Technische
Universität
München

HelmholtzZentrum münchen
German Research Center for Environmental Health

Computational modeling and model selection to reveal cellular mechanisms in tissue homeostasis

Lisa Bast

February, 2022

TECHNISCHE UNIVERSITÄT MÜNCHEN

Fakultät für Mathematik

Computational modeling and model selection to reveal cellular mechanisms in tissue homeostasis

Lisa Bast

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

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende:

Prof. Dr. Christina Kuttler

Prüfer der Dissertation:

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

Die Dissertation wurde am 07.04.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 18.10.2021 angenommen.

Acknowledgements

First I would like to thank Fabian Theis and Carsten Marr for giving me the opportunity to conduct my doctoral research in close collaboration with the Institute of Computational Biology (ICB), thus providing an excellent environment which gave me ideal conditions to scientifically grow as a researcher. It was a great honour to get to know both of you and get inspired by you. Especially I would like to thank you for supervising, advising and supporting me. Particular thanks go to Carsten Marr, who always had an open ear, especially during difficult phases, stood up for me whenever necessary and had proven more than once to be a master motivator. Without your support, I would have had a much harder time to succeed and would have certainly enjoyed the road tremendously less.

I would also like to thank my collaboration partners, especially Filippo Calzolari, Jovica Ninkovic, Michèle Buck, Katharina Götze and Robert Oostendorp for commonly working with me and Carsten Marr on exciting research projects. I enjoyed our meetings in particular because of the pleasant and respectful company and the numerous discussions in which I learned a multitude of things from you.

Additionally, I would like to thank the TUM Graduate School, the HELENA Graduate School, the IRTG of SFB 1243 and the Joachim Herz Foundation for the extensive soft skill course and seminar offer, and financial support throughout my doctorate, which gave me the opportunity to attend conferences, summer schools, workshops and to work abroad for two months. In this regard, I would also like to thank Sui Huang and Joseph Zhou for inviting me for a reasearch stay at the Institute for Systems Biology (ISB) in Seattle. It had been two scientifically inspiring months for me and I enjoyed our discussions and working at the ISB.

Of course I would like to thank Fabian Theis, Michael Flossdorf, Damien Hicks, and Thomas Höfer for reading and evaluating my dissertation and Christina Kuttler for chairing the examination committee.

Many thanks also go to my former colleagues at the Technical University Munich and the ICB for valuable input to my research projects, and for broadening my mathematical and biological knowledge. I enjoyed listening to your presentations at seminars and on retreats a lot. In addition I would like to thank Gero Friesecke, Johannes Petermeier, Norbert Krautenbacher, Sarah Rathbauer and Sebastian Schober for the great teamwork in teaching students. Many thanks also to Michael Strasser for supervising and advising the first 2 years of my PhD work and to everyone who supported me in maintaining a healthy work-life-balance by playing table tennis, volleyball, badminton, human kicker, running, dancing at concerts and alike. I am very happy that many of you became my friends over the years.

Lastly, I would like to thank my family for their unrestrained support not only during the time of my doctorate, but also before. Thank you Mom, Dad and Erik for teaching me all the useful soft skills (like how to organize myself and not give up easily, being curious, determined and eager to learn new things, communicate and solve problems) that were immensely helpful to master this challenge.

Abstract

Most mammalian tissues consist of millions of cells that strongly differ in their phenotype. To fulfil specific functions and collectively guarantee an organ's functionality, every tissue requires balanced cell populations, which is realised by a process called tissue homeostasis. So far it is only partly understood how this balance is maintained throughout life, and which mechanisms cause an imbalance that eventually leads to disease. In recent years, various elaborate experimental studies and sophisticated mathematical models investigating tissue homeostasis have been developed. Modelling approaches were however often not data-driven and thus led to rather qualitative than quantitative results. Furthermore they often focused on a small subset of the respective tissue's cell types and considered only one particular model hypothesis, hence only few provided a comprehensive picture of the tissue homeostasis process.

This thesis aims to gain mechanistic insights into two important adult tissue homoeostasis processes: neurogenesis and hematopoiesis. I thus developed mathematical models that describe homoeostasis as a chemical reaction network (i) stochastically on the single cell level by formulating a Markov jump process, deriving the chemical master equation, and using moment equations as an approximation and (ii) deterministically on the cell population level by using differential equations, which incorporate auxiliary states to realistically model transition times. For both approaches I derived the likelihood function under a suitable noise model assumption to infer unknown model parameters from experimental observations. To identify plausible and implausible division and differentiation mechanisms I derived several competing models of neurogenesis and hematopoiesis, respectively, with varying parameter complexity. For these competing models I analysed structural and practical parameter identifiability, and performed a quantitative comparison by using scoring methods such as the Bayesian and Akaike information criterion.

The developed neurogenesis model revealed that in aged mice, the emptying dormant stem cell pool is compensated by shifting the division mode towards almost exclusively asymmetric stem cell divisions, and by prolonging quiescent stem cell phases. This explains how neurogenesis gradually declines in late adulthood. For understanding human blood cell production I conducted a quantitative comparison between the classical lineage hierarchy model and nine alternative models. Overall, the classical model outperformed all others with respect to its selection criteria scores. Additionally, my approach revealed that hematopoietic stem cell and common myeloid progenitor proliferation decreases throughout life, which explains the previously observed decline in blood production with age. For patients suffering from the hematopoietic disorder myelodysplastic syndrome, I uncovered dysfunctions in proliferation, differentiation and cell death with cell type resolution, and a large heterogeneity between patients.

In conclusion, data-driven modelling and model selection is a powerful tool to reveal homoeostatic mechanisms of tissues, and age-related changes from time-resolved cell count data. The application to case-control data can uncover dysfunctional homeostatic mechanisms and their heterogeneity in diseased individuals.

Zusammenfassung

Die meisten Gewebe von Säugetieren bestehen aus Millionen von Zellen, die sich in ihrem Phänotyp funktionell stark unterscheiden. Um gemeinschaftlich die Funktionsfähigkeit von Organen zu gewährleisten, müssen die verschiedenen Zellpopulationen des Gewebes ausbalanciert sein. Dies wird durch den Prozess der sogenannten Gewebehomöostase realisiert. Bisher ist nur in Ansätzen bekannt, wie dieses Gleichgewicht über das gesamte Leben aufrecht erhalten wird und welche Mechanismen es im Krankheitsfall stören. In den letzten Jahren wurden diverse wohldurchdachte Experimente durchgeführt und ausgeklügelte mathematische Modelle hergeleitet, welche die Gewebehomöostase untersuchen. Solche Modellierungsansätze waren jedoch oft nicht datengetrieben und führten zu vornehmlich deskriptiven und weniger quantitativen Resultaten. Da sie zudem meist nur eine kleine Teilmenge der Zelltypen und eine bestimmte Modellhypothese in Betracht gezogen haben, lieferten nur wenige ein umfassendes Bild des homöostatischen Prozesses.

