this requires knowledge about the mean length of the $\mathrm{G_2M}$-phase, see section 3.2. Note that gate probabilities translate into gate frequencies for large numbers of cells, which, in practice, will always be the case for our flow cytometric measurements.

3.1.3 Intuitive limit of the primary equation

It is convenient to manipulate eq. 3.9 even further to develop a more intuitive understanding of this formalism. To this end, consider eq. 3.9 for small exponents, i.e. $c\,(\langle T_{\mathrm{G_2M}}\rangle - t_\mathrm{m}) \ll 1$. The Taylor expansion up to first order leads to the approximation

$$\Pr(\text{gate iii}) \approx 2ct_\mathrm{m} - 2c\langle T_{\mathrm{G_2M}}\rangle. \tag{3.10}$$

Additionally, the probability of a cell to be in $\mathrm{G_2M}$-phase at the beginning of the experiment is, by means of eq. 3.6, given by

$$p^{\mathrm{ss}}_{\mathrm{tnd}}(0 < \delta < \langle T_{\mathrm{G_2M}}\rangle) = \mathrm{e}^{c\langle T_{\mathrm{G_2M}}\rangle} - 1 \approx c\langle T_{\mathrm{G_2M}}\rangle =: f_{\mathrm{G_2M}}, \tag{3.11}$$

which, at the same time, is the fraction of cells being in $\mathrm{G_2M}$-phase at any time in steady state, particularly at $t = t_0$. Hence we can identify the factor in eq. 3.10 with $f_{\mathrm{G_2M}}$ and get

$$\Pr(\text{gate iii}) \approx 2ct_\mathrm{m} - 2f_{\mathrm{G_2M}}. \tag{3.12}$$

Formula 3.12 has a very intuitive interpretation. The first term describes all cells that have divided during the experiment. However, gate iii is formed by a subset of these cells. We have to subtract those cells that were in $\mathrm{G_2M}$-phase at the beginning of the experiment. Even though these cells also divide, they do not incorporate BrdU since their S-phase had taken place before the experiment started. The second term in the approximation formula represents these very cells. It should be emphasised that this last approximation only served the purpose of illustration, especially since the original eq. 3.9 can always be used.

3.1.4 Analogue consideration of age representation

Instead of working with the object $p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta)$, we can alternatively consider the age representation. The corresponding distribution is given by $p^{\mathrm{ss}}_{\mathrm{age}}(a)$ and captures the probability of cells having an age a in steady state. Then, in complete analogy to the previous formulations, the probability of cells to be found in gate iii at measurement time reads

$$\Pr(\text{gate iii}) = \int_0^\infty \mathrm{d}T_{\mathrm{G_2M}} \int_0^{\max\left(t_{\mathrm{m}}-T_{\mathrm{G_2M}},0\right)} \mathrm{d}a\, p^{\mathrm{ss}}_{\mathrm{G_2M}}(T_{\mathrm{G_2M}})\, p^{\mathrm{ss}}_{\mathrm{age}}(a), \tag{3.13}$$

where the boundaries of the inner integral refer to $t = t_{\mathrm{m}}$. Hence no further normalisation factor needs to be considered in this representation. Assuming peaked $\mathrm{G_2M}$ length distributions as before, we get

$$\Pr(\text{gate iii}) \approx \int_0^{\max\left(t_{\mathrm{m}}-\langle T_{\mathrm{G_2M}}\rangle,0\right)} \mathrm{d}a\, p^{\mathrm{ss}}_{\mathrm{age}}(a). \tag{3.14}$$

The exact form of $p^{\mathrm{ss}}(a)$ can be derived from the von Foerster equation, cf. [Von Foerster, 1959], employing steady state assumptions. Unlike for $p^{\mathrm{ss}}(\delta)$, we do not provide a derivation for $p^{\mathrm{ss}}(a)$, since the derivations are structurally very similar. For a complete derivation we refer instead to [Stukalin et al., 2013]. The analytical expression is given by

$$p^{\mathrm{ss}}_{\mathrm{age}}(a) = 2c\mathrm{e}^{-ca} \int_a^\infty \mathrm{d}a'\, D(T_{\mathrm{div}}{=}a') \tag{3.15}$$

and the boundary condition reads

$$2\int_0^\infty \mathrm{d}a\, D(T_{\mathrm{div}}{=}a)\,\mathrm{e}^{-ca} = 1, \tag{3.16}$$

which leads to the already mentioned relation 3.5. The probability of encountering a cell of age a that lies in the interval $[a_1, a_2]$, is given by

$$p^{\mathrm{ss}}_{\mathrm{age}}(a_1 < a < a_2) = \int_{a_1}^{a_2} \mathrm{d}a\, p^{\mathrm{ss}}_{\mathrm{age}}(a) = 2\left(\mathrm{e}^{-ca_1} - \mathrm{e}^{-ca_2}\right). \tag{3.17}$$

Again, we assumed the existence of a minimal division time T_{div} for which $D(T_{\mathrm{div}} < T_{\mathrm{div}}^{\mathrm{min}}) = 0$ holds, and consider ages that satisfy $a_1 < a_2 < T_{\mathrm{div}}$. Evaluating eq. 3.17 at the age configurations relevant for gate iii, we find

$$\Pr(\text{gate iii}) = 2\left(1 - \mathrm{e}^{-c\left(t_{\mathrm{m}} - \langle T_{\mathrm{G_2M}}\rangle\right)}\right), \tag{3.18}$$

which corresponds exactly to our result from eq. 3.5 in the time-to-next division representation. The two representations in terms of time-until-next division and age are equivalent. Expressions derived within one formalism can always be converted into expressions derived from the other formalism by simple algebraic manipulations. From now on, we use both representations together. To this end, we choose the representations in such a way that the derived equations are as compact as possible from the start.

3.2 Resolving cell cycle phases

We have found a way to calculate mean cell cycle lengths using eq. 3.9 and relation 3.5. However, eq. 3.9 depends on the mean G_2M length. Therefore, we still have to provide a complete system of equations. To continue, we must relate the mean G_2M length to another gate probability. For this purpose, gates i and vi, in which cells of tetraploid DNA content are found, could be used in principle (see Fig. 3.1B for gate notations). Conceptually, the same time t_{m} that we considered in eq. 3.9 can be used to establish additional gate probabilities. On average, the G_2M length is short compared to the overall cell cycle length. This means that if we measure long enough to accumulate enough cells in primary gate iii, gate vi will be empty and thus uninformative. Gate i, on the other hand, will not be empty, but we could not identify its cells reliably. This is because its cells will be continuously connected to the neighbouring gate ii, containing (late) S-phase cells. Hence, establishing the gate threshold along the DNA axis is difficult to achieve in practice. Therefore, we generally recommend inferring the fraction of G_2M-phase cells by evaluating DNA content histograms. For example, the *Watson pragmatic curve fitting algorithm* could be used for this purpose, see [Watson et al., 1987].

However, we find that using two distinct time points of measurement, an early one, $t_{\mathrm{m}}^{\mathrm{e}}$, and a later one, $t_{\mathrm{m}}^{\mathrm{l}}$, significantly improves the accuracy of our method. Furthermore, this allows calculating all other cell cycle phase lengths. Fig. 3.1C shows why the early time point is better suited to quantify the G_2M length: since all S-phase cells get BrdU-labelled, the unlabelled cells with tetraploid DNA content located in gate vi can be identified reliably. This holds for all time points that are smaller than the average G_2M length. Of course,

this choice of experimental design comes with the disadvantage of combining information from two mice and hence introduces some external variability. Nevertheless, we think this is a reasonable approach, especially since the G_2M-phase contribution to the overall cell cycle is rather small (in the context of T cells). In principle, one can also circumvent this problem by adding another DNA label to the experimental setup with a substance other than BrdU, e.g. EdU. One chooses the late time point $t_{\mathrm{m}}^{\mathrm{l}}$ as the time point of analysis and administers EdU shortly before the analysis at time $t_{\mathrm{m}}^{\mathrm{l}} - t_{\mathrm{m}}^{\mathrm{e}}$. In this way, information about the G_2M-phase is obtained from the same mouse. For simplicity, however, we will switch back to our exemplary experimental design from Fig. 3.1A.

As in the previous section, we can write,

$$\begin{aligned}\Pr(\text{gate vi}) &= \int_0^\infty \mathrm{d}T_{\mathrm{G_2M}} \int_0^{\max(T_{\mathrm{G_2M}}-t_{\mathrm{m}}^{\mathrm{e}},0)} \mathrm{d}\delta\, p_{\mathrm{G_2M}}^{\mathrm{ss}}(T_{\mathrm{G_2M}}) p_{\mathrm{tnd}}^{\mathrm{ss}}(\delta) \\ &\approx \mathrm{e}^{c\max\left(\langle T_{\mathrm{G_2M}}\rangle - t_{\mathrm{m}}^{\mathrm{e}},0\right)} - 1, \end{aligned} \tag{3.19}$$

where the same approximation scheme was applied as before. Here, our steady state formula was evaluated directly at measurement time. Together, equations 3.9 and 3.19 form a closed system. However, we are interested in resolving the full cell cycle in terms of all its phases. To this end, we choose to infer the average G_1 length in addition. For this purpose, it is advisable to consider gate iv at the early time point $t_{\mathrm{m}}^{\mathrm{e}}$. Analogously we conclude that

$$\begin{aligned}\Pr(\text{gate iv}) &= \int_0^\infty \mathrm{d}T_{\mathrm{G_2M}} \int_0^\infty \mathrm{d}T_{\mathrm{G_1}} \int_{\max(0,t_{\mathrm{m}}^{\mathrm{e}}-T_{\mathrm{G_2M}})}^{T_{\mathrm{G_1}}} \mathrm{d}a\; p_{\mathrm{G_2M}}^{\mathrm{ss}}(T_{\mathrm{G_2M}}) p_{\mathrm{G_1}}^{\mathrm{ss}}(T_{\mathrm{G_1}}) p_{\mathrm{age}}^{\mathrm{ss}}(a) \\ &\approx 2\left(\mathrm{e}^{-c\max\left(t_{\mathrm{m}}^{\mathrm{e}}-\langle T_{\mathrm{G_2M}}\rangle,0\right)} - \mathrm{e}^{-c\langle T_{\mathrm{G_1}}\rangle}\right). \end{aligned} \tag{3.20}$$

Altogether, this gives us the following closed system of coupled equations

$$\begin{aligned}\Pr(\text{gate iii}) &\approx 2\left(1 - \mathrm{e}^{-c\,\max\left(t_{\mathrm{m}}^{\mathrm{l}}-\langle T_{\mathrm{G_2M}}\rangle,0\right)}\right), \\ \Pr(\text{gate vi}) &\approx \mathrm{e}^{c\max\left(\langle T_{\mathrm{G_2M}}\rangle - t_{\mathrm{m}}^{\mathrm{e}},0\right)} - 1, \\ \Pr(\text{gate iv}) &\approx 2\left(\mathrm{e}^{-c\max\left(t_{\mathrm{m}}^{\mathrm{e}}-\langle T_{\mathrm{G_2M}}\rangle,0\right)} - \mathrm{e}^{-c\langle T_{\mathrm{G_1}}\rangle}\right), \\ \langle T_{\mathrm{S}}\rangle &= \langle T_{\mathrm{div}}\rangle - \langle T_{\mathrm{G_1}}\rangle - \langle T_{\mathrm{G_2M}}\rangle, \\ c &\approx \frac{\ln(2)}{\langle T_{\mathrm{div}}\rangle}, \end{aligned} \tag{3.21}$$

whose implications shall be discussed in more detail now. As we have presented our formalism in a problem-oriented manner, it is clear that it can be adapted to a variety of experiments. These include for example extensions to dual-pulse-chase experiments, involving two different labellings such as BrdU and EdU, see e.g. [Gitlin et al., 2015]. In the further discussion, however, we refer to the experimental setup from Fig. 3.1A.

It is important to note that the system of equations 3.21 does not depend on the full cell cycle distribution, nor on any relation between the individual phase length distributions. Likewise, the unknown BrdU degradation kinetics does not enter the equations. In addition, we take advantage of the fact that BrdU has only a short availability to uniquely classify cells in gate iii. This is due to labelled cells of diploid DNA content being separated from those cells being in S-phase through the gap. Note that the gap will be preserved even if gate iii cells re-enter S-phase and start synthesising DNA. In principle, one could even calculate pharmacokinetic parameters of BrdU by setting up gate probabilities for the middle gates ii and v. One of the reasons why our equations do not depend on the BrdU availability is related to the specific choice of our gate probabilities: gates ii and v indeed depend on the time-dependent BrdU labelling efficiency, as well as on the mean length of the S-phase. However, since we obtain the latter simply by covering the other phases and the overall cell cycle speed, we do not utilise these gates. Nevertheless, it should be kept in mind that these gates in principle provide information about the pharmacokinetics of BrdU, or the respective NA under consideration.

So far, we have not given any details on the choice of the experimental parameters $t_{\mathrm{m}}^{\mathrm{e}}$ and $t_{\mathrm{m}}^{\mathrm{l}}$. $t_{\mathrm{m}}^{\mathrm{e}}$ should be as small as possible to prevent the cells from leaving the gates iv and vi. Nevertheless, it is necessary to choose $t_{\mathrm{m}}^{\mathrm{e}}$ big enough so that S-phase cells have the opportunity to incorporate BrdU. We think that $t_{\mathrm{m}}^{\mathrm{e}} \approx 0.5$ h is an adequate choice of the early time point since respective cells are rather rapidly classified as positive during analysis, given BrdU was available. The later time point $t_{\mathrm{m}}^{\mathrm{l}}$ depends on the division behaviour of the cells under consideration. It is advisable to wait long enough such that a sufficient number of cells can enter gate iii. This also means that a sufficient number of mother cells can enter the S-phase after BrdU availability has ceased. This leads to better identification of gate iii cells due to the resulting gap. However, if the waiting time considerably exceeds the mean G_1 length, cells in gate iii will re-enter S-phase and hence will leave the gate upon DNA synthesis. Furthermore, the validity of our approximation scheme can no longer be guaranteed when evaluated at times bigger than $T_{\mathrm{div}}^{\mathrm{min}}$. In our case of very rapidly dividing cells, $t_{\mathrm{m}}^{\mathrm{l}} \approx 3.0$ h has proven to be a good experimental choice.

A final point to be addressed is the estimation of quantification errors. Our method is, partly by design, very sensitive to gate iii frequencies at time t_{m}. Therefore, the false positive classification of cells in this gate represents the largest error source. This misclassification is driven by limited sensitivity regarding DNA content measurements. Cells that have incorporated BrdU but have not divided during the experiment, could contaminate gate iii in principle. This especially holds for cells that have not synthesised enough DNA yet to be distinguished from labelled daughter cells. On the other hand, this number of cells will be small in case of many cells entering S-phase after availability of BrdU has ceased, i.e. for large time intervals $t_{\mathrm{m}} - t_{\mathrm{BrdU}}$. Visually, this is addressed in terms of the gap which should be sufficiently large to avoid misclassification of this kind. Given identifiability of the labelled daughter cells, cell cycle speed can be calculated separately for each data point of gate iii, while gate frequencies obtained from the early time points are averaged across mice. Here, the resulting standard error of the mean seems to be a justified measure of the error. However, in cases where a clear identification of gate iii is difficult due to large noise, for example, such an approach should be considered with caution. Resulting standard errors should rather be interpreted as lower error bounds. In the next sections, we will focus on other possible confounding effects, which we are going to address in terms of stochastic simulations.

3.3 Application to simulated data

Next, we want to simulate the experiment shown in Fig. 3.1A. This allows to determine the optimal experimental parameters as well as to check the accuracy of our method. Later we will verify specific assumptions of our formalism that we have made in section 3.1.

3.3.1 Stochastic simulation

First, we want to show exactly how the simulated data are obtained. Let $D_{\mathrm{div}}(T_{\mathrm{div}})$ and $D_i(T_i)$, $i \in \{\mathrm{G}_1, \mathrm{S}, \mathrm{G}_2\mathrm{M}\}$, be the distributions of new born cells, satisfying $a = 0$. To avoid confusion, we have added indices to the distributions. In the presence of cell death, new born cells are also assigned a death-time T_{death} according to $D_{\mathrm{death}}(T_{\mathrm{death}})$. In steady state, the system's kinetics is fully described in terms of $p^{\mathrm{ss}}(\delta, T_{\mathrm{div}}, \lambda)$, where λ is defined as the remaining life-time, $\lambda := T_{\mathrm{death}} - a$, following the notation of [Dowling et al., 2005]. Again, note that in the presence of cell death, λ will only be the remaining life-time if, and only if, $\lambda < \delta$. Otherwise, the cell will divide at δ and hence give rise to two daughter cells. The opposite argument holds for the interpretation of δ. In the following, we will not consider this phenomenon of *censorship* notationally.

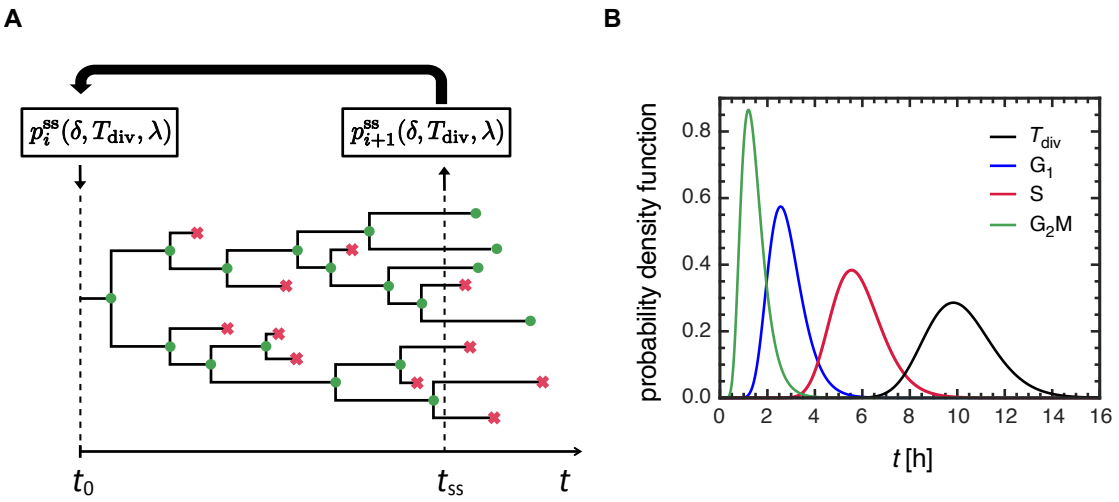

Figure 3.2: Simulation scheme to obtain equilibrium distributions by iteration. **(A)** All kinetic properties of the mother cell are initialised at time point t_0. Green nodes correspond to cell divisions and red crosses represent cell deaths. The system is propagated in time such that the distribution p^{ss} can be inferred at time t_{ss}. This distribution is then used to initialise mother cells in a subsequent simulation. Iteratively, the full steady state distribution is thus obtained. **(B)** Plot of the cell cycle phase distributions (coloured) and total length distribution (black) for parameter set 1 from Table 3.1.

To our knowledge, there is no analytical expression for the equilibrium distribution p^{ss} in the presence of cell division and death. Therefore, we must access this distribution via simulating whole trees of cells, which is schematically shown in Fig. 3.2A. Here, a mother cell's time-to-next division gets initialised with its division-time, while its remaining life-time gets initialised with its death-time, i.e. $p_0^{ss}(\delta, T_{div}, \lambda) = p^{ss}(\delta = T_{div}, T_{div}, \lambda = T_{death})$. In principal, any distribution type could be used here. For fixed parameters, the system is simulated until the desired point in time. Due to technical reasons, we find it easier to follow a division tree up to five generations and then use the full distribution thus obtained as the initialisation of the mother cells in the next simulation. Iteratively, the desired equilibrium distribution p^{ss} can be obtained in this way after only a small number of iterations.

The shape of the division-time and death-time distributions are assumed to be log-normal, as this distribution has been shown to describe *in vitro* data of proliferating lymphocyte populations accurately, cf. [Deenick et al., 2003; Dowling et al., 2014; Hawkins et al., 2009]. Throughout this analysis, we consider the exemplary set of parameters that have been estimated in [Dowling et al., 2014]. Here, the authors fitted a log-normal division-time model to *in vitro* data of stimulated OT-I T cells. Interestingly, the authors could also identify the variance of the inter-division-time distribution. Parameter set 1 in Table 3.1

parameter	set 1	est. 1	err. 1	set 2	est. 2	err. 2
$\mu(T_{\mathrm{div}})$	10.13 h	10.3 h	1.7 %	8.55 h	8.2 h	4.1 %
$\mu(T_{\mathrm{G_1}})$	2.84 h	2.9 h	2.1 %	3.11 h	2.8 h	10.0 %
$\mu(T_{\mathrm{S}})$	5.83 h	5.9 h	1.2 %	4.84 h	4.6 h	5.0 %
$\mu(T_{\mathrm{G_2M}})$	1.46 h	1.5 h	2.7 %	0.59 h	0.7 h	18.6 %
$\sigma(T_{\mathrm{div}})$	1.43 h			1.71 h		
$\sigma(T_{\mathrm{G_1}})$	0.76 h			1.03 h		
$\sigma(T_{\mathrm{S}})$	1.09 h			1.23 h		
$\sigma(T_{\mathrm{G_2M}})$	0.54 h			0.45 h		

Table 3.1: Simulation parameters (means and standard deviations of the respective phase length distributions) assuming log-normal distributions and their corresponding estimates using our method. Parameter set 1 was taken from [Dowling et al., 2014], while parameter set 2 stems from [Kretschmer et al., 2020]. Relative errors of quantification refer to the simulation parameters.

shows the estimated parameters. In contrast to the authors of this study, we assume that the individual cell cycle phases follow independent log-normal distributions. We decided to distribute the total variability of the cell cycle among the individual phases according to their means. Fig. 3.2B shows the division-time distribution as well as the individual phase distributions for parameter set 1 under consideration. Note that realistic distributions, such as this one, satisfy our basic assumption of a lag in the overall division-time distribution. We could allow even more variability here, which might be true for the *in vivo* case.

Neglecting cell death for the moment, Fig. 3.3 illustrates the transition of the system to steady state, based on the marginal distribution $p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta)$ for which the exact formula, eq. 3.3, was solved numerically. Normalised histograms based on $N = 10^5$ mother cells were plotted for different numbers of iterations, while $n = 0$ represents the initialisation step in terms of a log-normal division-time distribution. Already after a few divisions ($n = 1$), probability mass is shifted to smaller time-to-next divisions. After only three iterations, the discrepancy between exact and estimated distribution is small, while after ten iterations they completely coincide. The number of iterations needed for convergence depends mainly on the variability of the underlying distributions.

Now that we have established the correct starting distribution, we can proceed with the actual simulation of our DNA-labelling experiment. To this end, we first initialise a sufficient number of mother cells with δ, T_{div}, and λ. Then we assign them phase lengths drawn from the log-normal distributions $D_{\mathrm{G_1}}$, D_{S}, and $D_{\mathrm{G_2M}}$. Here we employ the con-

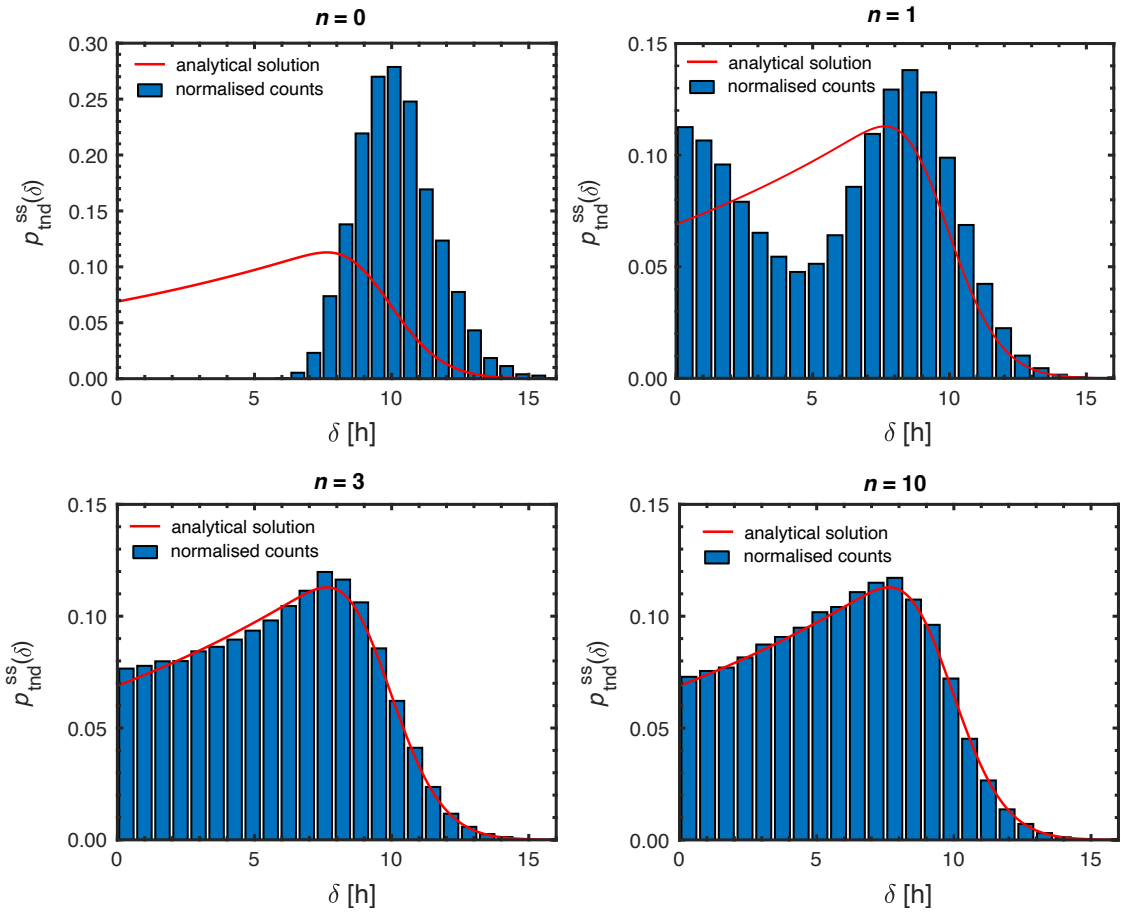

Figure 3.3: Simulation of the equilibrium distribution by iteration. Normalised histograms of the equilibrium distribution for different numbers of iterations (blue) and exact solution (red). $N = 10^5$ mother cells were initialised with parameter set 1 from Table 3.1, cell death was neglected.

straint of rejecting drawn values that result in a sum that exceeds the overall cell cycle length T_{div}. Hence the individual phase distributions are only approximately log-normal. The effective joint distribution $p(T_i, T_{\mathrm{div}})$ with enforcement of $T_i < T_{\mathrm{div}}$ is of the form

$$p(T_i, T_{\mathrm{div}}) = \begin{cases} \dfrac{p(T_i)p(T_{\mathrm{div}})}{\int_0^\infty \mathrm{d}T_i \int_{T_i}^\infty \mathrm{d}T_{\mathrm{div}}\, p(T_i)p(T_{\mathrm{div}})} & T_i < T_{\mathrm{div}} \\ 0 & \text{else,} \end{cases} \tag{3.22}$$

while T_i is for example given by $T_i = T_{\mathrm{G_2M}} + T_{\mathrm{S}}$. The resulting deviations from log-normal distributions are minimal in our case.

For $\delta < \lambda$, two daughter cells are born, where δ and λ are given by T_{div} and T_{death}, respectively. The individual phases are assigned in the same way as for the mother cells. In the simulation, we assume that finite overlap of a mother cell's S-phase with the BrdU availability window results in getting BrdU-labelled immediately. This label will then be inherited upon cell division. Furthermore, it is assumed that labelling efficiency follows a Heaviside function $\Theta(t_{\mathrm{BrdU}} - t)$, i.e. BrdU is fully available until t_{BrdU}, after which it becomes unavailable immediately. Now each initial cell configuration at t_0 can be assigned a particular gate (or eventually two different gates in case of different daughter cells) at t_{m}. Normalising the respective gate counts with the overall number of cells at t_{m} leads to the desired gate frequencies. In the stochastic limit, these gate frequencies become robust with respect to stochastic fluctuations and hence can be interpreted as gate probabilities. Finally, white Gaussian noise is added to the BrdU and the DNA signal to increase the similarity between simulated and measured plots. Note that we have assumed that DNA content gets synthesised linearly upon progression through S-phase. However, in reality, DNA replication might follow more complex dynamics. At the same time, BrdU intensity is discrete in our model. A cell is either positive or negative and the irreversible transition from negative to positive happens as a "jump". The apparent continuity on this axis is created exclusively by the added noise. This simulation was implemented in MATLAB (version R2020a). In addition, the *Flow Cytometry GUI* (version 1.3.0.0) was used to visually match the plots to the experimental FACS plots, cf. [Steinberg, 2021].

Fig. 3.4 shows the simulation results for a BrdU availability window of $t_{\mathrm{BrdU}} = 0.75$ h, considering $N = 15000$ mother cells at $t_0 = 0$ and different measurement time points t_{m}, ranging from 0.5 h to 5.0 h. It can be seen that the unlabelled cells in gate vi are rapidly leaving the gate. 2.0 h after BrdU administration, there are still too few cells in gate iii to allow for quantification. Only after 3.0 h, enough cells have been collected such that reliable statements can be made about the average cell cycle speed. Cells that entered S-phase after degradation of BrdU are found at the left edge of gate iv, i.e. just below the "gap". From 3.0 h to 5.0 h, cells continue to accumulate in this gate, as their mean $\mathrm{G_1}$-phase has not yet been completed. This type of visualisation supports the experimenters in choosing their measurement time points given prior knowledge about the kinetic behaviour of their cells, e.g. from the literature or a pilot experiment.

Another important parameter is the availability of BrdU. This can be manipulated experimentally by administering BrdU several times, e.g. every 0.5 h. Fig. 3.5 shows the

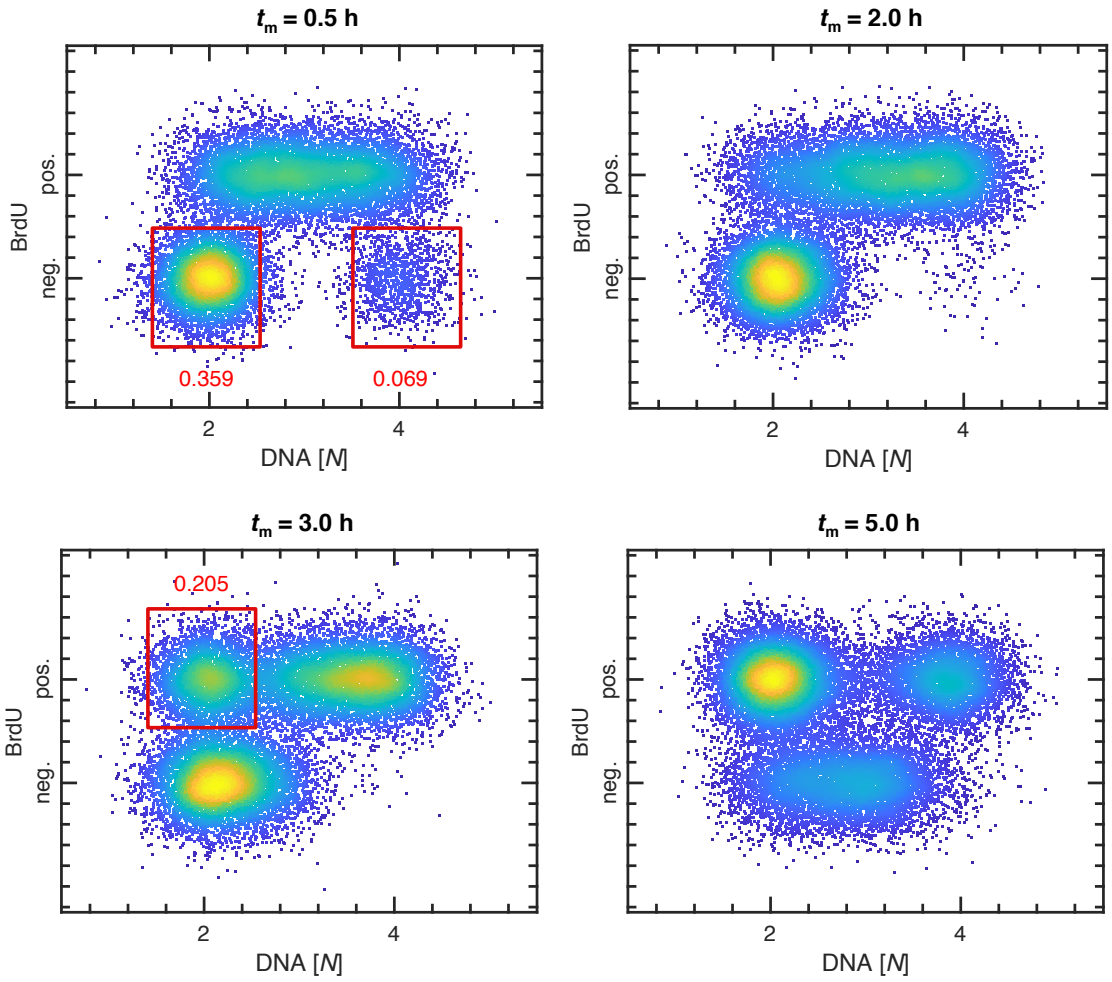

Figure 3.4: Simulated time series of BrdU/DNA profiles for different time points of analysis. Simulations are based on $N = 1.5 \cdot 10^3$ mother cells, neglecting cell death. Red gates identify those cells that contribute to the quantification procedure. Corresponding fractions result from larger simulations of $N = 10^6$ mother cells. BrdU availability was assumed to be 0.75 h. Simulation parameters are given by set 1 from Table 3.1.

BrdU/DNA profiles for a fixed measurement time of 3.0 h and different time intervals of BrdU availability. At a minimal availability of only 6 min, the "gap" is pronounced most. At an availability of 2.0 h, the labelled daughter cells in gate iii are no longer identifiable. At a permanent availability ($t_{\mathrm{BrdU}} \geq t_{\mathrm{m}}$), every cell in S-phase is reliably labelled. Since the G_2M cells in gate vi have already divided, any unlabelled cell is necessarily a mother cell in G_1-phase (gate iv). For permanent BrdU availability, this gate does not show any overlap with S-phase cells. Overall, we find that an availability between 0.5 h and 1.0 h fits our mouse data best. Therefore, we leave the availability for all subsequent simulations

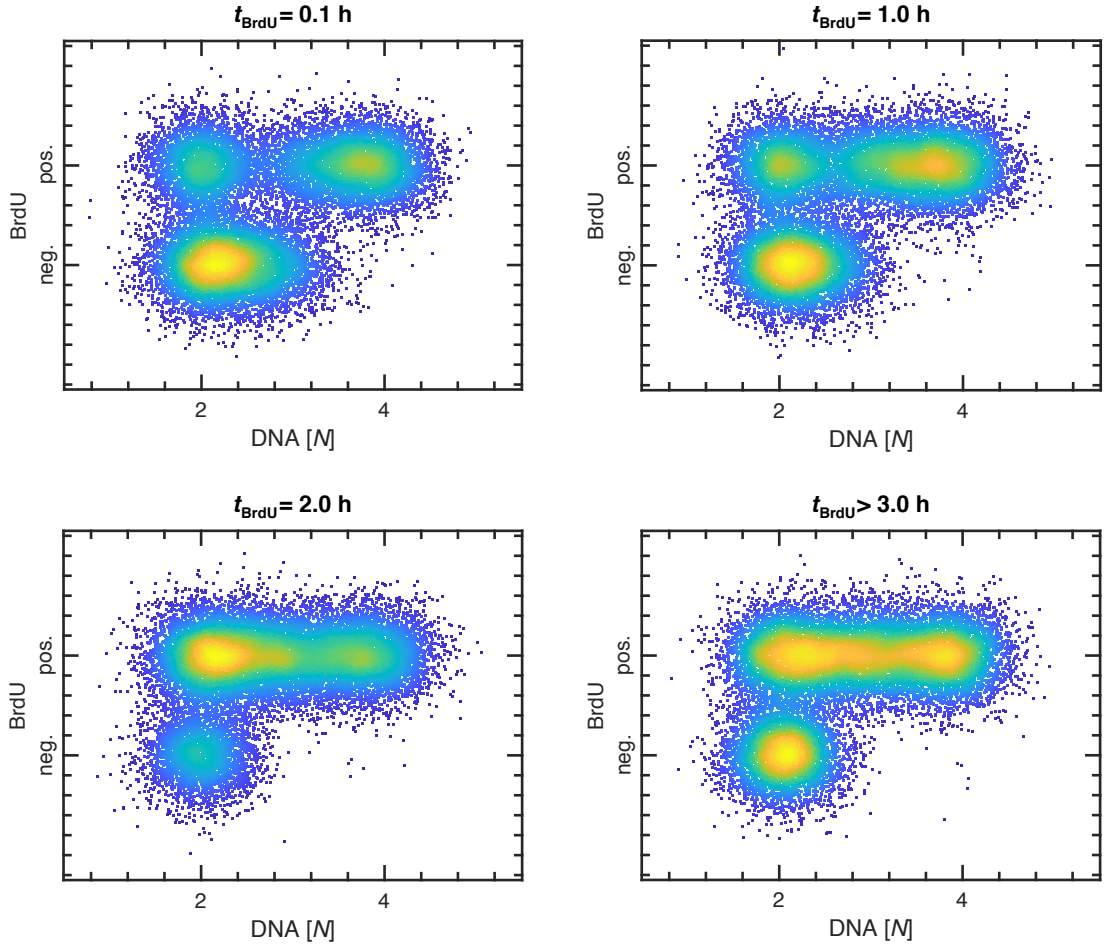

Figure 3.5: Simulated BrdU/DNA profiles for different assumptions about BrdU availability and fixed measurement time of 3.0 h post BrdU administration. The last choice of BrdU availability corresponds to unlimited availability during the measurement period. Simulations are based on $N = 1.5 \cdot 10^3$ mother cells, neglecting cell death. Simulation parameters are given by set 1 from Table 3.1.

at $t_{\mathrm{BrdU}} = 0.75$ h. Although this sounds like a relatively short time, this trend is consistent with results presented in [Matiasova et al., 2014]. In this study, the authors assessed the availability of BrdU after intraperitoneal injection in rodents, including mice. They concluded that BrdU is only fully available for about 15 min after administration to mice and then degrades rapidly. The pharmacokinetics of BrdU certainly depends on many other experimental parameters, e.g. injection dose and type of administration. However, this study reinforces our impression that BrdU is only capable of efficiently labelling cells a short time after administration.

3.3.2 Back-calculation of simulation parameters

Some approximations were made to quantify the mean division-time $\langle T_{\mathrm{div}} \rangle$. These include the assumption of a peaked G_2M length distribution and the existence of a minimum division-time $T_{\mathrm{div}}^{\mathrm{min}}$. It was also assumed that eq. 3.5 represents an accurate approximation for the relation between mean growth factor c and mean division-time $\langle T_{\mathrm{div}} \rangle$, even in the case of variable division-time distributions. Moreover, it was implicitly assumed that cell death effects do not play any major role in our estimation scheme. To determine the precision of our method, we solve the system of equations 3.21 using the simulated data shown in Fig. 3.4. The simulated gate frequencies (marked in red) have emerged from the stochastic limit based on parameter set 1 from Tab. 3.1. In addition, we applied this approach to parameter set 2, shown in the same table[3]. This second set stems from *in vivo* measurements quantified with the same method and was published in [Kretschmer et al., 2020]. Briefly, these are cell cycle parameters that origin from $CD62L^+$ OT-I memory T cells after immunisation. The total variance of the cell cycle was chosen arbitrarily here so that the CV of the distribution is 0.2, i.e. higher than the CV corresponding to parameter set 1. As with parameter set 1, this variance was then divided among the cell cycle phases according to their mean lengths.

Table 3.1 shows the quantification results as well: For both parameter sets, accurate estimates of the ground truth are obtained. Relative errors are computed with respect to the true parameter value, i.e. $\mathrm{err}(p) = |p_{\mathrm{true}} - p_{\mathrm{est.}}| \,/\, p_{\mathrm{true}}$. For the mean cell cycle length, we get relative errors of 1.7 % (set 1) and 4.1 % (set 2). The slightly worse accuracy for estimates corresponding to parameter set 2 is related to its significantly bigger CV of the division-time distribution. The estimation of the mean phase lengths, especially of the G_2M-phase, is significantly affected by the variability. In practice, the G_2M-phase is the least relevant phase in terms of contribution to the overall cell cycle length. The method thus proves to be accurate, especially for estimating mean cell cycle lengths. This is particularly important as the method was mainly designed for this purpose.

3.4 Stability analysis

Since the quantification formalism has proven successful with simulated data, several conceptual assumptions need to be revisited. It was outlined earlier that the accuracy of our method decreases with bigger variabilities in the division-time distribution. This was

[3]Simulated gate frequencies associated with parameter set 2 are given by 0.014 (gate vi, early time point), 0.430 (gate iv, early time point) and 0.361 (gate iii, late time point).

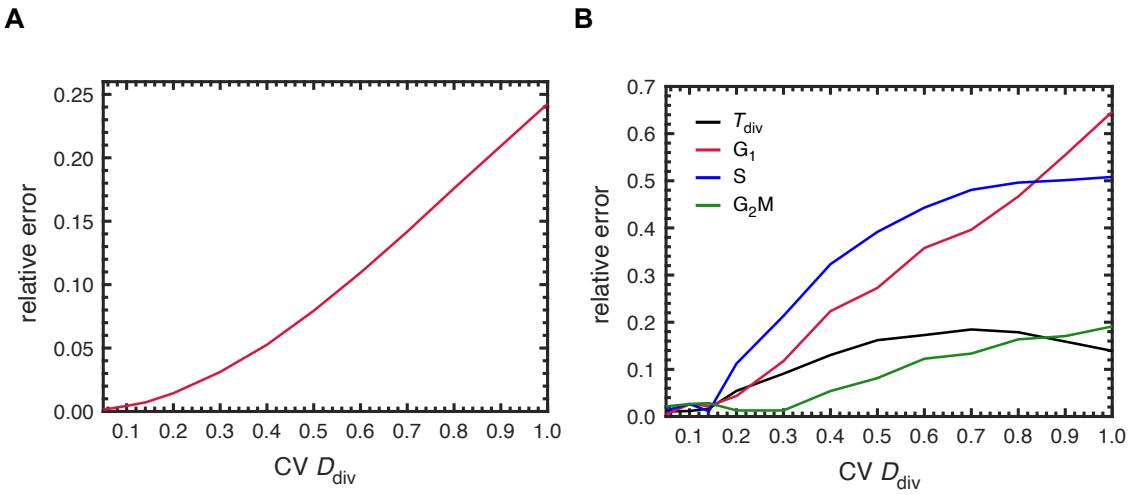

Figure 3.6: Stability analysis of the quantification procedure with respect to variability in the underlying distributions for parameter set 1 in Tab. 3.1. **(A)** Relative error of the approximative formula 3.5, relating mean growth factor and mean of division-time distribution for varying CVs of the division-time distribution and phase distributions. Errors were computed by numerically solving integral equation 3.4. **(B)** Relative error of cell cycle estimates were accessed via simulated data for varying CVs of the division-time distribution and phase distributions. As before, simulations were based on $N = 10^6$ mother cells and a BrdU availability of 0.75 h.

to be expected, especially since the equations are only exact in the case of non-variable distributions. Here we want to examine more systematically the stability with respect to variability in the underlying distributions. To this end parameter set 1 from Tab. 3.1 is used, i.e. the division-time distribution is assumed to be log-normal. First, we examine the variability regime in which formula 3.5 is a good approximation of the relation between c and the division-time distribution $D_{\mathrm{div}}(T_{\mathrm{div}})$. For this purpose, we make use of the exact relation given by eq. 3.4. The relevant integral equation for c is solved numerically, leading to a relative error of the approximative formula which only contains the mean of the overall distribution. Fig. 3.6A shows this error for increasing CVs of $D_{\mathrm{div}}(T_{\mathrm{div}})$. For CVs smaller than 0.1 the approximative formula is exact. The error non-linearly increases with increasing variability up to a CV of ~ 0.4. From there on, the error grows almost linearly with the variability.