Diese Dissertation verfolgt das Ziel, Einblicke in die Mechanismen zweier bedeutender Gewebehomöostaseprozesse, der Neurogenese und Hämatopoese, zu gewinnen. Dazu habe ich mathematische Modelle entwickelt, die Homöostase als Reaktionsnetzwerk beschreiben und zwar (i) stochastisch auf Einzelzellebene durch Herleitung eines Markov-Sprungprozesses, Formulierung der chemischen Mastergleichung und Approximation mit Momentengleichungen und (ii) deterministisch auf Zellpopulationsebene mit Hilfe von Differentialgleichungssystemen, die Behelfs-Kompartimente beinhalten um Übergangszeiten realistisch zu modellieren. Für beide Modellierungsansätze leitete ich die Likelihood-Funktion unter der Annahme eines geeigneten Fehlermodells her. Um plausible Zellteilungs- und Differenzierungsmechanismen zu inferieren, stellte ich verschiedene Modelle der Neurogenese bzw. der Hämatopoese mit variierender Komplexität auf. Für diese analysierte ich strukturelle und praktische Parameteridentifizierbarkeiten und führte einen quantitativen Vergleich mittels dem Akaiken und dem Bayesianischen Informationskriterium durch.

Die hergeleiteten Neurogenesemodelle deckten auf, dass bei älteren Mäusen der fast entleerte, ruhende Stammzellpool durch eine Verschiebung des Teilungsmodusses zu fast ausschließlich asymmetrischen Stammzellteilungen und durch eine Verlängerung der Ruhephasen kompensiert wird. Diese altersbedingten Veränderungen entschlüsseln mechanistisch, wie Neurogenese im späten Erwachsenenalter stetig abnimmt. Um menschliche Blutzellproduktion zu verstehen, führte ich einen quantitativen Vergleich zwischen dem klassischen Abstammungshierarchie-Modell und neun Alternativmodellen durch. Insgesamt übertraf das klassische Modell alle alternativen Modelle hinsichtlich der Werte zweier Informationskriterien. Darüber hinaus zeigte mein Ansatz, dass die Proliferationsaktivität der Stammzellen und der gemeinsamen myeloischen Vorläuferzellen mit dem Alter abnimmt. Dies erklärt den zuvor experimentell beobachteten altersbedingten Rückgang der Blutzellproduktion. Bei Patienten, die unter der hämatopoetischen Erkrankung Myelodysplastisches Syndrom leiden, deckte ich Fehlfunktionen im Proliferations-, Differenzierungs- und Zelltodverhalten der verschiedenen Zelltypen und eine diesbezüglich große Heterogenität zwischen den Patienten auf.

Der datengetriebene Modellierungs- und Modellvergleichsansatz erweist sich zusammenfassend als ein effektives Werkzeug, um Mechanismen der Gewebehomöostase sowie altersbedingte Veränderungen anhand von experimentellen zeitaufgelösten Zell-Anzahl-Daten aufzudecken. Die Anwendung auf Fall-Kontroll-Datensätze ermöglicht es Fehlfunktionen bei diesbezüglichen Erkrankungen und Unterschiede zwischen Patienten aufzudecken.

List of contributed articles

(i) L. Bast*, F. Calzolari*, M. K. Strasser, J. Hasenauer, F. J. Theis, J. Ninkovic^{+}, C. Marr^{+}. **Increasing Neural Stem Cell Division Asymmetry and Quiescence Are Predicted to Contribute to the Age Related Decline in Neurogenesis.** Cell Reports 25, 3231-3240 (2018).

My contribution was the analysis of data (together with F. Calzolari), derivation and implementation of a set of mathematical models, and the performance of parameter inference, model comparison, model averaging, and model simulation. I wrote the manuscript together with F. Calzolari, J. Ninkovic, and C. Marr.

(ii) L. Bast*, M. C. Buck*, J. S. Hecker, R. A. J. Oostendorp, K. S. Götze^{+}, C. Marr^{+}. **Computational modeling of stem and progenitor cell kinetics identifies plausible hematopoietic lineage hierarchies.** iScience 24, 102120 (2021).

My contribution was the derivation and implementation of mathematical models, the performance of parameter inference, model selection and in silico analysis. I analysed the results and wrote the manuscript together with M. C. Buck, K. S. Götze, and C. Marr.

(iii) M. C. Buck*, L. Bast*, J. S. Hecker, M. Rothenberg-Thureley, I. Andrä, I. Struzina, D. Wang, F. J. Theis, F. Bassermann, K. H. Metzeler, R. A. J. Oostendorp, C. Marr^{+}, K. S. Götze^{+}. **Computational Modeling of Stem and Progenitor Cell Kinetics Reveals Individual Changes in Clonal Hematopoiesis and Myelodysplastic Syndromes.** (drafted, not yet submitted)

My contribution was the analysis of experimental data (together with M. C. Buck), the derivation and implementation of mathematical models, the performance of parameter inference, and the analysis of modeling results. I wrote the manuscript together with M. C. Buck, C. Marr, and K.S. Götze.

*These authors contributed equally.
$^{+}$These authors are co-corresponding authors.

I am the shared first author of the articles listed above.

Contents

1 Introduction

Every living organism arises from a single cell. In complex, multicellular organisms such as humans, this cell is an embryonic stem cell [Chakraborty and Agoramoorthy, 2012]. Stem cells differ from other cells as they are able to divide and renewing themselves, and they are unspecialised but can give rise to more specialized, mature and functional cells of the various tissues of an organism, i.e. neurons, blood cells, or skin cells. These properties describe processes which are biologically termed cell proliferation and cell differentiation and they are necessary to form tissues and whole organisms during development, and to maintain a balanced cell population size within tissues during adulthood [Hui et al., 2011].

To gain insights about tissue forming and maintaining processes, but also to better understand diseases related to disturbed tissue formation or maintenance, biologists nowadays experimentally observe cells in a living organism *in vivo* or in a defined culture media *in vitro* and classify them according to their degree and type of specialization. The insights these experiments can reveal are however limited as not everything can be measured or observed and resources such as time, work forces and experimental costs are limited.

The research field of systems biology aims to overcome these obstacles by applying mathematical modelling to gain mechanistic insights about biological systems. The biological system of interest consists of several components, which can be molecules, cells, organs, organisms, or entire species. The mathematical model, which serves an abstraction of the respective biological process, ideally describes the dynamical behaviour of the system by capturing the most important underlying mechanisms to a certain degree. To construct such a model or a set of plausible models, findings from literature are employed to define the system components, or states of the components and possible transitions between them. By comparing the observations of the mechanistic model(s) to the respective experimentally measured observations, the model(s) can further be specified. Unknown model parameters can be determined and alternative models can be compared to each other in order to reject implausible ones. The identified model can then be used to observe the systems behaviour that cannot be investigated experimentally or at least not with the available resources. It thereby serves as a tool to test several hypotheses and to make predictions.

Systems biology is thereby exploiting the information observed in an experiment, which often brings light into the dark for many fundamental research questions. In recent years, systems biology has majorly contributed to understand the bigger picture in healthy and diseased individuals and is in general a promising approach to further decode the structural and functional organization of cells.

In my thesis, I use a systems biology approach to mechanistically understand cell division and differentiation processes and to decode differences that arise during ageing or in diseased organisms. In section 1.1, I will introduce the main principles and achievements of cell biology. Section 1.2 gives a short introduction about mechanistic mathematical models and sections 1.3 and 1.4 provide an overview of the research questions I address in my thesis and an outline of my dissertation.

1.1 Cell biology

For the past two centuries, many researchers have dedicated their work to understand how the body creates itself out of a single cell and the mechanisms by which a specific tissue renews itself throughout life. The progress in the field of molecular biology yielded many experimental techniques and devices to study and analyse cells. Nowadays it is not doubted that every organism consists of different organs which are itself assembled by cells and that these cells are specialized according to their function. This section gives an overview of the main principles and achievements in cell biology.

The smallest unit of a living thing was named 'cell' by Robert Hooke in 1665, who observed them under a microscope, but it took more than a century until the importance of cells was realized. In the 1830s, Matthias Schleiden and Theodor Schwann postulated that all plants and animals consist of one or more cells and that cells are the structural unit of life. Another tenet of cell theory was introduced by Rudolf Virchow two decades later, in 1855. It states that cells can arise only by division from a pre-existing cell [Karp, 2004]. In the following century, it has been revealed that every organism inherits a construction plan from its parents which is termed desoxyribose nucleic acid (DNA) and determines an organisms genotype [Alberts et al., 1989]. The DNA is stored in the nuclei of all its cells in eukaryotes and freely available in prokaryotes, which do not obtain a nucleus [Karp, 2004]. It is composed of many nucleotide sequences, which are termed genes and which can be accessed independently by the cell. To run a specific genetic program, a subset of genes can be expressed by first copying the respective sequence, which is termed RNA transcript, and then translating it into proteins which can then fulfil a specific function. If a gene is highly expressed it is said to be up-regulated and if not it is said to be down-regulated [Alberts et al., 1989]. Although all cells of an organism share the same genotype, every cell runs its own program to fulfil specific functions. As a result, some cells obtain different observable physical properties, which can be a different morphology such as shape or size, and is referred to the term phenotype. In order to better understand and analyse cell systems, biologists sought to classify cells into distinct cell types for more than 150 years [Regev et al., 2017]. These cell types can be defined based on a cells phenotype, its location within the tissue of the respective organism, its relationship to other cells, its functions and its molecular properties such as the expression of certain genes, which are in this context termed marker genes. Irrespective of the studied tissue, the respective cells can be classified as stem cells, progenitors or mature cells (see concept of tissue organisation in Figure 1.1).

In case a cell exhibits the property of self-renewal and if it in addition is able to produce any mature cell type of the respective tissue, it is called a stem cell. The process a cell undergoes when transitioning from a rather unspecialised cell type to a more specialized cell type is termed cell differentiation. Progenitors describe cells which are also able to self-renew but are neither fully differentiated nor undifferentiated. These fully differentiated cells are termed mature cells and are in general not able to self-renew.

According to its differentiation possibilities, every cell has a certain potency [Singh et al., 2016]. A cell is said to be uni-potent if it has exactly one and multi-potent if it has several possibilities to transit to a more specialized cell type. Stem cells have the highest potency and exhibit often but not always a multi-potent cell identity which means they can give rise to any cell type of the respective lineage, either directly or indirectly via progenitor cells. In comparison to stem cells, progenitors obtain a lower potency but can be both, uni-potent or multi-potent depending on the considered tissue. The various tissues in an organism all result from totipotent embryonic stem cells, which obtain the highest cell potency and can give rise to any cell type of the organism [Hima Bindu and Srilatha, 2011].

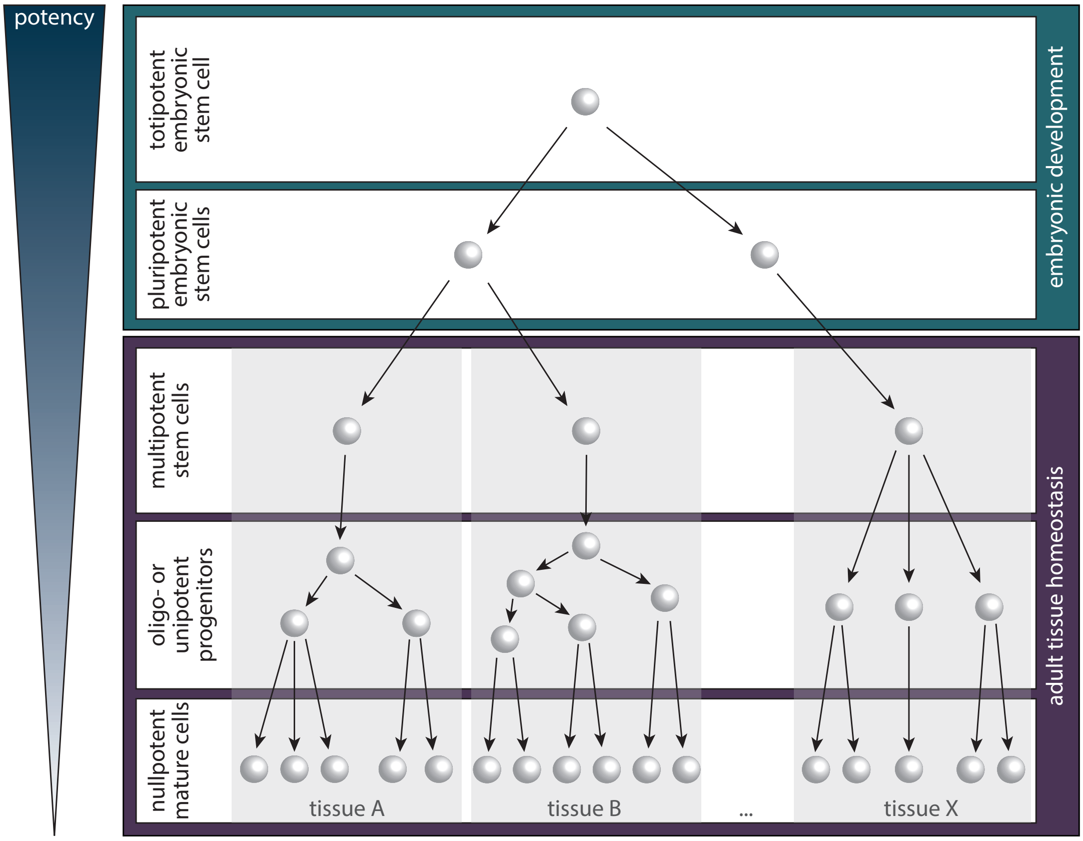


Figure 1.1: Tissue organization conceptually described by five distinct cell layers with decreasing potency. Adult tissue homeostasis is described by three distinct cell layers, which are multipotent stem cells, oligo- or unipotent progenitors and nullpotent mature cells.