Next, we investigate the accuracy of the calculated cell cycle parameters as a function of the variability in the underlying distributions. For this purpose, we simulate data for different variabilities in the total cell cycle length, analogous to the previous section, and back-calculate the underlying cell cycle parameters. By comparing the calculated parameters with the true parameters, a relative quantification error is obtained. Fig. 3.6B

parameter	**exponential (strong)**	**log-norm. (moderate)**	**log-norm. (strong)**
$\mu(T_{\text{death}})$	12.0 h	20.0 h	12.0 h
$\sigma(T_{\text{death}})$	12.0 h	5.0 h	3.0 h
$\mu(T_{\text{div}})$	10.3 h	9.9 h	10.7 h
$\mu(T_{\text{G}_1})$	2.9 h	2.8 h	2.9 h
$\mu(T_{\text{S}})$	5.9 h	5.6 h	6.4 h
$\mu(T_{\text{G}_2\text{M}})$	1.5 h	1.5 h	1.5 h

Table 3.2: Parameters of death-time distributions for strong exponential, and moderate or strong log-normal cell death, respectively. Shown are cell cycle estimates in the presence of these death mechanisms and for division parameters from Tab. 3.1 (set 1).

shows this error as a function of the CV of the division-time distribution. Notably, the error of the mean cell cycle length never exceeds 20 %. For extreme CVs close to one, the error even decreases. This can be explained by the fact that the total error is made up of different errors and that these can in principle cancel each other out. Even the error shown in Fig. 3.6A contributes here. The error of the G_2M-phase is of similar magnitude, although no such cancellation effects seem to occur here for extreme CVs. The errors of the S- and G_1-phase grow more rapidly with increasing variability. However, this is not relevant for the quantification of the mean cell cycle length, especially since the equations for gate iii and iv form an independent system of equations and hence do not depend on the mean G_1 or S length. Nevertheless, one should keep in mind that the individual phase lengths can be extremely error-prone in case of large variabilities. These findings also hold for parameter set 2 from Tab. 3.1 as we have already shown in [Kretschmer et al., 2020].

3.5 Cell death effects

At last, it needs to be shown that cell death has no particular influence on our quantification procedure. In the simulation, we can introduce arbitrary cell death mechanisms. In the following, we will switch on such mechanisms while keeping the same parameters for cell division, namely set 1 from Tab. 3.1. Table 3.2 shows three different possibilities of death-time distributions $D_{\text{death}}(T_{\text{death}})$, which govern the times-to-death of newborn cells. Hence, these distributions represent the uncensored distributions. We consider strong exponential, and moderate or strong log-normal cell death mechanisms with given parameters from Tab. 3.2. Log-normal cell death is conceptually the more relevant mechanism here, especially since live-cell imaging data of lymphocytes suggest heavy-tailed distribu-

tions; the major difference compared to division-time distributions seems to be the absence of a lag time, leading to more variability, see for example [Zhou et al., 2018].

The effects of cell death can alter the gate frequencies in principle and thus confound the quantification scheme. The main focus here is on a change in the marginal steady state distribution $p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta)$, which initialises experiments to measure BrdU/DNA profiles. Fig. 3.7 shows the marginal distributions $p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta)$ and $p^{\mathrm{ss}}_{\mathrm{rl}}(\lambda)$ in the presence of the above mentioned death-time distributions.

In case of exponential cell death, the steady state distribution of remaining life-time $p^{\mathrm{ss}}_{\mathrm{rl}}(\lambda)$ equals the death-time distribution $D_{\mathrm{death}}(T_{\mathrm{death}})$, see Fig. 3.7A. This is due to the memorylessness property of the exponential distribution. We have indeed shown in appendix A, that the inclusion of constant death rates into the von Foerster equation will always cancel out as soon as the equation is translated to the steady state level. Hence, this result was to be expected. Every other distribution than the exponential distribution naturally implies age-dependent death behaviour. For moderate log-normal cell death, the steady state distribution differs from the death-time distribution mainly by being shifted towards smaller ages (comparison not shown). This can be explained by the fact that cells in steady state have an average age bigger than zero and are therefore more likely to die than newborn cells. It should be mentioned that, to our knowledge, there is no solution to the general form of $p^{\mathrm{ss}}_{\mathrm{rl}}(\delta)$ in the presence of both cell division and death. For solutions considering log-normal cell death without cell division, we refer to [Dowling et al., 2005]. Nevertheless, the presence of moderate log-normal cell death has almost no effects on the analogous steady state distribution of cell division $p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta)$, see Fig. 3.7B. Finally, Fig. 3.7C clearly shows that in the case of strong log-normal cell death, only small deviations in the $p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta)$ distribution are obtained, especially for small to medium δ between about 0 h and 8 h. On the other hand, $p^{\mathrm{ss}}_{\mathrm{rl}}(\lambda)$ changes strongly compared to the corresponding distribution for newborn cells (not shown).

The finding that in the presence of cell death the distribution $p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta)$ and consequently also $p^{\mathrm{ss}}_{\mathrm{age}}(a)$ change only minimally, is of central importance, since they cannot be adjusted for. Additionally, to rule out any effects on the quantification scheme, data was simulated for all three conditions of cell death that were discussed above[4]. As previously, these simulated data were then used to calculate the underlying parameters. Table 3.2 shows the quantified parameters, with cell cycle parameters taken from set 1 from Table 3.1 as before.

[4]For exponential death we get the same gate fractions as for the absence of cell death. For moderate log-normal death the resulting gates give 0.069 (gate vi), 0.358 (gate iv) and 0.206 (gate iii). Strong log-normal death gives 0.066 (gate vi), 0.336 (gate iv) and 0.186 (gate iii)

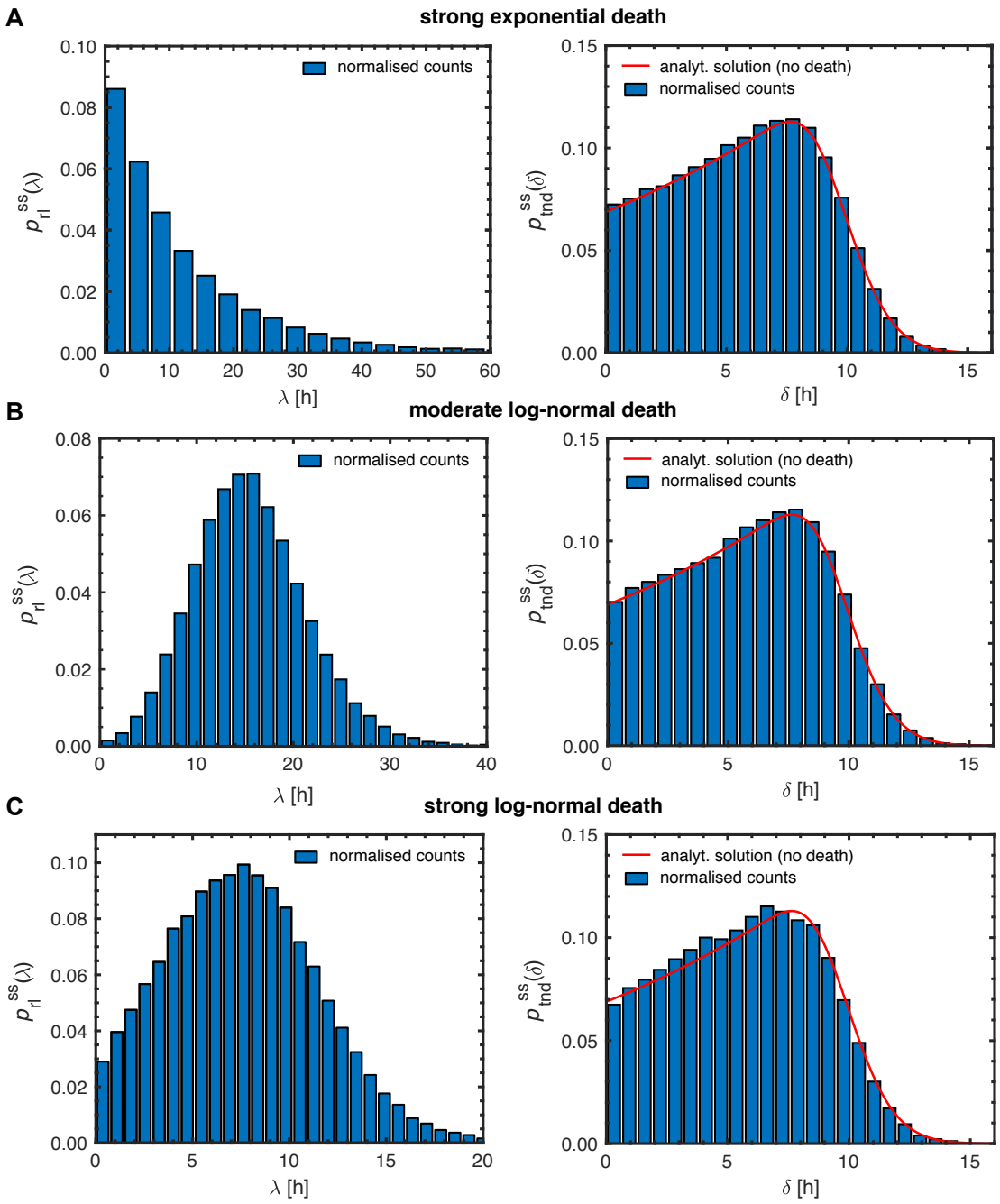

Figure 3.7: Remaining life-time (left) and time-until-next division distribution (right) during steady state in the presence of strong exponential **(A)**, and moderate **(B)** or strong **(C)** log-normal cell death. Simulations are based on division- and death-time distributions with parameters shown in Tab. 3.1 (set 1) and Tab. 3.2, respectively.

Exponential cell death, as strong as it may be, has no impact on the simulated gate frequencies, especially since the distribution of time-to-next division remains unchanged.

Hence, there is no change in the calculated parameters. Note that (exponential) cell death can of course have a great influence on the cell numbers and thus on reaching the stochastic limit of the simulation. Since we already started from a large number of initial cells, simulation parameters did not have to be readjusted here. For both moderate and strong log-normal death, we obtain deviations of less than 5 % for the mean cell cycle length and less than 10 % for the mean phase lengths.

3.6 Application to *in vivo* data

Now that all conceptual considerations have been clarified, the method can be used to quantify division speed of rapidly dividing T cells following acute infection. In this thesis, we deal with two types of T cells that differ in their TCR: OT-I T cells and P14 T cells, derived from OT-I and P14 transgenic mice, respectively. While OT-I cells form a high-affinity binding with the ovalbumin-derived peptide SIINFEKL, P14 cells bind to the GP33 epitope. The pathogens used in the experiments express one of these peptides, so their presentation induces a highly specific T cell response. In the case of OT-I cells, the wild-type pathogen must be genetically modified to additionally express SIINFEKL. This is taken into account by the additional suffix "OVA" in the pathogen's name. Here, we quantify cell cycle lengths of OT-I cells subject to *Listeria monocytogenes*–OVA (*L.m.*-OVA) and influenza-OVA infection. P14 cells were used in the context of lymphocytic choriomeningitis Armstrong (LCMV Arm) infection. All these infections are classified as acute infections. In this application, we will also briefly explain how additional kinetic properties, such as the presence of non-dividing G_0 cells or generation-resolved measurements, can be embedded in our formalism.

3.6.1 Cell cycle speed of P14 T cells during acute LCMV Arm infection

The experiments on which this subsection is based were conducted by my colleague Albulena Toska. Mice were infected with LCMV Arm and analysed at day 4.5 p.i. The experimental parameters were chosen such that the early measurement time point was at 0.5 h (two mice available) and the late one at 3.0 h (four mice available). In addition, the phenotype of the P14 cells was measured in terms of TCF1. TCF1 is a memory T cell marker and allows for classification between $TCF1^+$ memory progenitor (CMp) cells and effector-associated (EFF) cells. At day 4.5 after infection, one would expect that most of the cells are still dividing. This could indeed be confirmed by the analysis of scRNA-seq data. It was found that in this particular infection, about 90 % of the cells are still in division at day 4.5 p.i. In section 4.2, we will discuss the origin and the significance of

CMp			**EFF**		
gate iv	**gate vi**	**gate iii**	**gate iv**	**gate vi**	**gate iii**
0.393	0.010	0.317	0.259	0.010	0.495
0.373	0.027	0.289	0.251	0.008	0.449
		0.265			0.489
		0.222			0.462

Table 3.3: Measured gate fractions of P14 cells following acute LCMV Arm infection at day 4.5 p.i. The data for gates iv and vi (see Fig. 3.1B for gate notation) are based on an early measurement time of 0.5 h, while gate iii was considered at 3.0 h post BrdU injection. Phenotypic classification was performed using the memory marker TCF1. (Experiments conducted by Albulena Toska)

this result in more detail. Here we briefly show how such information can be integrated in our quantification method.

Suppose that in steady state there is a fraction of quiescent G_0 cells, q, where possible time-dependencies of q during measurement are neglected. There is no reliable way to distinguish those cells from dividing G_1 cells since both groups of cells will mix in gate iv. Let f_i be the gate fraction of gate i at the time point of analysis. We are interested in the gate fraction f_i' of cycling cells only. Hence we must adjust the normalisation N to $N' = (1-q)N$. The cell counts of the respective gates remain the same, except for gate iv, which includes non-cycling cells. This results in the following modification of the former gate fractions,

$$f_i' = \begin{cases} \dfrac{f_i - q}{1-q} & i \in \text{gate iv} \\ \dfrac{f_i}{1-q} & \text{else,} \end{cases} \tag{3.23}$$

which consistently ensures that $\sum_i f_i' = 1$.

Table 3.3 shows the raw data containing the uncorrected fractions f_i. Quiescent cells were considered by employing correction 3.23. Since we had no information about phenotype-specific quiescence behaviour, it was assumed that $q_{\mathrm{CMp}} = q_{\mathrm{EFF}}$. Since nearly all cells are in division at this early time point, we consider this approximation adequate. Throughout this work, gates iv and vi (see Fig. 3.1B for gate notation) are summarised by their mean frequencies (early time point). For each measurement of the gate iii frequency (late time point), the system of equations 3.21 is solved separately. The resulting parameters are then averaged. The standard error of the mean (SEM) is chosen as an estimate for the error of the total cell cycle length.

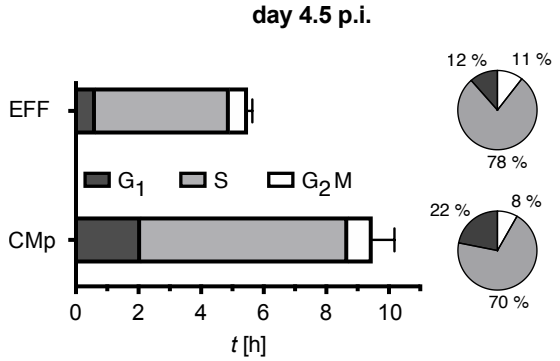

Figure 3.8: Cell cycle phase-resolved quantification results of cycling P14 T cells at day 4.5 post LCMV Arm infection. Based on the data in Table 3.3, this figure shows the mean phase lengths for CMp and EFF cells. Contributions of quiescent cells were removed by means of eq. 3.23. Standard error of the mean and composition of the overall cell cycle in terms of the individual phases are shown. (Experiments conducted by Albulena Toska)

Based on Tab. 3.3 and the correction for quiescent cells by means of eq. 3.23, mean cell cycle lengths were quantified. Fig. 3.8 shows the results at day 4.5 p.i. for CMp and EFF cells, respectively. The SEM of the overall cell cycle is depicted and the contribution of the individual phases to the overall cycling length is shown. While CMp cells on average need 9.5 h to fully complete a cycle, EFF cells only need 5.5 h. The discrepancy between these lengths and the lengths that would result if one assumes that all cells are cycling is only about 10 %. The relative G_1-phase contribution of the memory cells is larger by a factor of 1.8 compared to the effector cells. This percentage increase is mainly accompanied by a relative decrease of the S-phase, as the G_2M-phase remains short. The G_1-phase thus provides the largest percentage difference between the two phenotypes, or between rapidly and more slowly dividing cells. Nevertheless, it should be pointed out that the division speed of CMp cells is also considerably high. However, it is difficult to define a physiological limit for the maximum proliferation rate. To our knowledge, the fastest dividing mammalian cells were reported during early embryogenesis in mice. Here, average division times of 5 h must be reached in a certain time interval to explain the cell numbers counted from histology data, cf. [Snow, 1977]. In very specific spatial and temporal ranges, certain cells can even complete their cell cycles in 2 to 3 h on average, see [Mac Auley et al., 1993; Snow, 1977]. However, it should be added that these special cell divisions only involve DNA synthesis, not protein synthesis. The finding that memory cells divide more slowly than effector cells was already indicated in [Buchholz et al., 2013] and could be confirmed in [Kretschmer et al., 2020] using this same quantification method.

These cell cycle quantifications will be of great relevance later in chapter 4, especially since they represent an important component for the mathematical modelling approach. This is because these results define kinetic differences between memory and effector cells. Together with the stochasticity of the model, they play an important role in explaining the fate mapping data shown later.

3.6.2 Cell cycle speed of OT-I T cells during acute *L.m.*-OVA infection

Now the analogous situation for the *L.m.*-OVA case shall be considered. Since data at later time points are available for this infection, this system is suitable for studying the contraction phase. In particular, this means that we cannot further neglect the time-dependence of proliferation in the form of a time-averaged rate. Therefore, experiments were conducted by Albulena Toska in complete analogy to the previous section, now including the time point of 8 d p.i. in addition to the time point of 4.5 d p.i. Phenotypic classification of OT-I cells into CMp and EFF cells was performed using the memory marker CD62L. For this system, we even possessed data on phenotypic quiescence behaviour. Day 4.5 p.i. yielded $q_{\mathrm{CMp}}(4.5\ \mathrm{d}) = 2\ \%$ and $q_{\mathrm{EFF}}(4.5\ \mathrm{d}) = 5\ \%$, while day 8 gave $q_{\mathrm{CMp}}(8\ \mathrm{d}) = 60\ \%$ and $q_{\mathrm{EFF}}(8\ \mathrm{d}) = 73\ \%$. As before, we refer to section 4.2 for a comprehensive derivation of those results. Here, the previous thoughts from subsection 3.6.1 about the inclusion of quiescent cells become highly relevant, since by day 8 p.i. most cells have stopped dividing. We do not want to average kinetic parameters over non-cycling cells because those cells will be taken into account manifestly in terms of our model topology in chapter 4. Also, we would like to point out that in the case of time-dependent rates, we assume that the steady state condition is always satisfied. This is to be expected, especially since the dynamics of cell division occur on a much larger time scale than the transient phase to equilibrium. This is especially true for rapidly dividing cells, a comparatively slowly changing rate, and sufficiently asynchronous cell division. These criteria are always met in our *in vivo* setting.

Results for the average phase lengths are obtained using the data from Tab. 3.4 and phenotypic fractions of quiescent cells. Fig. 3.9A shows that the mean cell cycle lengths are given by 8.7 h for the CMp and 5.8 h for the EFF cells at day 4.5 p.i. Within the uncertainties, this corresponds to the calculated values during LCMV Arm infection (see previous subsection). During *L.m.*-OVA infection, the G_1-phase takes up a larger proportion of the total cell cycle length compared to the LCMV Arm case. For CMp cells, the G_1-phase, on average, is even larger than the S-phase. In addition, CMp cells divide more slowly than EFF cells. Again, the phenomenon that the slower CMp cells are characterised by a rel-

CMp			EFF		
gate iv	**gate vi**	**gate iii**	**gate iv**	**gate vi**	**gate iii**
0.597	0.073	0.274	0.506	0.048	0.411
		0.195			0.361
		0.238			0.480
		0.245			0.462
0.736	0.029	0.074	0.825	0.011	0.074
0.653	0.007	0.084	0.724	0.011	0.095
0.665	0.028	0.053	0.747	0.011	0.063
		0.082			0.092

Table 3.4: Measured gate frequencies of OT-I cells following acute *L.m.*-OVA infection at day 4.5 (top) and 8 p.i. (bottom). The data for gates iv and vi are based on an early measurement time of 0.5 h, while gate iii was considered at 3.0 h post BrdU injection. Phenotypic classification was performed using the memory marker CD62L. (Experiments conducted by Albulena Toska)

atively longer G_1-phase is evident. As before, the G_2M-phase occupies equal proportions between CMp and EFF cells.

Meanwhile, Fig. 3.9B shows the same parameters for the later time point of 8 days p.i. For the CMp cells, we obtain a mean cell cycle length of 11.9 h, while for the EFF cells we again get a shorter length of 8.9 h. Notably, both CMp and EFF cells slow down their proliferation rates compared to day 4.5 p.i., as would be expected given the lack of stimulatory signals on day 8 p.i. Again, the speed differences must be due to slowing down, as cell cycle dropout effects have already been taken into account. It appears as if EFF cells slow down more than CMp cells (cell cycle length ratios compared to day 4.5 p.i. given by 1.6 vs. 1.3). Assuming that the most significant difference between day 4.5 and 8 p.i. is antigen availability, this result was not expected: in [Kretschmer et al., 2020] it has been shown that the absence of antigen causes stronger effects on the CMp than the EFF cell cycle. However, this could be due to system-specific effects as there was no replicating pathogen present, and antigen was administered only once. Also, other differences than antigen availability may play a significant role when comparing day 8 with day 4.5 p.i., for example, cytokine concentrations, which are considered to have a bigger impact on effector cells.

Another striking result is the phase composition of the cell cycle at day 8 p.i. The largest proportion of the cell cycle is accounted for by the S-phase, while the G_1-phase does not contribute significantly (~ 5 %). The G_1-phase seems to be even shorter than the G_2M-

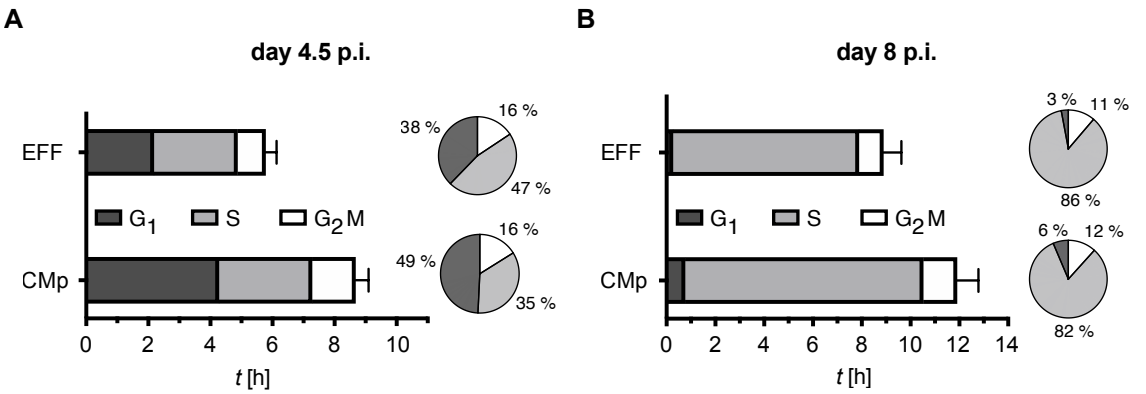

Figure 3.9: Cell cycle phase-resolved quantification results of OT-I T cells at day 4.5 **(A)** and day 8 **(B)** post *L.m.*-OVA infection. Mean phase lengths for CMp and EFF cells are shown. Gate frequencies in Tab. 3.4 have been corrected for contributions from quiescent cells. Standard error of the mean and individual phase contributions to the overall cell cycle are shown. (Experiments conducted by Albulena Toska)

phase, which is very unusual at first. We suspect that an overcorrection of the non-cycling cell contributions has taken place. If we assume that all cells are in division, then we obtain average cell cycle lengths of 30.0 h for the CMp cells and 33.7 h for the EFF cells. The corresponding G_1 fractions are then given by 61 % (CMp) and 70 % (EFF), since the cells in G_0 also contribute here. This illustrates once again that although such a correction is necessary, it also introduces additional variability, especially since the data for the G_0 proportions come from other mice than the actual gate measurements. In this sense, the mean division-time results in Fig. 3.9B may be slightly or even moderately underestimated. That is, cells may cycle even slower.

The cell cycle speed quantifications presented in this subsection will be used further in chapter 5. They not only demonstrate differences in phenotypic kinetics but also allow the use of dynamic proliferation rates. Moreover, they are absolutely necessary for the identification of death rates during contraction phase, as will be shown in chapter 5.

3.6.3 Cell cycle speed of OT-I T cells during acute influenza-OVA infection

As a last application example, we would like to show the special case of adding generation information to the data. The results in this section are based on data from Stephen J. Turner's lab (Monash University) obtained by Jasmine Li. The experimental setup differs only slightly from that shown in Fig. 3.1A: Before the OT-I cells are transferred, they are prepared to incorporate CellTrace^{TM} Violet (CTV). Similar to carboxyfluorescein

succinimidyl ester (CFSE), this is a fluorescent dye. During cell division, this dye dilutes and thus enables the assignment of cells to a specific generation according to the number of processed cell divisions. Typically, up to eight cell divisions can be tracked through flow cytometric analysis, cf. [De Boer et al., 2005; Heinzel et al., 2017]. In this way, generation-resolved BrdU/DNA profiles of OT-I cells from the spleen and mediastinal lymph nodes (MLNs) were collected after intranasal influenza-OVA infection at 2.5 days p.i. Again, 0.5 h and 3.0 h were chosen as early and late time points for FACS analysis after BrdU administration, assuming that all cells are dividing.

First, we will show how generation-dependent cell cycle lengths are calculated based on the experiment described above. For simplicity, the G_2M-phase shall be neglected, especially since it has already been established before that this phase hardly contributes to the total cell cycle length in the context of rapidly dividing $CD8^+$ T cells. The extension of the approximate formula 3.12 by the generation information and for neglected G_2M-phase gives

$$\Pr(\text{gate iii})^{(j)} n^{(j)} \approx 2c^{(j-1)} t_\mathrm{m} n^{(j-1)} \quad (j > 1). \tag{3.24}$$

Here, $n^{(j)}$ is the absolute number of cells in generation j at time t_m. $c^{(j)}$ denotes the constant growth rate of cells in generation j. The generation-dependent growth rates can be converted into mean division times by using eq. 3.5. By definition, the undivided cells belong to generation zero. The relation in eq. 3.24 is obvious because the number of cells in gate iii of generation j depends on the cell cycle speed of the previous generation $j-1$. In the limit of large cell numbers, we have $\Pr(\text{gate iii})^{(j)} = f^{(j)}_{\text{gate(iii)}}$, which is measured directly. The quotient $n^{(j)}/n^{(j-1)}$ is also available from the data, since it equals $f^{(j)}/f^{(j-1)}$, while $f^{(j)}$ denotes the fraction of cells in generation j among cells across all generations. This gives us a scheme of equations of the form

$$\begin{aligned}
f^{(1)}_{\text{gate(iii)}} \frac{f^{(1)}}{f^{(0)}} &= 2c^{(0)} t_\mathrm{m}, \\
f^{(2)}_{\text{gate(iii)}} \frac{f^{(2)}}{f^{(1)}} &= 2c^{(1)} t_\mathrm{m}, \\
&\vdots \\
f^{(M)}_{\text{gate(iii)}} \frac{f^{(M)}}{f^{(M-1)}} &= 2c^{(M-1)} t_\mathrm{m}, \\
f^{(>M)}_{\text{gate(iii)}} \frac{f^{(>M)}}{f^{(M)}} &= 2c^{(M)} t_\mathrm{m},
\end{aligned} \tag{3.25}$$

where M denotes the last generation up to which cells are reliably classifiable. The division speed in the last equation of system 3.25 becomes uninformative since the corresponding

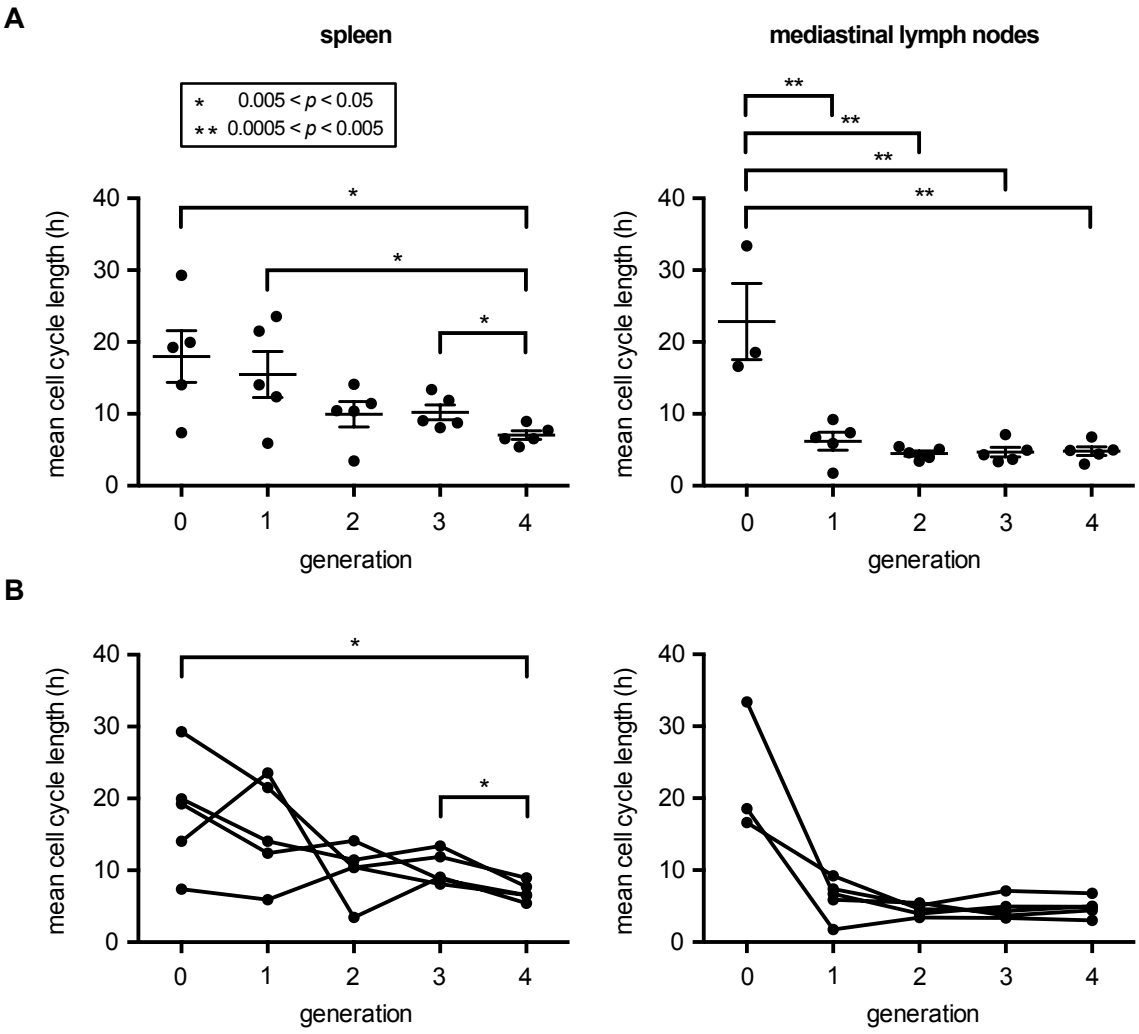

Figure 3.10: Generation-resolved cell cycle quantification results of OT-I T cells, 2.5 days after influenza-OVA infection. **(A)** Generation-dependent cell cycle lengths for cells harvested from the spleen (left) and mediastinal lymph nodes (right). The individual results per mouse, the mean of each generation, and the corresponding standard error are shown. Unpaired, two-tailed *t*-tests were used to decide whether mean differences between generations were significant. All significant differences are shown, see legend for *p*-value thresholds. **(B)** The same conditions as in (A), but here a paired *t*-test was used instead, assuming the data are not independent. (Experiments conducted by Jasmine Li and Stephen Turner)

data is only available for all generations bigger than M, denoted by $(> M)$. With a suitable choice of gating strategy, both factors of the left-hand sides of the equations are easily accessible. Hence the equations fully decouple and the $c^{(j)}$ can be calculated for all $j < M$.

In our case, we had $M = 5$ such that division speed up to generation four could be quantified based on five mice. Fig. 3.10A shows the results for spleen and MLN cells, respectively. Here, per-mouse quantifications are shown, together with the generation mean as well as its SEM. As mentioned before, generation zero cells correspond to those cells that have not yet divided at time t_{m}. All significant differences between those means are indicated, using unpaired t-tests. If we compare the spleen with the MLNs, we notice that the cells in the spleen generally cycle more slowly on average. In both organs, the first cell division is the slowest and the most variable, whereby this effect is most pronounced in the MLNs. While in the spleen it appears as if the cells divide faster with ascending generation, such an effect is not recognisable in the MLNs. Fig. 3.10B shows the same quantifications, whereby the assumption of statistical independence of the results within the respective generations was not made. Therefore, the respective mouse membership was additionally plotted for each set of results. A paired t-test was used to test for differences between generations. Even within a mouse, a downward trend in cell cycle duration across generations is now recognisable, but there are many exceptions to this, especially in the spleen between generations one and three.

The outlined results can be interpreted in different ways. To begin with, the observation that the first cell division is the longest corresponds exactly to our expectations. This has already been very well documented for T cells, see [Deenick et al., 2003; Lee et al., 2008]. If we first consider spleen cells only, the most direct hypothesis would be that cells gain speed during their first divisions. Alternatively, there could be high variability in cell cycle length, which is then inherited upon division. This would also lead to slower cells being found in smaller generations. A third possible explanation arises if one assumes that cells in lower generations have not yet undergone many cell divisions at the time of measurement because they were recruited late into the immune response. Then, one would further conclude that these late-recruited cells divide more slowly compared to cells that have been recruited earlier. At the same time, it could be true for the MLN cells that they divide rapidly despite late recruitment. Here one could speculate that this could be related to differences in antigen availability between spleen and MLNs. Furthermore, when looking at the results per mouse, one notices that the cell cycle lengths differ systematically between the mice. This could be due to variations in infectious doses, for example.

These results will not be used further in the course of this work. Nevertheless, they represent an interesting application example for our formalism. In particular, this application illustrates the flexibility of the presented method and its added value for the formulation of biological hypotheses, which can then be verified or falsified in subsequent experiments.

Chapter 4

Mechanistic insights into CD8$^+$ T cell responses to acute infection

In this chapter, we address the central question of this thesis: how do naïve T cells respond to acute infections in terms of their differentiation behaviour, and what are the dominating mechanisms that drive the observed variability at the single-cell level? We approach this question in an unbiased manner by analysing scRNA-seq data from CD8$^+$ T cells subject to acute infection. We conclude possible developmental trajectories, which then form the basis of data-driven mathematical modelling. Models are fitted to a variety of informative single-cell data. An important component of this data is given by population data derived from single naïve CD8$^+$ T cells, so-called *single-cell fate mapping* data. The *in vivo* data used in this work are based on two different acute infection models, namely *Listeria monocytogenes*–OVA (*L.m.*-OVA) infection and lymphocytic choriomeningitis Armstrong (LCMV Arm) infection. Our working hypothesis is that different acute infections obey the same model topology, whereby the underlying parameters can differ arbitrarily. Therefore, we never combine data across different infections when estimating parameters, but only transfer conceptual insights. In this chapter, model topologies are concluded from scRNA-seq data, subject to *L.m.*-OVA infected mice, while mathematical modelling is performed based on LCMV Arm data.

4.1 Extraction of model topologies

In this section, we conclude possible differentiation pathways from the analysis of single-cell transcriptomes of OT-I T cells subject to infection with *L.m.*-OVA. These bacteria lead to an acute infection of the host and, as they grow intracellularly, induce increased activation of the CD8$^+$ T cell pathway of the immune system. Our data are based on the droplet-based technology of 10X Genomics and include the time points of day 4.5, 8, and 12 after infection. All experiments were planned and conducted by my colleague Albulena Toska. As our major analysis tool, we used the Python package SCANPY, see [Wolf et al., 2018] and refer to section 2.2 for more technical details on the analysis and working flow.

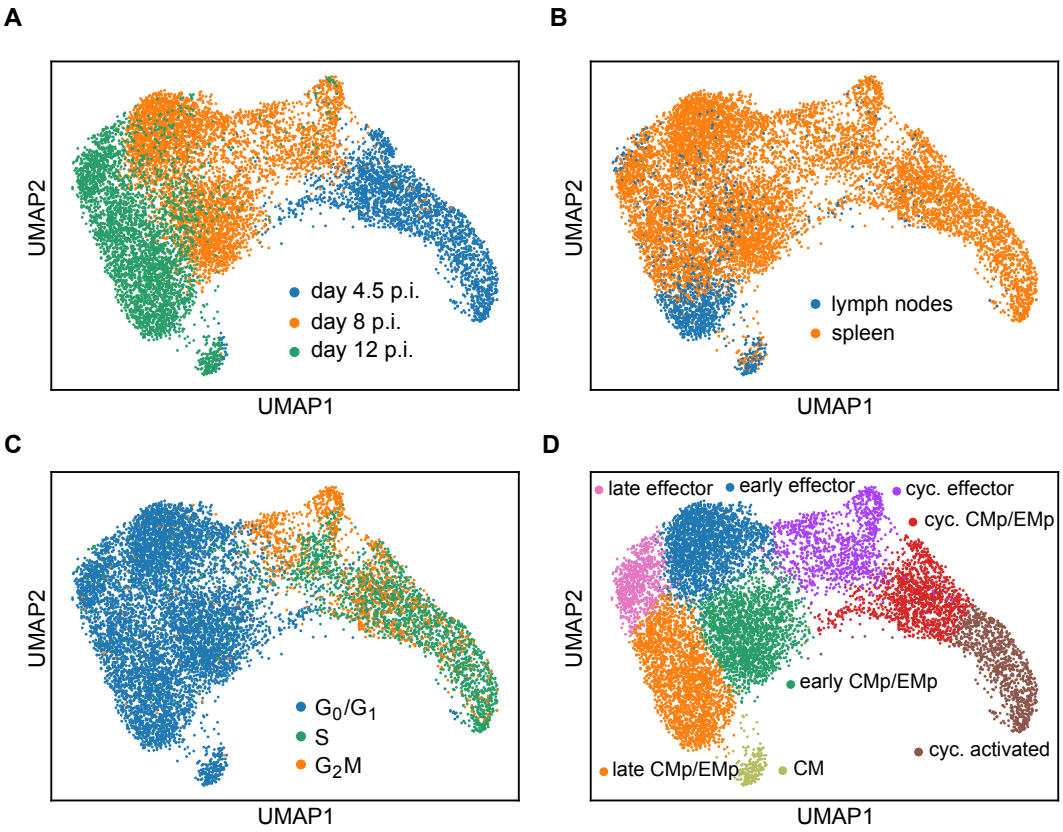

Figure 4.1: Uniform manifold approximation and projection (UMAP) of the data, indicating the individual time points after infection **(A)**, the respective organs **(B)**, the cell cycle stage **(C)**, and cluster annotations **(D)**. Clustering was performed using the Leiden algorithm. Annotations were concluded through comparison with the expression of known marker genes and by differential expression testing. Here, central memory precursors are denoted with *CMp* and effector memory precursors with *EMp*, respectively.

4.1.1 Dimensionality reduction and overview

We begin with an overview of the data, referring to the methods section 2.2 for all necessary details regarding the preprocessing. To this end, we use the uniform manifold approximation and projection (UMAP) method, which provides a descriptive visualisation of the data through robust dimensionality reduction, see [McInnes et al., 2018]. Fig 4.1A shows the resulting UMAP embedding with the batch label being plotted. The individual time points are lined up from right to left, with little overlap between the time points day 4.5 and 8, while day 8 and 12 have a more significant intersection in the embedding. It should be noted here that no batch correction was performed, especially since all batches are based on the same experiment and the same experimenter. Only fractions of mitochondrial gene counts were regressed out within the framework of a linear model. The fraction

of mitochondrial genes is a measure of cell stress or cell death. Unlike conventional mRNA, mitochondrial mRNA can be kept in the cell even in the case of porous cell membranes. In the course of the experiment, different waiting times occurred during the preparation of the batches, which likely affected the proportion of mitochondrial genes at the time of analysis. Fig. 4.1B shows the distribution of the cells with respect to the organ from which they originated. While the majority of the cells stem from the spleen, on day 12 p.i. there is a considerable proportion of lymph node cells. This is because only a few cells could be obtained from the lymph nodes at earlier time points experimentally. The classification of the cells concerning their cell cycle stage is depicted in Fig. 4.1C. It can be seen that at the early stage almost all cells are in division. At day 8 p.i. this is only true for a small part, while on day 12 p.i. almost all cells have become quiescent. Fig. 4.1D shows the annotated Leiden clusters subject to a resolution parameter of $r = 0.8$, while the concrete choice of annotations is justified below. For the given resolution, the cycling cells break down into three clusters that contain all relevant phenotypes as central memory precursors (CMp), effector-memory precursors (EMp), and effector cells. The center of the embedding is characterised by early and non-cycling CMp, EMp, and effector cells, while the left-hand side of the UMAP contains their corresponding matured subsets. Here, the terminology early vs. late is mostly based on the observation that day 8 cells are on average less differentiated compared to cells at day 12 p.i. Interestingly, we find a small disconnected cluster that mainly contains cells from the late lymph node, exhibiting a strong central memory (CM) signature.