A structured overview of the potency of tissue specific cell types is provided by the developmental history of differentiated cells which is called cell lineage or hierarchy. Cell lineages can change during development according to the change in cellular function [Stent, 1985]. While in the embryonic phase the tissue is developed, in adults the purpose shifts towards tissue homeostasis and repair [Karp, 2004, Passier, 2003]. Tissue homeostasis describes the process of maintaining a balanced cell population within the tissue, which is achieved by cells via balancing proliferation, differentiation, and cell death. Upon injury or in diseased tissues, this balance is disturbed. Especially in age-related diseases, cells of the respective tissue often acquired mutations, i.e. modifications in the DNA of cells, which give them a fitness advantage towards other competing cells in the tissue. This process is also referred to clonal dominance and thought to be the cause of an imbalanced homeostasis [Lee and Abdel-Wahab, 2014] (see Figure 1.2).

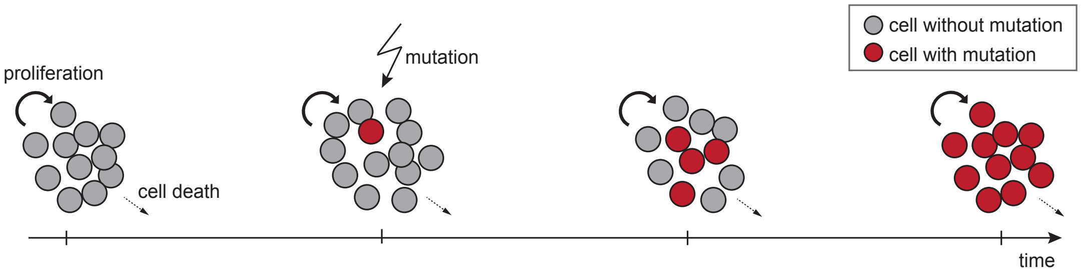


Figure 1.2: Scheme illustrating how a cell population of initially non-mutated (grey) cells (left) evolves over time, while cells are underlying proliferation and cell death and being subject to eventual mutations. Due to a fitness advantage of a single or a few mutated (red) cells (middle) clonal dominance can be achieved (right).

While modelling homeostasis of healthy and diseased tissues, proliferation, differentiation, and cell death are taken into account as these three processes influence the number of cells of the various cell types of the respective tissue. Pre-existing knowledge of the cell lineages can be incorporated as an assumption into the model. By assuming cell differentiation is coupled to cell division, one can distinguish between symmetric and asymmetric cell division [Blanpain and Simons, 2013, Greulich and Simons, 2016, Nowak et al., 2003, Watson et al., 2015, Yang et al., 2015]. If upon division, a stem or progenitor cell produces two daughter cells of the same cell type, a symmetric division takes place and if the daughter cells are of different cell types an asymmetric division occurs (see Figure 1.3).

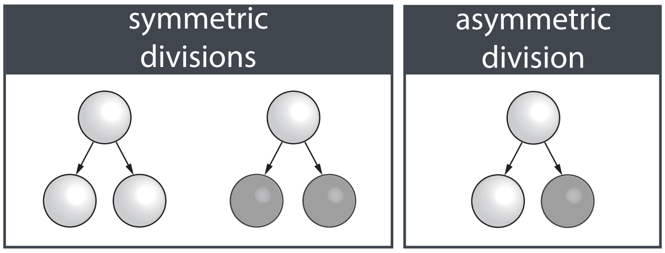


Figure 1.3: Under the assumption that a cell can only differentiate upon division, a mother cell (white) can divide according to one of the three shown division modes. It can divide symmetrically (left) into two daughter cells of the same cell type (white), or another cell type (grey), or alternatively divide asymmetrically (right) into a daughter cell of the same cell type (white) and one of another cell type (grey).

For some cell systems this assumption is not valid, i.e. for the differentiation of oligodendrocyte progenitor cells to oligodendrocytes, which are located in white and grey matter tissues in the central nervous system. For oligodendrocytes, differentiation and migration, or cell death has to occur before cells in the environment are able to fill the empty space with their progeny upon proliferation [Hughes et al., 2013].

While studying cell division and differentiation processes one aims to identify differences between cell types, individuals, or cohorts. The respective experimental data is always subject to technical noise but also contains biological variability on the single cell level. This cellular heterogeneity is present even if cells are genetically identical [Altschuler and Wu, 2010], as it is induced by interactions between cells or with the environment [Sutherland, 1988, Wills et al., 2017]. These molecular interactions lead to regulatory processes, i.e. up- or down-regulation of genes, which introduce transcriptomic differences between single cells and result in different phenotypes [Komin and Skupin, 2017]. In addition, cellular heterogeneity can also result from genetic modifications, i.e. accumulated mutations in diseased cells such as cancer stem cells. Dependent on the number and composition of mutations present within each stem cell, there exists a huge variability between these malignant clones originating from different stem cells [Altschuler and Wu, 2010].

For many research questions, the goal is to detect the biological variability between single cells or cell populations in order to gain mechanistic insights [Altschuler and Wu, 2010]. If one observes a population of cells on the single cell level, one can identify and measure the morphology, function or behaviour of single cells, or of several sub-populations which are each a relatively homogeneous group of cells, that obtain similarities. If one however studies a cell population in a bulk experiment, this single-cell behaviour is not observable as these experiments only represent the average behaviour of the cell population. While modelling cell populations that have been observed in bulk, one might thus not be able to model the biological process very detailed but instead only describe the mean population, or sub-population behaviour. One the one hand, the challenge in experimental design is to make the biological variability of interest detectable, i.e. design the experiment such that the biological signal is not less prominent in the data than technical noise resulting from measurement errors. On the other hand, the challenge in mathematical modelling is then to capture the most important details of the underlying biological process and to describe the technical noise resulting from the experiment realistically in order to detect the signal of interest, namely the biological variability between cells or groups of cells.

1.2 Mechanistic mathematical models

Biological processes are in general complex and it is often difficult to understand the global behaviour from experimental measurements that were observed locally from some of its parts. In addition, solving specific research questions is often challenging because not everything can be measured technically or accurately as experimental techniques are in practice limited. Testing each and every hypothesis about the system of interest experimentally is in most cases intractable as it would be expensive and very time consuming to perform each experiment.

Following the philosophy of Galileo Galilei, mechanistic mathematical models aim to make measurable what cannot be measured directly by experimentalists [Hasenauer, 2020]. They can be used as an abstract representation of a specific biological process in order to answer specific questions [Klipp, 2010]. The model contains the system's states, which are in the scope of my dissertation cell types or cell states, such as active and inactive regarding proliferation, and all possible transitions between the states. Model transitions represent the rate at which the system jumps from one state to another and are stated in cells per time unit. The states and transitions together define the model structure. If the model structure is unknown, one can derive a set of plausible models with different structures. Mechanistic models thereby incorporate established knowledge about the respective system of interest, but also allow to specify unknown model properties such as structure and parametrization of transition rates. The unknown model parameters can in principle be inferred by comparing the model to functions of experimental observations of some system states, but this highly depends on the model's structure and which states are measured and at which time points. If for instance measurements are uninformative or data quality is poor, parameter inference can be problematic. This can be studied by analysing structural and practical parameter identifiability of a model or a set of models (see 2.2.3). However, to identify the parameters for which model and experimental data agree best, a cost function needs to be specified and optimized. Also optimization can be challenging and highly depends on the optimization problems structure. If the optimization problem is non-convex, it is often only possible to solve it locally instead of globally and thus requires a multi-start procedure (see section 2.2.1.2). In case optimization is successful, the resulting parameter estimates will give insight about how fast the transitions occur, i.e. relating to cell differentiation processes how fast cells divide, differentiate, die, and get inactivated or activated or which cell division mode, i.e. symmetric or asymmetric division occurs most likely for a certain cell type. In order to quantitatively determine plausible model structures one can perform model selection or averaging (see section 2.3). Beyond answering biological questions, the resulting model can be used for simulations and serve to make predictions (see section 2.1.7). Therefore it first needs to be validated, ideally based on independent data, to ensure that it faithfully predicts the systems behaviour [Klipp, 2010] (see Figure 1.4).

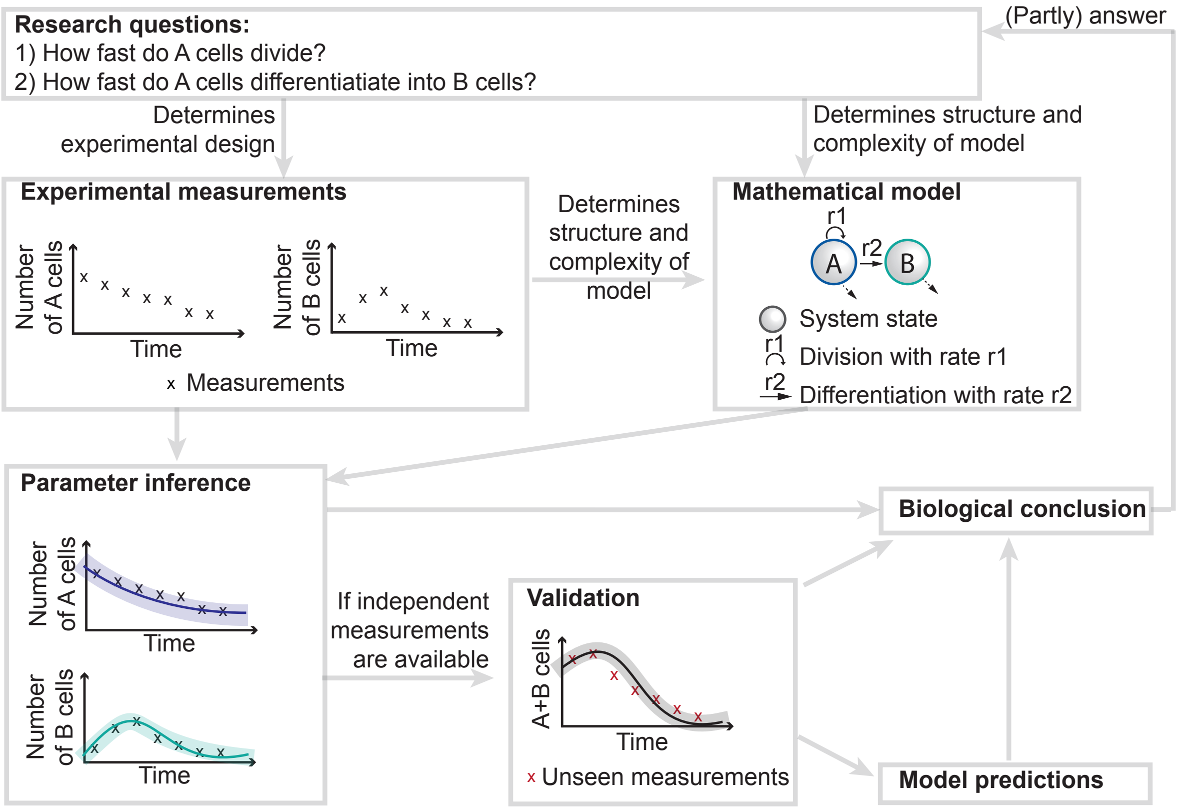

Research questions:
1) How fast do A cells divide?
2) How fast do A cells differentiatiate into B cells?
(Partly) answer
Determines experimental design
Determines structure and complexity of model
Experimental measurements
Number of A cells
Time
Number of B cells
Time
x Measurements
Determines structure and complexity of model
Mathematical model
r1
A
r2
B
System state
r1 Division with rate r1
r2 Differentiation with rate r2
Parameter inference
Number of A cells
Time
Number of B cells
Time
Biological conclusion
If independent measurements are available
Validation
A+B cells
Time
x Unseen measurements
Model predictions

Figure 1.4: Overview of a systems biology approach to solve a division and differentiation process related research question by inferring model dynamics from experimental data.

In the context of healthy and diseased tissue homeostasis, mathematical models have already proven to be a useful tool to investigate underlying mechanisms of intrinsic cell fate regulation [Greulich and Simons, 2016, Ritsma et al., 2014] or tumour growth [Lan et al., 2017, Roeder et al., 2006] for various kinds of experimental data, resulting for instance from time-lapse microscopy, fluorescence activated cell sorting (FACS), single-cell RNAseq or clonal lineage tracing analysis. Depending on the type of experimental data, different modelling approaches, such as chemical reaction networks (CRNs) [Watson et al., 2015], stochastic processes [Nowak et al., 2003, Roeder et al., 2006], gene regulatory networks (GRNs) [Herberg and Roeder, 2015, Kalmar et al., 2009], pseudo-temporal ordering [Trapnell et al., 2014] or factor graph models [Niederberger et al., 2015], were used to describe healthy and malignant homeostasis.