To justify the previously annotated Leiden clusters, we focus on the expression of some exemplary phenotypic marker genes, shown in Fig. 4.2A and B. Fig. 4.2A shows the log-expression of known memory genes, while Fig. 4.2B depicts important effector genes. The gene *Dapl1* is highly co-expressed with the more prominent memory marker *Sell* (related to the surface marker CD62L). Overall, the expression plots imply a vertical phenotypic gradient. This gradient even seems to hold for the cell cycle clusters based on the expression of *Slamf6* and *Cx3cr1*, although the cell cycle signature is dominating the overall transcriptome. When testing for differentially expressed genes across different Leiden clusters, this impression becomes more sound (results not shown). Given the annotations from Fig. 4.1, the dot plot shown in Fig. 4.2C confirms the choice of annotations: The early activated cluster differs from the other clusters regarding the expression of T cell activation markers. Interestingly, a slightly increased activation signature is also found in the dividing CMp/EMp cells. Both the cycling activated cluster and the cycling CMp/EMp cluster are characterised by a memory signature. The cycling effector cells on the other hand are clearly distinguished by an even stronger expression signature of genes as *Klrg1*, *Cx3cr1*, *Gzma*, and *Gzmb*. The last two genes are also expressed by memory cells. Together with

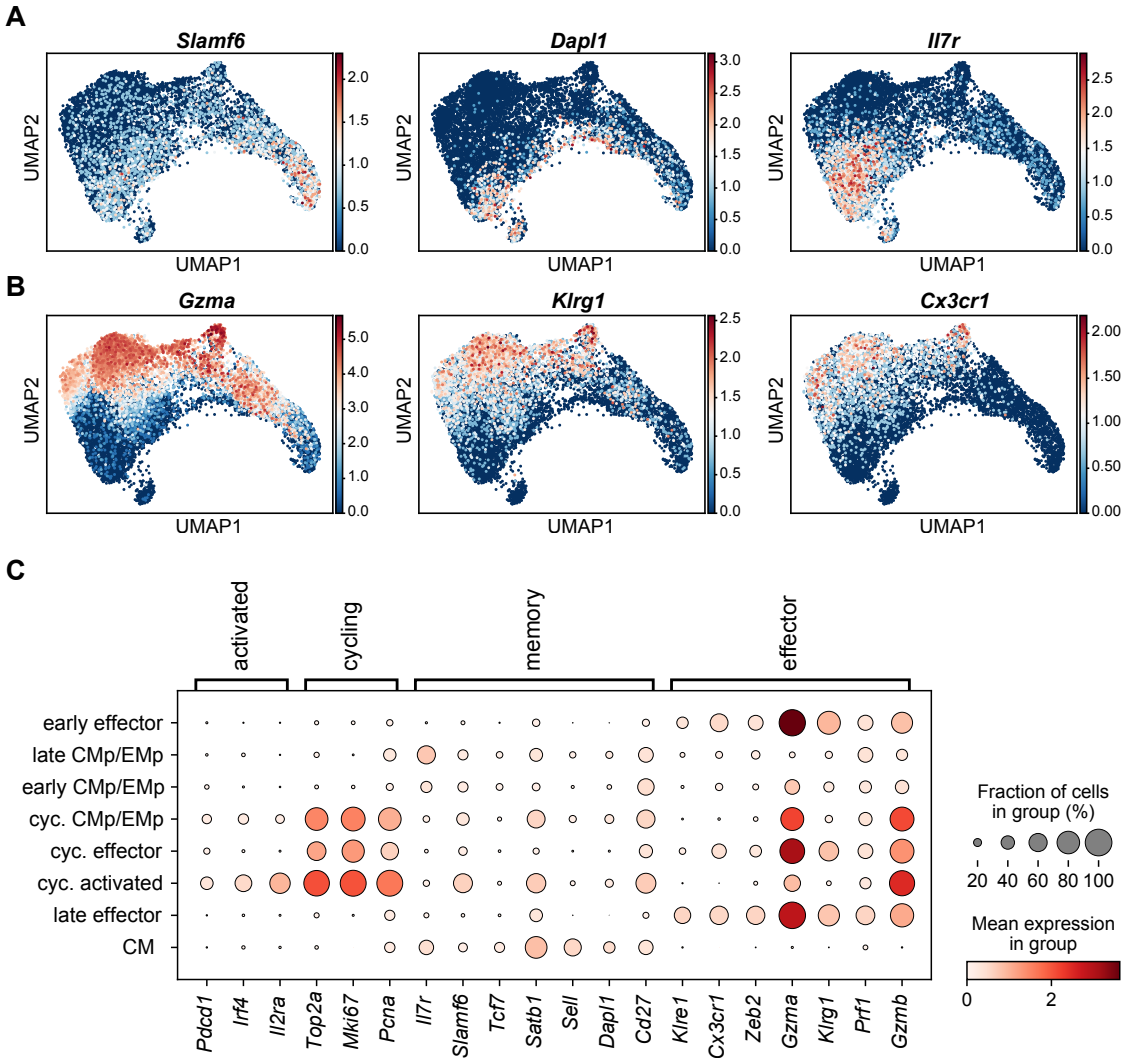

Figure 4.2: Log-expression of exemplary memory marker genes **(A)** and effector genes **(B)** mapped into the UMAP embedding. While the upper part of the UMAP is exclusively occupied by effector cells, the lower part is mainly occupied by memory and effector-memory cells. The phenotypic gradient thus runs from bottom to top. **(C)** Dot plot showing the expression of chosen genes related to $CD8^+$ T cell activation, cell cycle activity, and memory or effector signatures across the annotated Leiden clusters. Expression is based on unnormalised log-counts.

the observation that all cells express *Prf1*, this indicates that memory cells might also possess killing capacity, as it has been shown for instance in [Barber et al., 2003]. Early and late CMp/EMp cells can mainly be identified by their higher levels of *Cd27*, as well as their expression of *Il7r*. In particular, *Il7r* expression seems to increase strongly as the memory-like cells continue to mature from day 8 to day 12 p.i. As expected, the CM cells

from the lymph nodes exhibit the strongest memory signature. We observed that typical effector marker genes such as *Gzma* generally lead to good expression signals in scRNA-seq data, which makes the early and late effector cells identifiable more easily. Here the late effector cells at day 12 p.i. show a reduced *Gzma* expression compared to the day 8 effector cells. At day 12 p.i. the effector cells are subject to cell death. The decreased *Gzma* expression may reflect a reduced functionality and hence a reduced killing capacity of the cells through apoptosis. On the other hand, downregulation of *Gzma* at day 12 p.i. might also be linked to differentiation, which will be addressed in section 4.1.2. In section 4.2, we will provide arguments why we think that the cells that have been classified to be in G_0/G_1-phase are predominantly in G_0-phase.

Next, we utilised principal component analysis (PCA) to get a quantitative understanding of the various sources of heterogeneity in the data, cf. [Pearson, 1901]. Fig. 4.3 depicts the data after the transformation projected onto the first two principal components. Fig. 4.3A shows a line-up of the individual time points along the direction of biggest variance (PC1). Looking at the cell cycle phases (Fig. 4.3B), we can identify cell division as the major source of variance in the data. Cell cycle activity naturally correlates with the number of expressed genes (Fig. 4.3C). Although the number of total counts and thus the number of expressed genes can often be related to technical variance, we can confirm our previous statement by looking at the loadings of PC1 (not shown), which exhibits many cell cycle genes such as *Mki67*. The second biggest source of variance is given by phenotypic differences between the cells. This can be concluded from the loadings underlying PC2, which show typical immunological marker genes as *Gzma* and *Gzmb* among the top ten positively signed loadings. The negative loadings are mainly given by ribosomal genes which commonly dominate the signatures of memory cells. Fig. 4.3D shows the fraction of ribosomal gene counts per cell, thereby defining a phenotypic gradient.

4.1.2 Trajectory inference

Now we want to establish lineage relationships between the clusters just presented, which have already been identified with known phenotypes. To this end, we now present some methods that approach this fundamental question in an unbiased way. For conceptual as well as technical details regarding those methods we refer to subsection 2.2.3.

Inferring developmental pathways using RNA velocities

To infer developmental trajectories, we utilised RNA velocities, which were established in [La Manno et al., 2018]. This method takes advantage of the fact that cell differentiation

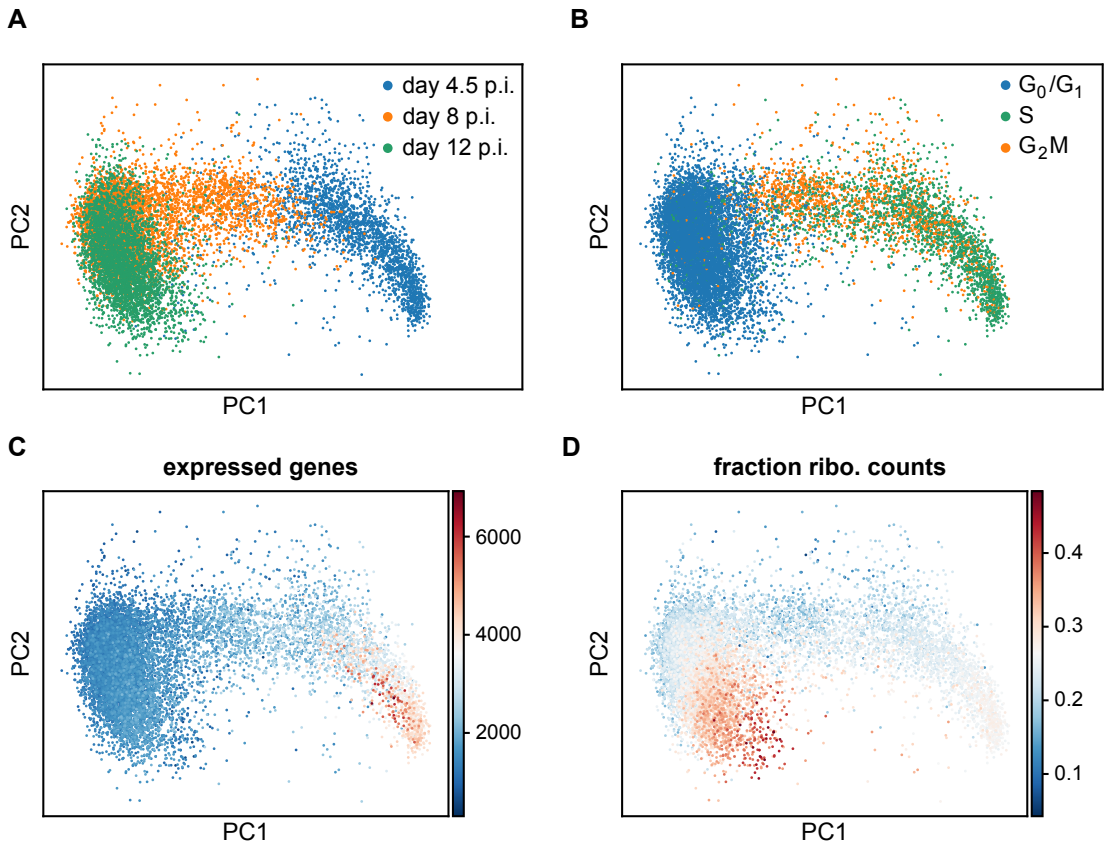

Figure 4.3: Principal component analysis (PCA) showing the time points **(A)**, the cell cycle class labels **(B)**, the number of expressed genes per cell **(C)**, and the fraction of ribosomal gene counts per cell **(D)**. Cell cycle activity represents the largest contribution to the overall variance in the data (variance ratio $\sim$11 %). The second-largest source of variance is given by phenotypic differences between the cells, with high levels of ribosomal gene counts being typical for memory cells (variance ratio $\sim$4 %).

processes occur on the same time scale as the life cycle of mRNA molecules. Based on the ratio of spliced to unspliced mRNA counts, each cell can be assigned its first time-derivative, i.e. its velocity, with respect to any expressed gene. Considering all genes and projecting the overall velocity into an embedding of choice, one thus obtains an interpretable extrapolation of the cell's state into its near future. Using the dynamic velocity model that was put forward in [Bergen et al., 2020], we computed RNA velocities for our data. This extended model captures the whole transcription dynamics and avoids the strong steady state assumption of the original model.

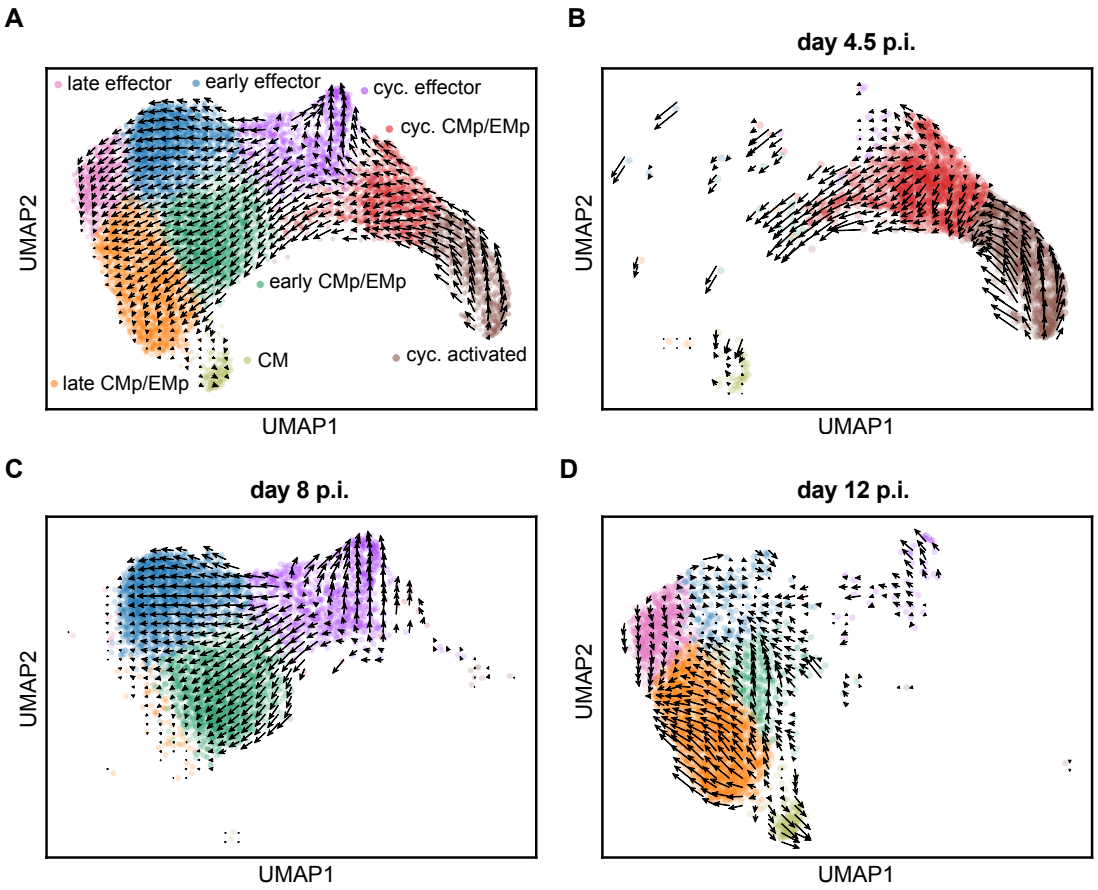

Figure 4.4: RNA velocities based on the dynamic velocity model for all cells **(A)** and the individual time points **(B, C, D)**. Velocities per time point were computed separately and projected into the joint embedding shown in (A).

Fig. 4.4A shows the velocities projected into the previous UMAP embedding. It can be seen that cells straightforwardly progress through the cell cycle clusters, thereby losing memory potential, e.g. in terms of *Slamf6* expression. Once the cells stop cycling, they move vertically with respect to the phenotypic gradient, i.e. they seem to maintain their phenotype. The left edge of the UMAP, where the late effectors and the late EMp cells intersect, we interpret as cell death. The dynamics of the small fraction of CM cells remain largely unresolved. Overall, it is remarkable that the cell cycle takes up a large amount of the overall dynamics. This is also reflected in the fact that the actual velocity genes, i.e. the genes that can in principle contribute to the vector field shown, contain a large number of cell cycle genes in addition to important regulatory genes for differentiation (velocity genes not shown).

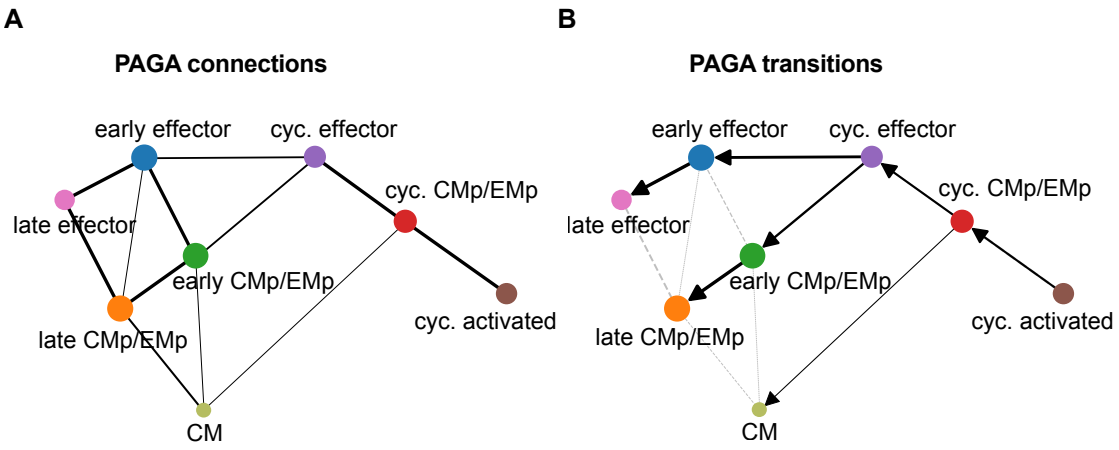

Figure 4.5: Partition-based graph abstraction (PAGA) mapping connectivities **(A)** and transitions **(B)** between annotated Leiden clusters. PAGA graphs are based on a threshold of $\alpha = 0.05$, while the transition matrix was gained by utilising RNA velocities computed from the dynamical model. Dashed lines indicate possible transitions that can neither be excluded nor confirmed.

Computing the velocities separately per batch and mapping them into their joint UMAP embedding, leads to the results shown in Fig. 4.4B, C, and D. Fig. 4.4B and C could be interpreted in such a way that the memory precursor cells exit the cell cycle sooner than the effector cells. Moreover, at day 8 p.i. almost all cells that are still cycling are characterised by a strong effector signature. The only major discrepancy with regard to the simultaneous calculation from A is found on day 12 p.i. While no differentiation takes place in the velocity calculation including all cells, progressive differentiation can be recognised in the individual day 12 computation (Fig. 4.4D). The terminal points of differentiation most likely indicate cell death. The reason for this discrepancy could be that at earlier time points than day 12 p.i., cell cycle effects mask the possibly more subtle effects of differentiation. At day 12 p.i., cell cycle genes are no longer found in the velocity genes, which could mean that the transitions then become visible. However, this is difficult to assess because differentiation and cell division are closely linked biologically. For this reason, we do not think that simply regressing out cell cycle effects is a meaningful approach for this data set. In section 4.2, we give another reason why we advise against this strategy.

Mapping transitions between clusters with PAGA

Partition-based graph abstraction (PAGA) addresses our question even more precisely, as it directly considers relationships between whole (Leiden) clusters, see [Wolf et al., 2019]. PAGA offers the possibility of mapping connectivities and transitions between the clusters. The latter application option uses the transition matrix from the (dynamic) velocity model. Applying PAGA on our data, we get the results shown in Fig. 4.5, which provide both the PAGA connectivities and the transitions for a rather small threshold parameter of $\alpha = 0.05$. Here, α sets a threshold, below which all edge connections are suppressed. A small choice for α thus also takes into account edges of lower connectivity to ensure that we do not miss possibly relevant transitions here. In terms of connectivities (Fig. 4.5A), the first notable feature is that the cell cycle super-cluster, as well as the quiescent cell super-cluster, are strongly connected within each super-cluster. This agrees well with the UMAP representation of the data, except for the connection cyc. CMp/EMp - CM, which is not present in the two-dimensional UMAP visualisation. On the other hand, this connection is found to be particularly weak by PAGA. Probably the biggest conceptual difference between PAGA and the UMAP visualisation is given by the absence of the connection between the cyc. CMp/EMp and early CMp/EMp clusters. This connection is clearly visible in the UMAP. Looking at the transitions, the overall picture is very similar to the RNA velocities projected into the UMAP, up to the discrepancies mentioned above. In particular, despite increased connectivity between quiescent cells of mixed phenotypes, there are only transitions between clusters of the same phenotype found by PAGA. Interestingly, PAGA shows a transition from cyc. CMp/EMp to the CM cluster. However, this transition is classified as very weak and disappears for default thresholds of $\alpha = 0.1$.

Inferring fate maps using CellRank

Considering the uncertainties of RNA velocities and taking into account the stochasticity of differentiation processes, the tool *CellRank* allows to analyse possible fate decisions more confidently, see [Lange et al., 2020]. Although this method uses RNA velocities[1], it avoids over-interpreting them and thus forms an important complement. Fig. 4.6 shows the results of CellRank applied to our data set. Fig. 4.6A depicts the three possible terminal states that were manually selected. These include late effectors, late CMp/EMp cells, and CMs. The highlighted cells represent cells most representative of their cluster

[1]Note that conceptually, CellRank only depends on a kernel in order to derive the transition matrix. This kernel does not necessarily need to rely on RNA velocities. Newer implementations of CellRank (version 1.3.0) also allow for estimation of transition matrices based on CytoTRACE scores for example.

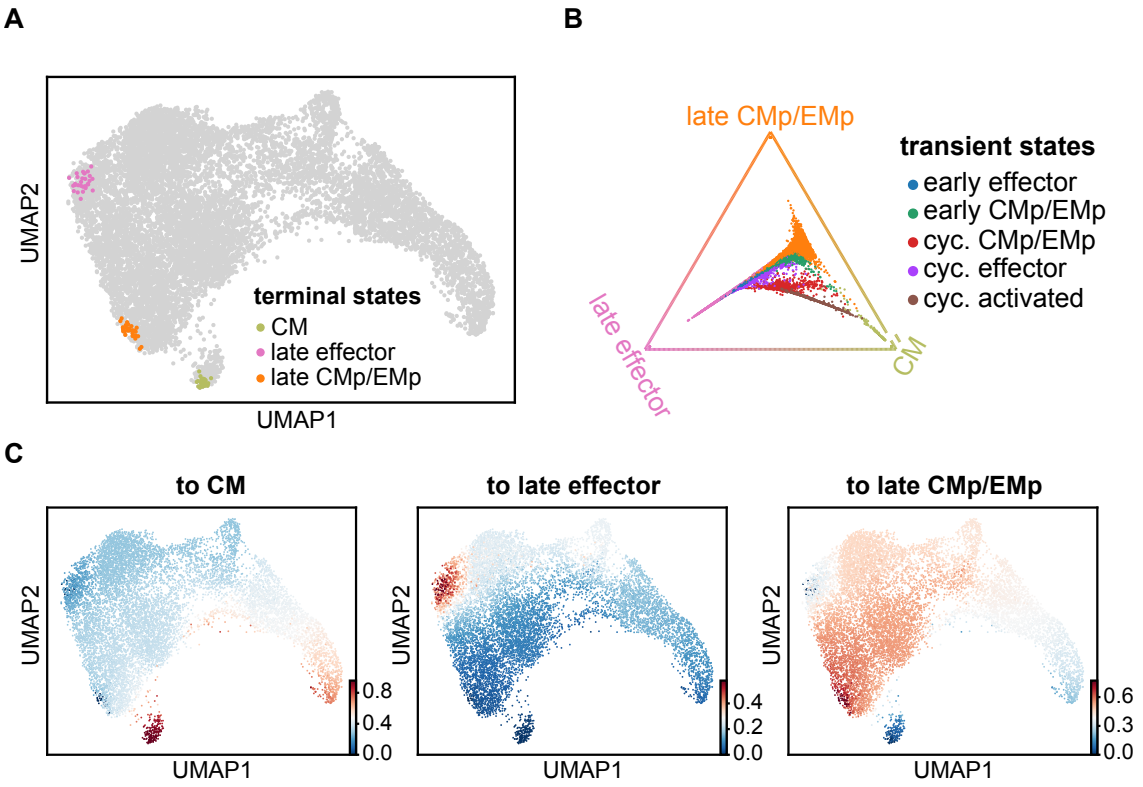

Figure 4.6: CellRank-based fate maps assigning absorption probabilities to each transient cell, that is, probabilities to transition into one of the terminal states. **(A)** Leiden clusters that have been selected as terminal states throughout this analysis. Shown are cells that are considered highly representative of the respective terminal states. **(B)** Fate map, indicating the absorption probabilities of the cells, i.e. the probability of a cell transitioning into one of the terminal states shown in (A). Here, the probabilities are translated into distances with respect to the three possible fates (corners of the triangle). **(C)** Absorption probabilities for the respective terminal states projected into the UMAP embedding.

according to CellRank. Fig. 4.6B represents a fate map: the distance of each transient cell to the respective terminal states (endpoints of the triangle) reflects the relative probability of transitioning into these states. It is noticeable that the cycling activated cluster is very close to the CMs, that the early CMp/EMp cells are very closely linked to the late CMp/EMp cells, and that the early effector cluster is connected to the late effector endpoint. In contrast, the cycling effectors appear to be able to transition into late effectors, as well as (albeit to a much lesser extent) into late CMp/EMp cells. A similar bifurcation appears to exist for the cycling CMp/EMp cells: these can transition either into late effectors or into CM cells. Fig. 4.6C shows the absolute absorption probabilities for the transient cells projected into the UMAP embedding. Although the overall impression is very similar to the RNA velocities projected into the UMAP embedding,

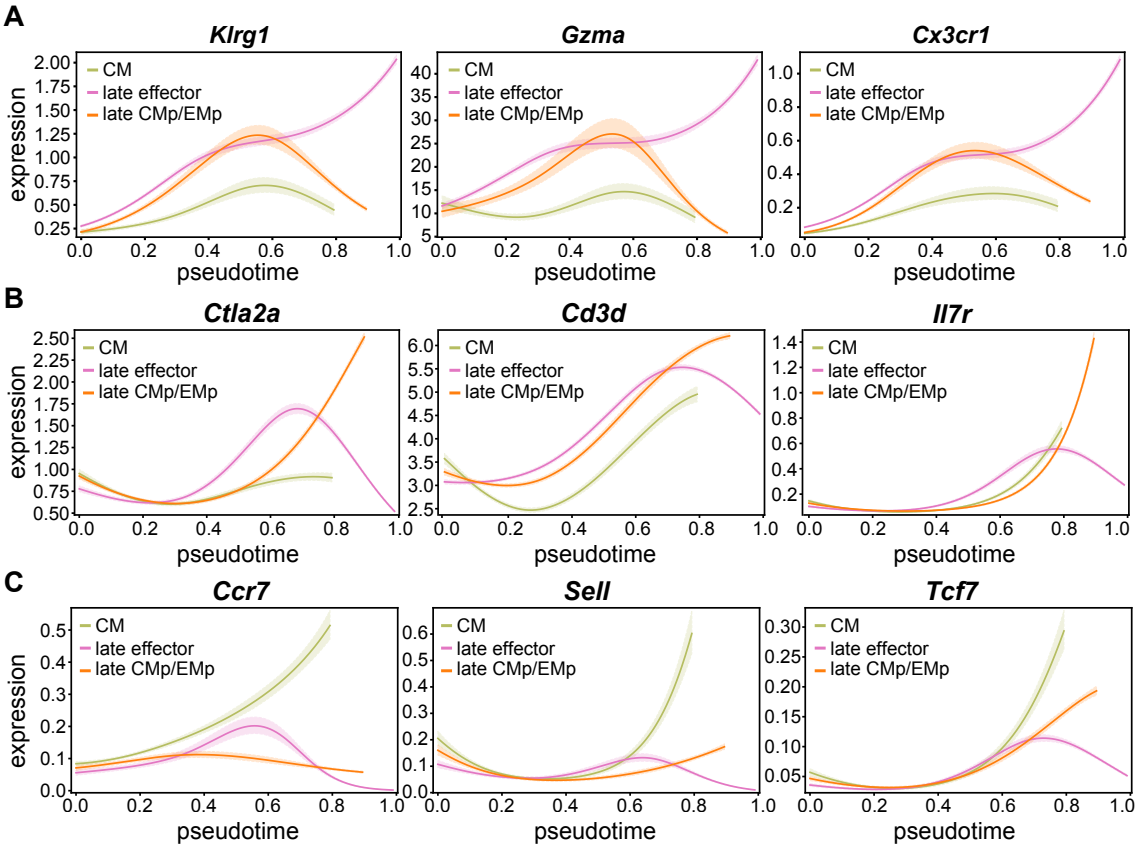

Figure 4.7: Expression profiles of identified lineage drivers for each of the terminal states shown in Fig. 4.6. Cells were ordered in ascending order according to their RNA velocity pseudo-time. Shown are driver genes that are key regulators involved in late effector **(A)**, late CMp/EMp **(B)**, and CM **(C)** fate decisions.

some aspects are better resolved with CellRank. For example, it is indicated that the very early activated cells become the later CMs. This also fits well with the expression of the memory genes *Slamf6* and *Satb1*, cf. Fig. 4.2. This is a very important finding as it suggests differentiation within the cell cycle cluster. Thus, differentiation starts from memory precursor cells that begin to divide upon activation. A small fraction of them exit the cell cycle early on to become central memory cells. That part that divides longer then loses memory potential and becomes effector-like. It should be noted that according to CellRank, dividing effector cells partially seem to have a re-differentiation potential to late CMp/EMp cells. However, a transition to CM cells is not supported by the method.

In addition to providing fate maps, CellRank also allows for the identification of driver genes that correlate strongly with the cells transitioning into one of the respective terminal states. Here, we selected three genes per terminal state that were among the top 10

regarding their positive correlation. Only for the CM cells, we additionally considered the important genes *Sell* and *Tcf7*, as the correlating genes for the memory fate have been mainly dominated by ribosomal genes. Fig. 4.7 shows the expression of such genes for the terminal states, while cells were ordered according to their RNA velocity pseudo-time. As a matter of fact, not all states reach the end of pseudo-time because their effective distance to the root in terms of a random walk on the velocity graph is not the same. Fig. 4.7A contains three genes that were identified as being lineage drivers for the late effector fate. As mentioned earlier, *Klrg1* and *Cx3cr1* are known to be terminal effector markers, while *Gzma* can also be expressed in memory-like cells, as they also seem to possess killing capacity. For late effectors, these genes are expressed in a strictly monotonically increasing manner, with a rapid increase for late pseudo-times. Similarly, among the CMp/EMp specific genes, *Il7r* can be identified. The gene *Cd3d* encodes a part of the T cell receptor (TCR). It seems that at least this component of the TCR is downregulated later by memory-like cells than by effector cells. A possible reason for this could be that CMp/EMp cells are even more dependent on the antigen stimulus mediated by the TCR, whereas late effectors, in contrast, are increasingly controlled by cytokine signals, cf. [Kretschmer et al., 2020]. *Ctla2a* is not further known as an important immunological marker gene. However, it is reported to be a tissue-specific secretory factor, see [Zhang et al., 2015]. As classical memory genes we show *Ccr7*, *Sell*, and *Tcf7*, where *Ccr7* could be identified by CellRank. All three genes reconfirm the central memory properties of the CM cluster. Here, *Ccr7* and *Sell* are important homing genes that regulate cell migration, while *Tcf7* encodes a known transcription factor involved in longevity.

Relative stemness scores from CytoTRACE

As an alternative approach to the differentiation question, we have used *CytoTRACE*, see [Gulati et al., 2020]. This method utilises the number of expressed genes per cell, which the authors have identified as a surprisingly good correlate for stemness. A normalised score is calculated from those genes that highly correlate with the number of expressed genes per cell. This score takes values between zero (relatively differentiated) and one (relatively undifferentiated). Fig. 4.8A shows the number of expressed genes per cell projected into the UMAP embedding. Generally, dividing cells are expected to express a larger number of genes. Within the dividing cells, however, the highest values are reached by cells in the center of the cluster. Within the non-dividing cells no statements can be made. Fig. 4.8B shows the corresponding CytoTRACE scores. As expected, cells with the highest scores are located at the bottom of the cycling activated cluster. Notably, the gradient progresses directly through the cell cycle clusters. Cells seem to exit the division

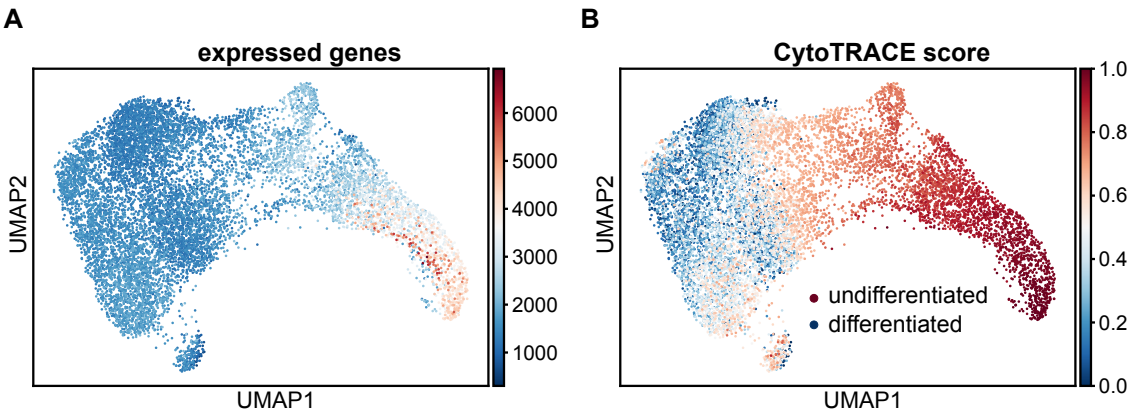

Figure 4.8: CytoTRACE scores indicating the directionality of differentiation in an unbiased manner. **(A)** Number of expressed genes per cell as the starting point in the calculation of CytoTRACE scores. **(B)** CytoTRACE score of the data set based on the number of expressed genes per cell, quantifying relative degrees of differentiation between the cells. Cells can undergo developmental changes moving from high (most undifferentiated) to low (terminal states) scores.

cluster parallelly, i.e. no phenotypic transitions are recognisable on day 8 p.i. At day 12 p.i., however, there is a clear differentiation gradient from late CMp/EMp cells to late effectors. Nevertheless, in our view, a gradient in CytoTRACE scores does not necessarily translate into differentiation. Rather, high scores are indicative of higher differentiation potential, as the CytoTRACE scores were designed to quantify the stemness of the cells. Interestingly, the majority of the CM cells are also assigned increased scores. This is consistent with the stem cell character of mature memory cells during homeostasis.

Conclusion of model topologies

Next, we condense the obtained results into a basic model topology. Overall, it is striking that an essential part of the differentiation dynamics is determined by the exit of cells from the cell cycle. Since it is not clear per se to what extent proliferation and differentiation are coupled, we manifestly incorporate this mechanism into the model scheme. So far, we have been able to gather some indications that the earliest activated cells have memory-like gene signatures. These cells might partially stop dividing very early on to mature and become central memory cells. The other part of the cells seems to continue cycling, thereby differentiating progressively to acquire effector functions. Generally, cells seem to be able to stop proliferating at any developmental stage. By doing so, they mainly retain their phenotype. A part of the dividing effector cells constitutes an exception. Those cells still have the potential to acquire memory-like (mainly effector-memory-like) properties.

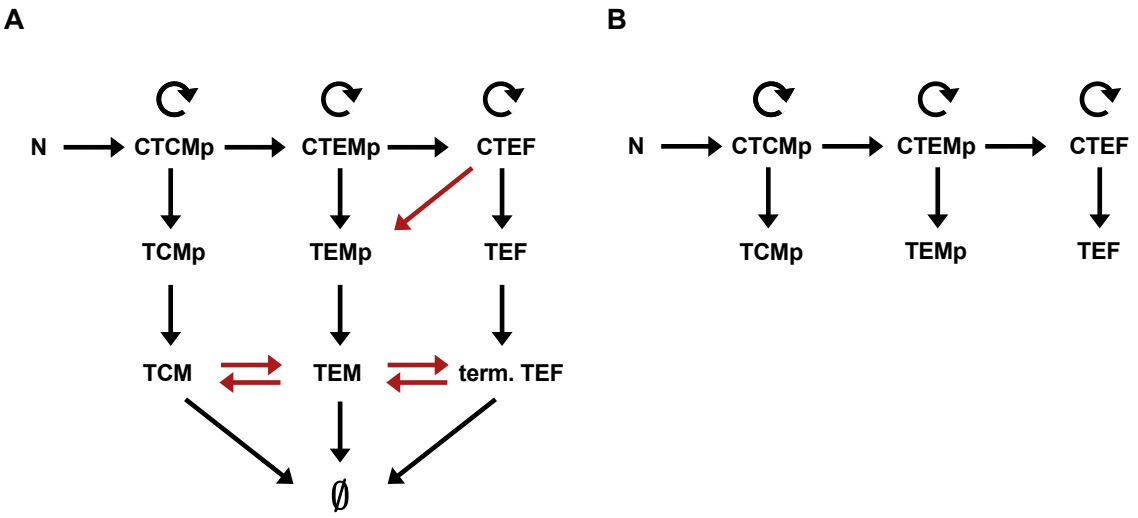

Figure 4.9: Basic model topologies derived from scRNA-seq-based trajectory inference. **(A)** Full scheme of the cell cycle dropout model including the contraction phase, motivated by time series data until day 12 p.i. Black arrows mark reactions that are more strongly indicated by the data. Red arrows, on the other hand, represent directions of differentiation that could potentially occur. Cell death was additionally introduced since our fate mapping data shows a significant decrease in population size between day 8 and 12 p.i., see section 5.3. **(B)** Simplest possible model variant to describe the expansion phase of an acute infection until day 8 p.i.

Likewise, it cannot be ruled out that transitions across phenotypes can still take place at a later stage, such as day 12 p.i. If this is the case, then these transitions tend to be progressive (Fig. 4.4D and 4.8B). Nevertheless, we do not want to exclude the possibility of back-differentiation, especially since weak differentiation processes may not be resolved for the reasons already mentioned above.

Fig. 4.9A depicts a graphical summary of these considerations until day 12 p.i. Here, red arrows stand for uncertain transitions or transitions that could not be ruled out. On the other hand, black arrows mark transitions that could be inferred more confidently from the analysis in this section. Fig. 4.9B provides the simplest imaginable model topology aiming to capture the dynamics until day 8 p.i. Later we will state that small extensions of this model suffice to describe all heterogeneity observed in the (single-) cell fate mapping data and measured differences in cell cycle speed between the subsets.

Finally, we want to address the relationship between gene expression level and protein level. This is an important point as the classification of the population data used to fit our mathematical models is based exclusively on protein markers. The markers chosen here differ depending on the infection model. For LCMV Arm, TCF1 (the protein associated with the gene *Tcf7*) and CX3CR1 are used to classify cells, while in the *L.m.*-OVA

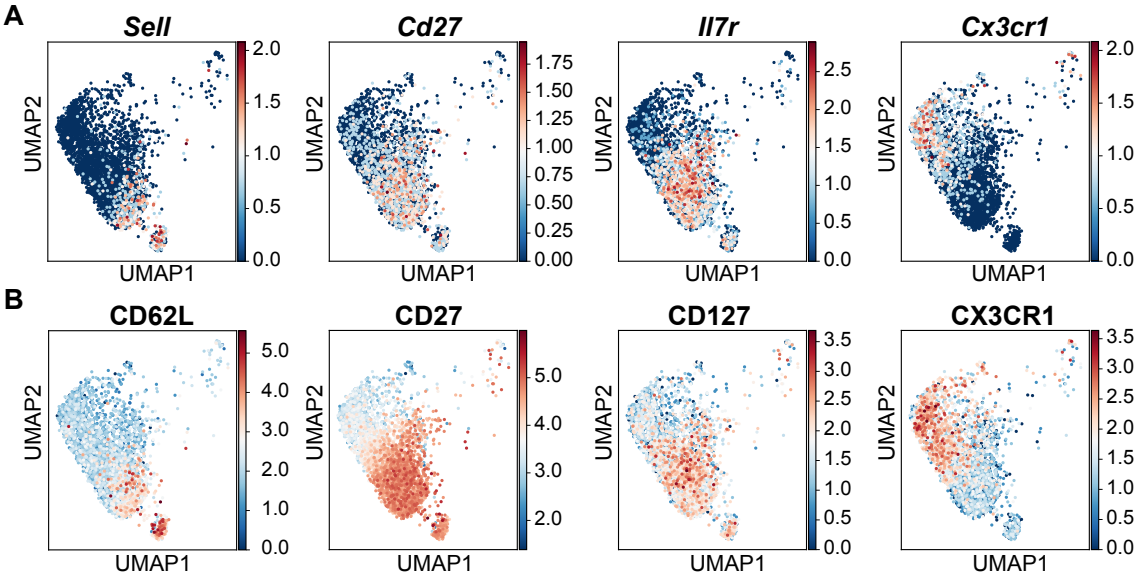

Figure 4.10: Correspondence between mRNA and (surface) protein level for cells at day 12 p.i. projected into the joint UMAP embedding. **(A)** Log-counts of mRNA abundance for important marker genes used for phenotypic classification of cells in the FACS. **(B)** Respective log-counts of the corresponding surface protein levels based on antibody-derived tags (ADTs). (Experiments conducted by Albulena Toska)

case, CD62L, and CD27 are used respectively. The concrete phenotyping scheme is the following: TCMp ($TCF1^+/CX3CR1^-$ or $CD62L^+/CD27^+$), TEMp ($TCF1^-/CX3CR1^-$ or $CD62L^-/CD27^+$) and TEF ($TCF1^-/CX3CR1^+$ or $CD62L^-/CD27^-$). It should be added that this is a standard scheme in the field, cf. e.g. [Hikono et al., 2007; Martin et al., 2018]. We have observed in our data that CD62L positive cells often represent a subset of TCF1 positive cells, i.e. CD62L seems to be a more exclusive memory marker. However, at very early time points after infection (around 3 d), CD62L is a problematic marker, as the surface protein is initially cleaved upon activation; the biological significance of this effect could be that it is favourable for the host if early activated cells are not directed into the lymph nodes, but move straight to the site of infection, see [Yang et al., 2011]. However, the mRNA level is not affected by CD62L "shedding".