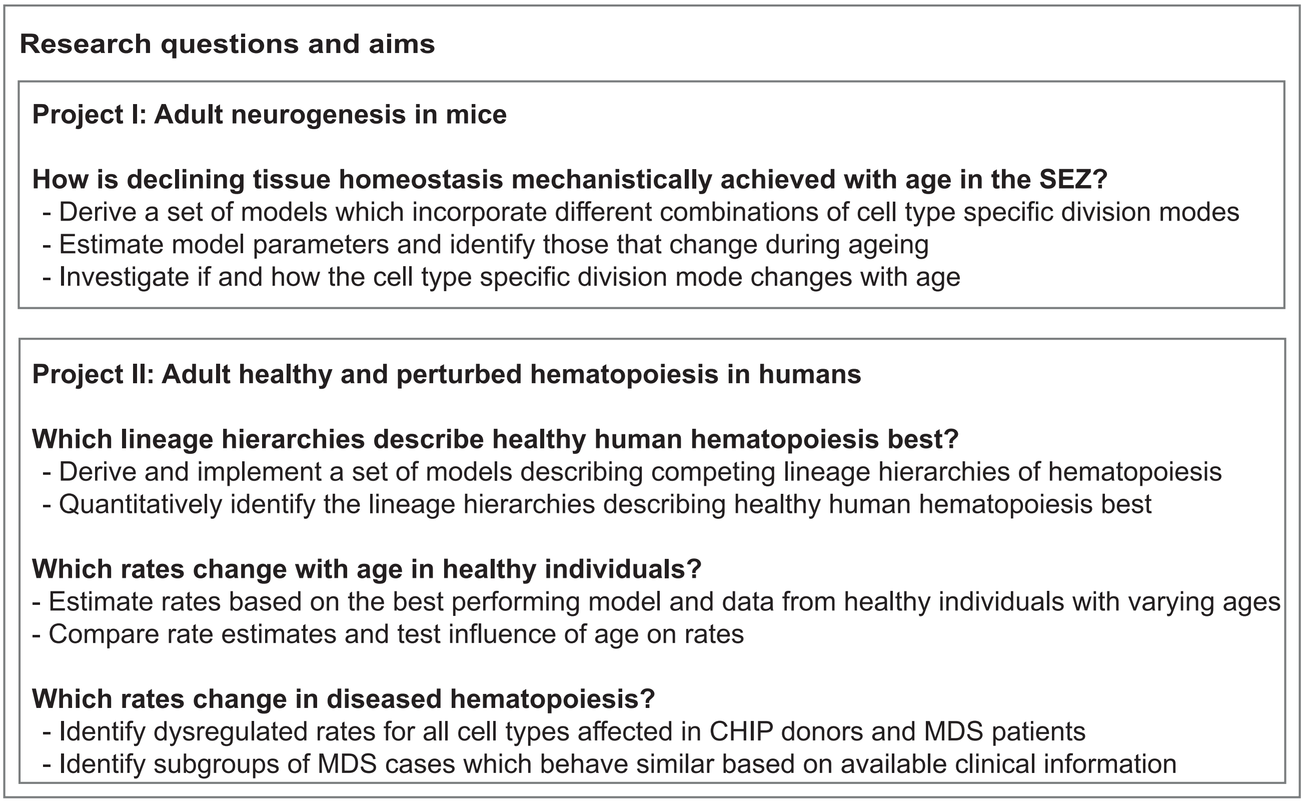
As model assignment is not unique, but instead always depends on the problem, the purpose, and the intention of the investigator [Klipp, 2010], existing models can only answer some specific questions and are tailored to a specific tissue and region. Often these models consider only some cell types of the analysed lineage but do not provide a comprehensive picture of the homeostatic process of the respective lineage.

In this thesis, I develop a computational modelling and model selection approach which allows for the comprehensive study of tissues on the macroscopic cellular level as it can be tailored and applied to an arbitrary number of cell types of any tissue of interest from which single cell or bulk (several cells) count measurements were derived. In case of single cell measurements, the process is described by a stochastic model (see section 2.1.2), whereas in case of bulk measurement deterministic models (see section 2.1.3) are the mathematical approach of choice. As two applications I will derive models to describe adult neurogenesis and hematopoiesis to solve the biologically driven research questions introduced in section 1.3.

1.3 Research questions

Cell division and differentiation processes can be studied on various biological scales and for various tissues and regions. In the scope of this dissertation an approach for the data-driven modeling and model selection for cell division and differentiation processes on the macroscopic level was developed. This mathematical approach was applied to two cell systems, namely adult neurogenesis and adult healthy and malignant hematopoiesis.

While studying adult neurogenesis in mice, I focused on a particular brain region, called the subependymal zone (SEZ). About this region it is known that less neurons are produced in adult mice, i.e. neurogenesis is declining with age. The main goal is to answer the question: how is this declining tissue homeostasis mechanistically achieved with age in the SEZ (see Figure 1.5)? To reach this goal I want to derive a set of models considering all possible combinations of cell type specific division modes. The first subaim is to identify plausible and implausible models to identify which division mode is used when stem- and progenitor cells divide. In addition, I want to estimate all model parameters that change during ageing from experimental quasi time-resolved cell count data and to investigate if the division mode (see Figure 1.3) changes with age for the proliferating cell types. The experimental data used for parameter estimation stem from a collaboration with experimentalists who repeatedly measured the progeny of single neural stem cells *in vivo* in the SEZ of young and aged adult mice.



Research questions and aims

Project I: Adult neurogenesis in mice

How is declining tissue homeostasis mechanistically achieved with age in the SEZ?

- Derive a set of models which incorporate different combinations of cell type specific division modes
- Estimate model parameters and identify those that change during ageing
- Investigate if and how the cell type specific division mode changes with age

Project II: Adult healthy and perturbed hematopoiesis in humans

Which lineage hierarchies describe healthy human hematopoiesis best?

- Derive and implement a set of models describing competing lineage hierarchies of hematopoiesis
- Quantitatively identify the lineage hierarchies describing healthy human hematopoiesis best

Which rates change with age in healthy individuals?

- Estimate rates based on the best performing model and data from healthy individuals with varying ages
- Compare rate estimates and test influence of age on rates

Which rates change in diseased hematopoiesis?

- Identify dysregulated rates for all cell types affected in CHIP donors and MDS patients
- Identify subgroups of MDS cases which behave similar based on available clinical information

Figure 1.5: Overview of projects, research questions and aims of this dissertation.