By considering the mRNA level, we systematically anticipate the classification state obtained in FACS measurements. Hence, we would like to briefly show that our previous conclusions remain legitimate. To this end, Fig. 4.10 shows different marker genes both at the mRNA level (Fig. 4.10A) and the surface protein level (Fig. 4.10B). In both cases, the logarithmised counts are shown as an example for the day 12 time point in the joint UMAP representation. When comparing the two levels, it is immediately apparent that

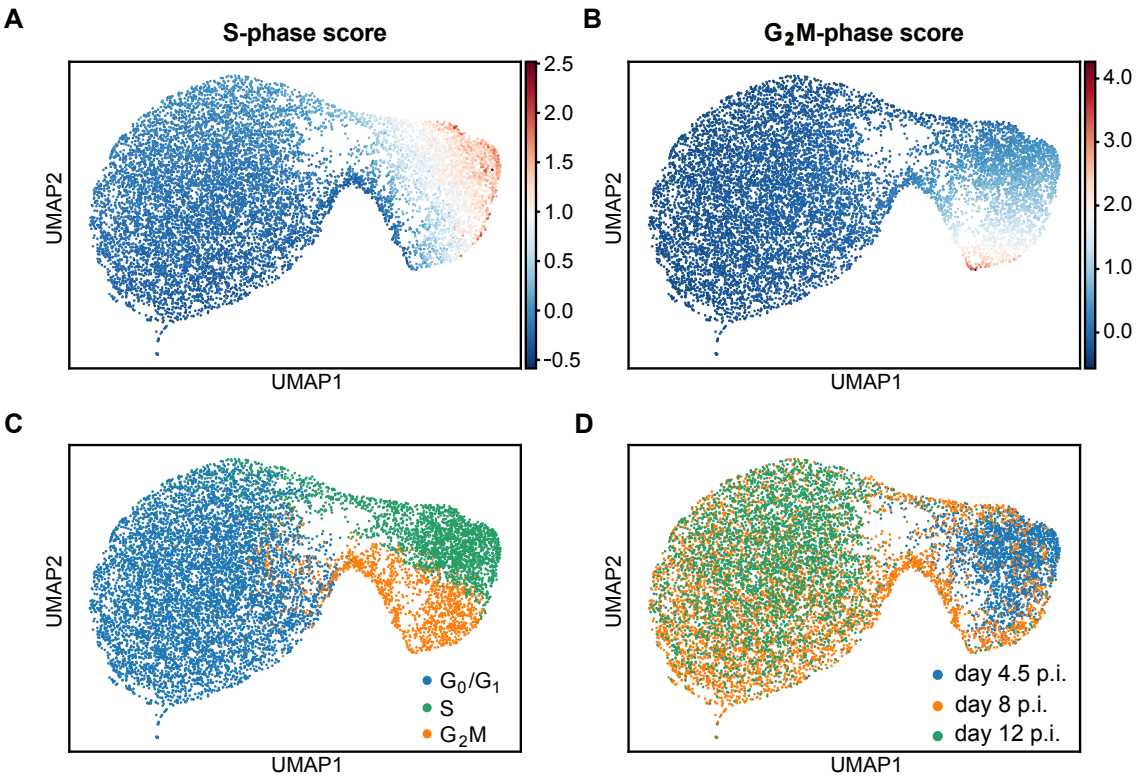

Figure 4.11: UMAP embedding based on 96 cell cycle genes from [Tirosh et al., 2016], showing the phase scores **(A, B)**, the resulting classification label **(C)**, and the time point after infection **(D)**.

there is sufficiently good agreement between mRNA and protein levels. This indicates that the conclusions previously drawn at the mRNA level could be transferred at least approximately to the protein level for our considerations.

4.2 Inferring quiescence from scRNA-seq data

So far, we have interpreted the data as if the cells classified as G_0/G_1 were mainly in G_0-phase. This would consequently mean that almost all of the G_1 cells would have been classified incorrectly as S- or G_2M-phase cells. In the following, this point shall be discussed in more detail. To this end, let us consider the same data in a reduced feature space that is spanned by 96 cell cycle genes from [Tirosh et al., 2016], i.e. in *cell cycle space*. The same set of S- and G_2M-phase genes was used to compute the previous phase scores. Recomputing the UMAP embedding based on these genes leads to Fig. 4.11. As expected, in this embedding the cells cluster with respect to their cell cycle phase label

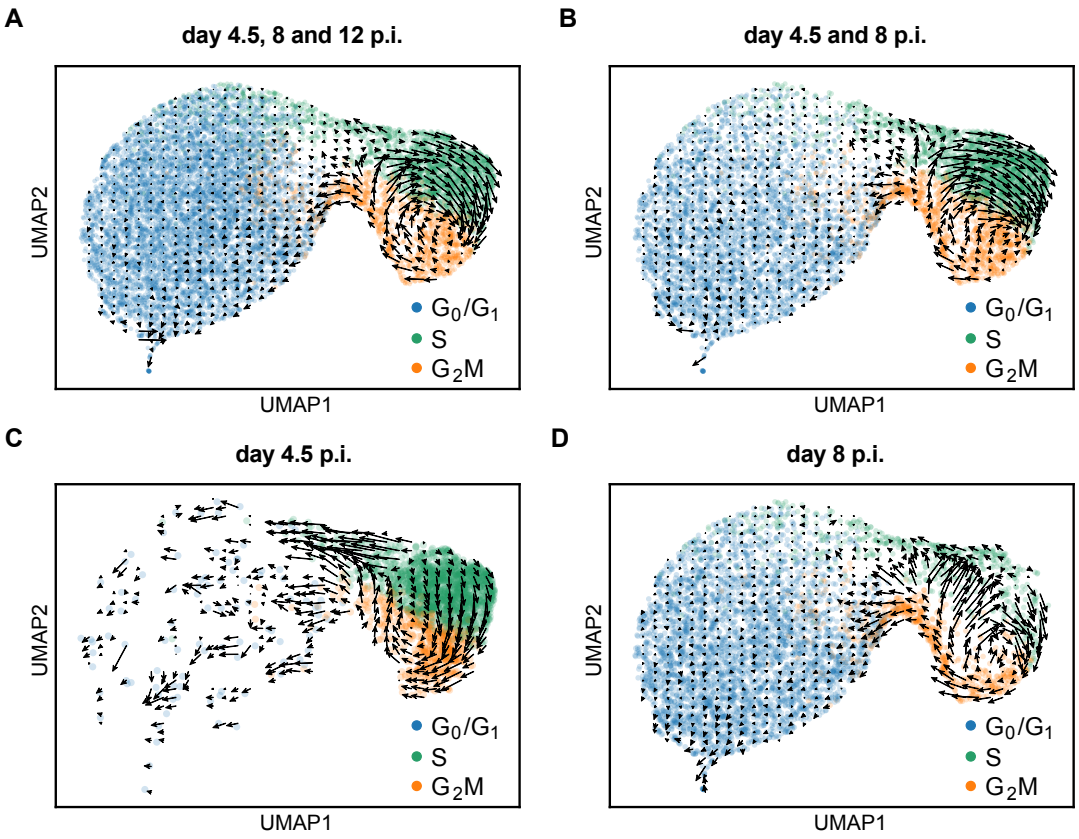

Figure 4.12: RNA velocities from the dynamic velocity model computed exclusively on cell cycle genes from [Tirosh et al., 2016] and projected into the UMAP embedding showing cell cycle phase labels. Velocities were computed based on all cells **(A)**, a combination of batches **(B)**, and on cells from individual batches **(C, D)**. In all cases, velocities were projected into the joint UMAP representation. Differentiation in this subspace of genes indicates progression through cell cycle stages, whereas vanishing velocities imply quiescence.

(Fig. 4.11C), which is determined by the combination of S- and G_2M scores (Fig. 4.11A and 4.11B). Overall, the S and G_2M cells form a subcluster which mainly consists of day 4.5, and to a lesser extent, of day 8 cells (Fig. 4.11D).

Next, we recomputed RNA velocities based on these same cell cycle genes. Fig. 4.12A shows the results for the inclusion of cells from all time points into the velocity calculation. We recognise closed differentiation trajectories in the smaller subcluster, descriptively mapping progression through the cell cycle. This is because non-zero velocities in this particular space are likely linked to cycling activity. Therefore, we conclude that the actual G_1 cells are probably contained in the smaller subcluster and thus must have been classified incorrectly. On the other hand, the cells classified as G_0/G_1 predominantly consist of quiescent G_0 cells. Note, that this argument is not affected by the fact that

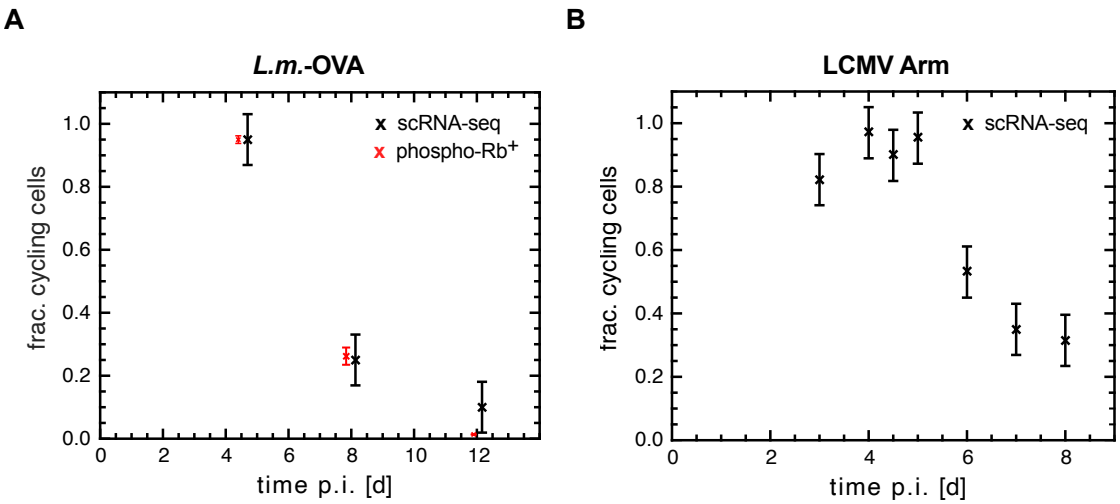

Figure 4.13: Fraction of cycling cells for different time-points of infection inferred from scRNA-seq data. Here, cells classified as G_0/G_1 were interpreted as G_0 cells based on RNA velocities in cell cycle space, i.e. representations as shown in Fig 4.12. **(A)** Fractions of cycling cells inferred from scRNA-seq data from OT-I T cells during acute *L.m.*-OVA infection were confirmed measuring phosphorylated retinoblastoma protein (phospho-Rb). The scRNA-seq based results and the phospho-Rb measurements were conducted at the same time points after infection. **(B)** Fractions of cycling cells derived from multiple scRNA-seq data sets of P14 cells during acute infection with LCMV Arm. If more than one data set per time point was available, the mean is shown. The analysis was performed by Albulena Toska based on her own and published data. (*L.m.*-OVA related experiments conducted by Albulena Toska; LCMV Arm data from [Kurd et al., 2020] (day 3, 4, 6 and 7 p.i.), [Yao et al., 2019] (day 4.5 and 7 p.i.), Albulena Toska (day 6 p.i.) and [Chen et al., 2019] (day 8 p.i.))

the gene list from [Tirosh et al., 2016] only considers S and G_2M genes, which indeed may not be strongly expressed during G_1-phase. However, low expression of S and G_2M genes does not imply that there may be no regulation of these genes during G_1-phase. This regulation (first derivative) is exactly quantified by the RNA velocities. Also, we know from the results shown in Fig. 3.6 that the G_1-phase is definitely relevant in terms of average length. Fig. 4.12B, C, and D show the corresponding RNA velocities when calculated on subsets of time points. The closed-loop structure within the cycling cells remains preserved. Overall, these results are consistent with the study conducted in [Schwabe et al., 2020]. Here, the authors have shown that progression of cells through the cell cycle happens on cyclic-planar trajectories in gene expression space; all changes in the transcriptome related to other processes than cell cycle were shown to occur orthogonally to these dimensions.

The previous considerations allow for extracting fractions of quiescent cells from the data by simply counting the cells that were classified as G_0/G_1 for the individual time points. These fractions are depicted in Fig. 4.13A (black). The error was estimated within our mathematical modelling framework to be discussed in section 4.5. My colleague Albuena Toska was able to validate these conclusions by FACS-based measurements of phosphorylated retinoblastoma protein (phospho-Rb) (red) at the same time points after infection. The standard error of the mean was used to quantify the uncertainties of the measurement. Here, a cell is positive for phospho-Rb if, and only if, it is cycling. Compared to other often used markers, such as KI67 (gene name *Mki67*), phospho-Rb has the advantage that it is degraded more rapidly and thus negative cells with respect to this marker are more reliably in G_0-phase, cf. [Gookin et al., 2017].

Fig. 4.13B shows the corresponding fractions of cycling cells for seven time points during acute LCMV Arm infection. The results are based on 11 published scRNA-seq data sets and one data set generated by Albulena Toska herself. The associated publications are given by [Yao et al., 2019], [Chen et al., 2019], and [Kurd et al., 2020]. Comparing the two kinetics shown in Fig. 4.13A and B, it is noticeable that in both cases, the majority of the cells seem to exit their cell cycle in a relatively narrow time interval. The precise knowledge of the proportion of dividing cells is not only important for calculating average cell cycle speed, as we have already outlined in section 3.6, but will also play a vital role in explaining the heterogeneity of the population data introduced in the following section.

Finally, we would like to emphasise that the possibly incorrect classification of the actual G_1 cells in the scRNA-seq data could be related to the rapid division speed of the cells and thus be T cell specific. Likewise, it cannot be ruled out that the cells can enter the S-phase cluster very quickly after having completed their G_1-phase. Such rapid transitions might not be resolved by UMAP. However, the high level of agreement between our outlined interpretation and the subsequent phospho-Rb measurements (Fig. 4.13A) rather indicates an incorrect classification of the majority of the G_1 cells in the T cell context. It should also be noted that the contribution of the G_1-phase to the average cell cycle length was shown to be relatively large during *L.m.*-OVA infection at day 4.5 p.i. (cf. Fig. 3.9A from section 3.6.2).

4.3 Summarising the data

At this point, we would like to give a brief overview of the data used to fit our mathematical models in this chapter. The data essentially consists of three parts: the phenotypically

classified fate mapping data, cell cycle quantifications from BrdU/DNA experiments, and the dropout kinetics derived from the scRNA-seq analyses. In this chapter, all data for the modelling refers to the LCMV Arm infection. This infection is particularly suitable for modelling because, compared to the *L.m.*-OVA infection, it has some additional features that severely constrain model selection, as will be explained later.

The fate mapping data originate from the laboratory of Veit Buchholz and were collected by Theresa Busch and Lorenz Kretschmer. Single naïve P14 cells were transferred into C57BL/6 ("black 6") recipient mice. Mice were infected with LCMV Arm, which causes activation of these P14 cells. The activated cells begin to divide and differentiate, thereby inheriting their congenic marker that uniquely determines their clone membership. The experimental basis was already established in [Buchholz et al., 2013]. Importantly, the special design of the congenic matrix allows the use of up to seven individual naïve cells per recipient. At day 8 p.i., the spleens of the mice were FACS-analysed and clonal sizes and phenotypic compositions were determined. The previously described markers TCF1 and CX3CR1 were used for this purpose, defining the subsets TCMp, TEMp, and TEF. These experiments resulted in a total of 40 recovered clones on day 8 p.i. The added value of these data is that a recovered clone is based on exactly one cell, i.e. the initial condition is precisely known. We have summarised these data in terms of subset-specific moments of the cell number distributions. Here, we considered the means, coefficients of variations, and pairwise Pearson correlation coefficients. In addition to the single-cell fate mapping data just described, Albulena Toska also collected fate mapping data from bulk transfers at day 8, 7, 6, and 4.5 after infection. In each case, 130, 1000, 5000, and 50000 naïve cells were transferred, respectively. Relative subset sizes were then determined from these data.

The second important input for the modelling are the cell cycle quantification results from section 3.6. The results belonging to LCMV Arm have already been shown in Fig. 3.8. Unfortunately, these quantifications cannot be adopted directly for the modelling because the fate mapping data and the BrdU/DNA experiments were carried out with different virus strains. The newer strain used for the cell cycle speed quantifications is characterised especially by higher absolute division speed and prolonged cycling activity of the cells, leading to higher progeny numbers. However, we assume here that ratios of the rates between subset fractions are comparable across virus strains. Therefore, we only include the calculated ratio of rates, $p_{\mathrm{CMp}}/p_{\mathrm{EFF}}$, on day 4.5 after infection into the fitting procedure. This also seems reasonable to us, particularly since we work with time-independent division rates in the model and only this one time point has been available. Later in chap-

ter 5, we will look at time-dependencies of division rates in more detail by considering absolute quantification results.

The third and last important component of the modelling is the cell cycle dropout kinetics, which in the LCMV Arm case is based exclusively on scRNA-seq data and has already been shown in Fig. 4.13. This is essential since we manifestly integrate the exit of cells from the cell cycle into the model topology. A reliable estimation of the parameters would be impossible in such a model without these data.

4.4 Scope of the progressive model of differentiation

Before we implement the exit of cells from the cell cycle into our model topology, let us first consider an effective description within the framework of the progressive model of differentiation. We do this for two reasons: first, this model has already proven reliable in the past for describing the acute phase of $CD8^+$ T cell responses, cf. [Buchholz et al., 2013], and second, it is included as a sub-model in our proposed model topology, see Fig. 4.22. Since we strongly follow [Buchholz et al., 2013] in the choice of summary statistics of the data and the formalism of stochastic modelling in general, we find it very instructive to look at this particular limit. Within the progressive model, naïve cells differentiate directly into long-lived TCMp cells upon activation. TCMp cells then further transition into TEMp cells, which in turn differentiate into short-lived TEFs. Furthermore, fitting the model to data from acute *L.m.*-OVA infection revealed in [Buchholz et al., 2013] that memory cells divide more slowly on average than the subsequent compartments with effector potential.

The progressive model allows for six elementary reactions to happen in a stochastic fashion, three of which are given by differentiation and three by proliferation. All reactions are described with constant rates, thereby assuming exponential waiting time distributions. Mathematically, moment equations for the means and the covariance matrix of the cell number distributions are solved for the initial condition of one naïve cell at $t = 0$. Parameter estimation was done by weighted least-squares fitting with errors of the moments being estimated via bootstrapping, see [Efron et al., 1993]. Errors of the relative subset sizes were estimated in terms of pooled standard deviations. The model was only fitted to the fate mapping data described in the previous section. For details, we refer to the methods section 2.1. The results associated with the best model fit are shown in Fig. 4.14. Here, naïve cells were not included in the calculation of the moments (Fig. 4.14A, B, and C), but were added to the TCMp compartment when considering relative population sizes. Overall, the progressive model can describe the means, CVs, and pairwise correlations, as

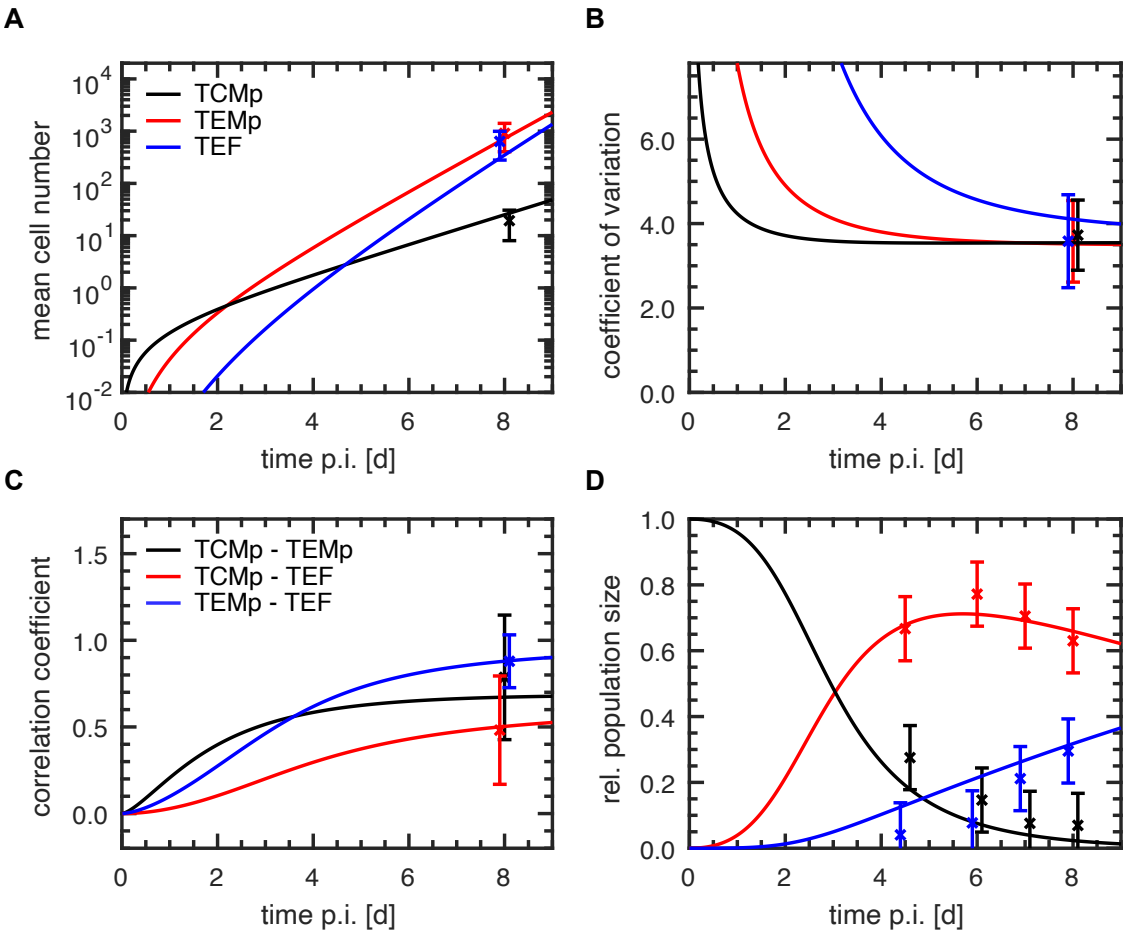

Figure 4.14: Best fit results for the progressive model of differentiation fitted to single-cell progeny and population data from acute LCMV Arm infection. Corresponding parameter values are given in Tab. 4.1. Shown are the dynamics of the moments **(A, B, C)** and relative population sizes **(D)** for the respective compartments. All data used for estimating parameters are shown. (Single-cell fate mapping experiments conducted by Veit Buchholz, Theresa Busch, and Lorenz Kretschmer; relative subset sizes measured by Albulena Toska)

well as the phenotypic composition at earlier time points. Especially the remarkably high CVs, which are around ~ 4 for all subsets, are reproduced correctly.

Although the model curves are very well fitted to the data, the model explains the observed stochasticity in the wrong way. This becomes clear when looking at the underlying model parameters of this best fit, shown in Table 4.1, including their profile likelihood-based confidence bounds, see [Venzon et al., 1988]. In contrast to [Buchholz et al., 2013], it is noticeable that the effective proliferation rates of all subsets do not differ from each other within their 95 % confidence intervals. The ratio of differentiation rates

parameter	best fit	95 % CI
p_{TCMp}	1.12 d^{-1}	[0.78; 2.22] d^{-1}
p_{TEMp}	1.22 d^{-1}	[0.46; 1.82] d^{-1}
p_{TEF}	1.23 d^{-1}	[0.00; 1.48] d^{-1}
$r_{\mathrm{N}\to\mathrm{TCMp}}$	0.10 d^{-1}	[0.06; 0.21] d^{-1}
$r_{\mathrm{TCMp}\to\mathrm{TEMp}}$	0.46 d^{-1}	[0.20; 1.63] d^{-1}
$r_{\mathrm{TEMp}\to\mathrm{TEF}}$	0.07 d^{-1}	[0.03; 0.60] d^{-1}

Table 4.1: Best fit parameters for the progressive model of differentiation, as it was used in [Buchholz et al., 2013]. The model was fitted to single-cell progeny data subject to acute LCMV Arm infection at day 8 p.i. 95 % confidence intervals (CIs) were computed in terms of the profile likelihood method. Corresponding model fits are depicted in Fig. 4.14.

$r_{\mathrm{TEMp}\to\mathrm{TEF}}/r_{\mathrm{TCMp}\to\mathrm{TEMp}}$, however, agrees well with [Buchholz et al., 2013] (0.15 vs. 0.20). On the other hand, the recruitment rate $r_{\mathrm{N}\to\mathrm{TCMp}} \approx 0.1\ \mathrm{d}^{-1}$ corresponds to average waiting times of ~ 10 d, being even bigger than the actual time point of measurement. In forward simulations until day 8 p.i. based on these parameters, a substantial proportion of the naïve cells are thus not activated at all (simulations not shown). This is in clear contradiction to experimental observations from [Van Heijst et al., 2009]. Here, the authors transferred uniquely labelled naïve cells and measured the number of recruited cells during infection under various conditions (e.g. pathogen and infection dose). They concluded that nearly all cells get recruited upon acute infection. Also in [Schlub et al., 2010] the authors could show that even for large numbers of transferred cells, the fraction of unrecruited cells always stayed smaller than 0.1 %. Moreover, R. Ahmed and colleagues measured viral kinetics during LCMV Arm infection in the spleen and showed that at the peak of the T cell response, most of the virus had already been eradicated, see [Lau et al., 1994]. Therefore, it is very unlikely that a substantial part of the naïve cells could still be recruited in the absence of antigen. The model thus incorrectly explains the increased variability in the data (increased compared to the *L.m.*-OVA data from [Buchholz et al., 2013]) by the recruitment of the cells into the immune response. Also, the analysis of the scRNA-seq data in section 4.1 has shown that a major contribution to the overall variability in the data arises from cell cycle effects, suggesting model topologies as depicted in Fig. 4.9. In the next section, we address the variability of the single-cell fate mapping data in more detail and show how to explain it in a biologically consistent way.

4.5 The cell cycle dropout model

The previous considerations have shown that the progressive model of differentiation highly overemphasises the process of recruitment and first division when explaining the observed variability. Based on the scRNA-seq data analyses, it appears that cell cycle dropout effects play a major role in the overall dynamics of the response. We want to include irreversible division stops in our model topology by considering each of the three phenotypes TCMp, TEMp, and TEF being either cycling or non-cycling. First, we focus on the simplest possible scheme that does not contradict the previous trajectory inference analyses from section 4.1.2. The resulting model scheme has already been shown in Fig. 4.9B.

Next, we would like to counteract the model's tendency to explain the entire variability by the initial recruitment rate $r_{\mathrm{N}\to\mathrm{TCMp}}$. To do this, we replace the exponential waiting time distribution with a heavy-tailed Erlang distribution to limit the variability of this process. Conceptually, this seems even more reasonable when considering that the first cell division is known to take significantly longer than all subsequent divisions, cf. [De Boer et al., 2005; Deenick et al., 2003; Lee et al., 2008]. As a matter of fact, the available data do not allow reliable discrimination between recruitment and first cell division. Accordingly, suppressing arbitrarily short waiting times for the combination of these two processes makes sense biologically. Indeed, there is experimental *in vitro* evidence that time to (first) division is not exponentially distributed but rather follows skewed distributions as log-normal or Gamma-distributions, see [Dowling et al., 2014; Hawkins et al., 2009; Robert et al., 2021]. Even *in vivo*, mathematical modelling based on carboxyfluorescein succinimidyl ester (CFSE) could show that the time to first division is unlikely to be the main driver of variability in individual T cell responses, see [Gerlach et al., 2013]. Mathematically, we can realise the Erlang distribution by introducing n naïve auxiliary compartments being linked by $n-1$ transitions of identical rates, i.e. $r_{\mathrm{N}_1\to\mathrm{N}_2} = ... = r_{\mathrm{N}_{n-1}\to\mathrm{N}_n}$. The CV of the effective waiting time distribution is given by $1/\sqrt{n-1}$ with mean waiting time $(n-1)/r_{\mathrm{N}_1\to\mathrm{N}_2}$. In our case, we choose $n = 5$ and thus halve the CV of the original exponential distribution. The transition to the dividing memory precursors is then set to be instantaneous, i.e. $r_{\mathrm{N}_5\to\mathrm{CTCMp}} \to \infty$, whereby two cells flow into the CTCMp compartment (first cell division). It should be noted that the choice of $n = 5$ has proven successful in our case, especially as this achieves average waiting times that are in line with biological expectations, see later. Note that model selection alone does not necessarily address this question, as unrealistically high variabilities are not penalised as we have seen previously. In this sense, the choice of $n = 5$ should be understood as rather pragmatical.

All rates in the model are constant in time throughout this chapter, except for the cell cycle dropout rate $r_{\mathrm{C}\to\mathrm{Q}}$. This time-dependence must be introduced especially in an exponential model, as cells would otherwise leave the cycling compartments earlier than observed (see Fig. 4.13). For the time-dependence, we choose a simple step function of the form

$$r^{(i)}_{\mathrm{C}\to\mathrm{Q}}(t) = \begin{cases} 0 & t < T^{(i)}_{\mathrm{dropout}} \\ \\ \tilde{r}^{(i)}_{\mathrm{C}\to\mathrm{Q}} & t \geq T^{(i)}_{\mathrm{dropout}}, \end{cases} \tag{4.1}$$

while $i \in \{\mathrm{TCMp}, \mathrm{TEMp}, \mathrm{TEF}\}$. Cells are thus initially kept in the cycling compartments, where they can exit at a constant rate after reaching the subset-specific time $T^{(i)}_{\mathrm{dropout}}$. To further reduce model complexity, we make some assumptions for now. First, we set the cell cycle dropout behaviour of all phenotypes equal, i.e. $T^{(i)}_{\mathrm{dropout}} = T_{\mathrm{dropout}}$ and $\tilde{r}^{(i)}_{\mathrm{C}\to\mathrm{Q}} = \tilde{r}_{\mathrm{C}\to\mathrm{Q}}$ for all i. Although this seems a questionable assumption, we will see later in section 4.7 that this does not necessarily mean that all subsets share the same fractions of cycling cells at times bigger than T_{dropout}. Also, for simplicity, we assume that TEMp and TEF cells divide with the same rate. We denote the proliferation rates as p_{CMp} and p_{EFF}, respectively. Also we neglect cell death during the acute phase of the infection.

4.5.1 Best fit results

We now fit the model described above to the fate mapping data, to cell cycle quantifications, and to the dropout kinetics inferred from the scRNA-seq data. In addition to the previously mentioned parameters, we additionally fit an absolute and time-independent error for the fractions of cycling cells since some of the data points are based on one data set only. Overall, our model is parametrised in terms of eight parameters. Technically, we proceed in the same way[2] as in section 4.4, additionally adjusting the objective function with respect to the error contributions. The resulting best fit parameters and their confidence bounds are shown in Tab. 4.2. Let us first consider the corresponding model fits of the moments, shown in Fig. 4.15. The dynamics of the mean cell numbers (Fig. 4.15A) show a reduced increase at about 5 d p.i., as the cells then begin to stop cycling. The memory cells are most affected by these cell cycle dropout effects. Their growth rate almost becomes zero, as the net efflux rate from the CTCMp compartment is bigger than its proliferation rate. The naïve cells were again counted with the TCMp cells, since literature predominantly states that cells are $\mathrm{TCF1}^+$ during their first few cell divisions, see for instance [Lin et al., 2016]. The high CVs measured are correctly described

[2]Importantly, the inclusion of time-dependent rates does not require an adjustment of the formalism since all derivation steps in section 2.1.1 still hold and the Markovian property is preserved.

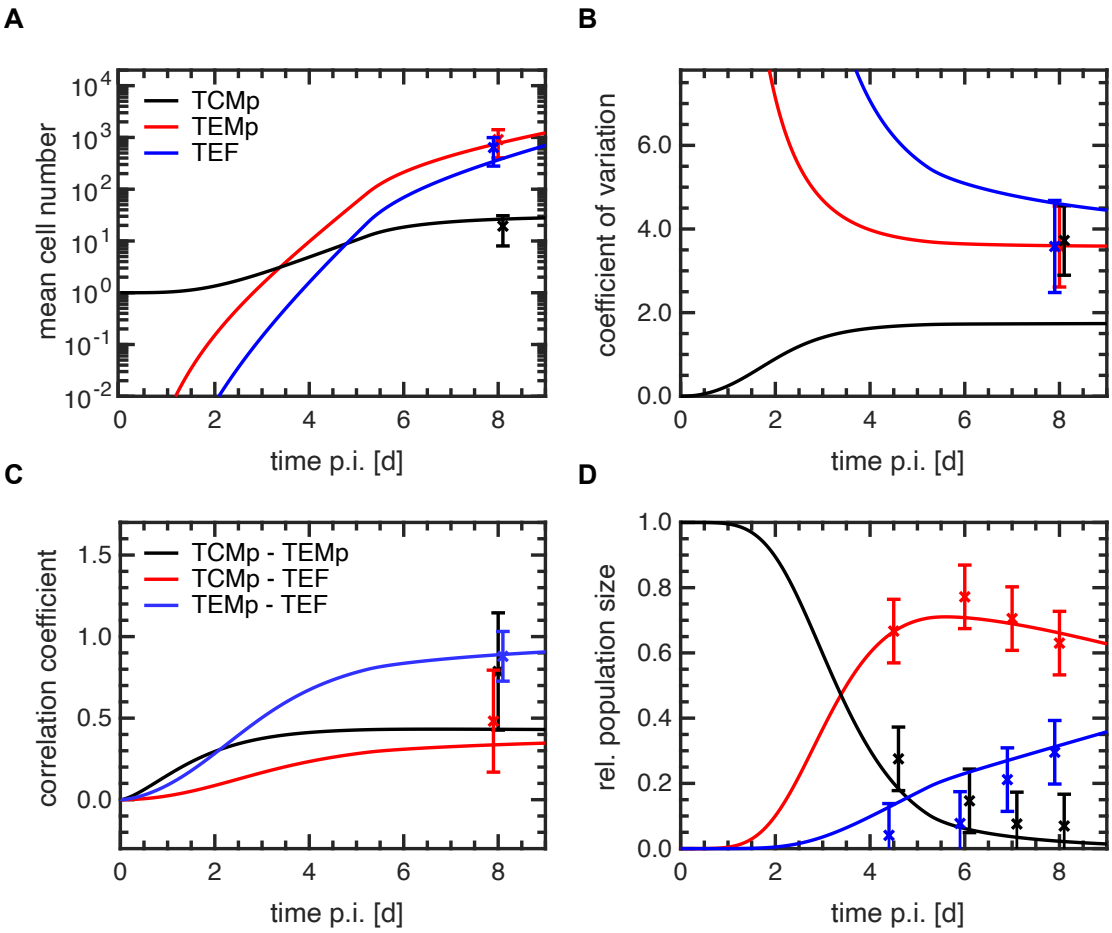

Figure 4.15: Best model fit of the cell cycle dropout model, corresponding to the parameters given in Table 4.2. Shown are the estimated moments of the cell number distribution over time after LCMV Arm infection and the summary statistics of the single-cell progeny data used for fitting **(A, B, C)**. In addition, bulk transfer-based relative population sizes from earlier time points were included in the fitting procedure **(D)**. (Single-cell fate mapping experiments conducted by Veit Buchholz, Theresa Busch, and Lorenz Kretschmer; relative subset sizes measured by Albulena Toska)

(Fig. 4.15B), although the CV of the TCMp cells is underestimated by the model. In our model, this is closely related to the fact that differentiation starts with two dividing memory precursor cells by having combined recruitment and first division. This considerably reduces variability. Nevertheless, we consider this approach to be useful and believe that there may well be other early mechanisms that generate variability at this point. These could be based on asymmetric cell division or irregularities in the first cell division, for example, and will not be investigated further in this work. The measured correlations of TCMp-TEMp and TCMp-TEF are overall subject to large uncertainties (Fig. 4.15C).

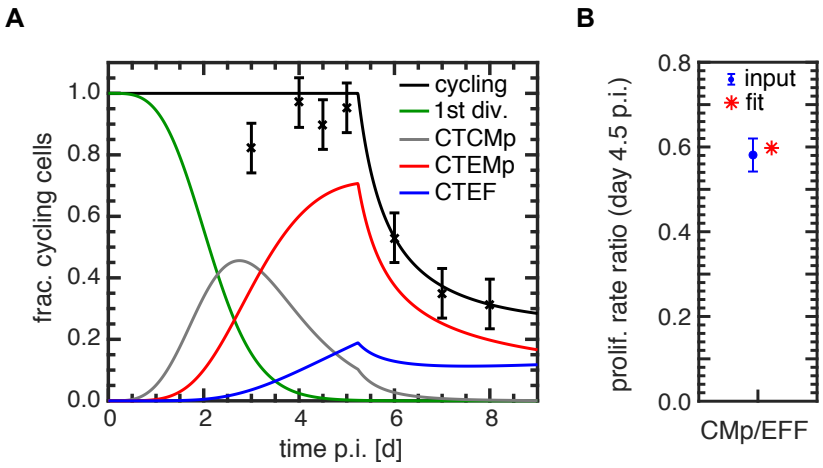

Figure 4.16: Best model fit of the cell cycle dropout model, with underlying parameters given in Table 4.2. Shown are the best model fits related to cell cycle kinetics of the LCMV Arm infection model and the data used for fitting. **(A)** Fitted fraction of cycling cells across all cells (black) and contributions of the respective cycling compartments (coloured) to the overall kinetics. Uncertainties of the scRNA-seq based dropout fractions were estimated during parameter estimation, assuming an absolute and constant error. **(B)** CMp/EFF ratio of proliferation rates at day 4.5 p.i. inferred from our cell cycle quantification method (blue) and the resulting model outcome (red). (Experiments conducted by Albulena Toska)

The TEMp-TEF correlation, on the other hand, could be determined relatively sharply from the data, as higher cell numbers were available for these subsets. From day 2 p.i. onwards, this model correlation is larger than the other two. Such dynamics was to be expected, as the correlation will always be high if one subset differentiates into another subset with a similar (or smaller) proliferation rate. This can be seen by considering that large proliferation rates of the target compartment make the influx due to differentiation less relevant. Therefore, after a certain waiting time, the subsets will decorrelate. By construction, the TEMp and TEF proliferation rates are identical, which leads to an increased correlation of these two cell types after a short time. The relative population sizes (Fig. 4.15D) are also well described by the model.

Let us consider the second part of the model fits related to cell cycle kinetics. Fig. 4.16A shows the fitted cell cycle activity for all cells (black) and the respective contributions of the dividing cell types (coloured). The plotted fractions refer to all cells in the system, i.e. the total curve is the sum of all individual contributions. It can be seen that the naïve starting cell has on average transitioned fully into the CTCMp compartment at day 4 p.i. This is consistent with the mean waiting time of activation and first cell division of approx. 3.8 d. The progressive differentiation hierarchy can be seen in the kinetics. At

parameter	best fit	95 % CI
$r_{N_1 \to N_2}$	1.06 d^{-1}	[0.82; 1.52] d^{-1}
p_{CMp}	1.05 d^{-1}	[0.85; 1.25] d^{-1}
p_{EFF}	1.76 d^{-1}	[1.46; 1.99] d^{-1}
$r_{CTCMp \to CTEMp}$	0.28 d^{-1}	[0.17; 0.49] d^{-1}
$r_{CTEMp \to CTEF}$	0.08 d^{-1}	[0.05; 0.12] d^{-1}
$\tilde{r}_{C \to Q}$	1.33 d^{-1}	[0.95; 1.84] d^{-1}
$T_{dropout}$	5.24 d	[4.74; 5.73] d
error(frac. cycling)	0.08	[0.05; 0.16]

Table 4.2: Best fit parameters for the cell cycle dropout model fitted to data from acute LCMV Arm infection. Corresponding fits and data are shown in Fig. 4.15 and 4.16. 95 % confidence intervals (CIs) were calculated based on the profile likelihood method.

day 5.2 p.i., the subsets then enter their respective quiescent states at a rate of 1.3 d^{-1}, or continue to divide or differentiate. In the model, most of the cells that are still dividing at day 8 p.i. are accounted for by CTEMp cells. In the scRNA-seq analyses, on the other hand, the majority of these cells were determined by CTEF cells. This discrepancy could indicate e.g. difficult discrimination of these two subsets in the scRNA-seq data or even a limp comparison between bulk and single-cell level. The direct comparison of the different infection models could also play a role here and should possibly be omitted. The absolute error of the data was estimated to be approx. 8 %. Overall, the kinetics are described remarkably well. It is noticeable that at the earliest time point on day 3 p.i., the fraction of dividing cells is smaller than on the following days. This could be related to extrinsic variability, e.g. mouse-to-mouse variability. On the other hand, it could be related to a very early exit of individual clones or the TCMp subset as a whole. We will discuss such scenarios in more detail in section 4.6. Fig. 4.16B shows the ratio of quantified proliferation rates of CMp to EFF at day 4.5 after infection and the corresponding model fit. The estimated rates are 1.1 d^{-1} for the CMp cells and 1.8 d^{-1} for the EFFs. We conclude that the model has some freedom here and that the ratio of the rates seems to be a fairly independent data point that can be easily described by the model.

4.5.2 Simulating clone size distribution and composition

So far, we have found a model that explains the (single-) cell fate mapping data, the measured proliferation differences between the subsets, and the dropout kinetics with biologically reasonable parameters. In the following, we show Gillespie simulations based on these parameters, see [Gillespie, 1977]. This examination is of great importance, as good model fits do not necessarily imply consistency of forward simulations with the data, which

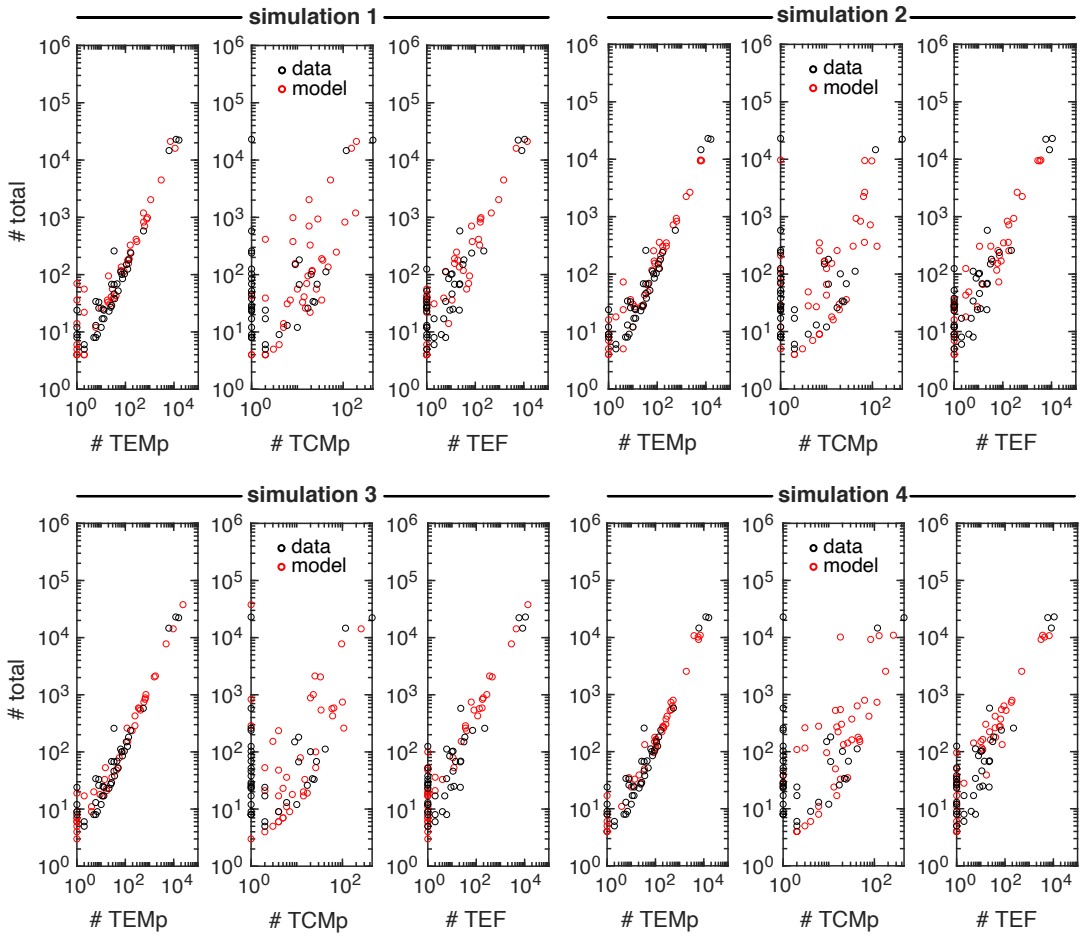

Figure 4.17: Representative Gillespie simulations for the cell cycle dropout model based on the best fitting parameters shown in Tab. 4.2. Here, the absolute size of a clone was plotted against the absolute number of TCMp, TEMp, and TEF cells, respectively. The number of simulated clones corresponds to the number of measured clones. (Experiments conducted by Veit Buchholz, Theresa Busch, and Lorenz Kretschmer)

has already been illustrated in section 4.4. Fig. 4.17 shows four representative simulation results evaluated at day 8 p.i., where the number of simulated clones corresponds to the number of measured clones ($n = 40$). The absolute size of a clone was plotted over the absolute number of cells of the respective phenotype in a log-log plot. Zeros were plotted on the axes. First, let us focus on the data (black): In the distribution of clone sizes, it is immediately apparent that three clones are strongly distinguished from the rest by their increased cell numbers ("giants" vs. "dwarfs"). In this sense, three clones out of 40 account for about 95.8 % of the total cell numbers. This dichotomy is a key

characteristic of the LCMV Arm data set and is not found to this extent in the *L.m.*-OVA data analysed in [Buchholz et al., 2013], see chapter 5. Furthermore, one of these three giants is characterised by the fact that it has no TCMp cells. These are highly constraining features of the data that were not explicitly considered in their summary statistics during parameter estimation.

Given the four simulation results (red), one observes that the dichotomy pattern is reproduced extremely well. The simulations are even able to produce giants that have no memory cells on day 8 p.i. The mechanistic basis of a giant in the model is mainly given by particularly early completion of the first cell division. The mean value of the associated waiting time distribution is 3.8 d, while the standard deviation is 1.9 d. This can give very early activated cells a proliferative advantage of more than two days over average cells. However, there is a second independent effect that modulates clone size: after activation, those clones become large that make the transition from slower CTCMp cells to the faster CTEMp compartment very rapidly. That this transition occurs immediately is rather rare, since $r_{\mathrm{CTCMp}\to\mathrm{CTEMp}}/(r_{\mathrm{CTCMp}\to\mathrm{CTEMp}}+p_{\mathrm{CMp}}) \approx 0.21$, i.e. a CTCMp cell will divide much more often than differentiate into the faster CTEMp compartment. However, if such a transition occurs immediately (for both CTCMp cells), then this results in a clone having no TCMp cell at all. At later time points, which are characterised by increased cell numbers, the inherent model stochasticity is only of minor relevance for the question of whether a clone becomes a giant or not.

We want to emphasise that small differences in division-time (or division speed) can cause enormously large differences in cell numbers under exponential growth. However, the experimentally observed gap in clone size between about 10^2 and 10^4 cells is probably a consequence of the small number of measured clones. In the model, there is no general mechanism preventing clones from acquiring cell numbers in this range. Indeed, the simulations show that the complete continuum of clone sizes can in principle be occupied. This prediction would have to be checked in future experiments.

The simulated correlation patterns of the dwarfs are also consistent with the data. However, the model is not able to explain the observed continuity in the size of small clones without any TCMp cells, particularly between 10^1 and 10^2 cells (see cells on the y-axis of the middle plots of Fig. 4.17). This is because the model too rarely generates clones without any TCMp cells. This in turn is closely linked to the fact that the first cell division directly generates two CTCMp cells. However, if this first cell division is removed from the model, the variabilities of the TEMp and TEF compartments exceed the measured CVs enormously (results not shown). We suspect that modelling of stochasticity in

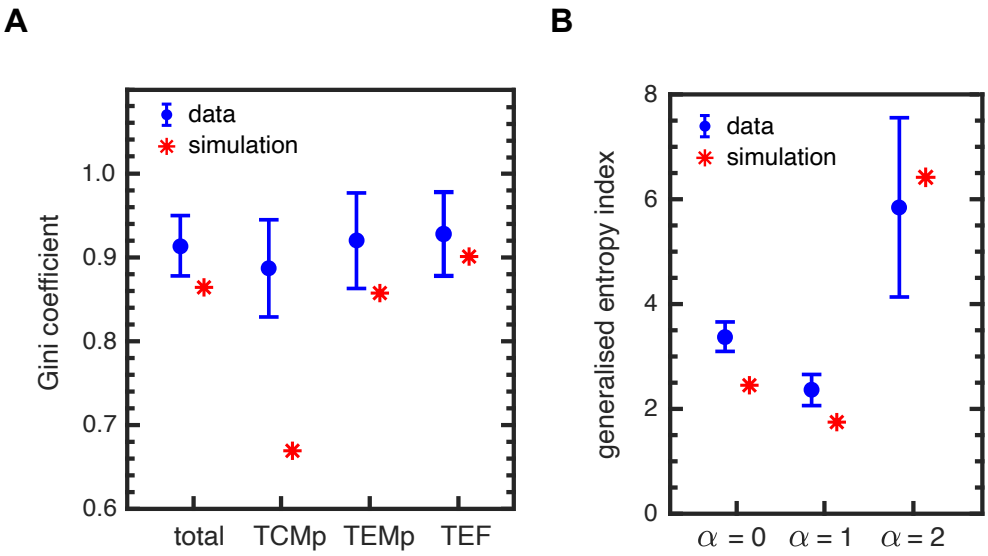

Figure 4.18: Inequality measures to quantify the dichotomy observed in the clone size distribution of the LCMV Arm single-cell fate mapping data at day 8 p.i. Simulation results are based on parameters shown in Tab. 4.2 and $N = 10^4$ simulated clones. **(A)** Gini coefficients of the total clone size and compartment size distributions. **(B)** Generalised entropy index of the total clone size distribution for typical choices of α. Bigger α increase the sensitivity of the measure with respect to the occurrence of giants. (Experiments conducted by Veit Buchholz, Theresa Busch, and Lorenz Kretschmer)

the context of Markov processes could be the reason for this behaviour, since exponential waiting times systematically overestimate the variabilities.

A robust way of quantifying the earlier described dichotomy in the data is provided by the Gini coefficient G, cf. [Gini, 1912], defined as

$$G = \frac{\sum_{i,j}^{N} |x_i - x_j|}{2N^2 \langle x \rangle}, \tag{4.2}$$

where x denotes the distributed quantity of interest and x_i its observations. The Gini coefficient takes values between zero and one, while $G = 1$ corresponds to maximal inequality. In our case, we are interested in the clone size distribution. Fig. 4.18A shows the Gini coefficient of the data and the simulation ($N = 10^4$ clones), with the uncertainties of the data being calculated by the bootstrap method. It can be seen that there is a good agreement for the total cell number distribution, as well as for the TEMp and TEF cell number distributions. For the TCMp cells, however, we encounter the same problem that

we have explained before: here we lack variability, which is reflected in the fact that the model underestimates the CV of the TCMp cells at day 8 p.i., see Fig. 4.15B.

Another inequality or redundancy measure is given by the generalised entropy index GE(α), see [Cowell, 2000], defined by

$$\mathrm{GE}(\alpha) = \begin{cases} -\frac{1}{N}\sum_i^N \ln\left(\frac{x_i}{\langle x\rangle}\right) & \alpha = 0 \\ \frac{1}{N}\sum_i^N \frac{x_i}{\langle x\rangle}\ln\left(\frac{x_i}{\langle x\rangle}\right) & \alpha = 1 \\ \frac{1}{N\alpha^2 - N}\sum_i^N \left[\left(\frac{x_i}{\langle x\rangle}\right)^\alpha - 1\right] & \alpha \notin \{0,1\}, \end{cases} \tag{4.3}$$

where the parameter α offers a possibility to make the index more sensitive to the occurrence of smaller or larger clones. Here, the measure for bigger α ($\alpha > 1$) becomes more sensitive to the occurrence of giants. The range of possible index values is given by $[0, \infty]$, where zero stands for maximum equality, i.e. a δ_{D}-peaked distribution of clone sizes. The resulting entropy indices for typical choices of α are depicted in Fig. 4.18B, where we restrict ourselves to the total clone size. In the case of $\alpha = 0$, the index measures the mean log deviation. Here we have the biggest discrepancy between simulations and data (uncertainties based on the bootstrap method). For $\alpha = 1$, the measure equals the Theil index, see [Theil, 1967], and $\alpha = 2$ is proportional to the squared CV of the distribution. The latter is particularly sensitive to large clone sizes and here we have the best agreement between data and model. One reason for this is that the CV was directly considered for model fitting. Furthermore, the other measured moments are of course also dominated by the three giants in the data. This leads to a strong overemphasis of the large clones during parameter estimation and is reflected in increasingly better matches between simulation and data for bigger α.

4.5.3 Identifiability of estimated parameters

In this section, we briefly comment on the identifiability of all estimated model parameters. Fig. 4.19A shows an overview of the local minima found in the optimisation. Here, the fits were plotted over their respective χ^2_{min}-values and sorted in descending order according to their negative log-likelihoods. The *waterfall* plot is based on $N = 10^3$ fits, with the initial parameter values chosen using the Latin hypercube method, cf. [McKay et al., 2000]. The lowest value of $\chi^2_{\mathrm{min}} \approx 112$ is obtained by a substantial subset of possible starting values, indicating that this fit might be associated with the global minimum. The parameters

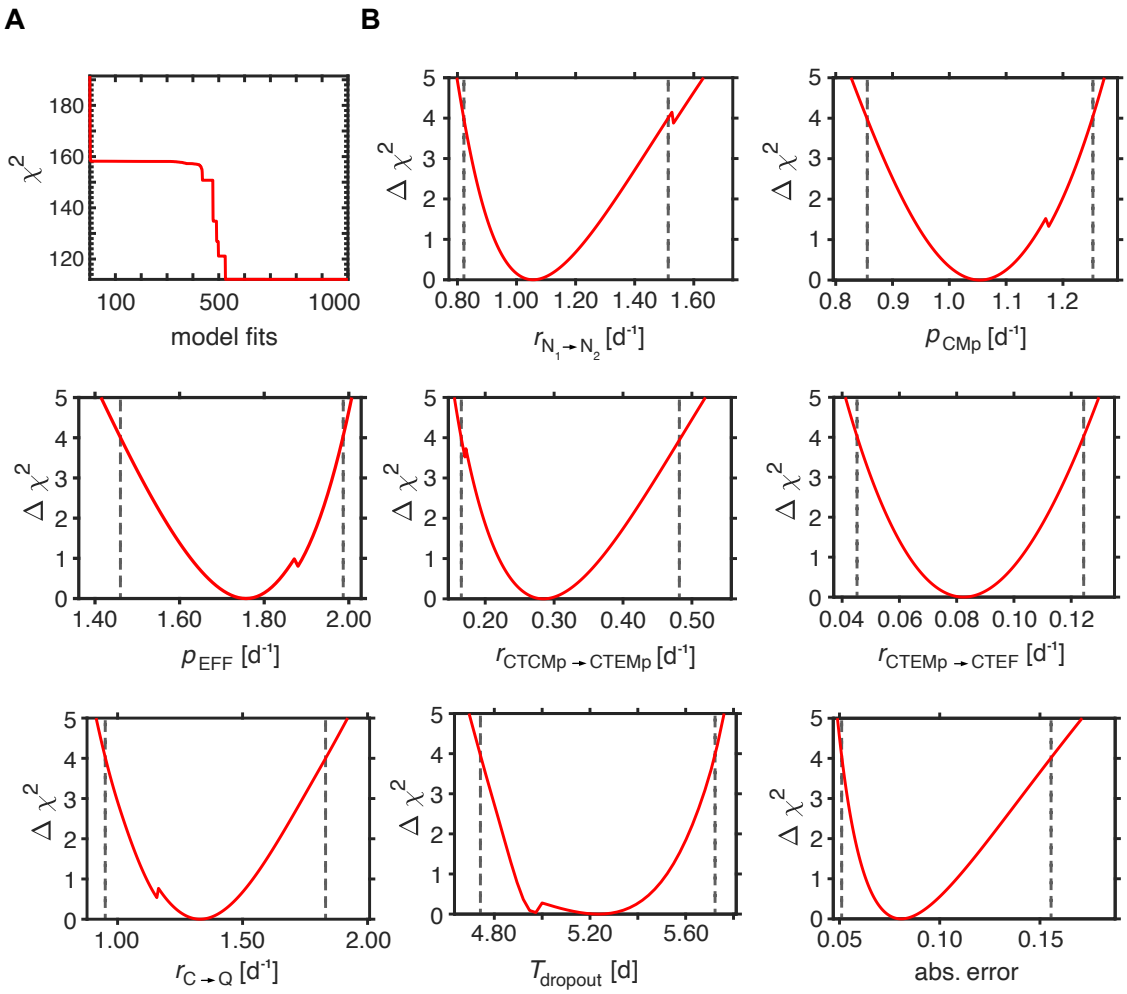

Figure 4.19: Visualising the identifiability of the parameters for the simplest variant of the cell cycle dropout model. **(A)** Waterfall plot based on $N = 10^3$ starting points in parameter space subject to Latin hypercube sampling. **(B)** Profile likelihoods leading to the 95 % confidence intervals shown in Tab. 4.2. Differences in χ^2 are measured with respect to χ^2_{min}, i.e. the value at the global minimum shown in (A). All eight model parameters are identifiable in this way, where dashed lines indicate the confidence bounds.

of this best fit and their 95 % confidence intervals have already been shown in Table 4.2. The confidence bounds were determined using the profile likelihood method, cf. [Venzon et al., 1988]. The profile likelihoods of all eight parameters underlying the cell cycle dropout model, including the estimated error for the scRNA-seq-based fractions of the cycling cells, are shown in Fig. 4.19B. The dashed lines mark the parameter bounds to which $\Delta\chi^2 = 4$ applies. All parameters have both an upper and a lower bound and are thus identifiable.

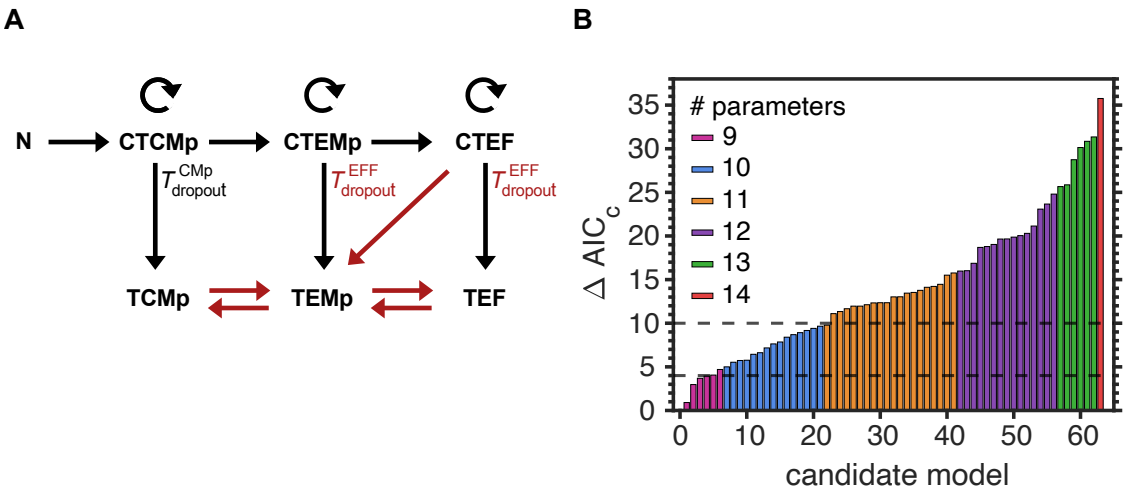

Figure 4.20: Systematic comparison between extensions made to the basic cell cycle dropout model scheme from section 4.5. **(A)** Most complex model variant of the cell cycle dropout model. Here, five additional transitions are considered (red). Furthermore, different dropout times between memory-like and effector-associated cells are allowed. Together with the simplest model scheme, this results in 64 possible candidate models to be tested. **(B)** Differences in corrected AIC-values with respect to the basic model for all 64 possible combinations. The candidate models have been sorted by their AIC difference in ascending order and their number of parameters is shown. Dashed lines represent differences in corrected AIC-values of four and ten, respectively.

4.6 Model selection

We have previously argued that the simplest variant of the cell cycle dropout model is to be favoured over more complex models. We now want to justify this in a systematic way. To this end, we compare it with modified versions. Our approach is guided primarily by the fact that any possible model extension must be consistent with the scRNA-seq data analyses. Therefore, these represent a strong constraint to be met in the model selection process. Since all models are fitted to the same data, a formal comparison using the (corrected) Akaike information criterion (AIC_c) is suitable, cf. [Akaike, 1974].

4.6.1 Extensions of the basic model scheme

There are two possibilities to extend the model scheme shown in Fig. 4.9 that are consistent with the scRNA-seq trajectory inference from section 4.1.2. First, additional transitions between the compartments can be inserted and second, the dropout rate can be considered phenotype-specific. Reasonable hypothetical transitions between the sub-

sets include CTEF → TEMp, TCMp → TEMp, TEMp → TEF, TEF → TEMp, and TEMp → TCMp. Fig. 4.20A graphically depicts these additional transitions. For the parametrisation of the time-dependent dropout rate, we could assign T_{dropout} and $\tilde{r}_{\text{C}\to\text{Q}}$ a respective subset label. These two parameters are both well identifiable. Within their 95 % confidence bounds, they are highly dependent on each other, since big rates $\tilde{r}_{\text{C}\to\text{Q}}$ imply short times T_{dropout} and vice versa. That is why we restrict ourselves only to the dropout time for simplicity, i.e. $T_{\text{dropout}} = T^{(i)}_{\text{dropout}}$, $i \in \{\text{CMp}, \text{EFF}\}$. Altogether, this formally results in $2^6 = 64$ different candidate models (Fig. 4.20A).

As a result, the simplest model has the smallest AIC_c-value of 137.1 for eight parameters and a total of 25 data points. Hence, all other models are compared to this best model. We follow the convention that candidate models with differences of $\Delta\text{AIC}_\text{c} < 4$ cannot be rejected and models leading to $\Delta\text{AIC}_\text{c} > 10$ can be rejected, see [Burnham et al., 2002]. For models in between, no clear decision can be made. Fig. 4.20B graphically depicts the results of all model comparisons with respect to the best model. Model indices were assigned in ascending order with increasing $\Delta\text{AIC}_\text{c}$. The colour encodes the number of model parameters, ranging from $n_{\text{par.}} = 9$ to $n_{\text{par.}} = 14$. The dashed lines indicate differences of delta $\Delta\text{AIC}_\text{c} = 4$ and delta $\Delta\text{AIC}_\text{c} = 10$, respectively. The models are grouped according to their complexity without exception. Simple models with one additional parameter cannot be distinguished from the best model, with one exception. For all models with two additional parameters, no clear decision can be made. Models with three or more additional parameters can be rejected with one exception (model with index 22 and $\Delta\text{AIC}_\text{c} = 8.4$). This model contains two additional transitions that are progressive.

In summary, minor extensions of the proposed cell cycle dropout model in terms of two additional parameters cannot be rejected. First, this means that simple models are systematically preferred, and second, that models of the same complexity class are not discriminated by the data. Furthermore, the simple models lead to acceptable model fits. We conclude that the scRNA-seq-based inference of differentiation topologies has already functioned as an efficient preselection process. In the following, we will focus on the simplest model variant, which has already been discussed in more detail.

4.6.2 Comparison of different time-dependencies

So far, we have used the step function proposed in eq. 4.1 as the time-dependence of the cell cycle dropout rate $r_{\text{C}\to\text{Q}}$. Since this concrete choice seems arbitrary, we want to test some other conceivable time-dependencies for comparison. For this purpose, we focus on sigmoid functions, whose derivatives are non-negative at every point in time. We set the

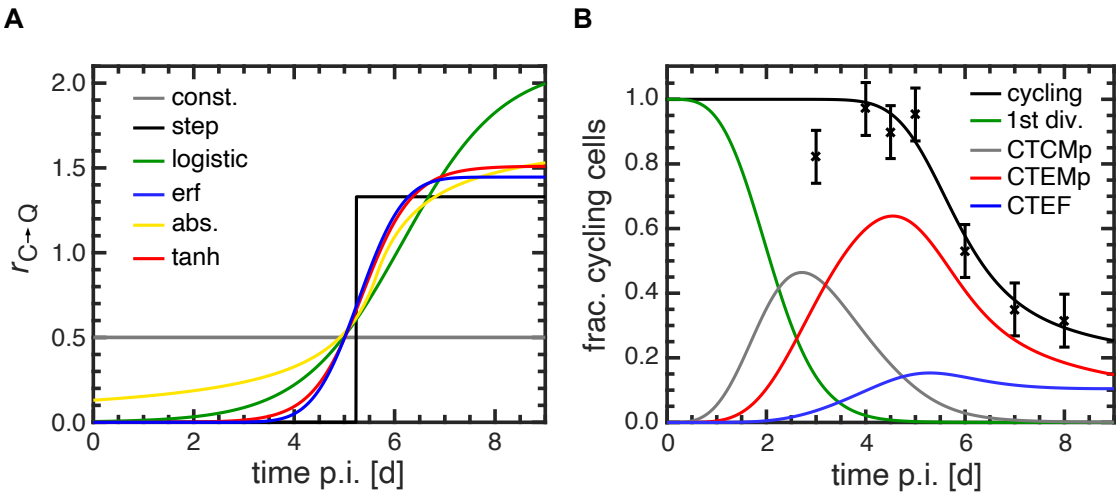

Figure 4.21: Implementation of sigmoidal functions as possible choices for the underlying time-dependence of the cell cycle dropout rate. **(A)** Best fit results of the time-dependent dropout rate for a selection of rescaled and shifted sigmoid functions (coloured) with two parameters each. Also shown are the step function used so far (black) and a constant rate (grey). **(B)** Same plot as in Fig. 4.16A for the choice of the error function instead of the step function shown in (A). The resulting dropout kinetics is now differentiable at any point in time.

image of the functions in the range $[0; \tilde{r}_{C\to Q}]$ and shift the functions in time such that a point symmetry of the curves with respect to the point $(T_{dropout}, 1/2\,\tilde{r}_{C\to Q})$ is achieved. All curves are thus characterised by two parameters, which, however, cannot directly be compared to each other. Defining $\tau := t - T_{dropout}$, our choice of functions reads

$$r^{\text{logistic}}_{C\to Q}(\tau) = \frac{1}{2}\left(1 + \frac{1}{1+e^{-\tau}}\right)\tilde{r}_{C\to Q}, \tag{4.4}$$

$$r^{\text{erf}}_{C\to Q}(\tau) = \frac{1}{2}\left(1 + \text{erf}(\tau)\right)\tilde{r}_{C\to Q} = \frac{1}{2}\left(1 + \frac{2}{\sqrt{\pi}}\int_0^\tau dz\, e^{-z^2}\right)\tilde{r}_{C\to Q}, \tag{4.5}$$

$$r^{\text{abs}}_{C\to Q}(\tau) = \frac{1}{2}\left(1 + \frac{\tau}{1+|\tau|}\right)\tilde{r}_{C\to Q}, \tag{4.6}$$

$$r^{\text{tanh}}_{C\to Q}(\tau) = \frac{1}{2}\left(1 + \tanh(\tau)\right)\tilde{r}_{C\to Q} = \frac{1}{2}\left(1 + \frac{e^{\tau} - e^{-\tau}}{e^{\tau} + e^{-\tau}}\right)\tilde{r}_{C\to Q}. \tag{4.7}$$

Fig. 4.21A shows the time-dependent dropout rate of the respective best fit result for the different functions (coloured) and the previous step function (black). In addition, a model was chosen that ignores the time-dependence of the rate (grey). As a result, our previous step function model has the overall smallest AIC_c-value. In the following,

parameter	best fit	95 % CI
$r_{N_1 \to N_2}$	1.08 d^{-1}	[0.83; 1.54] d^{-1}
p_{CMp}	1.07 d^{-1}	[0.87; 1.25] d^{-1}
p_{EFF}	1.78 d^{-1}	[1.48; 2.00] d^{-1}
$r_{CTCMp \to CTEMp}$	0.28 d^{-1}	[0.17; 0.48] d^{-1}
$r_{CTEMp \to CTEF}$	0.08 d^{-1}	[0.04; 0.12] d^{-1}
$\tilde{r}_{C \to Q}$	1.45 d^{-1}	[1.01; 2.05] d^{-1}
$T_{dropout}$	5.28 d	[4.65; 5.92] d
error(frac. cycling)	0.08	[0.05; 0.16]

Table 4.3: Best fit parameters for the cell cycle dropout model fitted to data from acute LCMV Arm infection. In contrast to the parameters shown in Tab. 4.2, the underlying model is based on a differentiable choice for the dropout rate given by the error function, cf. eq. 4.5. Note that the parametrisation of the dropout rate in terms of $\tilde{r}_{C \to Q}$ and $T_{dropout}$ has a different meaning now and cannot directly be compared to the estimations in the previous table. 95 % confidence intervals (CIs) are based on profile likelihoods.

the other candidate models are compared to this same model. The constant model is associated with $\Delta AIC_c = 17.4$. The error of the scRNA-seq-based fractions of cycling cells is estimated to be 0.27. If the error is bounded to 0.1, the result is $\Delta AIC_c = 45.0$. Thus, the constant model can be clearly rejected.

For the other models we get $0.1 \leq \Delta AIC_c \leq 3.4$. The best model in terms of AIC is the error function model with $\Delta AIC_c = 0.1$. All models lead to comparable-looking fits. Table 4.3 shows the estimated parameters of the best fit for the error function model as a representative of the four sigmoid functions shown. The corresponding dropout kinetics are shown in Fig. 4.21B, exhibiting differentiable kinetics around $T_{dropout}$. A comparison with the parameters of the step function model shows that the parameter sets differ only slightly. The data, therefore, do not offer any possibility to distinguish between the different choices of time-dependencies. We therefore retain the step function used, as it allows an easier interpretation of the results obtained. However, we will return to the error function model in section 4.7.

4.6.3 Clonal cell cycle dropout model

Next, we would like to investigate the extent to which the overall variability in the data may be due to kinetic variability between different clones. There is evidence from *in vitro* studies that a major contribution to heterogeneity in dividing T and B cell populations can be explained by the concept of "division destiny", see [Heinzel et al., 2017; Marchingo et al., 2014; Turner et al., 2008]. In its most recent formulation, an activated cell is assigned a

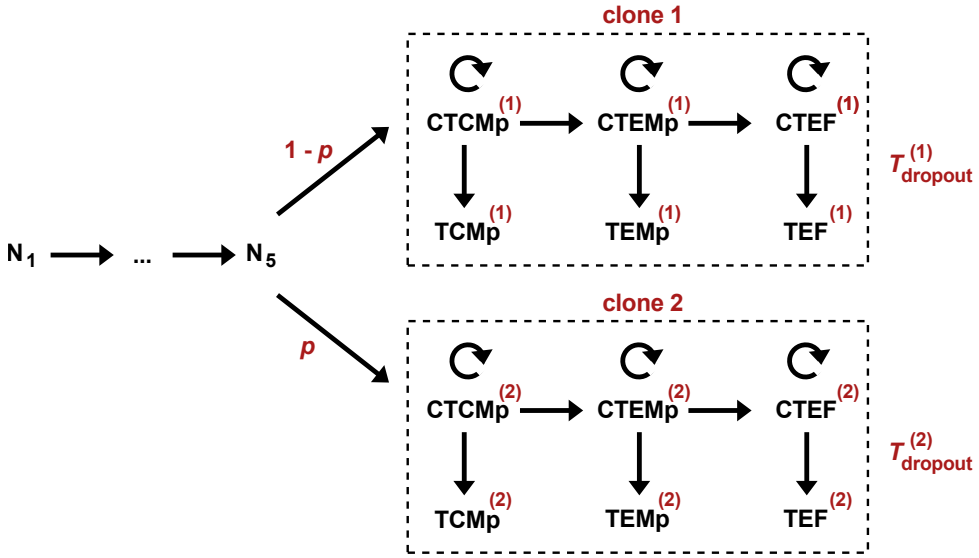

Figure 4.22: Topology of a clonal dropout model in which the two clones exit the cell cycle at fixed but different times. The additional source of variability is controlled via the dichotomy parameter p. In the limit of $p \to 0$, the original model is obtained. By construction, all other parameters between the two model branches are identical.

maximum time to divide before dying or becoming quiescent. This time is distributed, with the underlying parameters of the distribution depending on cell type, stimulus conditions, etc. In our case, although all clones exit the cell cycle at the same time, they do not enter it at the same time: due to the highly variable recruitment at the beginning (standard deviation of waiting time given by 1.9 d), there can be considerable differences between clonal cycling times. We now want to emphasise these clonal differences to a larger extent and study the resulting model response.

To this end, we retain the basic model topology but extend it such that recruited cells enter one of the two submodels shown in Fig. 4.22. A possible biological factor influencing this decision could be given, for example, by the initial antigen stimulus that the naïve cells are receiving. The only difference between the two submodels is the dropout time. Therefore, the dropout time is assigned a clonal label, i.e. $T_{\mathrm{dropout}} = T^{(i)}_{\mathrm{dropout}}$, $i \in \{1, 2\}$. Without loss of generality, we set $T^{(1)}_{\mathrm{dropout}} < T^{(2)}_{\mathrm{dropout}}$. The parameter p represents the proportion with which the cells after recruitment belong to the clone whose cells divide longer. The underlying idea is that the dichotomy in the data could be better represented by two extremes in the cycling times: for example, cells can either become giants with a probability of p or they become dwarfs with a probability of $1 - p$. In the limit of $p \to 0$, our previous model is obtained.

In the best fit, the essential parameter p is estimated to be 78 %, with a difference in dropout times of $\Delta T_{\text{dropout}} = 1.5$ d. Thus, the model does not represent the dichotomy property of the data in a desirable way. In the data, only $3/40 = 7.5$ % of the clones are accounted for by giants. The parameters p and $T^{(i)}_{\text{dropout}}$ are not determined by the strong inequality of clone sizes. Applied to this particular case, this model does not add any conceptual value and, compared to our previous model, results in $\Delta\text{AIC}_\text{c} = 8.5$.

In case of not estimating the dichotomy parameter p, but setting it directly to $p = 7.5$ %, the result is $\Delta T_{\text{dropout}} = 0.4$ d, i.e. the clones hardly differ from each other and we obtain almost the same results as in the basic model. This indicates that the introduction of an additional clonal source of variability in terms of the specific choice depicted in Fig. 4.22 is not further necessary. It seems as if there is already enough variability in the recruitment process, even though the CV of the associated waiting time distribution has already been halved. This assumption can be checked by removing all variability from the recruitment process, i.e. replacing the Erlang distribution with a delta distribution. In this case, we obtain a complete recruitment/first cell division at 2.9 d p.i. with dropout times of $T^{(1)}_{\text{dropout}} = 5.0$ d and $T^{(2)}_{\text{dropout}} = 6.7$ d, i.e. a difference of about 1.7 d. Interestingly, the dichotomy parameter is now estimated to be $p = 9$ %. This is exactly in the expected range, especially considering that the observed value of 7.5 % is subject to a Poisson error. The 95 % confidence interval for p lies between 0.2 % and 23.5 %. The lower bound is so close to zero because small values of p always produce large values of $T^{(2)}_{\text{dropout}}$ close to 8 d p.i. This allows a very small proportion of recruited cells to be in exponential growth for a very long time, thus producing a high dichotomy level.

Overall, we find that some form of clonal variability in the cycling time is needed to explain the observed heterogeneity consistently. Our data cannot discriminate between this time being at the beginning of the expansion phase (recruitment and first cell division) or the end of it (division destiny). However, with our basic model we have provided enough clonal variability so that adding more variability does not improve the fits. Of course, by considering only two different clonal dynamics, our tested model has been rather extreme. In principle, these considerations could be extended by discretising the dropout time to fit whole histograms to the data. However, this would lead to expensive calculations within the moment equations approach. We also want to emphasise that a similar skewness in the clone size distribution could also be created by a clonal distribution of cell cycle speed. Nevertheless, it would most probably not be possible to distinguish between these two mechanisms based on the given data.

4.7 Predicting phenotypic cell cycle dropout dynamics

The observations so far have strengthened the impression that the cell cycle dropout model reliably captures the essential degrees of freedom relevant to the dynamics during the early phase of infection. Indeed, all model parameters have been identifiable. In order to test the predictive power of this model, we focus on phenotype-specific cell cycle dropout effects. To this end, we must recall the following assumption we made: at time T_{dropout}, all cells start to become quiescent at the same rate $\tilde{r}_{\mathrm{C}\to\mathrm{Q}}$, regardless of their current phenotypes. Within our model, we can now ask how the proportions of cycling cells per phenotype behave for times $t > T_{\mathrm{dropout}}$, as we initially had no available data to address this question.

Fig. 4.23A shows these same fractions over time. Up to $t = T_{\mathrm{dropout}}$, all cells are cycling by definition. After this time has passed, the fractions decrease in a subset-specific manner, which needs to be explained: Despite our assumption about the phenotype-unspecific dropout behaviour, cells proliferate and differentiate further after T_{dropout} is exceeded. Consequently, the net influx and efflux into the respective cycling compartments are different for each subset. For example, no cells differentiate into the cycling memory compartment for times beyond T_{dropout}, but some cells differentiate out. Although they keep on dividing, we have $p_{\mathrm{CMp}}/(\tilde{r}_{\mathrm{C}\to\mathrm{Q}} + r_{\mathrm{CTCMp}\to\mathrm{CTEMp}} + p_{\mathrm{CMp}}) \approx 0.40$. Therefore, the dividing fraction decreases exponentially. In contrast, the CTEMp cells receive an influx from the CTCMp compartment, proliferate more strongly, but pass on a proportion to the CTEF compartment. The CTEF compartment has the biggest influx and the smallest efflux, as being located at the end of the differentiation topology. This explains the hierarchy of decays in Fig. 4.23A. For day 6 p.i. (dashed line), a prediction was made for this hierarchy.

This same prediction was verified by my colleague Albulena Toska in a subsequent experiment. For this purpose, she measured phenotypic proportions of dividing cells at day 6 after LCMV Arm infection using phospho-Rb. However, since she used a different virus strain for this purpose, which has turned out to lead to faster and longer-lasting cell divisions, only ratios of these proportions can be compared with the model prediction. This comparison is shown in Fig. 4.23B. The errors of the data (blue) were estimated using the standard error of the mean, while the errors of the model prediction (red) were calculated using prediction profile likelihoods, cf. [Kreutz et al., 2012]. The hierarchy of decrease in the cycling compartments is thus correctly predicted by the model, although there is an overlap of the ratios within the 68 % confidence intervals. However, this was to be expected, especially since the exit of the cells from the cell cycle at $T_{\mathrm{dropout}} = 5.2$ d was too close to the time of the validation experiment. For slightly later time points, e.g.

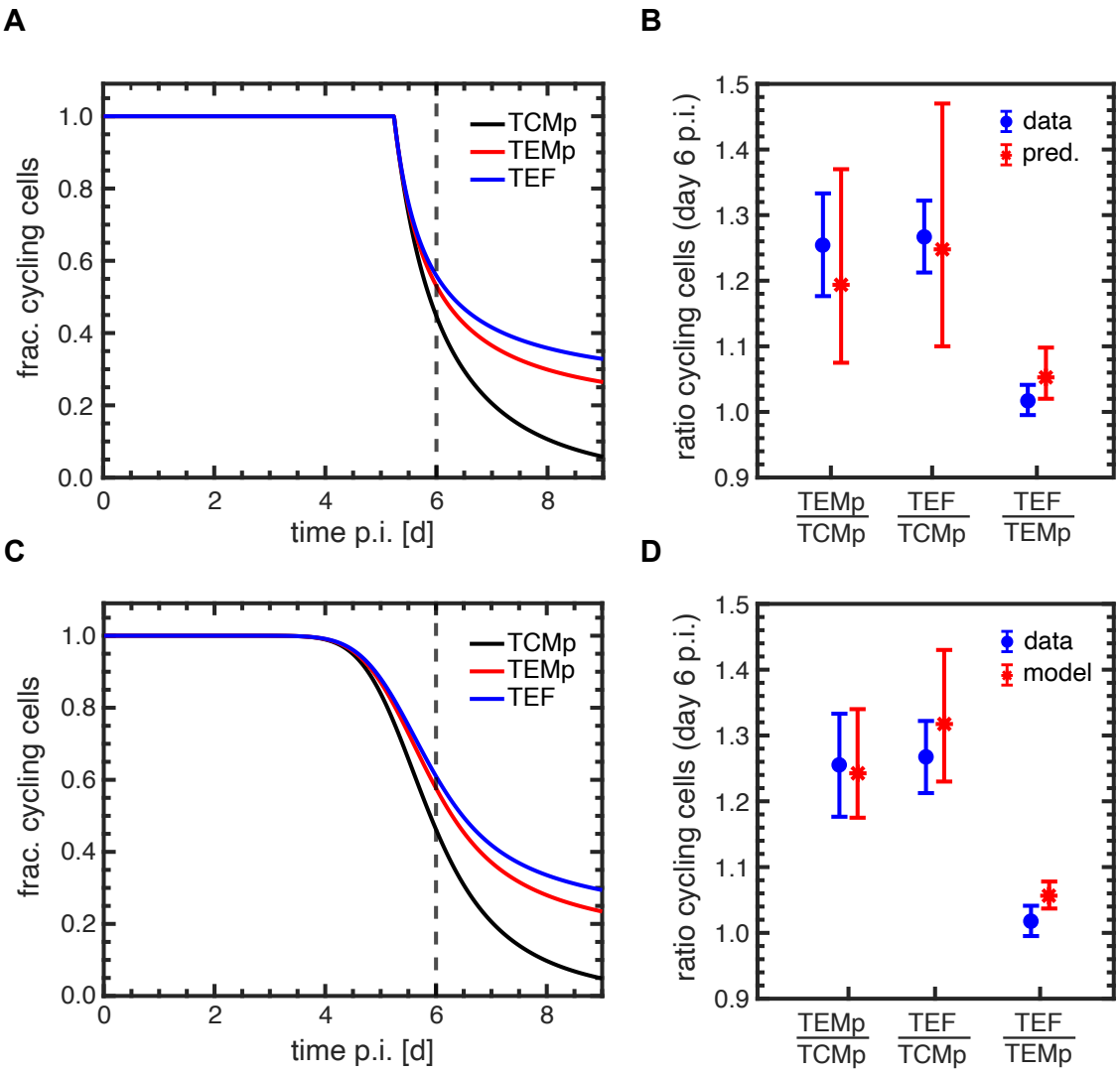

Figure 4.23: Model predictions of phenotype-specific cell cycle dropout dynamics. **(A)** Fraction of cycling cells per phenotype (across all cells of the same phenotype) over time. The dashed line indicates day 6 p.i., i.e. the time point associated with the model prediction. **(B)** Ratios of modelled kinetics from (A) evaluated at day 6 p.i. including 68 % confidence intervals (red) and phospho-Rb-based measurements of these same ratios (blue). **(C)** Same as in (A) but using an error-function-shaped time-dependence of the cell cycle dropout rate. Associated parameters have been shown in Tab. 4.3. **(D)** Model results including 68 % confidence bounds, corresponding to the kinetics from (C). This is not a model prediction, as this result was only obtained after the measurement. (Experiments conducted by Albulena Toska)

day 7 or 8 p.i., there is no overlap in the model predictions. Therefore, we recommend re-measuring on one of these two days.

Models with continuous time-dependent dropout rates allow cells to stop dividing earlier. Therefore, they might address the above problem, related to the early measurement at day 6 p.i., better. The previously discussed model with the error function as the chosen time-dependence shall be used as a representative example of such a model, see Tab. 4.3 for the estimated model parameters. Fig. 4.23C confirms the presumed increased distance between the TCMp curve and the TEMp/TEF curves at day 6 p.i. The resulting ratios are shown in Fig. 4.23D. Here we see a significant difference in the dropout hierarchy between the TCMp cells and the effector-associated cells within the 68 % confidence intervals. These results were obtained after the actual measurement and therefore do not represent proper predictions. Nevertheless, this analysis confirms that the timing of 6 d p.i. was merely chosen too early.

4.8 Analysis of variability contributions

As a final point, we would like to investigate to what extent the overall variability of our stochastic model can be related to individual variability-generating mechanisms. Overall, there are three major sources of variability that need to be distinguished: the inherent stochasticity of the model resulting from exponentially distributed waiting times of the individual reactions (i), clonal variability caused by recruitment and first cell division (ii), and variability arising from differences in cell cycle speed between memory and effector subsets (iii). All three mechanisms have already been discussed in detail. However, we would like to better understand their individual contribution to the coefficients of variation. To this end, we remove the variability of each of the mechanisms (ii) and (iii) separately, re-fit the model and compare the CVs with those of the full model. When removing the clonal variability, the underlying Erlang distribution is replaced by a Dirac-function, with the location of the peak being estimated. The variability associated with cell cycle differences is obtained by setting the two proliferation rates equal. Furthermore, the recruitment rate is set to the best fit result of the full model so that the model cannot compensate for the loss of variability in terms of smaller recruitment rates.