In the adult hematopoiesis project I addressed several research questions with my modeling and parameter inference approach (see Figure 1.5). Over the past years, the hematopoietic lineage has often been debated and several alternative direct differentiation transitions for mice and a few for human have been suggested. As there is no established ground truth, the first question is: which lineage hierarchies describe healthy human hematpoiesis best? Thus my goal is to derive and implement a set of models describing competing lineage hierarchies as highly resolved as possible. Subsequently, I want to quantitatively identify the lineage hierarchies, that describe healthy human hematopoiesis best and the ones that can be rejected based on experimental time-resolved cell count data. The experimental data used in this project stem from a collaboration with clinicians who measured the progeny of thousands of human bone marrow cells of healthy donors *in vitro* (i.e. in a bulk cell culture) repeatedly. The best performing model is then used to answer the question: which proliferation, differentiation, and cell death rates change with age in healthy individuals? Finally, I want to study perturbed hematopoiesis and analyse experimental cell count data of cultured bone marrow cell samples of donors with a clonal hematopoiesis of indeteminate potential (CHIP) and of patients suffering from myelodysplastic syndromes (MDS). MDS is an age-related hematopoietic disorder, which is characterized by ineffective hematopoiesis and peripheral cytopenia (i.e. lack of mature blood cells in the peripheral blood) and can lead to acute myeloid leukaemia. Mutations in hematopoietic stem cells are thought to be causative events in MDS but are also found in aged individuals without evidence of an hematological disease, a clinical entity known as CHIP. I want to solve the question: which rates are disturbed in CHIP donors and MDS patients compared to healthy age-matched individuals? I want to identify if stem or progenitor cells of the bone marrow or both are affected to investigate how clonal dominance (see Figure 1.2) could be achieved. Therefore, I want to uncover which cell types and which cell differentiation, proliferation and death processes are changed in comparison to healthy for every CHIP/ MDS case individually. Subsequently, I wanted to identify subgroups of MDS cases which behave similar based on available clinical information.

1.4 Overview of this thesis

In my dissertation I want to solve the research questions listed in section 1.3. My goal is to derive mathematical models which describe how cells of a specific tissue are organized and maintained throughout life, and which also allow to analyse how they are changed during ageing or upon perturbation, e.g. in diseased individuals.

In this thesis I provide an approach in which the respective cell division and differentiation process can be modelled with stochastic or deterministic biochemical reaction network models (see section 2.1), thereby allowing to describe time resolved single cell and bulk cell count measurements. In both modelling approaches, I will consider cell types or cell states as the systems components and introduce parameters to describe the transition between them with rates. By fitting sets of derived models to time-resolved cell count measurements stemming from an experiment, I want to gain mechanistic and structural insights about the underlying homeostatic process of the respective tissue. I will introduce existing parameter estimation techniques, discuss the pitfalls of parameter inference (see section 2.2) in this context and introduce and apply state of the art model uncertainty and identifiability analysis techniques (see section 2.2.3), which are important to investigate the (non-)uniqueness, precision and accuracy of resulting parameter estimates and the consequential uncertainty of model simulations. Moreover, I will introduce existing strategies to perform model selection and model averaging (see section 2.3) and explain how these can be used to assess the probability of several biologically plausible mathematical models and how these strategies can be used to either rank by or combine competing models.

As specific applications, I studied adult murine neurogenesis (see chapter 3) and adult human healthy and perturbed hematopoiesis (see chapter 4). For these two tissue homeostasis processes I derived and implemented mathematical models, performed parameter inference, identifiability analysis, model selection, in silico analysis, and the analysis of data and modelling results. These applications thereby answer the research questions introduced in section 1.3 and resulted in the publications listed in the beginning of my dissertation. The results from chapters 3 and 4 will be discussed and possible future directions will be outlined at the end of my dissertation (see chapter 5).

2 Methods

The goal of this thesis is to design models which are detailed enough to describe the available data, but are also a simplified representation of the underlying process and thus provide a framework to draw general conclusions. The model specification highly depends on the choice of the modelling approach, which should be in line with the experimental set-up used to observe the data. In addition, each model should ideally be designed in such a way that the parameters are identifiable and parameter results should not only contain the optimal values but also the estimates uncertainty resulting from the optimization analysis. Because prior knowledge about the process is often limited and over-fitting should be prevented, it can often be useful to define a set of possible models with varying complexity, instead of considering only one, and quantitatively compare them.
In this chapter, I introduce both, stochastic (see section 2.1.2) and deterministic (see section 2.1.3) mechanistic modelling approaches and the algorithms which can be used for model simulations (see section 2.1.7). The main goal of this thesis is to answer specific biological questions that can be formulated as parameter estimation problems and partly require the comparison of different models. Hence, parameter inference methods (see section 2.2), parameter identifiability and model uncertainty (see section 2.2.3), and model selection and averaging approaches (see section 2.3) are introduced in this chapter.

2.1 Model specification and simulation

To describe cell state transitions in continuous time and for discrete (in case of cell counts) or continuous (in case of cell concentrations) response, compartmental models [Burnham and Anderson, 2003] are a reasonable choice. One distinguishes between stochastic compartmental models (see section 2.1.2), which consider stochastic influences on single cells, and deterministic compartmental models (see section 2.1.3), which describe the mean behaviour of a cell population.
In order to describe a system's behaviour, one first needs to define the system's states, which represent a snapshot of the system at a given time by a set of state variables [Klipp, 2010] e.g. cell types or cell states of a particular lineage. Subsequently, all possible transitions between the states are defined by constants or functions which can be either parametrized or fixed according to prior knowledge. This leads to the following definition of a mechanistic model:

Definition 2.1. *Mechanistic models describe the evolution of one or several state variables* $\boldsymbol{x} = (x_1, x_2, ..., x_{n_s})$ *over time* $\boldsymbol{t} = (t_1, ..., t_{n_t})$*, which depend on unknown parameters* $\boldsymbol{\theta} = (\theta_1, ...\theta_{n_\theta}) \in \mathcal{P}$ *and known constants* $\boldsymbol{k} = (k_1, ..., k_{n_k}) \in \mathcal{C}$*. A particular mechanistic model* $\mathcal{M}(\boldsymbol{\theta})$ *consists of model dynamics* $\dot{\boldsymbol{x}} = f(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{k})$ *and model observables* $y^{\mathcal{M}} = h(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{k})$*, which were experimentally measured.*

Example 2.2. *Let us consider a simple compartmental model describing four possible transitions of two cell types* $x_1 = [A]$ *and* $x_2 = [B]$*:*

(i) the proliferation of cell type A,

(ii) the differentiation of cell type A to B,

(iii) the death of cell type A,

(iv) the death of cell type B,

which is schematically depicted in Figure 2.1.

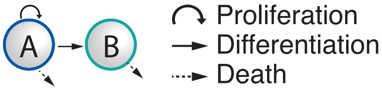


Figure 2.1: Model scheme describing transitions of a simple two cell-type model.

This particular model $\mathcal{M}(\boldsymbol{\theta})$ is defined by the observables, e.g. the measured cell abundances of cell type A and/or B and the dynamics, e.g. the evolution of cell abundances $x_1(t) = [A]$ and $x_2(t) = [B]$ over time.