Fig. 4.24A shows the resulting CV-dynamics for the case where both sources (ii) and (iii) have been removed. These must be compared with the CVs of the original model shown in Fig. 4.24D. The estimated and constant time of recruitment is 2.9 d. The inherent stochasticity of the model is by far not sufficient to compensate for the other two mechanisms, despite the highly variable exponential waiting time distributions. If we again consider differences in cell cycle speed, the CVs of the TEMp and TEF compartments increase. In contrast, the CV of the TCMp cells decreases since their division speed is

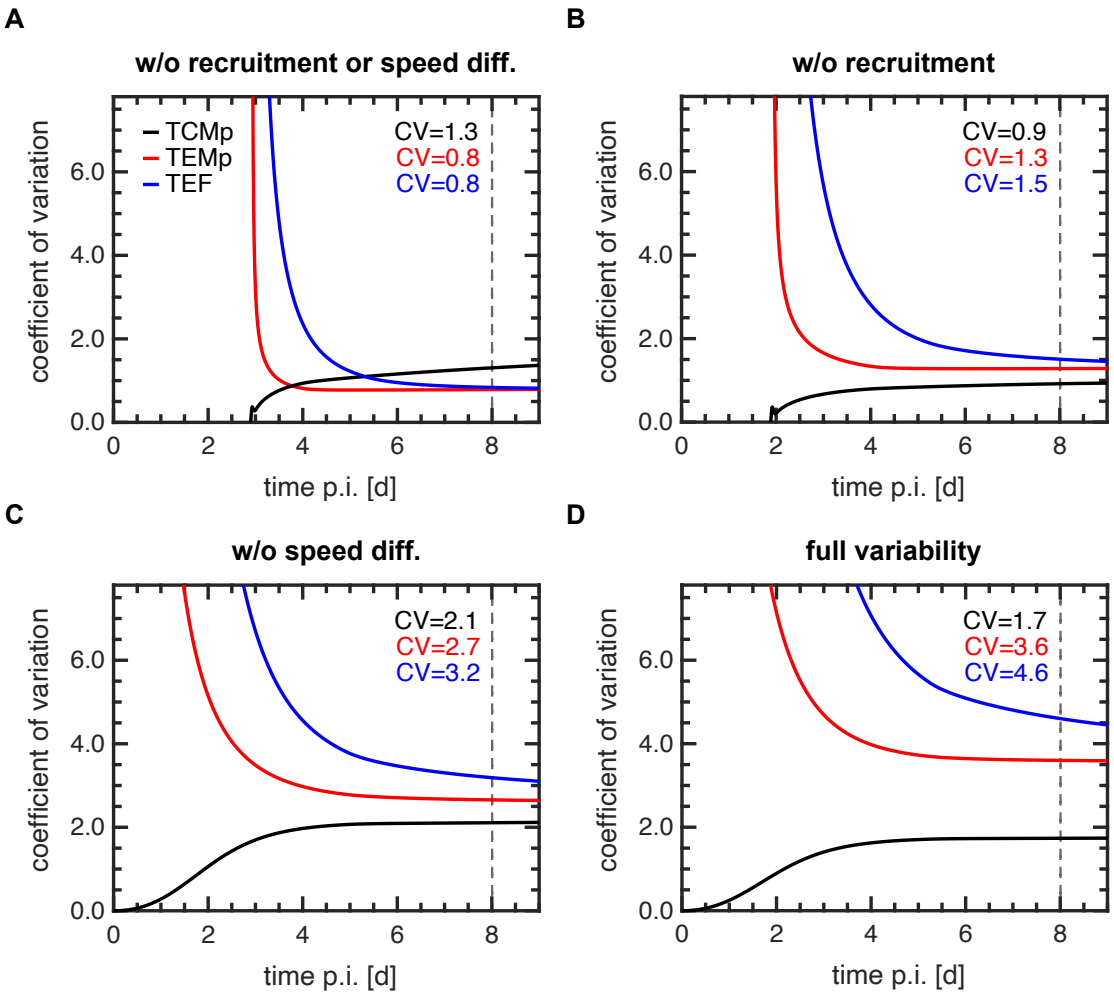

Figure 4.24: Effects of different variability-generating mechanisms of the cell cycle dropout model on the coefficients of variation. The model was re-fitted upon removal of individual mechanisms. CVs at day 8 p.i. are shown in numbers. **(A)** Clonal variability related to recruitment and first cell division, as well as variability caused by differences in cell cycle speed across subsets, were removed. **(B)** Variability caused by initial recruitment and first cell division was removed. **(C)** Proliferation rates of CMp and EFF subsets were set equal, thereby eliminating variability caused by differences in cell cycle speed. **(D)** Original model with full variability for reference.

now correctly described. In the previous case, division speed of the TCMp cells was highly overestimated, which lead to an increased CV. The reason for this is that an overestimation of the TCMp proliferation rate is not as significant as an underestimation of the TEMp and TEF rate. This is because at day 8 p.i. the latter compartments are much stronger represented in number. If we consider the opposite case of variable recruitment and equal

division speeds, we can conclude that recruitment makes a larger contribution to the variability, see Fig. 4.24C. An additional consideration of different proliferation rates strongly separates the CV of the TCMp cells from that of the effector-associated cells.

We find that the mechanism of recruitment, despite considering non-exponential waiting times, makes the largest contribution to the overall variability. This is particularly interesting as this source of variability is clonal. At the same time, it should be noted that the different division speeds of the subsets act as a strong amplifier of this very first source of stochasticity. Especially in a model in which the cells cannot be in division for an arbitrarily long time, the early differentiation from the slower CTCMp compartment into the much faster CTEMp compartment plays a decisive role. The model CVs for equal proliferation rates are not sufficient to explain the strong dichotomy (see section 4.5.2) of the data. However, precise quantification of the variability contributions is not possible in this framework, as the compared CVs of the different conditions are not additive.

Chapter 5

Model extensions beyond the acute phase of infection

In the previous chapter, we have presented a model that correctly explains the heterogeneity of individual $CD8^+$ T cell responses during the expansion phase of infection. This model is now extended to the contraction phase of the response. In the first part of this chapter, we restrict all considerations to the early phase of contraction. Then, we give an outlook on the full contraction phase, at the end of which the memory phase is initiated. To this end, we consider immune responses following *L.m.*-OVA infection since data points are available at late time points. This also gives us the opportunity, to apply our cell cycle dropout model to a different acute infection. It should be emphasised again that our central working hypothesis is that acute infections follow the same mechanisms, although the particular parameter values describing them can generally be very different.

5.1 Available data and model extensions

The structure of the data is the same as in the previous chapter. Single-cell fate mapping data from OT-I T cells during *L.m.*-OVA infection collected at day 8 and 12 p.i. are available. Cells were classified using the surface markers CD62L and CD27. In this way, $n = 76$ clones could be measured at day 8 and $n = 93$ clones at day 12 after infection. In addition, relative population sizes collected at day 1, 2, 3, 4, 6, 8, and 12 p.i. were available. All these data were obtained by Veit Buchholz and Dirk Busch and were published in [Buchholz et al., 2013]. Fractions of dividing cells at day 4.5, 8, and 12 after infection corresponding to the *L.m.*-OVA case have already been shown in Fig. 4.13. Here, the scRNA-seq-based inference has given results very similar to the subsequent FACS validation by measuring phospho-Rb. For parameter estimation, we opt for the FACS measurements here, as these were based on a larger number of mice. Likewise, the division speed quantifications of CMp and EFF subsets on day 4.5 and 8 p.i. are highly relevant. These have already been shown in Fig. 3.9 and are based on experiments by Albulena Toska.

In the following, we leave the model topology unchanged. In section 4.1.2, we have seen that in the course of the immune response the respective phenotypes continue to mature, while the cells mainly retain their phenotypes in terms of expression of relevant marker genes as *Slamf6*, *Cd27*, *Gzma/b*, and *Klrg1*, see Fig. 4.2. We do not consider it necessary to add another layer of compartments (e.g. by introducing the subsets TCM, TEM, and late TEF). Especially because we do not have any data that distinguish these more mature states from their respective precursor states. Based on the data mentioned above, we must consider some necessary model extensions. To start with, we account for our measurements that the proliferation rate of both CMp and EFF cells slows down from day 4.5 to 8 p.i. First, we describe theses dynamics in terms of exponential decays. However, since a basic exponential model overestimates the rate for early time points, we choose the following form, given by

$$p_i(t) = \begin{cases} c_i & t < \tau_i \\ \\ c_i \mathrm{e}^{-k_i(t-\tau_i)} & t \geq \tau_i, \end{cases} \tag{5.1}$$

i.e. cells initially divide at a constant rate c_i until time τ_i. After that, the rate drops exponentially with rate k_i, while i denotes the subset label. A slowing down of cell cycle speed together with the exit of cells from the cell cycle can, in extreme cases, stop cell growth completely. However, since cell numbers decrease on average between day 8 and day 12 p.i., death rates d_i must additionally be introduced. In the most general case we have $i \in \{\mathrm{TCMp}, \mathrm{TEMp}, \mathrm{TEF}\}$, assuming that only non-dividing subsets are affected by cell death. The kinetics of the model parametrised in this way has the advantage that one does not have to choose or estimate the onset of contraction (marked by the time at which cell death becomes relevant). This is because cell death is naturally suppressed for times smaller than T_{dropout}.

5.2 Determination of kinetic properties

In this section, it shall be investigated which model can best describe the initial contraction phase. For this purpose, we rely on the model selection results from section 4.6 as far as differentiation is concerned. Here, the emphasis is put on possible kinetic variants in terms of cell division and cell death. In the previous section, the kinetically most complex model was outlined. All subsets were assigned different constant death rates d_i and different proliferation rates p_i, each parametrised by the three parameters c_i, τ_i, and

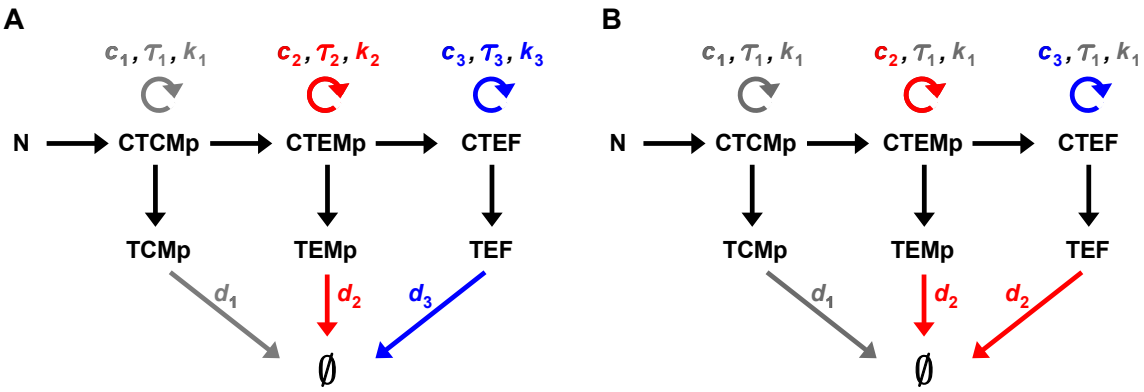

Figure 5.1: Overview of possible kinetic variants of the cell cycle dropout model extended to the early contraction phase. **(A)** Most general parametrisation scheme in which all three subsets show different proliferative dynamics in terms of eq. 5.1 and different death rates. **(B)** Output model of the forward selection procedure, in which all three compartments differ in their initial proliferation rate c_i and effector-associated cells have different death rates d_i compared to memory cells. The dynamics of proliferation in terms of τ and k is the same for all subsets.

k_i, see equation 5.1 for the choice of parametrisation. Fig. 5.1A depicts this model in its most general form.

To determine the model with the smallest $\mathrm{AIC_c}$-value, we use the forward selection method, see [Fröhlich et al., 2019; Hastie et al., 2009]. We start from the simplest model in which all three subsets have the same kinetics. This model has eight parameters less than the previously described most complex variant. In case a separate parameter is assigned for the CTEMp or the CTEF subset, the parameter of the respective other effector-associated compartment is set to the same value. As a result, the forward selection is terminated in the fourth step. This leads (in the order of the selected parameters) to an inclusion of the parameters c_{TEMp}, d_{TEMp}, and c_{TEF}, with $\mathrm{AIC_c} = 114.9$ for the possibly[1] best model, see Fig. 5.1B. This means that the resulting model considers differences in death rates between memory and effector-associated cells as well as differences in proliferation rates between all three subsets. A difference in death rates between CMp and EFF cells was to be expected, especially as it ensures the known short-lived nature of the latter cells. It is also interesting that all subsets seem to slow down their cycling speed simultaneously. We discuss the exact implications of this model in the next section when the best fit results are shown.

[1]Forward selection is a greedy algorithm. Therefore, identification of the most appropriate model with the smallest $\mathrm{AIC_c}$-value among all candidate models is not guaranteed.

Before we turn to the best model, let us consider other possible time-dependencies of the proliferation dynamics. So as not to to overestimate the rate for early time points, we retain our basic form of initially constant rate c up to time τ and subsequent k-dependent decay. Defining $t - \tau =: \tilde{\tau}$, this gives $p(\tilde{\tau} \leq 0) = c$. We have tested the following three alternative time-dependencies, given by

$$p^{\text{cosech}}(\tilde{\tau} > 0) = \frac{2c}{\mathrm{e}^{k\tilde{\tau}} + \mathrm{e}^{-k\tilde{\tau}}}, \tag{5.2}$$

$$p^{\text{hyperb}}(\tilde{\tau} > 0) = \frac{c}{\tilde{\tau}^k + 1}, \tag{5.3}$$

$$p^{\text{Gauss}}(\tilde{\tau} > 0) = c\,\mathrm{e}^{-(k\tilde{\tau})^2}. \tag{5.4}$$

Comparing these models to the exponential decay model from above, we get $\Delta\text{AIC}_\text{c} = 0.2$ (hyperbolic cosecant), $\Delta\text{AIC}_\text{c} = 2.9$ (hyperbole), and $\Delta\text{AIC}_\text{c} = 0.0$ (Gaussian). Hence, these different choices cannot be discriminated by the data.

5.3 Mechanisms of the early contraction phase

Next, we address the model obtained by forward selection using the exponentially decreasing proliferation rate that has been outlined in the previous section. Fig. 5.2 shows the best fit of this model in terms of the moments of the cell number distributions. It can be seen that on average the cell numbers decrease for all subsets between day 8 and 12 p.i. (Fig. 5.2A). Thereby, the TCMp subset already experiences a decrease in mean cell number at the estimated dropout time $T_\text{dropout} = 4.5$ d (cf. Tab. 5.1 for a complete list of all estimated parameters including their 95 % CIs). The cycling part of the TCMp compartment can no longer compensate for the outflow given by differentiation and death. On the other hand, the TEF compartment expands the longest, as it is the last state within the linear differentiation hierarchy. Consequently, the TEMp and TEF cells only decrease in number after their proliferation rates start to decrease exponentially at $\tau = 7.4$ d and hence cell death becomes dominant. Here, the death rate of TEMp/TEF cells is significantly higher than that of TCMp cells (0.7 d^{-1} vs. 0.2 d^{-1}). The peak of the modelled CD8^+ T cell response turns out to be located at $t = 7.9$ d p.i. with a mean cell number of $1.6 \cdot 10^4$. This result is in line with measurements carried out in [Porter et al., 2006], where the authors report peak times between 8 d to 9 d after *L.m.*-OVA infection.

In contrast, the CVs and correlation coefficients show little dynamics between day 8 and 12 p.i. In particular, the CV of the TCMp cells is underestimated at day 8 p.i. (Fig. 5.2B), whereas it is correctly described at day 12 p.i. Mechanistically, this is also due to the fact

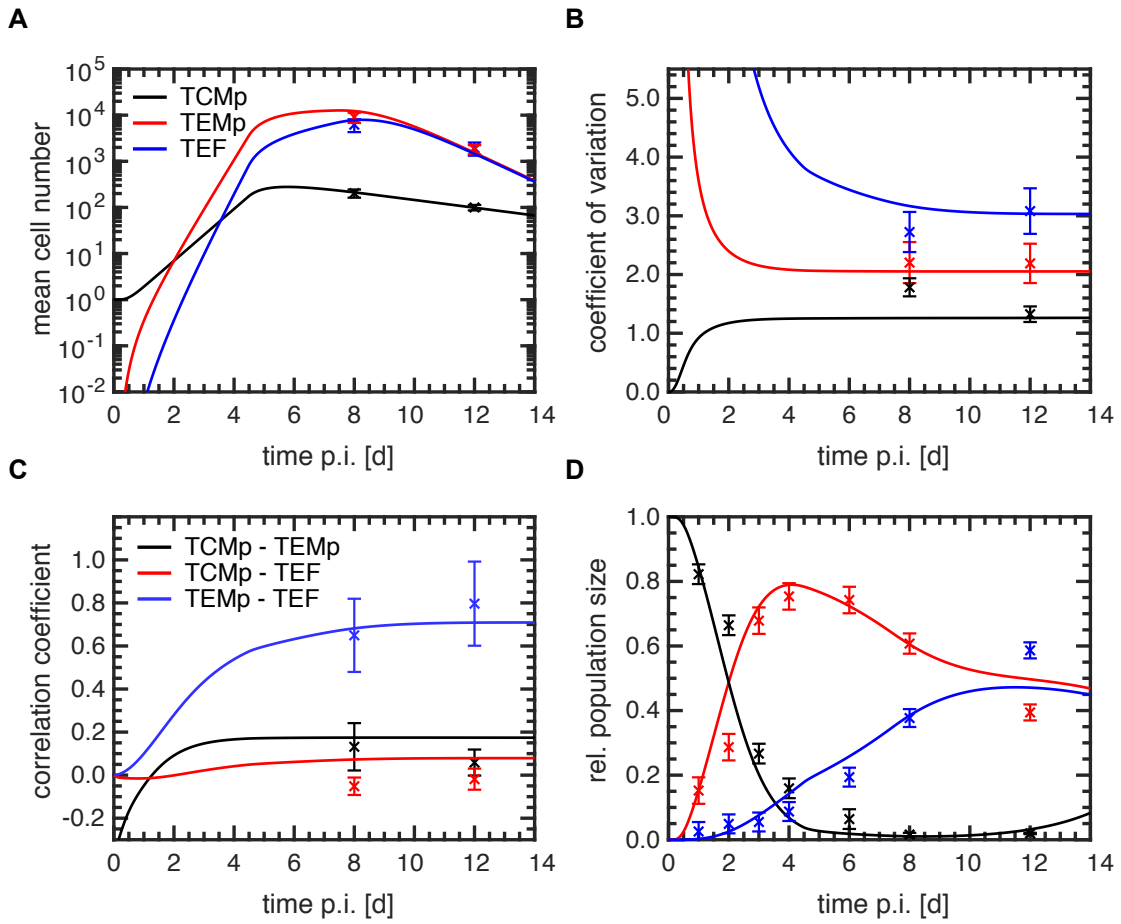

Figure 5.2: Best fit results for the extended cell cycle dropout model fitted to single-cell progeny data from day 8 and 12 post acute *L.m.*-OVA infection and kinetic data on cell cycle activity and speed. Associated parameter values are given in Tab. 5.1. Shown are the dynamics of the moments **(A, B, C)** and relative population sizes **(D)** of the respective compartments. Relative compartment sizes were not used for model fitting. (Data from [Buchholz et al., 2013])

that the initial proliferation rate of the TCMp cells ($c_{\mathrm{TCMp}} = 1.5\ \mathrm{d}^{-1}$) is much higher than the differentiation rate into the TEMp compartment ($r_{\mathrm{CTCMp}\to\mathrm{CTEMp}} = 0.2\ \mathrm{d}^{-1}$). This leads to less stochasticity in the initial response, which in turn leads to less variability in the size of the mean TCMp cell numbers. In addition, the waiting time for recruitment is relatively short, only about half a day on average, limiting the variability even further. Except for the TCMp-TEF correlation (Fig. 5.2C), the model reproduces the stochasticity of the data well. One way to reduce the TCMp-TEMp/TEF correlations within our model would be to put more emphasise on the proliferative differences between TCMp and TEMp/TEF cells. As already explained in section 4.5.1, the differentiation of a slow

parameter	best fit	95 % CI
$r_{N_1 \to N_2}$	7.50 d^{-1}	[5.55; 13.06] d^{-1}
$r_{CTCMp \to CTEMp}$	0.23 d^{-1}	[0.18; 0.35] d^{-1}
$r_{CTEMp \to CTEF}$	0.04 d^{-1}	[0.03; 0.06] d^{-1}
$T_{dropout}$	4.48 d	[4.43; 4.52] d
$\tilde{r}_{C \to Q}$	2.50 d^{-1}	[2.27; 2.67] d^{-1}
c_{TCMp}	1.53 d	[1.48; 1.59] d^{-1}
c_{TEMp}	2.48 d	[2.43; 2.60] d^{-1}
c_{TEF}	2.71 d	[2.63; 2.85] d^{-1}
τ	7.40 d	[6.88; 7.69] d
k	0.28 d^{-1}	[0.23; 0.41] d^{-1}
d_{CMp}	0.19 d^{-1}	[0.13; 0.29] d^{-1}
d_{EFF}	0.71 d^{-1}	[0.63; 0.91] d^{-1}

Table 5.1: Best fit parameters for the extended cell cycle dropout model, fitted to data from acute *L.m.*-OVA infection. Corresponding model fits are shown in Fig. 5.2 and 5.3, respectively. 95 % confidence intervals are based on profile likelihoods.

dividing into a fast dividing subset leads to a decorrelation of these two subsets. In our case, the only reasonable option to increase division speed differences between TCMp and TEMp/TEF compartments would be given by an even higher effector proliferation rate. In contrast, the absolute cell numbers would not be consistent with such a scenario. We conclude that either the dynamics of the proliferation rate is not ideally chosen or that dividing (effector) cells might already be affected by cell death as well.

Fig. 5.2D shows the relative population sizes obtained from bulk transfers. These were not used during parameter estimation. Until day 8 p.i., the relative subset kinetics of the phenotypes are surprisingly well described by the model. However, at day 12 p.i., relative sizes of TEMp and TEF cells are not in line with the data. It should be noted again, that the very early time points ($t < 4$ d) are affected by CD62L shedding, i.e. CD62L is cleaved upon activation of naïve cells, see [Yang et al., 2011].

The kinetic characteristics related to cell cycle are shown in Fig. 5.3. Between day 4.5 and 8 p.i. the majority of the cells stop dividing (Fig. 5.3A, black curve). By day 12 p.i., almost all cells become quiescent. The coloured curves show the phenotypic contributions and mainly reflect the progressive hierarchy of differentiation. The estimated dropout time of 4.5 d has a strikingly narrow confidence interval of 0.1 d in length. Regarding the fit of the proliferation rates (Fig. 5.3B), one can see that the effector rates are well described. As the results show, the TEFs have a slightly increased rate compared to the TEMp cells. In contrast, the TCMp rate is underestimated at day 4.5 p.i. As described above, one

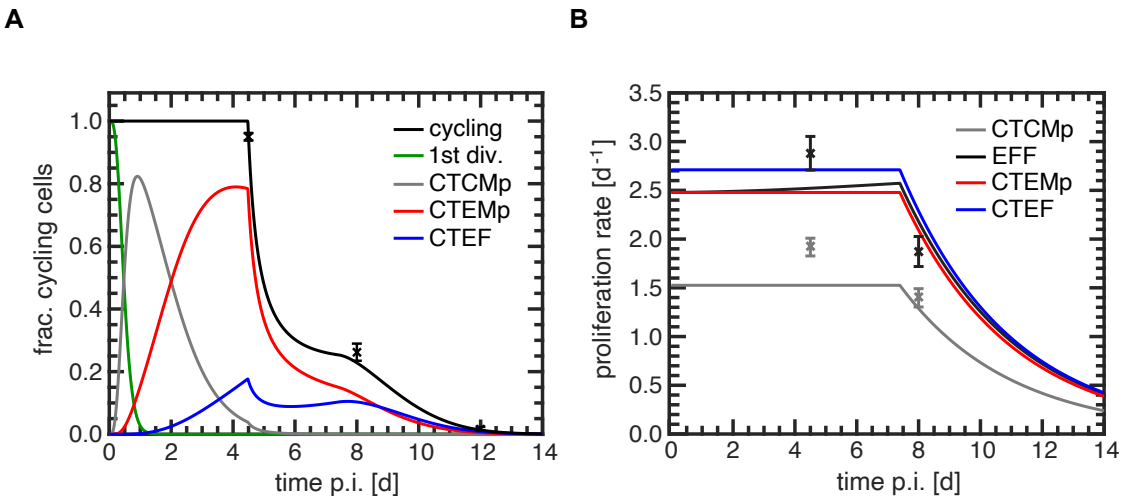

Figure 5.3: Best model fits of the cell cycle kinetics based on parameters shown in Table 5.1. **(A)** Model curves of the cycling activity of all dividing compartments (coloured) and their sum (black), with fractions referring to all cells. Data were gained in terms of phospho-Rb measurements. **(B)** Model fits of the time-dependent proliferation rates for cycling memory cells (grey) and effector cells (black). At the estimated time τ, proliferation rates begin to decrease exponentially with decay rate k. Both parameters are identical for all compartments. (Experiments conducted by Albulena Toska)

explanation is that for the low correlations between TCMp-TEMp/TEF there must be a sufficiently large proliferative difference. However, this was not measured to the same extent and could therefore indicate an inadequacy of the model. With regard to kinetics, it is noticeable that the used data are highly informative as all kinetic parameters can be determined with very narrow confidence intervals. In this way, even the death rates underlying the early contraction phase can be identified, which otherwise would be difficult to access experimentally.

Overall, we find that the dropout of cells from the cell cycle seems to be more relevant in regulating cell numbers during the course of the immune response than a slowing down of division speed. This is shown in terms of $T_{\text{dropout}} < \tau$, i.e. at the time of slowing down division speed, most cells are already no longer dividing. Consequently, this leads to a less identifiable decay rate k compared to the other parameters (see Tab. 5.1).

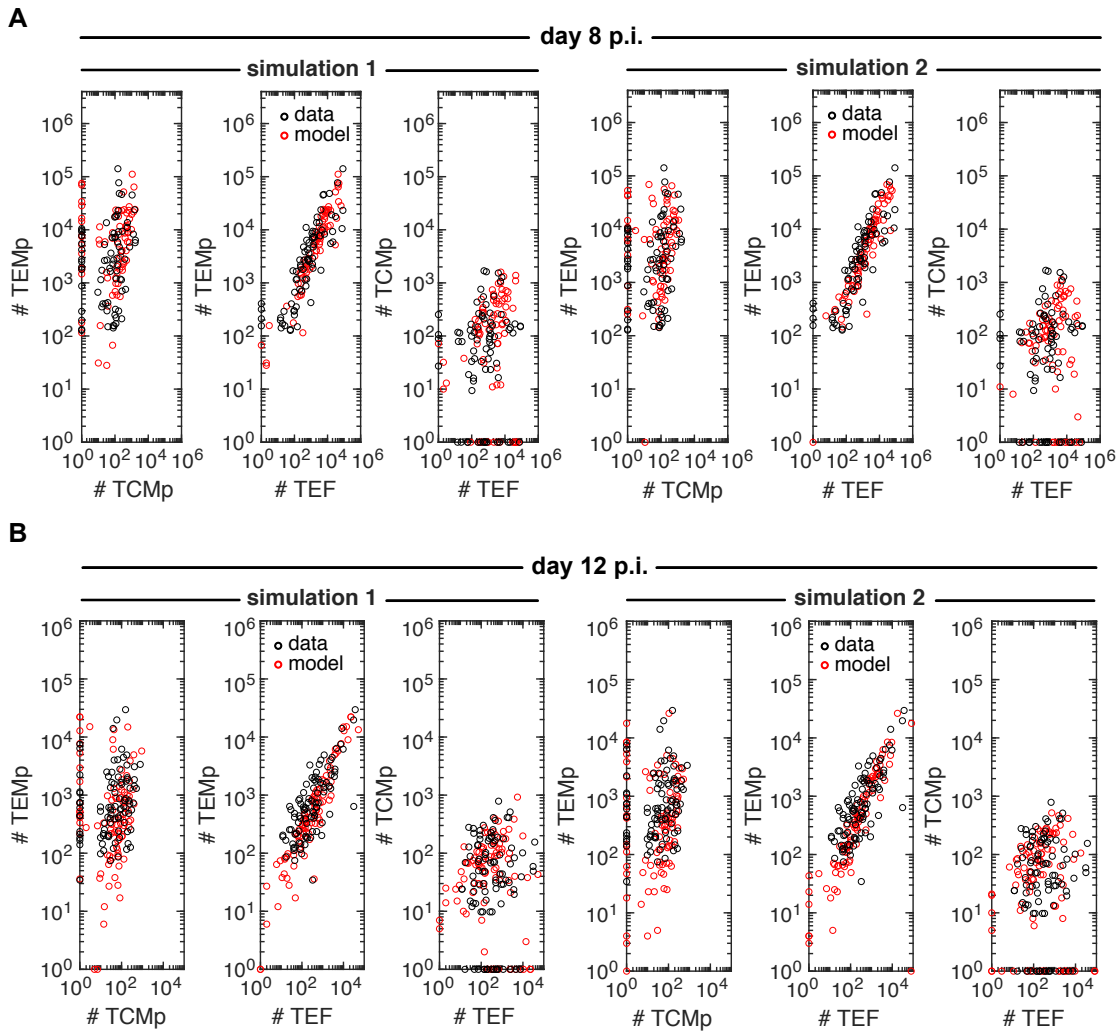

Figure 5.4: Representative Gillespie simulations of the extended cell cycle dropout model, depicting the correlation structure of the respective phenotypes at day 8 and 12 p.i. Simulations are based on the parameters shown in Tab. 5.1, where the number of simulated clones corresponds to the number of measured clones. (Data from [Buchholz et al., 2013])

5.4 Simulation of clones and their correlation structure

Based on the parameters from Tab. 5.1, clones and their phenotypic composition are simulated according to Gillespie's algorithm, cf. [Gillespie, 1977]. Simulations are evaluated at day 8 and 12 p.i. in order to compare them with the corresponding single-cell fate mapping data. For day 8 p.i., $n = 76$ clones could be recovered experimentally, while 12 days after infection $n = 93$ clones were measured. Compared to the LCMV Arm fate mapping data from section 4.5.2, the clones in the *L.m.*-OVA case are larger and less variable in

clone size. Despite cells exiting the cell cycle at about 5 d p.i. in both infections, cells are recruited on average more than three days earlier during *L.m.*-OVA infection. This also leads to a significant reduction of variability in the resulting clone sizes during *L.m.*-OVA infection. The full correlation structure of the data and two representative simulations per time point can be found in Fig. 5.4. At day 8 p.i., the correlation patterns between simulation and data agree well (Fig. 5.4A). Only when looking at the absolute TEMp versus TEF cell numbers (middle plots) there is a slight discrepancy. This shortcoming is further emphasised during the transition to day 12 p.i. (Fig. 5.4B). Except for this feature, the simulations robustly reproduce the observed correlation structure of the experiments.

Our analyses showed that for the TEMp-TEF correlation of absolute cell numbers at day 12 p.i., the discrepancy between simulation and data is difficult to solve, especially since the Pearson correlations are already in good agreement (see Fig. 5.2C). At the simulation level, however, we have more flexibility to deal with this. It turns out that a clonal variability in the transition rate $r_{\mathrm{CTEMp}\to\mathrm{CTEF}}$ solves this problem, while still describing the moments correctly. Here, we consider a normal distribution of this transition rate with mean of $\langle r_{\mathrm{CTEMp}\to\mathrm{CTEF}}\rangle = 0.23\ \mathrm{d}^{-1}$ (best fit result) and CV $= 0.5$. It seems to be important that the rate must be distributed from the beginning of the response. Considering a distributed rate from day 8 p.i. onwards is not sufficient to reduce the discrepancy between data and simulation. Fig. 5.5A shows the associated simulation results for day 8 p.i. These indicate an improvement in the TEMp-TEF correlation without impairing the other correlations. In particular, at day 12 p.i. (Fig. 5.5B), the data are now consistent with the simulation results. Interestingly, considering other model parameters to be distributed is not useful in this regard. The biological relevance of this observation could be related to increased clonal heterogeneity within the TEMp compartment. Thus, there might be clones whose TEMp cells rapidly differentiate into the TEF compartment, whereas other clones are more prone to accumulate TEMp cells. The latter might then contract less strongly in size, for example, and mature more into sustainable TEM cells. We will discuss this aspect later in chapter 6. In the following, we again consider the contraction model without distribution of the effector differentiation rate.

As in section 4.5.2, let us analyse the distribution of clone sizes using the Gini coefficient, see [Gini, 1912], considering the original set of parameters with constant $r_{\mathrm{CTEMp}\to\mathrm{CTEF}}$ rate. To this end, we simulated $N = 10^4$ clones based on the parameters from Tab. 5.1 and computed the Gini coefficient using eq. 4.2. At day 8 p.i., the Gini coefficient of total clone sizes in the data is given by $G = 0.75$, which is correctly described by the model, see Fig. 5.6A. The trend in Gini coefficients for the different phenotypes is correctly reproduced by the model, although the coefficient is slightly underestimated in the case

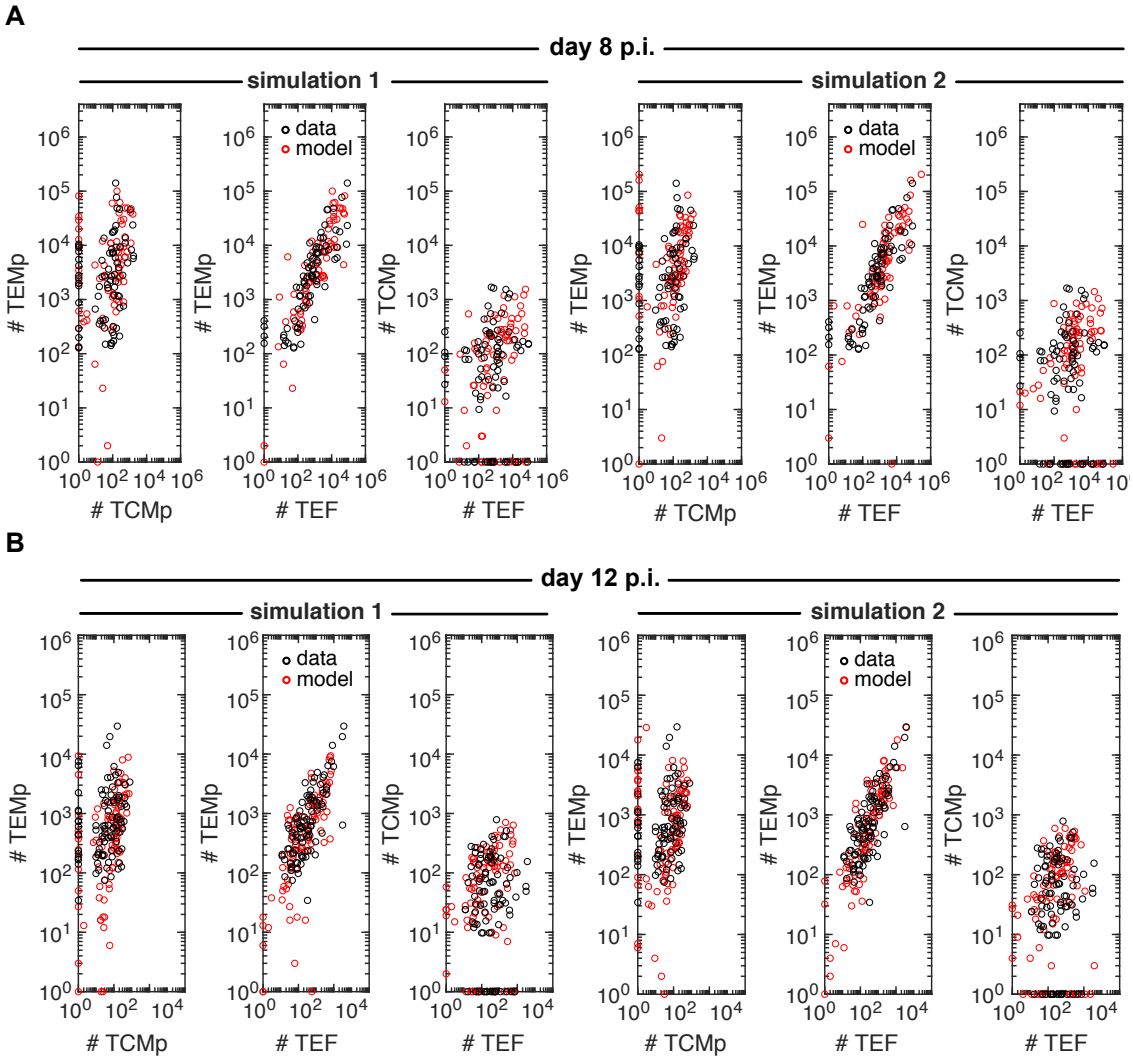

Figure 5.5: As in Fig. 5.4 but with distributed transition rate CTEMp $\rightarrow$ CTEF. For this purpose, a normal distribution with CV $= 0.5$ was chosen, where the mean corresponds to the best fit result of the rate from Tab. 5.1. (Data from [Buchholz et al., 2013])

of the TCMp and TEF compartments. At day 12 p.i., the observed Gini coefficient is given by $G = 0.73$. This overall coefficient, as well as the phenotype-specific coefficients, are correctly described by the model (Fig. 5.6B). Notably, the variability of clone sizes in the *L.m.*-OVA case is not overestimated by the model. This has not been clear yet as the used model is based on a more detailed analysis of the LCMV Arm fate mapping data, which were characterised by a much more unequal distribution ($G = 0.91$). Thus, no form of bias is detectable in the model topology itself that would reduce its applicability in the context of an *L.m.*-OVA infection.

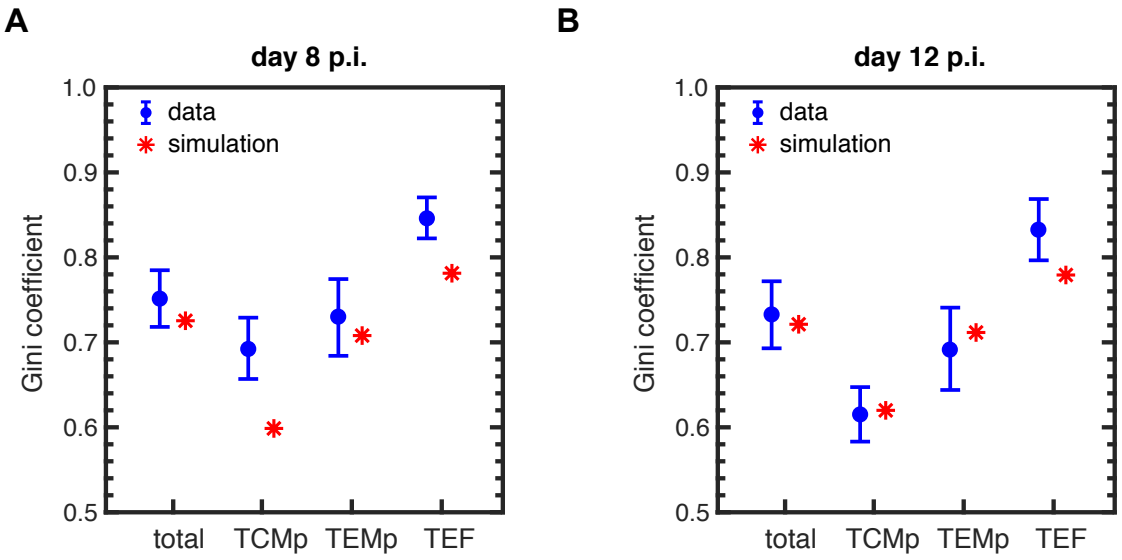

Figure 5.6: Gini coefficients of the total clone size and compartment size distributions at day 8 and 12 p.i. Simulations are based on the extended cell cycle dropout model for fixed, i.e. not distributed, effector differentiation rate CTEMp $\rightarrow$ CTEF. $N = 10^4$ clones were simulated with parameter values shown in Tab. 5.1. (Data from [Buchholz et al., 2013])

5.5 Late contraction and transition to memory phase

We have seen in the previous section that the extended cell cycle dropout model was indeed capable of kinetically resolving the early contraction phase. From now on, we will refer to this model extension as the *early contraction phase model*. All parameters of this model were identifiable with narrow confidence intervals. Now, implications of these parameters for the late contraction and entry into the memory phase are to be investigated. To this end, the model is slightly extended, which will lead to the *late contraction phase model*. Since no data at later time points are available, we have to rely on data from the literature.

We consider data from [Schlub et al., 2010] to be the most suitable for our purpose. Here, the authors measured relative OT-I T cell population sizes from blood after attenuated and *actA*-deficient (Att.) *L.m.*-OVA infection, classified in terms of the memory marker CD62L. The *actA* (actin assembly-inducing protein) gene is a virulent factor that modulates intracellular growth of the bacteria; *actA*-deficiency in this context decreases virulence by three orders of magnitude and inhibits cell-to-cell spread, see [Brundage et al., 1993]. The immune response stays qualitatively comparable to the usual *L.m.*-OVA infection, but is characterised by lower cell numbers and a shifted peak of the immune response to day 6 to 7 p.i., see [Porter et al., 2006; Schlub et al., 2010]. We extend our

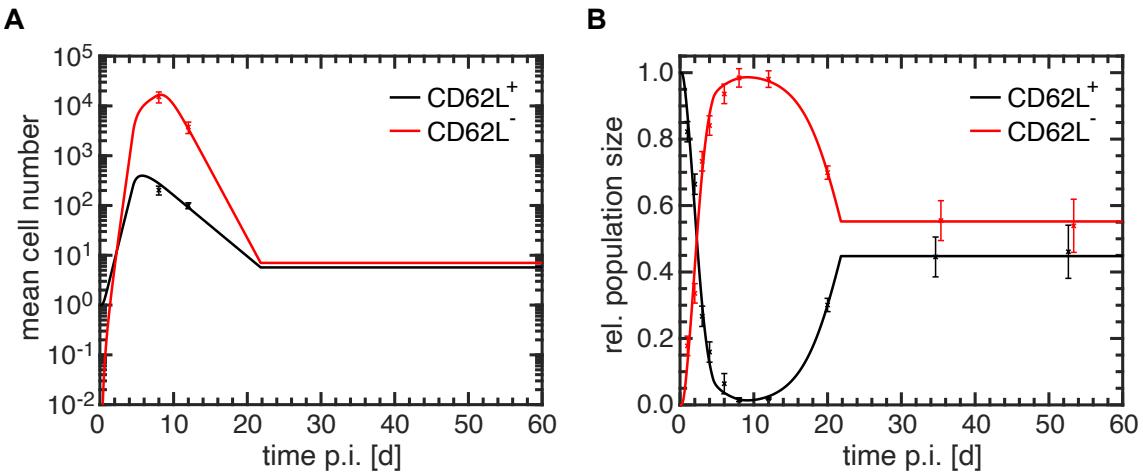

Figure 5.7: Kinetics of absolute **(A)** and relative **(B)** compartment sizes up to the early memory phase after infection with (Att.) *L.m.*-OVA within the late contraction phase model. The last three data points in (B) are based on infections with Att. *L.m.*-OVA, where classification is performed using CD62L. (Mean cell numbers and relative subsets sizes up to day 12 p.i. from [Buchholz et al., 2013] and relative subset sizes at day 20, 35, and 53 p.i. from [Schlub et al., 2010])

relative population data from Fig. 5.2D to include three additional data points measured at day 20, 35, and 53 p.i. Interestingly, the relative subset sizes of CD62L positive and negative cells at day 35 and 53 p.i. are both given by approximately 50 %. We assume that despite the mentioned differences between the infection variants at early time points, the long-term behaviour of the response remains at least approximately comparable. Furthermore, we assume that the two organs blood and spleen can be compared to each other. In [Crauste et al., 2017] the authors measured absolute numbers of $CD8^+$ T cell responses in the spleen after vaccinia virus infection. Cells were also classified in terms of the CD62L marker. Strikingly, the authors find similar long-term kinetics regarding the percentages of CD62L positive and negative cells: between day 40 and 90 post immunisation, the relative subset sizes are roughly given by 50 % each. Therefore, we think that the combination of data from the spleen and blood could be suitable for a first tentative analysis.

Interestingly, the first data point at day 20 p.i. is correctly described by our previous model fit above (fit only shown until day 14 p.i.) with parameters given in Tab. 5.1. However, for later time points, the model results become non-interpretable, especially as the cell numbers incorrectly go towards zero due to continued cell death. Therefore, the model must be further modified. An intuitive adjustment is the introduction of dynamic death rates, i.e. $d_i = d_i(t)$. In the simplest case, this dynamics is given by switching off constant cell death at time τ_{death} to be estimated. This parameter is initially chosen to

be the same for both subsets ($CD62L^+$ and $CD62L^-$). Furthermore, we assume that cells do not start dividing again until day 53 p.i., which means that only programmed cell death regulates the absolute cell number (for which we have no reference data) during the contraction phase. The reason we neglect cell cycle re-entry at this point is that turnover rates during the absence of antigen are reported to be small compared to the death rate estimates above. For example, in [Parretta et al., 2008], an expected lifetime (inverse of the daily turnover rate) of 90 d is estimated for $CD8^+$ T cells in homeostatic equilibrium, with 95 % confidence interval of [64; 133] d. This estimate is based on BrdU measurements, evaluated under the steady state assumption $p = d$. As noted in chapter 1, BrdU is a reliable method for measuring division speed of slowly dividing cells during equilibrium. A clear compilation of comparable study results regarding this question, including mathematical evaluation methods, can be found in [De Boer et al., 2013].

Utilising all relative population data available, the best fit parameters shown in Tab. 5.1 do not change significantly. This is because the relative kinetics up to day 12 p.i. were also well described before, without considering them during model fitting. In addition, the late time points (measured at $t > 20$ d p.i.) are relatively independent of the earlier ones and taken into account by the parameter τ_{death}. The model fits for the absolute and relative kinetics for $CD62L^{+/-}$ cells are shown in Fig. 5.7. It turns out that cell death stops at time $\tau_{\text{death}} = 21.8$ d p.i. The corresponding 95 % confidence interval is given by [20.5; 23.1] d. In the case of modelling the decrease of death rates continuously, τ_{death} would naturally be bigger. If one introduces phenotype-specific endpoints of death, then in the best fit these are estimated to be identical. However, it is questionable whether the data even allow for discrimination in this respect. A second interesting observation is that at the end of cell death, there are about 13 cells left (Fig. 5.7A). These consist both of CD62L positive (TCM) and negative (TEM) cells. This means that the number of cells after contraction could have increased by about one order of magnitude compared to the beginning of the immune response. Since we did not use any absolute cell numbers later than day 12 p.i., this is an indirect model prediction that could be tested experimentally. It should be noted again that the memory phase is expected to be dominated by $CD62L^+$ TCM cells. Since the Att. *L.m.*-OVA infection data from [Schlub et al., 2010] explicitly challenge this expectation, it remains open to what extent this result is affected by considering this particular infection.

Chapter 6

Discussion

In this thesis, we dissected cytotoxic T cell responses to acute infection in terms of their differentiation pathways and their kinetics during expansion and contraction phase. Our resultant cell cycle dropout model is a consequence of the joint consideration of various *in vivo* data sets at the single-cell level. These include scRNA-seq data, single-cell fate mapping data, and FACS-based quantification results of proliferation rates. The latter originate from a formalism developed specifically for this purpose, which has applications far beyond our problem formulation. Utilising these data, we could mechanistically explain the dynamics of individual T cell responses by stochastic differentiation models with identifiable parameters.

Quantitative aspects of cell division

We have developed a method to quantify mean cell cycle lengths of rapidly dividing cells *in vivo*. More precisely, the average length of each individual cell cycle phase can be computed. The method is based on data retrieved from labelling cells with a nucleoside analogue such as BrdU and subsequent measurement of their DNA content. The associated mathematical formalism does not require precise knowledge of the underlying distribution of inter-division times, nor of pharmacokinetic properties of BrdU. With the help of stochastic simulations, the accuracy of the method and its basic assumptions could be validated. In particular, it could be shown that the method is valid even during circumstances of moderate division-time variability and of severe cell death. Likewise, the formalism is flexible regarding experimental extensions and the incorporation of additional kinetic properties, such as information about cell generation or the amount of quiescent cells within the population.

The well-known Smith-Martin model, see [Smith et al., 1973], considers the S-phase of the cell cycle to be almost constant, while all stochasticity is attributed to the G_1-phase. Using appropriate reporter mice and (single) cell imaging methods, (lag-) exponentially distributed waiting times could be excluded in the context of B and T lymphocytes after *in vitro* stimulation, see [Dowling et al., 2014; Pham et al., 2018]. Recently, P. D. Hodgkin and colleagues were even able to show that a "converse" Smith-Martin model with constant

G_1-phase and variable S-phase describes data from dividing B lymphoma cell lines better than the original Smith-Martin model, cf. [Pham et al., 2018]. In our model of the cell cycle, we assume that all phase lengths are distributed according to general and unknown distributions. In this sense, our model represents a generalisation of these approaches.

Our formalism can easily be extended to dual-pulse-chase experiments. Here, an additional nucleoside analogue such as EdU is used, which is administered shortly before analysis, see e.g. [Gitlin et al., 2015]. This would have the advantage that all information required for quantifying average phase lengths could be obtained from the same mouse. The authors of the study conducted in [Akinduro et al., 2018] use such a dual-pulse-chase experiment to infer cell cycle speed of haematopoietic progenitor cells. In their model, the authors assume that BrdU and EdU, immediately after administration, are no longer sufficiently available to label S-phase cells. A measurement of the bio-availability of BrdU in [Matiasova et al., 2014] concluded that BrdU is no longer available approx. 15 min after intraperitoneal injection in mice and after 60 min in rats. Our analyses in mice have shown that BrdU is no longer available after approx. 30 to 60 min after administration. Therefore, the assumption made in [Akinduro et al., 2018] seems plausible at first but should be considered critically in the case of rapidly dividing cells, as a significant fraction of cells might enter S-phase during the experiment and thus the number of double-positive cells could be underestimated.

Another interesting proposal by P. D. Hodgkin and colleagues is based on the assumption that individual cell cycle phases are "stretched", see [Dowling et al., 2014]. Here, the total cell cycle length follows a distribution, e.g. a log-normal distribution, where the relative proportions of the individual phases are constant for each cell. In the same study, these proportions are then estimated from *in vitro* data of dividing lymphocyte populations. The modelling assumptions made in this study are in clear contradiction to ours, as we assume independence of the individual phase lengths. We are not aware of any other study that suggests a strong correlation of phase lengths. On the contrary, work by [Chao et al., 2019] shows that the individual phase lengths obey statistically independent Erlang distributions. However, this study did not consider populations of dividing lymphocytes. It also remains open to what degree *in vitro* findings in this context apply to the *in vivo* case.

A limitation of our method is that it does not allow for estimating variances of phase length distributions. In [Weber et al., 2014], the authors also considered simultaneous measurements of BrdU and DNA content to infer cell cycle speed. Here, the authors assumed that individual phases obey independent delayed-exponential distributions. Since experiments were carried out *in vitro*, BrdU could be removed from the experiment after

30 min of labelling, thus circumventing the unknown BrdU pharmacokinetics. Under these assumptions, the authors were able to derive equations for the different gate frequencies in steady state, depending on their parametrisation scheme. By adding another nucleoside analogue, even the variance of the cell cycle length distribution could be estimated. Although the assumptions made all seem realistic at first, it is not clear to what extent they can be transferred to the *in vivo* level. Especially strong assumptions on the exact shape of the distribution have to be viewed critically.

Our method showed that cells divide on average with a rate of $p \approx 3\ \mathrm{d}^{-1}$ at day 4.5 p.i. during the exponential growth phase. Here, CMp cells divide significantly slower ($p_{\mathrm{CMp}} = 1.9\ \mathrm{d}^{-1}$) than EFF cells ($p_{\mathrm{EFF}} = 2.9\ \mathrm{d}^{-1}$). Interestingly, these results were identical for both *L.m.*-OVA and LCMV Arm infection. In [De Boer et al., 2001], the authors estimated by mathematical modelling the overall proliferation rate in the LCMV Arm case to be $p = 3\ \mathrm{d}^{-1}$. Given their model, this number refers to the mean growth rate in the interval between about 1 d and 6 d after infection, thus being consistent with our quantification. For *L.m.*-OVA, we were able to calculate a slowing down of the mean division time to $p = 1.4\ \mathrm{d}^{-1}$ for the CMp cells and to $p_{\mathrm{EFF}} = 1.9\ \mathrm{d}^{-1}$ for the EFFs at day 8 p.i., where these estimates refer to cycling cells only. Overall, in this infection, it appears that the exit of cells from the cell cycle, rather than the slowing down of division speed, regulates cell numbers during the immune response. In [Kretschmer et al., 2020], using our method for another model system, it was shown that just the opposite is the case. Here, however, the entire amount of antigen was initially injected and no antigen replication took place. Likewise, cells stopped dividing much later after immunisation. Therefore, we conclude that the contribution of these two mechanisms in regulating cell number responses depends on the particular infection under consideration.

Differentiation pathways of CD8$^+$ T cells

Analysis of time-resolved CD8$^+$ T cell transcriptomes during acute infection allowed the identification of cell clusters with known immunological subsets. Furthermore, the application of various published methods lead to the inference of possible differentiation pathways between these clusters. Overall, we found that a cell cycle dropout model with mainly progressive differentiation during division is most consistent with the data. Here, cells seemed to primarily retain their phenotype upon exit from the cell cycle. From these same data, we could further conclude the fractions of cells undergoing division at each time point. These could then be validated in subsequent FACS experiments through phospho-Rb measurements. Using these dropout kinetics, quantifications of division speed, and stochastic moments of single-cell fate mapping data, we were able to fit stochastic models

of T cell diversification during LCMV Arm infection that were consistent with the inferred topologies above. Model selection revealed that simpler models were preferred by the data. The simplest of these models, in which all subsets began to turn quiescent at the same time and at the same rate, had the best AIC_c-value.

Overall, our results do not support the decreasing potential model, see [Crotty et al., 2004]. In this model, effector cells continue to differentiate through accumulated stimuli and lose the potential to generate sustainable memory cells thereby. In contrast, our data suggest that cells at the root of differentiation express memory genes such as *Slamf6*, *Satb1*, and *Cd27* and in turn give rise to effector-like cells. These findings are consistent with the study from [Lin et al., 2016]. Here, the authors were able to show that the first few cell divisions were characterised by TCF1 positivity of the cells. Only after several cell divisions did the cells lose this early memory marker. Therefore, our data are also not consistent with the linear model of differentiation, cf. [Opferman et al., 1999; Wherry et al., 2003], but best fit the progressive differentiation model, see [Buchholz et al., 2013; Gattinoni et al., 2017]. However, it must be added that progressive differentiation seems to take place mainly during cell division. After cells have become quiescent, it is difficult to decide whether differentiation still occurs at all. Moreover, the model proposed in [Buchholz et al., 2013] ignores the exit of cells from the cell cycle, which seems to be a highly relevant source of variability in our data. An important addition to our analyses can be found in [Pace et al., 2018], where chromatin dynamics were considered in addition to transcriptional data. The authors conclude that naïve cells first acquire an early activated state of highly proliferative capacity. These cells then either transition into early memory precursor cells or down-regulate their stem cell characteristics and become effector cells. Furthermore, they conclude that these early memory precursors can return to the early activated state upon re-stimulation with antigen, emphasising their stemness properties.

Interestingly, in our analyses, we have also found a non-negligible proportion of transitions from dividing TEF to quiescent TEMp cells. The TEF cells showed a strong *Klrg1* and *Cx3cr1* signature, i.e. they most likely corresponded to terminally differentiated effector cells. Cells that followed this developmental pathway could, according to our data, no longer become TCMp cells. Transitions from TEF $\rightarrow$ TEMp were already discussed in [Buchholz et al., 2013] and could not be rejected. In [Herndler-Brandstetter et al., 2018] the authors were able to show that a significant part of the long-lived memory cells had once expressed high levels of KLRG1, i.e. originated from $KLRG1^{high}$ cells. This proportion of so-called *ExKLRG1* cells became gradually smaller the more the memory properties became predominant. The authors found that about 40 % of the $KLRG1^{int}$ TEM compartment (intermediate intensity of KLRG1 expression) were based on such

ExKLRG1 cells. Also during our model selection process, the TEF → TEMp transition could not be excluded.

As mentioned above, our analyses have ruled out a linear differentiation model. In such a model, naïve cells adopt an early effector phenotype directly after recruitment. The majority of these cells die during contraction, while a small proportion of memory cells remain. Since this scheme has not only been favoured in the past but still enjoys acceptance today, we would like to comment on it briefly. More recent models of this kind can be found for example in [Crauste et al., 2017; Terry et al., 2012]. In [Crauste et al., 2017], the authors studied $CD8^+$ T cell responses during acute infection up to the memory phase. Their naïve → early effector → late effector → memory model describes the measured absolute cell numbers well. Furthermore, in [Crauste et al., 2017], the authors also tested the progressive model of differentiation from [Buchholz et al., 2013], which was proposed to capture the expansion phase of the response. In [Crauste et al., 2017] this model was fitted to contraction phase data where it showed poor performance. However, the comparison of the models was done without a suitable adjustment of the progressive model, e.g. by introducing death rates. In [Miles et al., 2019], a meta-study of progressive vs. linear differentiation models was conducted. The authors conclude that progressive models systematically outperform linear models when fitted to population data. Interestingly, they observed that in the case of high numbers of transferred cells, linear models increasingly perform better. One reason for this could be competition effects, which is why the authors recommend rescaling the data. The authors mention that rescaling, however, often leads to unrealistic parameters, such as mean lifetimes of naïve cells of more than 10 d (see below for discussion of reasonable recruitment times). In [Crauste et al., 2017], indeed $N = 2 \cdot 10^5$ cells were transferred, whereas in [Miles et al., 2019; Schlub et al., 2010] considerable competition effects can be identified for transferred cell numbers of already $N \sim 10^3$ compared to smaller transfer numbers. These effects include kinetic shifts in terms of phenotypic composition and overall size.

Heterogeneity in single T cell responses to acute infection

We have shown that the simplest variant of the cell cycle dropout model could explain the expansion phase during acute infection and the resulting heterogeneity of single-cell fate mapping data, as well as all available kinetic data on cell cycle activity and speed. The model parameters were shown to be identifiable in terms of their confidence intervals. Based on these parameters, forward simulations could reproduce the dichotomy of the LCMV Arm clone size distribution and correctly reflected phenotypic composition of the clones. Although all subsets started to become quiescent at the same time and rate,

our model showed phenotype-specific cycling activity following the onset of cell cycle dropout. This model prediction was confirmed in a subsequent experiment. Overall, the heterogeneity between individual T cell responses in the model was generated by clonal variability in the recruitment of cells, by subset-specific differences in division speed, and by the inherent model stochasticity, with clonal variability making the largest contribution here.

S. A. Teichmann and colleagues were able to identify early bifurcation of cells into T helper 1 or T follicular helper cells in the case of differentiating $CD4^+$ T cells during malaria infection, see [Lönnberg et al., 2018]. Using TCR sequences, these bifurcations could be observed in scRNA-seq data even at the clonal level. This is relevant for our work as early bifurcation of T cells could also lead to a strongly dichotomous clone size distribution (Gini coefficient of $G = 0.91$ for our LCMV Arm data). Concerning our data, the rare and large clones would then be produced by one model branch, with the other branch producing the much more abundant smaller clones. Such classes of models we have tested (results not shown). They were indeed able to explain the strong dichotomy in the data. The correlation patterns, on the other hand, were very poorly captured. The reason for this is that for small clones the proportion of slower dividing TCMp cells was greatly overestimated, while for large clones this proportion was underestimated, respectively. Moreover, we could not find any evidence for such early bifurcations in our scRNA-seq data. In the most extreme bifurcation model, a naïve cell gives rise to either only memory-like cells or only effector cells. However, this view can already be rejected by the single-cell fate mapping data itself, since these show that a clone generally contains cells of all phenotypes, cf. [Buchholz et al., 2013; Gerlach et al., 2013; Stemberger et al., 2007]. Regarding weakened forms of the bifurcation model, especially in combination with a first asymmetric cell division, see [Chang et al., 2007], we cannot make any statement because we have not tried models of this kind.

The stochasticity of our mathematical models integrates cell-intrinsic and cell-extrinsic factors. Here, we have no way to distinguish these two sources of variability. However, for clinical considerations, environmental influences on T cells and their modulation of the immune response play an important role. This includes in particular stimulatory signals caused by antigen, cytokines, and co-stimulus of the T cells, see [Curtsinger et al., 1999; Kaech et al., 2001; Watts et al., 2003]. However, the extent to which these signals quantitatively affect proliferation and differentiation behaviour is yet unknown. In [Gerlach et al., 2013], the authors have also been able to establish the existence of small and large clones by fate mapping data. They show that memory cells mainly arise from smaller clones. In contrast, our modelling results suggested that, in absolute terms, all

clone sizes contributed rather equally to the memory (precursor) pool, as indicated by low correlations between total clone size and the absolute number of TCMp cells. In relative terms, however, small clones showed a greater proportion of TCMp cells compared to large clones. D. Zehn and colleagues have investigated the $CD8^+$ T cell response after infection with respect to different TCR ligand affinities, cf. [Zehn et al., 2009]. They observed that even weak affinities caused robust activation. Moreover, they found that smaller affinities led to faster progression through the differentiation program and to an earlier onset of contraction. Here, strong antigen stimuli guaranteed stable long-term expansion during response. Regarding our model scheme, this could mean that especially the absolute proliferation rate of TCMp cells could strongly depend on the initial antigen stimulus. This would explain on the one hand the higher cell numbers after higher affinity TCR stimulation (*L.m.*-OVA vs. LCMV Arm) and on the other hand the effective differentiation rate of the cells: even with stimulus-independent differentiation rates, differentiation into the effector compartment would still occur less frequently in the case of high TCMp proliferation rates. The sensitivity of the TCMp proliferation rate with respect to early antigen availability has already been demonstrated in [Kretschmer et al., 2020].

We have noted that our model extends the progressive model of differentiation from [Buchholz et al., 2013] by explicitly accounting for the exit of cells from the cell cycle. Ignoring this important mechanism and trying to explain the heterogeneity of the data without it resulted in a mean waiting time for recruitment of about 10 d (95 % confidence interval given by [5; 17] d) during LCMV Arm infection. This long waiting time did not even take into account the first cell division. The model thus explained the entire variability in the data through the initial recruitment process of naïve cells. Consequently, this would mean that on day 8 p.i. a substantial part of the cells would not have been recruited into the response yet. This is in direct conflict with experimental data: In [Van Heijst et al., 2009], for example, it was shown in mice that almost all naïve cells were recruited during acute infection. Also in [Schlub et al., 2010] it could be shown that even for high numbers of transferred cells, less than 0.1 % of the cells remained unrecruited. Since the peak of the $CD8^+$ T cell response during LCMV Arm infection is already reached at about day 8 p.i., cf. [De Boer et al., 2003], cell recruitment must take place faster on average. Assuming a deterministic compartment model of naïve, activated, and memory cells, R. J. de Boer and colleagues estimated the mean recruitment time following LCMV Arm infection to be between 1 to 2 d p.i., see [De Boer et al., 2001]. Even if one would consider the lower bound of our estimated confidence interval for the mean recruitment time of 5 d, then the measured CVs could no longer be described by the model. Hence, the progressive model was not capable of explaining the observed heterogeneity in clone sizes.

The data presented in [Bousso et al., 1999] identified the time to recruitment as a driver of heterogeneity of a (polyclonal) T cell response, which is in line with our findings. At the same time, the authors showed that clonal differences in cell cycle speed were rather small. Precisely because we have found that most of the variability in our model is clonal, the division destiny concept, proposed by P. D. Hodgkin and colleagues, is of particular relevance to us. Here, after activation of the cell, a division destiny timer is programmed to determine when the cell becomes quiescent; this follows a parameterised distribution, which in turn depends on the specific stimulatory conditions of the cells, see [Heinzel et al., 2017; Marchingo et al., 2014; Turner et al., 2008]. We were able to show that increasing clonal variability in our model did not provide any improvement, as the inherent stochasticity of the model, together with compartments dividing at different rates, produced enough variability to describe the data. It may therefore be that P. D. Hodgkin and colleagues were slightly overestimating the clonal variability of the immune response because of not considering differentiation processes between the cells. Conceptually, this would mean that a significant part of the variability arises during the immune response and is not already programmed during initial stimulus.

Proliferation, death, and differentiation processes in our model are based on exponentially distributed waiting times. Experimental *in vitro* data has shown that both cell division and cell death obey log-normal distributions, see [Deenick et al., 2003; Dowling et al., 2014; Hawkins et al., 2009]. Differentiation processes are certainly not memoryless either. In [Robert et al., 2021] it was demonstrated for cell divisions that these could also be described sufficiently well in terms of Gamma-distributions. Moreover, *in vitro* experiments have shown that the first cell division takes longer than all subsequent divisions, cf. [De Boer et al., 2005; Deenick et al., 2003; Lee et al., 2008]. Assuming that these findings can be transferred directly to the *in vivo* level, it would be desirable for future considerations to describe all model reactions with skewed waiting time distributions. This could be achieved, for example, by replacing exponential distributions with Gamma-distributions, as we have already done to describe recruitment and first cell division more realistically. Here the problem arises that the introduction of Gamma-distributions for all reactions is on the one hand computationally expensive and on the other hand variances will most likely not be identifiable. A different approach to addressing these considerations is provided by the *cyton model*, see e.g. [Hawkins et al., 2007]. Here, waiting times until division or death, which are assumed to be independent, are described by log-normal distributions. Furthermore, first cell divisions are considered separately to account for the evidence mentioned above. Nevertheless, it is not clear a priori, how cell differentiation could be included within this formalism. Of course, it also remains unclear whether variances could be reliably estimated or not.

Mechanisms of T cell contraction

We have extended the resultant cell cycle dropout model beyond the expansions phase of the immune response to the early phase of contraction. To this end, we introduced dynamic proliferation rates and constant death rates, assuming that only quiescent cells are affected by cell death. Here, possible model variants were fitted to dropout kinetics, quantified proliferation rates, and stochastic moments of single-cell fate mapping data in the context of *L.m.*-OVA infection. The concrete form of kinetics was determined by model selection. In the best model, all subsets were initially dividing with different rates, and then slowed down proliferation in the same way. In addition, memory-like cells died at a different rate than effector cells. The resulting model was consistent with all available data up to day 12 p.i., with all parameters being identifiable. The relative subset sizes were correctly described without having used them for parameter estimation. Forward simulations reproduced both the clone size distribution and the correlation structure of the data at day 8 and 12 p.i., except for the TEMp-TEF correlation at day 12 p.i. It was found that distributing the effector differentiation rate CTEMp $\rightarrow$ CTEF between clonal families lead to significant improvement in this regard. Next, we considered the full contraction phase by adding published data of relative subset kinetics. We were able to obtain a time estimate for the end of cell death, as well as an estimate of the number of cells entering the memory phase.

In our model, we did not have to artificially introduce the onset of cell death and the associated start of contraction within the framework of a two-phase model, as it was done in [De Boer et al., 2003], for example. We estimated death rates of $d_{\mathrm{CMp}} = 0.2\ \mathrm{d}^{-1}$ and $d_{\mathrm{EFF}} = 0.7\ \mathrm{d}^{-1}$, with the latter rate driving the contraction phase because of the much larger proportion of effector cells. R. J. de Boer and colleagues estimated an overall death rate of $d = 0.5\ \mathrm{d}^{-1}$ for $\mathrm{CD8}^+$ T cells subject to LCMV Arm infection, cf. [De Boer et al., 2001]. This roughly fits our result, especially when taking into account that *L.m.*-OVA induces larger T cell responses than LCMV Arm. Furthermore, we found that cell death would have to stop around 22 d p.i. to explain the data consistently. Based on one recruited cell, this leads to a mean cell number of the order of 10^1 upon entry into the memory phase. When considering the lower bound for the cessation time of cell death and taking into account the CV of the mean total cell number, we would obtain a maximum of about 1 % of cells surviving the contraction phase. This is contrary to the consensus in the literature, which lies between 5 % to 10%, see e.g. [De Boer et al., 2001; Harty et al., 2008]. Therefore, we suspect that this discrepancy may result from considering data from the two organs of blood and spleen together. Likewise, having combined the infections *L.m.*-OVA and Att. *L.m.*-OVA could also play a role here. The major difference between these two

infections is that Att. *L.m.*-OVA is characterised by reduced virulence due to deficiency of the *actA* gene, being involved in intracellular growth of the bacteria, see [Brundage et al., 1993]. This infection resulted in approximately equal fractions of memory cells and effector cells during late contraction and early memory phase, see [Schlub et al., 2010]. In any case, absolute cell numbers from the spleen after *L.m.*-OVA infection would be helpful to make more precise statements. In addition, even later time points (later than 60 d p.i.) would be highly interesting, as these, together with proliferation measurements, would allow for quantitative description of the homeostatic memory phase in our model.

The TEMp-TEF correlation of the forward simulations evaluated at day 12 p.i. could be significantly improved when having considered a distributed TEMp to TEF differentiation rate. This could indicate that the TEMp compartment has been more heterogeneous under the CD62L/CD27 classification scheme than expected. This might be in accordance with the heterogeneity of the TEM compartment observed in [Milner et al., 2020]. Here, the authors considered $CD62L^{low}$ memory cells, which could be divided into two different subtypes with respect to CD127. $CD62L^{low}/CD127^{low}$, so-called *terminal* TEMs and $CD62L^{low}/CD127^{high}$, so-called *redefined* TEMs. Terminal TEM cells showed, among other characteristics, a higher upregulation of granzyme molecules and a reduced recall proliferation capacity compared to redefined TEM cells. For our case, this could mean that terminal TEMs differentiate into the TEF compartment at a higher rate, whereas the redefined TEMs are more likely to be retained in their compartment and develop enhanced homing potential. Thus, for future considerations, it would be highly interesting to generate analogous data with additional measurement of CD127. Another neglected possibility in our approach could be coupling effects, as described, for example, in [Bocharov et al., 2011]. Here, the authors fitted differentiation models to data, where the compartment size by construction had a negative feedback on the differentiation rates of the previous downstream compartments. The authors stated that local resource considerations of expanding cell clusters may be a biological cause here.

In this thesis, a variety of *in vivo* single-cell data were combined to infer mechanisms of T cell responses to acute infection. The new insights form an important quantitative contribution to the question of how immunological T cell memory is formed and thus have implications for clinical considerations, such as improvement of vaccination strategies or adoptive T cell therapies. Furthermore, the results presented in this work may provide a basis for better characterisation of mechanisms underlying T cell responses during chronic infection or cancer.

Bibliography

Akaike, H. (1974). "A new look at the statistical model identification". *IEEE Trans. Automat. Contr.* **19**.6, pp. 716–723.

Akinduro, O., Weber, T. S., Ang, H., Haltalli, M. L., Ruivo, N., Duarte, D., Rashidi, N. M., Hawkins, E. D., Duffy, K. R., and Lo Celso, C. (2018). "Proliferation dynamics of acute myeloid leukaemia and haematopoietic progenitors competing for bone marrow space". *Nat. Commun.* **9**.1.

Andrew, S. M., Baker, C. T., and Bocharov, G. A. (2007). "Rival approaches to mathematical modelling in immunology". *J. Comput. Appl. Math.* **205**.2, pp. 669–686.

Angerer, P., Simon, L., Tritschler, S., Wolf, F. A., Fischer, D., and Theis, F. J. (2017). "Single cells make big data: New challenges and opportunities in transcriptomics". *Curr. Opin. Syst. Biol.* **4**.July, pp. 85–91.

Badovinac, V. P., Porter, B. B., and Harty, J. T. (2002). "Programmed contraction of CD8+ T cells after infection". *Nat. Immunol.* **3**.7, pp. 619–626.

Barber, D. L., Wherry, E. J., and Ahmed, R. (2003). "Cutting Edge: Rapid In Vivo Killing by Memory CD8 T Cells". *J. Immunol.* **171**.1, pp. 27–31.

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". *J. R. Stat. Soc. Ser. B* **57**.1, pp. 289–300.

Bergen, V., Lange, M., Peidli, S., Wolf, F. A., and Theis, F. J. (2020). "Generalizing RNA velocity to transient cell states through dynamical modeling". *Nat. Biotechnol.* **38**.12, pp. 1408–1414.

Bernstein, N. J., Fong, N. L., Lam, I., Roy, M. A., Hendrickson, D. G., and Kelley, D. R. (2020). "Solo: Doublet Identification in Single-Cell RNA-Seq via Semi-Supervised Deep Learning". *Cell Syst.* **11**.1, 95–101.e5.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). "Fast unfolding of communities in large networks". *J. Stat. Mech. Theory Exp.* **2008**.10, p. 10008.

Bocharov, G., Quiel, J., Luzyanina, T., Alon, H., Chiglintsev, E., Chereshnev, V., Meier-Schellersheim, M., Paul, W. E., and Grossman, Z. (2011). "Feedback regulation of proliferation vs. differentiation rates explains the dependence of CD4 T-cell expansion on precursor number". *Proc. Natl. Acad. Sci. U. S. A.* **108**.8, pp. 3318–3323.

Bousso, P., Levraud, J. P., Kourilsky, P., and Abastado, J. P. (1999). "The composition of a primary T cell response is largely determined by the timing of recruitment of individual T cell clones". *J. Exp. Med.* **189**.10, pp. 1591–1600.

Brundage, R. A., Smith, G. A., Camilli, A., Theriot, J. A., and Portnoy, D. A. (1993). "Expression and phosphorylation of the Listeria monocytogenes actA protein in mammalian cells". *Proc. Natl. Acad. Sci. U. S. A.* **90**.24, pp. 11890–11894.

Buchholz, V. R., Flossdorf, M., Hensel, I., Kretschmer, L., Weissbrich, B., Gräf, P., Verschoor, A., Schiemann, M., Höfer, T., and Busch, D. H. (2013). "Disparate Individual Fates Compose Robust CD8+ T Cell Immunity". *Science* **340**.6132, pp. 630–635.

Buchholz, V. R., Schumacher, T. N., and Busch, D. H. (2016). "T Cell Fate at the Single-Cell Level". *Annu. Rev. Immunol.* **34**, pp. 65–92.

Burnham, K. P. and Anderson, David, R. (2002). *Model selection and multimodel inference - A practical information-theoretic approach.* 2nd Edit. New York: Springer.

Chang, J. T., Palanivel, V. R., Kinjyo, I., Schambach, F., Intlekofer, A. M., Banerjee, A., Longworth, S. A., Vinup, K. E., Mrass, P., Oliaro, J., Killeen, N., Orange, J. S., Russell, S. M., Weninger, W., and Reiner, S. L. (2007). "Asymmetric T lymphocyte division in the initiation of adaptive immune responses". *Science* **315**.5819, pp. 1687–1691.

Chao, H. X., Fakhreddin, R. I., Shimerov, H. K., Kedziora, K. M., Kumar, R. J., Perez, J., Limas, J. C., Grant, G. D., Cook, J. G., Gupta, G. P., and Purvis, J. E. (2019). "Evidence that the human cell cycle is a series of uncoupled, memoryless phases". *Mol. Syst. Biol.* **15**.3.

Chen, Z., Ji, Z., Ngiow, S. F., Manne, S., Cai, Z., Huang, A. C., Johnson, J., Staupe, R. P., Bengsch, B., Xu, C., Yu, S., Kurachi, M., Herati, R. S., Vella, L. A., Baxter, A. E., Wu, J. E., Khan, O., Beltra, J. C., Giles, J. R., Stelekati, E., McLane, L. M., Lau, C. W., Yang, X., Berger, S. L., Vahedi, G., Ji, H., and Wherry, E. J. (2019). "TCF-1-Centered Transcriptional Network Drives an Effector versus Exhausted CD8 T Cell-Fate Decision". *Immunity* **51**.5, 840–855.e5.

Cowell, F. A. (2000). *Measuring Inequality.* 3rd Edit. Oxford University Press.

Crauste, F., Mafille, J., Boucinha, L., Djebali, S., Gandrillon, O., Marvel, J., and Arpin, C. (2017). "Identification of Nascent Memory CD8 T Cells and Modeling of Their Ontogeny". *Cell Syst.* **4**.3, 306–317.e4.

Crotty, S. and Ahmed, R. (2004). "Immunological memory in humans". *Semin. Immunol.* **16**.3, pp. 197–203.

Curtsinger, J. M., Schmidt, C. S., Mondino, A., Lins, D. C., Kedl, R. M., Jenkins, M. K., and Mescher, M. F. (1999). "Inflammatory cytokines provide a third signal for activation of naive CD4+ and CD8+ T cells." *J. Immunol.* **162**.6, pp. 3256–62.

De Boer, R. J., Homann, D., and Perelson, A. S. (2003). "Different Dynamics of CD4+ and CD8+ T Cell Responses During and After Acute Lymphocytic Choriomeningitis Virus Infection". *J. Immunol.* **171**.8, pp. 3928–3935.

De Boer, R. J., Oprea, M., Antia, R., Murali-Krishna, K., Ahmed, R., and Perelson, A. S. (2001). "Recruitment Times, Proliferation, and Apoptosis Rates during the CD8+ T-Cell Response to Lymphocytic Choriomeningitis Virus". *J. Virol.* **75**.22, pp. 10663–10669.

De Boer, R. J. and Perelson, A. S. (2005). "Estimating division and death rates from CFSE data". *J. Comput. Appl. Math.* **184**.1, pp. 140–164.

– (2013). "Quantifying T lymphocyte turnover". *J. Theor. Biol.* **327**, pp. 45–87.

Deenick, E. K., Gett, A. V., and Hodgkin, P. D. (2003). "Stochastic Model of T Cell Proliferation: A Calculus Revealing IL-2 Regulation of Precursor Frequencies, Cell Cycle Time, and Survival". *J. Immunol.* **170**.10, pp. 4963–4972.

Dormand, J. R. and Prince, P. J. (1986). "A reconsideration of some embedded Runge-Kutta formulae". *J. Comput. Appl. Math.* **15**.2, pp. 203–211.

Dowling, M. R., Kan, A., Heinzel, S., Zhou, J. H., Marchingo, J. M., Wellard, C. J., Markham, J. F., and Hodgkin, P. D. (2014). "Stretched cell cycle model for proliferating lymphocytes". *Proc. Natl. Acad. Sci. U. S. A.* **111**.17, pp. 6377–6382.

Dowling, M. R., Milutinović, D., and Hodgkin, P. D. (2005). "Modelling cell lifespan and proliferation: Is likelihood to die or to divide independent of age?" *J. R. Soc. Interface* **2**.5, pp. 517–526.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Fischer, D. S., Fiedler, A. K., Kernfeld, E. M., Genga, R. M., Bastidas-Ponce, A., Bakhti, M., Lickert, H., Hasenauer, J., Maehr, R., and Theis, F. J. (2019). "Inferring population

dynamics from single-cell RNA-sequencing time series data". *Nat. Biotechnol.* **37**.4, pp. 461–468.

Fröhlich, F., Loos, C., and Hasenauer, J. (2019). "Scalable Inference of Ordinary Differential Equation Models of Biochemical Processes". *Gene Regulatory Networks: Methods and Protocols*. Ed. by G. Sanguinetti and A. Huynh-Thu. New York: Springer New York, pp. 385–422.

Ganusov, V. V. and De Boer, R. J. (2013). "A mechanistic model for bromodeoxyuridine dilution naturally explains labelling data of self-renewing T cell populations". *J. R. Soc. Interface* **10**.78, p. 20120617.

Garrod, K. R., Moreau, H. D., Garcia, Z., Lemaître, F., Bouvier, I., Albert, M. L., and Bousso, P. (2012). "Dissecting T Cell Contraction In Vivo Using a Genetically Encoded Reporter of Apoptosis". *Cell Rep.* **2**.5, pp. 1438–1447.

Gattinoni, L., Speiser, D. E., Lichterfeld, M., and Bonini, C. (2017). "T memory stem cells in health and disease". *Nat. Med.* **23**.1, pp. 18–27.

Gerlach, C., Rohr, J. C., Perié, L., Van Rooij, N., Van Heijst, J. W., Velds, A., Urbanus, J., Naik, S. H., Jacobs, H., Beltman, J. B., De Boer, R. J., and Schumacher, T. N. (2013). "Heterogeneous differentiation patterns of individual CD8+ T cells". *Science* **340**.6132, pp. 635–639.

Gillespie, D. T. (1977). "Exact stochastic simulation of coupled chemical reactions". *J. Phys. Chem.* **81**.25, pp. 2340–2361.

Gini, C. (1912). *Variability and Mutability: Contribution to the Study of Distributions and Statistical Relationships (in Italian)*. Bologna: Tipografia di Paolo Cuppini.

Gitlin, A. D., Mayer, C. T., Oliveira, T. Y., Shulman, Z., Jones, M. J., Koren, A., and Nussenzweig, M. C. (2015). "T cell help controls the speed of the cell cycle in germinal center B cells". *Science* **349**.6248, pp. 643–646.

Gookin, S., Min, M., Phadke, H., Chung, M., Moser, J., Miller, I., Carter, D., and Spencer, S. L. (2017). "A map of protein dynamics during cell-cycle progression and cell-cycle exit". *PLoS Biol.* **15**.9.

Graef, P., Buchholz, V. R., Stemberger, C., Flossdorf, M., Henkel, L., Schiemann, M., Drexler, I., Höfer, T., Riddell, S. R., and Busch, D. H. (2014). "Serial Transfer of Single-

Cell-Derived Immunocompetence Reveals Stemness of CD8+ Central Memory T Cells". *Immunity* **41**.1, pp. 116–126.

Grassmann, S., Mihatsch, L., Mir, J., Kazeroonian, A., Rahimi, R., Flommersfeld, S., Schober, K., Hensel, I., Leube, J., Pachmayr, L. O., Kretschmer, L., Zhang, Q., Jolly, A., Chaudhry, M. Z., Schiemann, M., Cicin-Sain, L., Höfer, T., Busch, D. H., Flossdorf, M., and Buchholz, V. R. (2020). "Early emergence of T central memory precursors programs clonal dominance during chronic viral infection". *Nat. Immunol.* **21**.12, pp. 1563–1573.

Griffiths, J. A., Scialdone, A., and Marioni, J. C. (2018). "Using single-cell genomics to understand developmental processes and cell fate decisions". *Mol. Syst. Biol.* **14**.4.

Gulati, G. S., Sikandar, S. S., Wesche, D. J., Manjunath, A., Bharadwaj, A., Berger, M. J., Ilagan, F., Kuo, A. H., Hsieh, R. W., Cai, S., Zabala, M., Scheeren, F. A., Lobo, N. A., Qian, D., Yu, F. B., Dirbas, F. M., Clarke, M. F., and Newman, A. M. (2020). "Single-cell transcriptional diversity is a hallmark of developmental potential". *Science* **367**.6476, pp. 405–411.

Harty, J. T. and Badovinac, V. P. (2008). "Shaping and reshaping CD8+ T-cell memory". *Nat. Rev. Immunol.* **8**.2, pp. 107–119.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. 2nd Edit. New York: Springer.

Hawkins, E. D., Markham, J. F., McGuinness, L. P., and Hodgkin, P. D. (2009). "A single-cell pedigree analysis of alternative stochastic lymphocyte fates". *Proc. Natl. Acad. Sci. U. S. A.* **106**.32, pp. 13457–13462.

Hawkins, E. D., Turner, M. L., Dowling, M. R., Van Gend, C., and Hodgkin, P. D. (2007). "A model of immune regulation as a consequence of randomized lymphocyte division and death times". *Proc. Natl. Acad. Sci. U. S. A.* **104**.12, pp. 5032–5037.

Heinzel, S., Binh Giang, T., Kan, A., Marchingo, J. M., Lye, B. K., Corcoran, L. M., and Hodgkin, P. D. (2017). "A Myc-dependent division timer complements a cell-death timer to regulate T cell and B cell responses". *Nat. Immunol.* **18**.1, pp. 96–103.

Herndler-Brandstetter, D., Ishigame, H., Shinnakasu, R., Plajer, V., Stecher, C., Zhao, J., Lietzenmayer, M., Kroehling, L., Takumi, A., Kometani, K., Inoue, T., Kluger, Y., Kaech, S. M., Kurosaki, T., Okada, T., and Flavell, R. A. (2018). "KLRG1+ Effector

CD8+ T Cells Lose KLRG1, Differentiate into All Memory T Cell Lineages, and Convey Enhanced Protective Immunity". *Immunity* **48**.4, 716–729.e8.

Hikono, H., Kohlmeier, J. E., Takamura, S., Wittmer, S. T., Roberts, A. D., and Woodland, D. L. (2007). "Activation phenotype, rather than central-or effector-memory phenotype, predicts the recall efficacy of memory CD8+ T cells". *J. Exp. Med.* **204**.7, pp. 1625–1636.

Hinkley, D. (1979). "Predictive Likelihood". *Ann. Stat.* **7**.4, pp. 718–728.

Höfer, T. and Flossdorf, M. (2019). "Modeling and inference of cell population dynamics". *Systems Immunology - An Introduction to Modeling Methods for Scientists*. Ed. by J. Das and C. Yayaprakash. CRC Press, Taylor & Francis Group. Chap. 15, pp. 251–264.

Kaech, S. M. and Ahmed, R. (2001). "Memory CD8+ T cell differentiation: Initial antigen encounter triggers a developmental program in naïve cells". *Nat. Immunol.* **2**.5, pp. 415–422.

Kalia, V., Sarkar, S., and Ahmed, R. (2010). "CD8 T-Cell Memory Differentiation during Acute and Chronic Viral Infections". *Memory T Cells*. Ed. by M. Zanetti. New York: Springer, pp. 79–95.

Khan, S. H. and Badovinac, V. P. (2015). "Listeria monocytogenes: a model pathogen to study antigen-specific memory CD8 T cell responses". *Semin. Immunopathol.* **37**.3, pp. 301–310.

King, C. G., Koehli, S., Hausmann, B., Schmaler, M., Zehn, D., and Palmer, E. (2012). "T Cell Affinity Regulates Asymmetric Division, Effector Cell Differentiation, and Tissue Pathology". *Immunity* **37**.4, pp. 709–720.

Kranich, J., Chlis, N. K., Rausch, L., Latha, A., Schifferer, M., Kurz, T., Foltyn-Arfa Kia, A., Simons, M., Theis, F. J., and Brocker, T. (2020). "In vivo identification of apoptotic and extracellular vesicle-bound live cells using image-based deep learning". *J. Extracell. Vesicles* **9**.1.

Kretschmer, L., Flossdorf, M., Mir, J., Cho, Y. L., Plambeck, M., Treise, I., Toska, A., Heinzel, S., Schiemann, M., Busch, D. H., and Buchholz, V. R. (2020). "Differential expansion of T central memory precursor and effector subsets is regulated by division speed". *Nat. Commun.* **11**.1.

Kreutz, C., Raue, A., and Timmer, J. (2012). "Likelihood based observability analysis and confidence intervals for predictions of dynamic models". *BMC Syst. Biol.* **6**.1.

Kurd, N. S., He, Z., Louis, T. L., Milner, J. J., Omilusik, K. D., Jin, W., Tsai, M. S., Widjaja, C. E., Kanbar, J. N., Olvera, J. G., Tysl, T., Quezada, L. K., Boland, B, J. T., Huang, W. J., Murre, C., Goldrath, A. W., Yeo, G. W., and Chang, J. T. (2020). "Early precursors and molecular determinants of tissue-resident memory CD8+ T lymphocytes revealed by single-cell RNA sequencing". *Sci. Immunol.* **5**.47.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., Bruggen, D. van, Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S., and Kharchenko, P. V. (2018). "RNA velocity of single cells". *Nature* **560**.7719, pp. 494–498.

Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M., Lickert, H., Ansari, M., Schniering, J., Schiller, H. B., Pe'er, D., and Theis, F. J. (2020). "CellRank for directed single-cell fate mapping". *bioRxiv*.

Lau, L. L., Jamieson, B. D., Somasundaram, T., and Ahmed, R. (1994). "Cytotoxic T-cell memory without antigen". *Nature* **369**.6482, pp. 648–652.

Lee, H. Y. and Perelson, A. S. (2008). "Modeling T cell proliferation and death in vitro based on labeling data: Generalizations of the Smith-Martin cell cycle model". *Bull. Math. Biol.* **70**.1, pp. 21–44.

Lestas, I., Paulsson, J., Ross, N. E., and Vinnicombe, G. (2008). "Noise in gene regulatory networks". *IEEE Trans. Automat. Contr.* **53**.Special Issue, pp. 189–200.

Lin, W. H. W., Nish, S. A., Yen, B., Chen, Y. H., Adams, W. C., Kratchmarov, R., Rothman, N. J., Bhandoola, A., Xue, H. H., and Reiner, S. L. (2016). "CD8+ T Lymphocyte Self-Renewal during Effector Cell Determination". *Cell Rep.* **17**.7, pp. 1773–1782.

Lönnberg, T., Svensson, V., James, K. R., Fernandez-ruiz, D., Sebina, I., Montandon, R., Soon, M. S. F., Fogg, L. G., Nair, A. S., Liligeto, U. N., Stubbington, M. J. T., Ly, L.-h., Bagger, F. O., Zwiessele, M., Lawrence, N. D., Souza-fonseca-guimaraes, F., Bunn, P. T., Engwerda, C. R., Heath, W. R., Billker, O., Stegle, O., Haque, A., and Teichmann, S. A. (2018). "Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves TH1/TFH fate bifurcation in malaria". *Sci. Immunol.* **3**.21.

Luecken, M. D. and Theis, F. J. (2019). "Current best practices in single-cell RNA-seq analysis: a tutorial". *Mol. Syst. Biol.* **15**.6.

Lun, A. T. L., Bach, K., and Marioni, J. C. (2016). "Pooling across cells to normalize single-cell RNA sequencing data with many zero counts". *Genome Biol.* **17**.1.

Mac Auley, A., Werb, Z., and Mirkes, P. E. (1993). "Characterization of the unusually rapid cell cycles during rat gastrulation". *Development* **117**.3, pp. 873–883.

Marchingo, J. M., Kan, A., Sutherland, R. M., Duffy, K. R., Wellard, C. J., Belz, G. T., Lew, A. M., Dowling, M. R., Heinzel, S., and Hodgkin, P. D. (2014). "Antigen affinity, costimulation, and cytokine inputs sum linearly to amplify T cell expansion". *Science* **346**.6213, pp. 1123–1127.

Martin, M. D. and Badovinac, V. P. (2018). "Defining memory CD8 T cell". *Front. Immunol.* **9**.11.

Matiasova, A., Sevc, J., Mikes, J., Jendzelovsky, R., Daxnerova, Z., and Fedorocko, P. (2014). "Flow cytometric determination of 5-bromo-2-deoxyuridine pharmacokinetics in blood serum after intraperitoneal administration to rats and mice". *Histochem. Cell Biol.* **142**.6, pp. 703–712.

McInnes, L., Healy, J., and Melville, J. (2018). "UMAP: Uniform manifold approximation and projection for dimension reduction". *arXiv*.

McKay, M. D., Beckman, R. J., and Conover, W. J. (2000). "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code". *Technometrics* **42**.1, pp. 55–61.

Miles, A. S., Hodgkin, P. D., and Duffy, K. R. (2019). "Inferring differentiation order in adaptive immune responses from population level data". *arXiv*.

Milner, J. J., Nguyen, H., Omilusik, K., Reina-Campos, M., Tsai, M., Toma, C., Delpoux, A., Boland, B. S., Hedrick, S. M., Chang, J. T., and Goldrath, A. W. (2020). "Delineation of a molecularly distinct terminally differentiated memory CD8 T cell population". *Proc. Natl. Acad. Sci. U. S. A.* **117**.41, pp. 25667–25678.

Murphy, K. and Weaver, C. (2017). *Janeway's Immuobiology*. 9th Edit. New York: Garland Science, Taylor & Francis Group, LLC.

Neeve, S. C., Robinson, B. W., and Fear, V. S. (2019). "The role and therapeutic implications of T cells in cancer of the lung". *Clin. Transl. Immunol.* **8**.8.

Obar, J. J., Khanna, K. M., and Lefrançois, L. (2008). "Endogenous Naive CD8+ T Cell Precursor Frequency Regulates Primary and Memory Responses to Infection". *Immunity* **28**.6, pp. 859–869.

Opferman, J. T., Ober, B. T., and Ashton-Rickardt, P. G. (1999). "Linear differentiation of cytotoxic effectors into memory T lymphocytes". *Science* **283**.5408, pp. 1745–1748.

Owen, J. A., Punt, J., Stranford, S. A., and Jones, P. (2013). *Kuby immunology*. 7th Edit. New York: W. H. Freeman and Company.

Pace, L., Goudot, C., Zueva, E., Gueguen, P., Burgdorf, N., Waterfall, J. J., Quivy, J. P., Almouzni, G., and Amigorena, S. (2018). "The epigenetic control of stemness in CD8+ T cell fate commitment". *Science* **359**.6372, pp. 177–186.

Parretta, E., Cassese, G., Santoni, A., Guardiola, J., Vecchio, A., and Di Rosa, F. (2008). "Kinetics of In Vivo Proliferation and Death of Memory and Naive CD8 T Cells: Parameter Estimation Based on 5-Bromo-2-Deoxyuridine Incorporation in Spleen, Lymph Nodes, and Bone Marrow". *J. Immunol.* **180**.11, pp. 7230–7239.

Pearson, K. (1901). " LIII. On lines and planes of closest fit to systems of points in space ". *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **2**.11, pp. 559–572.

Pham, K., Kan, A., Whitehead, L., Hennessy, R. J., Rogers, K., and Hodgkin, P. D. (2018). "Converse Smith-Martin cell cycle kinetics by transformed B lymphocytes". *Cell Cycle* **17**.16, pp. 2041–2051.

Porter, B. B. and Harty, J. T. (2006). "The onset of CD8+ T-cell contraction is influenced by the peak of Listeria monocytogenes infection and antigen display". *Infect. Immun.* **74**.3, pp. 1528–1536.

Powell, E. O. (1956). "Growth Rate and Generation Time of Bacteria, with Special Reference to Continuous Culture". *J. Gen. Microbiol.* **15**.3, pp. 492–511.

Reuter, B., Weber, M., Fackeldey, K., Ro, S., and Garcia, M. E. (2018). "Generalized Markov State Modeling Method for Nonequilibrium Biomolecular Dynamics : Exemplified on Amyloid β Conformational Dynamics Driven by an Oscillating Electric Field". *J. Chem. Theory Comput.* **14**.7, pp. 3579–3594.

Robert, P. A., Kunze-Schumacher, H., Greiff, V., and Krueger, A. (2021). "Modeling the Dynamics of T-Cell Development in the Thymus". *Entropy* **23**.4.

Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). "A comparison of single-cell trajectory inference methods". *Nat. Biotechnol.* **37**.5, pp. 547–554.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A. (2015). "Spatial reconstruction of single-cell gene expression data". *Nat. Biotechnol.* **33**.5, pp. 495–502.

Schlub, T. E., Badovinac, V. P., Sabel, J. T., Harty, J. T., and Davenport, M. P. (2010). "Predicting CD62L expression during the CD8+ T-cell response in vivo". *Immunol. Cell Biol.* **88**.2, pp. 157–164.

Schlub, T. E., Venturi, V., Kedzierska, K., Wellard, C., Doherty, P. C., Turner, S. J., Ribeiro, R. M., Hodgkin, P. D., and Davenport, M. P. (2009). "Division-linked differentiation can account for CD8+ T-cell phenotype in vivo". *Eur. J. Immunol.* **39**.1, pp. 67–77.

Schwabe, D., Formichetti, S., Junker, J. P., Falcke, M., and Rajewsky, N. (2020). "The transcriptome dynamics of single cells during the cell cycle". *Mol. Syst. Biol.* **16**.11.

Smith, J. A. and Martin, L. (1973). "Do cells cycle?" *Proc. Natl. Acad. Sci. U. S. A.* **70**.4, pp. 1263–1267.

Snow, M. H. (1977). "Gastrulation in the mouse: growth and regionalization of the epiblast". *J. Embryol. Exp. Morphol.* **42**.2, pp. 293–303.

Steinberg, N. (2021). "Flow Cytometry GUI for Matlab". *MATLAB Central File Exchange*.

Stemberger, C., Huster, K. M., Koffler, M., Anderl, F., Schiemann, M., Wagner, H., and Busch, D. H. (2007). "A Single Naive CD8+ T Cell Precursor Can Develop into Diverse Effector and Memory Subsets". *Immunity* **27**.6, pp. 985–997.

Stukalin, E. B., Aifuwa, I., Kim, J. S., Wirtz, D., and Sun, S. X. (2013). "Age-dependent stochastic models for understanding population fluctuations in continuously cultured cells". *J. R. Soc. Interface* **10**.85.

Subramanian, V. G., Duffy, K. R., Turner, M. L., and Hodgkin, P. D. (2008). "Determining the expected variability of immune responses using the cyton model". *J. Math. Biol.* **56**.6, pp. 861–892.

Terry, E., Marvel, J., Arpin, C., Gandrillon, O., and Crauste, F. (2012). "Mathematical model of the primary CD8 T cell immune response: Stability analysis of a nonlinear age-structured system". *J. Math. Biol.* **65**.2, pp. 263–291.

Thattai, M. and Van Oudenaarden, A. (2001). "Intrinsic noise in gene regulatory networks". *Proc. Natl. Acad. Sci. U. S. A.* **98**.15, pp. 8614–8619.

Theil, H. (1967). *Economics and information theory*. Amsterdam: North-Holland Publishing Company.

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., Rotem, A., Rodman, C., Lian, C., Murphy, G., Fallahi-Sichani, M., Dutton-Regester, K., Lin, J. R., Cohen, O., Shah, P., Lu, D., Genshaft, A. S., Hughes, T. K., Ziegler, C. G., Kazer, S. W., Gaillard, A., Kolb, K. E., Villani, A. C., Johannessen, C. M., Andreev, A. Y., Van Allen, E. M., Bertagnolli, M., Sorger, P. K., Sullivan, R. J., Flaherty, K. T., Frederick, D. T., Jané-Valbuena, J., Yoon, C. H., Rozenblatt-Rosen, O., Shalek, A. K., Regev, A., and Garraway, L. A. (2016). "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq". *Science* **352**.6282, pp. 189–196.

Traag, V. A., Waltman, L., and Eck, N. J. van (2019). "From Louvain to Leiden: guaranteeing well-connected communities". *Sci. Rep.* **9**.1.

Turner, M. L., Hawkins, E. D., and Hodgkin, P. D. (2008). "Quantitative Regulation of B Cell Division Destiny by Signal Strength". *J. Immunol.* **181**.1, pp. 374–382.

Van Heijst, J. W., Gerlach, C., Swart, E., Sie, D., Nunes-Alves, C., Kerkhoven, R. M., Arens, R., Correia-Neves, M., Schepers, K., and Schumacher, T. N. (2009). "Recruitment of antigen-specific CD8+ T cells in response to infection is markedly efficient". *Science* **325**.5945, pp. 1265–1269.

Venzon, D. J. and Moolgavkar, S. H. (1988). "A Method for Computing Profile-Likelihood-Based Confidence Intervals". *J. R. Stat. Soc.* **37**.1, pp. 87–94.

Von Foerster, H. (1959). "Some remarks on changing populations". *The kinetics of cellular proliferation*. Ed. by J. F. Stohlman. New York: Grune and Stratton, pp. 382–407.

Watson, J. V., Chambers, S. H., and Smith, P. J. (1987). "A pragmatic approach to the analysis of DNA histograms with a definable G1 peak". *Cytometry* **8**.1, pp. 1–8.

Watts, T. H., Bertram, E. M., Bukczynski, J., and Wen, T. (2003). "T cell costimulatory molecules in anti-viral immunity: Potential role in immunotherapeutic vaccines". *Can. J. Infect. Dis.* **14**.4, pp. 221–229.

Weber, T. S., Jaehnert, I., Schichor, C., Or-Guil, M., and Carneiro, J. (2014). "Quantifying the Length and Variance of the Eukaryotic Cell Cycle Phases by a Stochastic Model and Dual Nucleoside Pulse Labelling". *PLoS Comput. Biol.* **10**.7.

Wherry, E. J., Teichgräber, V., Becker, T. C., Masopust, D., Kaech, S. M., Antia, R., Andrian, U. H. von, and Ahmed, R. (2003). "Lineage relationship and protective immunity of memory CD8 T cell subsets". *Nat. Immunol.* **4**.3, pp. 225–234.

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

Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F. J. (2019). "PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells". *Genome Biol.* **20**.59.

Yang, S., Liu, F., Wang, Q. J., Rosenberg, S. A., and Morgan, R. A. (2011). "The shedding of CD62L (L-selectin) regulates the acquisition of lytic activity in human tumor reactive T lymphocytes". *PLoS One* **6**.7.

Yao, C., Sun, H. W., Lacey, N. E., Ji, Y., Moseman, E. A., Shih, H. Y., Heuston, E. F., Kirby, M., Anderson, S., Cheng, J., Khan, O., Handon, R., Reilley, J., Fioravanti, J., Hu, J., Gossa, S., Wherry, E. J., Gattinoni, L., McGavern, D. B., O'Shea, J. J., Schwartzberg, P. L., and Wu, T. (2019). "Single-cell RNA-seq reveals TOX as a key regulator of CD8+ T cell persistence in chronic infection". *Nat. Immunol.* **20**.7, pp. 890–901.

Zehn, D., Lee, S. Y., and Bevan, M. J. (2009). "Complete but curtailed T-cell response to very low-affinity antigen". *Nature* **458**.7235, pp. 211–214.

Zhang, J., Ahn, J., Suh, Y., Hwang, S., Davis, M. E., and Lee, K. (2015). "Identification of CTLA2A, DEFB29, WFDC15B, SERPINA1F and MUP19 as Novel Tissue-Specific Secretory Factors in Mouse". *PLoS One* **10**.5.

Zhou, J. H., Markham, J. F., Duffy, K. R., and Hodgkin, P. D. (2018). "Stochastically timed competition between division and differentiation fates regulates the transition from B lymphoblast to plasma cell". *Front. Immunol.* **9**.9.

Appendix A

Steady state solutions to the von Foerster equation

This supplementary information contains an analytical derivation of the exact form of $p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta)$, that was used in section 3.1. $p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta)$ is the distribution of times-until-next division in steady state of a continuously growing population of cells. Derivation and notation were adopted from [Dowling et al., 2005]. For a derivation of the analogous distribution in age representation, $p^{\mathrm{ss}}_{\mathrm{age}}(a)$, we refer to [Stukalin et al., 2013].

Let us consider a population of asynchronously dividing cells with general division-time distribution $D(T_{\mathrm{div}})$. For a given cell, its total division-time T_{div} is a sum of its age a and its time-to-next division δ. Here, the population density P depends on δ and T_{div}, i.e. $P = P(\delta, T_{\mathrm{div}}, t)$. At a given time t, its normalisation $N(t)$ gives the current cell number of the population

$$N(t) = \int_0^\infty \mathrm{d}\delta \int_0^\infty \mathrm{d}T_{\mathrm{div}}\, P(\delta, T_{\mathrm{div}}, t). \tag{A.1}$$

For our purpose it is helpful to consider the marginal distribution $P_{\mathrm{tnd}}(\delta, t)$, given by

$$P_{\mathrm{tnd}}(\delta, t) = \int_0^\infty \mathrm{d}T_{\mathrm{div}}\, P(\delta, T_{\mathrm{div}}, t). \tag{A.2}$$

The actual distribution of time-to-next division is then defined via

$$p_{\mathrm{tnd}}(\delta, t) = \frac{P_{\mathrm{tnd}}(\delta, t)}{N(t)}. \tag{A.3}$$

In order to derive a partial differential equation (PDE) for $P_{\mathrm{tnd}}(\delta, t)$ to first order, we start from the total differential

$$\mathrm{d}P_{\mathrm{tnd}}(\delta, t) = \frac{\partial P_{\mathrm{tnd}}(\delta, t)}{\partial t}\mathrm{d}t + \frac{\partial P_{\mathrm{tnd}}(\delta, t)}{\partial \delta}\mathrm{d}\delta. \tag{A.4}$$

From eq. A.4 it can be concluded that

$$\frac{\mathrm{d}P_{\mathrm{tnd}}(\delta,t)}{\mathrm{d}t} = \frac{\partial P_{\mathrm{tnd}}(\delta,t)}{\partial t} - \frac{\partial P_{\mathrm{tnd}}(\delta,t)}{\partial \delta} =: f(\delta,t), \tag{A.5}$$

where we have used that $\mathrm{d}\delta/\mathrm{d}t = -1$. The function $f(\delta,t)$ captures the left-hand side of the equation. In the presence of age-dependent division and age-independent cell death, f takes the form[1]

$$f(\delta,t) = 2D(T_{\mathrm{div}}{=}\delta)P_{\mathrm{tnd}}(\delta{=}0,t) - dP_{\mathrm{tnd}}(\delta,t). \tag{A.6}$$

This is because $2P_{\mathrm{tnd}}(\delta{=}0,t)$ takes into account all newly generated cells at time t. Weighting these with $D(\delta)$, gives the division-induced influx of cells with time-until-next division of δ. On the other hand, the second term represents exponential cell death occurring with rate $d \neq d(\delta)$. It should be emphasised that in the presence of cell death many quantities have a different interpretation. For example, δ can no longer be interpreted as the time-until-next division in case $T_{\mathrm{div}} > T_{\mathrm{death}}$, i.e. if a cell dies before it divides. For the purpose of simplicity, however, we refrain from adjusting the notation. Altogether, this leads to a time-evolution of P, given by

$$\frac{\partial P_{\mathrm{tnd}}(\delta,t)}{\partial t} = 2D(\delta)P_{\mathrm{tnd}}(0,t) + \frac{\partial P_{\mathrm{tnd}}(\delta,t)}{\partial \delta} - dP_{\mathrm{tnd}}(\delta,t). \tag{A.7}$$

Eq. A.7 has a similar structure as the von Foerster equation, which was formulated in terms of time and age, see [Von Foerster, 1959]. In steady state, solutions to eq. A.7 have the form

$$P_{\mathrm{tnd}}(\delta,t) = N(t)p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta), \tag{A.8}$$

[1]Note that here we deviate from [Dowling et al., 2005] by incorporating also cell death into the evolution equation of P_{tnd}.

i.e. the cell number captures the full time-dependence of the solution. The exact shape of $N(t)$ must be obtained. To this end, an equation for the normalisation N is derived, starting from its definition A.1

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}N(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\int_0^\infty \mathrm{d}\delta\, P_{\mathrm{tnd}}(\delta,t) = \int_0^\infty \mathrm{d}\delta\, \frac{\partial P_{\mathrm{tnd}}(\delta,t)}{\partial t} \\
&\overset{(A.7)}{=} \int_0^\infty \mathrm{d}\delta \left[2D(\delta)P_{\mathrm{tnd}}(0,t) + \frac{\partial P_{\mathrm{tnd}}(\delta,t)}{\partial \delta} - dP_{\mathrm{tnd}}(\delta,t)\right] \\
&= 2P_{\mathrm{tnd}}(0,t)\int_0^\infty \mathrm{d}\delta\, D(\delta) + \int_0^\infty \mathrm{d}\delta\, \frac{\partial P_{\mathrm{tnd}}(\delta,t)}{\partial \delta} - d\int_0^\infty \mathrm{d}\delta\, P_{\mathrm{tnd}}(\delta,t) \\
&= 2P_{\mathrm{tnd}}(0,t) + P_{\mathrm{tnd}}(\delta,t)\Big|_{\delta=0}^{\delta=\infty} - dN(t),
\end{aligned} \tag{A.9}$$

where from the third to the fourth line we used the known normalisation properties of D and P_{tnd}. Furthermore, considering that a properly defined density must satisfy $\lim_{\delta\to\infty} P_{\mathrm{tnd}}(\delta,t) = 0$ and using eq. A.3, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}N(t) = P_{\mathrm{tnd}}(0,t) - dN(t) = \left[p_{\mathrm{tnd}}(0,t) - d\right]N(t). \tag{A.10}$$

Employing the steady state condition given by eq. A.8 on eq. A.10, we obtain the equation that captures the dynamics of the cell number in steady state, i.e.

$$\frac{\mathrm{d}}{\mathrm{d}t}N(t) = \left[p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta{=}0) - d\right]N(t). \tag{A.11}$$

Defining the constant parameter $c := p^{\mathrm{ss}}_{\mathrm{tnd}}(0)$ and the initial condition $N_0 := N(t{=}0)$, eq. A.11 is solved by

$$N(t) = N_0\,\mathrm{e}^{(c-d)t}. \tag{A.12}$$

Eq. A.12 shows that although an age-dependent division-time distribution was assumed, the long-term behaviour of the population size is characterised by exponential growth with effective rate $c - d$. Now, the full ansatz of the steady state solution to eq. A.7 can be expressed as

$$P_{\mathrm{tnd}}(\delta,t) = N(t)p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta) = N_0\,\mathrm{e}^{(c-d)t}p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta). \tag{A.13}$$

Inserting this result into eq. A.7 gives

$$\frac{\partial}{\partial t}\left(N(t)p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta)\right) = 2D(\delta)N(t)p^{\mathrm{ss}}_{\mathrm{tnd}}(0) + \frac{\partial}{\partial \delta}\left(N(t)p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta)\right) - dN(t)p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta). \tag{A.14}$$

Evaluating the derivatives and dividing the equation by $N(t) \neq 0$ (for all t), we obtain the equation for $p^{\mathrm{ss}}_{\mathrm{tnd}}$, that is

$$\frac{\mathrm{d}p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta)}{\mathrm{d}\delta} = cp^{\mathrm{ss}}_{\mathrm{tnd}}(\delta) - 2cD(\delta). \tag{A.15}$$

Note that in eq. A.15 the death rate d cancels out, which only holds for age-independent death mechanisms. Eq. A.15 is solved by

$$p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta) = c\mathrm{e}^{c\delta}\left(1 - 2\int_0^\delta \mathrm{d}\delta' D(\delta')\mathrm{e}^{-c\delta'}\right). \tag{A.16}$$

Eq. A.16 depends on the full division-time distribution and the mean growth factor c. From calculating the derivative it follows that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\delta}p^{\mathrm{ss}}_{\mathrm{tnd}}(\delta) &= \frac{\mathrm{d}}{\mathrm{d}\delta}c\mathrm{e}^{c\delta} - \frac{\mathrm{d}}{\mathrm{d}\delta}\left[2c\mathrm{e}^{c\delta}\int_0^\delta \mathrm{d}\delta' D(\delta')\mathrm{e}^{-c\delta'}\right] \\
&= c^2\mathrm{e}^{c\delta} - 2c^2\mathrm{e}^{c\delta}\int_0^\delta \mathrm{d}\delta' D(\delta')\mathrm{e}^{-c\delta'} - 2c\mathrm{e}^{c\delta}\frac{\mathrm{d}}{\mathrm{d}\delta}\int_0^\delta \mathrm{d}\delta' D(\delta')\mathrm{e}^{-c\delta'} \\
&= c^2\mathrm{e}^{c\delta}\left(1 - 2\int_0^\delta \mathrm{d}\delta' D(\delta')\mathrm{e}^{-c\delta'}\right) - 2c\mathrm{e}^{c\delta}D(\delta)\mathrm{e}^{-c\delta} \\
&= cp^{\mathrm{ss}}_{\mathrm{tnd}}(\delta) - 2cD(\delta),
\end{aligned} \tag{A.17}$$

i.e. eq. A.16 indeed represents a solution to eq. A.15.

Appendix B

Moment equations of the cell cycle dropout model

In this appendix, we provide the second-order moment equations for the cell cycle dropout model specified in chapter 4. The given equations thus complement section 2.1.1 in which the corresponding equations for the mean state vector x are provided. Hence, the same considerations apply here. By using equation 2.9 all necessary equations for the symmetric covariance matrix Σ_{ij} are generated. The respective equations for the variances are given by

$$
\begin{aligned}
\dot{\Sigma}_{1,1} &= r_{1\to2}\langle x_1\rangle - 2\Sigma_{1,1}r_{1\to2} \\
\dot{\Sigma}_{2,2} &= 2\Sigma_{1,2}r_{1\to2} - 2\Sigma_{2,2}r_{2\to3} + r_{1\to2}\langle x_1\rangle + r_{2\to3}\langle x_2\rangle \\
\dot{\Sigma}_{3,3} &= 2\Sigma_{2,3}r_{2\to3} - 2\Sigma_{3,3}r_{3\to4} + r_{2\to3}\langle x_2\rangle + r_{3\to4}\langle x_3\rangle \\
\dot{\Sigma}_{4,4} &= 2\Sigma_{3,4}r_{3\to4} - 2\Sigma_{4,4}r_{4\to5} + r_{3\to4}\langle x_3\rangle + r_{4\to5}\langle x_4\rangle \\
\dot{\Sigma}_{5,5} &= 2\Sigma_{4,5}r_{4\to5} - 2\Sigma_{5,5}r_{5\to6} + r_{4\to5}\langle x_4\rangle + r_{5\to6}\langle x_5\rangle \\
\dot{\Sigma}_{6,6} &= 4\Sigma_{5,6}r_{5\to6} + p_6\langle x_6\rangle + 4r_{5\to6}\langle x_5\rangle + r_{6\to7}\langle x_6\rangle + r_{6\to9}\langle x_6\rangle - 2\Sigma_{6,6}(r_{6\to7} - p_6 + r_{6\to9}) \\
\dot{\Sigma}_{7,7} &= 2\Sigma_{6,7}r_{6\to7} + p_7\langle x_7\rangle + r_{6\to7}\langle x_6\rangle + r_{7\to8}\langle x_7\rangle + r_{7\to10}\langle x_7\rangle - 2\Sigma_{7,7}(r_{7\to8} - p_7 + r_{7\to10}) \\
\dot{\Sigma}_{8,8} &= 2\Sigma_{7,8}r_{7\to8} + p_8\langle x_8\rangle + r_{7\to8}\langle x_7\rangle + r_{8\to11}\langle x_8\rangle + 2\Sigma_{8,8}(p_8 - r_{8\to11}) \\
\dot{\Sigma}_{9,9} &= 2\Sigma_{6,9}r_{6\to9} + d_9\langle x_9\rangle + r_{6\to9}\langle x_6\rangle + r_{9\to10}\langle x_9\rangle - 2\Sigma_{9,9}(d_9 + r_{9\to10}) \\
\dot{\Sigma}_{10,10} &= 2\Sigma_{7,10}r_{7\to10} + 2\Sigma_{9,10}r_{9\to10} + d_{10}\langle x_{10}\rangle + r_{7\to10}\langle x_7\rangle + r_{9\to10}\langle x_9\rangle + r_{10\to11}\langle x_{10}\rangle - \\
&\quad - 2\Sigma_{10,10}(d_{10} + r_{10\to11}) \\
\dot{\Sigma}_{11,11} &= 2\Sigma_{8,11}r_{8\to11} - 2\Sigma_{11,11}d_{11} + 2\Sigma_{10,11}r_{10\to11} + d_{11}\langle x_{11}\rangle + r_{8\to11}\langle x_8\rangle + r_{10\to11}\langle x_{10}\rangle.
\end{aligned}
\tag{B.1}
$$

The equations for the off-diagonal elements, i.e. the covariances, read

$$
\begin{aligned}
\dot{\Sigma}_{1,2} &= \Sigma_{1,1}r_{1\to2} - \Sigma_{1,2}r_{1\to2} - \Sigma_{1,2}r_{2\to3} - r_{1\to2}\langle x_1\rangle \\
\dot{\Sigma}_{1,3} &= \Sigma_{1,2}r_{2\to3} - \Sigma_{1,3}r_{1\to2} - \Sigma_{1,3}r_{3\to4} \\
\dot{\Sigma}_{1,4} &= \Sigma_{1,3}r_{3\to4} - \Sigma_{1,4}r_{1\to2} - \Sigma_{1,4}r_{4\to5}
\end{aligned}
$$

$$\dot{\Sigma}_{1,5} = \Sigma_{1,4}r_{4\to5} - \Sigma_{1,5}r_{1\to2} - \Sigma_{1,5}r_{5\to6}$$
$$\dot{\Sigma}_{1,6} = 2\Sigma_{1,5}r_{5\to6} - \Sigma_{1,6}r_{1\to2} - \Sigma_{1,6}(r_{6\to7} - p_6 + r_{6\to9})$$
$$\dot{\Sigma}_{1,7} = \Sigma_{1,6}r_{6\to7} - \Sigma_{1,7}r_{1\to2} - \Sigma_{1,7}(r_{7\to8} - p_7 + r_{7\to10})$$
$$\dot{\Sigma}_{1,8} = \Sigma_{1,7}r_{7\to8} - \Sigma_{1,8}r_{1\to2} + \Sigma_{1,8}(p_8 - r_{8\to11})$$
$$\dot{\Sigma}_{1,9} = \Sigma_{1,6}r_{6\to9} - \Sigma_{1,9}r_{1\to2} - \Sigma_{1,9}(d_9 + r_{9\to10})$$
$$\dot{\Sigma}_{1,10} = \Sigma_{1,7}r_{7\to10} - \Sigma_{1,10}r_{1\to2} + \Sigma_{1,9}r_{9\to10} - \Sigma_{1,10}(d_{10} + r_{10\to11})$$
$$\dot{\Sigma}_{1,11} = \Sigma_{1,8}r_{8\to11} - \Sigma_{1,11}r_{1\to2} - \Sigma_{1,11}d_{11} + \Sigma_{1,10}r_{10\to11}$$
$$\dot{\Sigma}_{2,3} = \Sigma_{1,3}r_{1\to2} + \Sigma_{2,2}r_{2\to3} - \Sigma_{2,3}r_{2\to3} - \Sigma_{2,3}r_{3\to4} - r_{2\to3}\langle x_2\rangle$$
$$\dot{\Sigma}_{2,4} = \Sigma_{1,4}r_{1\to2} - \Sigma_{2,4}r_{2\to3} + \Sigma_{2,3}r_{3\to4} - \Sigma_{2,4}r_{4\to5}$$
$$\dot{\Sigma}_{2,5} = \Sigma_{1,5}r_{1\to2} - \Sigma_{2,5}r_{2\to3} + \Sigma_{2,4}r_{4\to5} - \Sigma_{2,5}r_{5\to6}$$
$$\dot{\Sigma}_{2,6} = \Sigma_{1,6}r_{1\to2} - \Sigma_{2,6}r_{2\to3} + 2\Sigma_{2,5}r_{5\to6} - \Sigma_{2,6}(r_{6\to7} - p_6 + r_{6\to9})$$
$$\dot{\Sigma}_{2,7} = \Sigma_{1,7}r_{1\to2} - \Sigma_{2,7}r_{2\to3} + \Sigma_{2,6}r_{6\to7} - \Sigma_{2,7}(r_{7\to8} - p_7 + r_{7\to10})$$
$$\dot{\Sigma}_{2,8} = \Sigma_{1,8}r_{1\to2} - \Sigma_{2,8}r_{2\to3} + \Sigma_{2,7}r_{7\to8} + \Sigma_{2,8}(p_8 - r_{8\to11})$$
$$\dot{\Sigma}_{2,9} = \Sigma_{1,9}r_{1\to2} - \Sigma_{2,9}r_{2\to3} + \Sigma_{2,6}r_{6\to9} - \Sigma_{2,9}(d_9 + r_{9\to10})$$
$$\dot{\Sigma}_{2,10} = \Sigma_{1,10}r_{1\to2} - \Sigma_{2,10}r_{2\to3} + \Sigma_{2,7}r_{7\to10} + \Sigma_{2,9}r_{9\to10} - \Sigma_{2,10}(d_{10} + r_{10\to11})$$
$$\dot{\Sigma}_{2,11} = \Sigma_{1,11}r_{1\to2} - \Sigma_{2,11}d_{11} - \Sigma_{2,11}r_{2\to3} + \Sigma_{2,8}r_{8\to11} + \Sigma_{2,10}r_{10\to11}$$
$$\dot{\Sigma}_{3,4} = \Sigma_{2,4}r_{2\to3} + \Sigma_{3,3}r_{3\to4} - \Sigma_{3,4}r_{3\to4} - \Sigma_{3,4}r_{4\to5} - r_{3\to4}\langle x_3\rangle$$
$$\dot{\Sigma}_{3,5} = \Sigma_{2,5}r_{2\to3} - \Sigma_{3,5}r_{3\to4} + \Sigma_{3,4}r_{4\to5} - \Sigma_{3,5}r_{5\to6}$$
$$\dot{\Sigma}_{3,6} = \Sigma_{2,6}r_{2\to3} - \Sigma_{3,6}r_{3\to4} + 2\Sigma_{3,5}r_{5\to6} - \Sigma_{3,6}(r_{6\to7} - p_6 + r_{6\to9})$$
$$\dot{\Sigma}_{3,7} = \Sigma_{2,7}r_{2\to3} - \Sigma_{3,7}r_{3\to4} + \Sigma_{3,6}r_{6\to7} - \Sigma_{3,7}(r_{7\to8} - p_7 + r_{7\to10})$$
$$\dot{\Sigma}_{3,8} = \Sigma_{2,8}r_{2\to3} - \Sigma_{3,8}r_{3\to4} + \Sigma_{3,7}r_{7\to8} + \Sigma_{3,8}(p_8 - r_{8\to11})$$
$$\dot{\Sigma}_{3,9} = \Sigma_{2,9}r_{2\to3} - \Sigma_{3,9}r_{3\to4} + \Sigma_{3,6}r_{6\to9} - \Sigma_{3,9}(d_9 + r_{9\to10})$$
$$\dot{\Sigma}_{3,10} = \Sigma_{2,10}r_{2\to3} - \Sigma_{3,10}r_{3\to4} + \Sigma_{3,7}r_{7\to10} + \Sigma_{3,9}r_{9\to10} - \Sigma_{3,10}(d_{10} + r_{10\to11})$$
$$\dot{\Sigma}_{3,11} = \Sigma_{2,11}r_{2\to3} - \Sigma_{3,11}d_{11} - \Sigma_{3,11}r_{3\to4} + \Sigma_{3,8}r_{8\to11} + \Sigma_{3,10}r_{10\to11}$$
$$\dot{\Sigma}_{4,5} = \Sigma_{3,5}r_{3\to4} + \Sigma_{4,4}r_{4\to5} - \Sigma_{4,5}r_{4\to5} - \Sigma_{4,5}r_{5\to6} - r_{4\to5}\langle x_4\rangle$$
$$\dot{\Sigma}_{4,6} = \Sigma_{3,6}r_{3\to4} - \Sigma_{4,6}r_{4\to5} + 2\Sigma_{4,5}r_{5\to6} - \Sigma_{4,6}(r_{6\to7} - p_6 + r_{6\to9})$$
$$\dot{\Sigma}_{4,7} = \Sigma_{3,7}r_{3\to4} - \Sigma_{4,7}r_{4\to5} + \Sigma_{4,6}r_{6\to7} - \Sigma_{4,7}(r_{7\to8} - p_7 + r_{7\to10})$$
$$\dot{\Sigma}_{4,8} = \Sigma_{3,8}r_{3\to4} - \Sigma_{4,8}r_{4\to5} + \Sigma_{4,7}r_{7\to8} + \Sigma_{4,8}(p_8 - r_{8\to11})$$
$$\dot{\Sigma}_{4,9} = \Sigma_{3,9}r_{3\to4} - \Sigma_{4,9}r_{4\to5} + \Sigma_{4,6}r_{6\to9} - \Sigma_{4,9}(d_9 + r_{9\to10})$$
$$\dot{\Sigma}_{4,10} = \Sigma_{3,10}r_{3\to4} - \Sigma_{4,10}r_{4\to5} + \Sigma_{4,7}r_{7\to10} + \Sigma_{4,9}r_{9\to10} - \Sigma_{4,10}(d_{10} + r_{10\to11})$$
$$\dot{\Sigma}_{4,11} = \Sigma_{3,11}r_{3\to4} - \Sigma_{4,11}d_{11} - \Sigma_{4,11}r_{4\to5} + \Sigma_{4,8}r_{8\to11} + \Sigma_{4,10}r_{10\to11}$$
$$\dot{\Sigma}_{5,6} = \Sigma_{4,6}r_{4\to5} + 2\Sigma_{5,5}r_{5\to6} - \Sigma_{5,6}r_{5\to6} - 2r_{5\to6}\langle x_5\rangle - \Sigma_{5,6}(r_{6\to7} - p_6 + r_{6\to9})$$

$$
\begin{aligned}
\dot{\Sigma}_{5,7} &= \Sigma_{4,7}r_{4\to5} - \Sigma_{5,7}r_{5\to6} + \Sigma_{5,6}r_{6\to7} - \Sigma_{5,7}(r_{7\to8} - p_7 + r_{7\to10}) \\
\dot{\Sigma}_{5,8} &= \Sigma_{4,8}r_{4\to5} - \Sigma_{5,8}r_{5\to6} + \Sigma_{5,7}r_{7\to8} + \Sigma_{5,8}(p_8 - r_{8\to11}) \\
\dot{\Sigma}_{5,9} &= \Sigma_{4,9}r_{4\to5} - \Sigma_{5,9}r_{5\to6} + \Sigma_{5,6}r_{6\to9} - \Sigma_{5,9}(d_9 + r_{9\to10}) \\
\dot{\Sigma}_{5,10} &= \Sigma_{4,10}r_{4\to5} - \Sigma_{5,10}r_{5\to6} + \Sigma_{5,7}r_{7\to10} + \Sigma_{5,9}r_{9\to10} - \Sigma_{5,10}(d_{10} + r_{10\to11}) \\
\dot{\Sigma}_{5,11} &= \Sigma_{4,11}r_{4\to5} - \Sigma_{5,11}d_{11} - \Sigma_{5,11}r_{5\to6} + \Sigma_{5,8}r_{8\to11} + \Sigma_{5,10}r_{10\to11} \\
\dot{\Sigma}_{6,7} &= 2\Sigma_{5,7}r_{5\to6} + \Sigma_{6,6}r_{6\to7} - r_{6\to7}\langle x_6\rangle - \Sigma_{6,7}(r_{6\to7} - p_6 + r_{6\to9}) - \\
&\quad - \Sigma_{6,7}(r_{7\to8} - p_7 + r_{7\to10}) \\
\dot{\Sigma}_{6,8} &= 2\Sigma_{5,8}r_{5\to6} + \Sigma_{6,7}r_{7\to8} + \Sigma_{6,8}(p_8 - r_{8\to11}) - \Sigma_{6,8}(r_{6\to7} - p_6 + r_{6\to9}) \\
\dot{\Sigma}_{6,9} &= 2\Sigma_{5,9}r_{5\to6} + \Sigma_{6,6}r_{6\to9} - r_{6\to9}\langle x_6\rangle - \Sigma_{6,9}(r_{6\to7} - p_6 + r_{6\to9}) - \Sigma_{6,9}(d_9 + r_{9\to10}) \\
\dot{\Sigma}_{6,10} &= 2\Sigma_{5,10}r_{5\to6} + \Sigma_{6,7}r_{7\to10} + \Sigma_{6,9}r_{9\to10} - \Sigma_{6,10}(r_{6\to7} - p_6 + r_{6\to9}) - \\
&\quad - \Sigma_{6,10}(d_{10} + r_{10\to11}) \\
\dot{\Sigma}_{6,11} &= 2\Sigma_{5,11}r_{5\to6} - \Sigma_{6,11}d_{11} + \Sigma_{6,8}r_{8\to11} + \Sigma_{6,10}r_{10\to11} - \Sigma_{6,11}(r_{6\to7} - p_6 + r_{6\to9}) \\
\dot{\Sigma}_{7,8} &= \Sigma_{6,8}r_{6\to7} + \Sigma_{7,7}r_{7\to8} - r_{7\to8}\langle x_7\rangle + \Sigma_{7,8}(p_8 - r_{8\to11}) - \Sigma_{7,8}(r_{7\to8} - p_7 + r_{7\to10}) \\
\dot{\Sigma}_{7,9} &= \Sigma_{6,7}r_{6\to9} + \Sigma_{6,9}r_{6\to7} - \Sigma_{7,9}(r_{7\to8} - p_7 + r_{7\to10}) - \Sigma_{7,9}(d_9 + r_{9\to10}) \\
\dot{\Sigma}_{7,10} &= \Sigma_{6,10}r_{6\to7} + \Sigma_{7,7}r_{7\to10} + \Sigma_{7,9}r_{9\to10} - r_{7\to10}\langle x_7\rangle - \Sigma_{7,10}(r_{7\to8} - p_7 + r_{7\to10}) - \\
&\quad - \Sigma_{7,10}(d_{10} + r_{10\to11}) \\
\dot{\Sigma}_{7,11} &= \Sigma_{6,11}r_{6\to7} - \Sigma_{7,11}d_{11} + \Sigma_{7,8}r_{8\to11} + \Sigma_{7,10}r_{10\to11} - \Sigma_{7,11}(r_{7\to8} - p_7 + r_{7\to10}) \\
\dot{\Sigma}_{8,9} &= \Sigma_{6,8}r_{6\to9} + \Sigma_{7,9}r_{7\to8} + \Sigma_{8,9}(p_8 - r_{8\to11}) - \Sigma_{8,9}(d_9 + r_{9\to10}) \\
\dot{\Sigma}_{8,10} &= \Sigma_{7,8}r_{7\to10} + \Sigma_{7,10}r_{7\to8} + \Sigma_{8,9}r_{9\to10} + \Sigma_{8,10}(p_8 - r_{8\to11}) - \Sigma_{8,10}(d_{10} + r_{10\to11}) \\
\dot{\Sigma}_{8,11} &= \Sigma_{7,11}r_{7\to8} - \Sigma_{8,11}d_{11} + \Sigma_{8,8}r_{8\to11} + \Sigma_{8,10}r_{10\to11} - r_{8\to11}\langle x_8\rangle + \Sigma_{8,11}(p_8 - r_{8\to11}) \\
\dot{\Sigma}_{9,10} &= \Sigma_{6,10}r_{6\to9} + \Sigma_{7,9}r_{7\to10} + \Sigma_{9,9}r_{9\to10} - r_{9\to10}\langle x_9\rangle - \Sigma_{9,10}(d_9 + r_{9\to10}) - \\
&\quad - \Sigma_{9,10}(d_{10} + r_{10\to11}) \\
\dot{\Sigma}_{9,11} &= \Sigma_{6,11}r_{6\to9} - \Sigma_{9,11}d_{11} + \Sigma_{8,9}r_{8\to11} + \Sigma_{9,10}r_{10\to11} - \Sigma_{9,11}(d_9 + r_{9\to10}) \\
\dot{\Sigma}_{10,11} &= \Sigma_{7,11}r_{7\to10} - \Sigma_{10,11}d_{11} + \Sigma_{8,10}r_{8\to11} + \Sigma_{9,11}r_{9\to10} + \Sigma_{10,10}r_{10\to11} - r_{10\to11}\langle x_{10}\rangle - \\
&\quad - \Sigma_{10,11}(d_{10} + r_{10\to11}).
\end{aligned}
\tag{B.2}
$$

Together with the equations for the first moment (cf. eq. 2.10 in section 2.1.1), this gives a set of 77 coupled ODEs in total, which were numerically solved using MATLAB's standard ODE-solver *ode45*.