The dynamics should be described stochastically or deterministically, dependent on the number of cells involved.

2.1.1 Chemical reaction networks

We will consider the system of interest as chemical reaction network (CRN), which allows us to formulate all possible transitions as reactions in terms of mass-action kinetics.

Definition 2.3. *A CRN can be defined as a triple $(\mathcal{S}, \mathcal{C}, \mathcal{R})$, where $\mathcal{S}$ is the set of chemical species, $\mathcal{C} \subseteq \mathbb{R}_+^{\mathcal{S}}$ is the set of complexes and $\mathcal{R}$ is the set of reactions in the network. The set of reactions describes all possible transitions and contain as elements relations on $\mathcal{C}$ denoted by $c \rightarrow c'$ that must satisfy the following three conditions:*

(i) $\mathcal{C}$ cannot contain elements of the form $c \rightarrow c$,

(ii) for any $c \in \mathcal{C}$, there exists some $c' \in \mathcal{C}$ such that either $c \rightarrow c'$ or $c' \rightarrow c$ holds,

(iii) the union of the supports of all $c \in \mathcal{C}$ is $\mathcal{S}$,

where $\mathbb{R}_+^{\mathcal{I}}$ denotes the set of formal sums $s = \sum_{i \in \mathcal{I}} s_i \cdot i$ in which all s_i are strictly positive and $\mathcal{I}$ is an arbitrary finite set and the support of an element $s \in \mathbb{R}^{\mathcal{I}}$ is defined by $supp(s) = \{i \in \mathcal{I} : s_i \neq 0\}$ [Craciun and Pantea, 2008].

Example 2.4. *In the simple two cell type example 2.1, species are cell states ($\mathcal{S} = \{A, B\}$), complexes are $\mathcal{C} = \{A, B, 2A, \emptyset\}$, reaction rate constants are $k_r = (\alpha, \beta, \gamma_A, \gamma_B)$, and the reactions*

$\mathcal{R}$ are defined by

$$\begin{aligned} R_1 :&\quad \mathrm{A} \xrightarrow{\alpha} \mathrm{B} \textit{ (differentiation)},\\ R_2 :&\quad \mathrm{A} \xrightarrow{\beta} 2\,\mathrm{A} \textit{ (proliferation)},\\ R_3 :&\quad \mathrm{A} \xrightarrow{\gamma_A} \emptyset \textit{ (cell death)},\\ R_4 :&\quad \mathrm{B} \xrightarrow{\gamma_B} \emptyset \textit{ (cell death)}, \end{aligned}$$

where $\emptyset$ is the empty set.

2.1.2 Stochastic compartmental models

Cellular dynamics are often driven by external and internal mechanisms, which are either not explicitly known and cannot be captured by the model, or would in most cases be too detailed and beyond the scope of the analysed system and are therefore on purpose not modelled [Bachar et al., 2013]. However, neglecting these influences while analysing bulk data is in most cases reasonable as they will disappear on average but for the analysis of single cell data they may drastically affect the analysis, as every cell runs its own program [Klipp, 2010].
To model cell division and differentiation processes, one can consider all possible transitions (proliferation, differentiation or cell death) as individual random events and model the probabilistic evolution of the state variable, i.e. the cell abundance of a particular state, over time as a stochastic process.

Definition 2.5. *A stochastic process is a sequence of random variables $\{X_i(t_i)\}, i = 0, 1, ..., n$. Each of these random variables takes values from the same state space $\mathcal{X}$ and the system is completely described by a set of joint probability density $P(X_0(t_0), X_1(t_1), ..., X_n(t_n))$ [Bachar et al., 2013, Gardiner, 2009].*

One distinguishes between discrete-time and continuous-time and between continuous and discrete state space stochastic processes [Gardiner, 2009, Klipp, 2010]. Moreover, one can assume the random variables to be completely independent such that the joint probability density is equal to the product of the individual probability densities of the random variables:

$$P(X_0(t_0), ..., X_n(t_n)) = \prod_{i=0}^{n} P(X_i(t_i)), \tag{2.1}$$

or alternatively assume some sort of dependency between the random variables as it is for instance assumed for the Markov process [Gardiner, 2009, Klipp, 2010].

2.1.2.1 Markov jump process

The Markov process assumes that the present system's state determines its future random behaviour [Gardiner, 2009, Klipp, 2010]. The Markov assumption is formulated in terms of the conditional probabilities

$$P(X_{n+1}(t_{n+1})|X_0(t_0), ..., X_n(t_n)) = P(X_{n+1}(t_{n+1})|X_n(t_n)). \tag{2.2}$$

Thus, transition probabilities between the present and the future state do not depend on states in the past [Klipp, 2010].

A continuous-time Markov process can be used to describe the biochemical random processes of a reaction network $(\mathcal{S}, \mathcal{C}, \mathcal{R})$. Within a short time interval $[t, t + \Delta t]$, each possible event r occurs with probability $p_{(r)}$, that depends on the current state of the system and the reaction and is approximated by the so called propensity $p_{(r)}$ of reaction $R_r \in \mathcal{R}, r = 1, ..., n_r$:

$$p_{(r)} \approx a_{(r)}(\boldsymbol{x}).$$

In this case the process is called a Markov jump process, because whenever an event occurs, the system jumps from a particular state to the next one according to R_r [Klipp, 2010]. This mathematical formulation leads to the definition of the Chemical master equation.

2.1.2.2 Chemical master equation

The stochastic evolution of the state vector $\boldsymbol{x}$ can be described by the chemical master equation (CME) [Gillespie, 1992, Klipp, 2010]. In detail, the CME describes the change of the probability to be in a certain state at a certain time point $P(\boldsymbol{x}|t)$ over time and can be written as

$$\frac{dP(\boldsymbol{x}|t)}{dt} = \sum_{r=1}^{n_r} a_{(r)}(\boldsymbol{x} - \nu_{(\cdot,r)})P(\boldsymbol{x} - \nu_{(\cdot,r)}|t) - a_{(r)}(\boldsymbol{x})P(\boldsymbol{x}|t), \tag{2.3}$$

where $\nu_{(\cdot,r)}$ indicates the rth column of the stoichiometric matrix and $a_{(r)}$ the rth entry of the propensity vector. Both can be derived from the reactions $R_r \in \mathcal{R}$ of the CRN $(\mathcal{S}, \mathcal{C}, \mathcal{R})$:

$$R_r : \sum_{i=1}^{n_s} \nu_{i,r}^- \cdot \mathrm{S_i} \xrightarrow{k_r} \sum_{i=1}^{n_s} \nu_{i,r}^+ \cdot S_i, r = 1, ..., n_r, \tag{2.4}$$

where $\nu_{i,r} = \nu_{i,r}^+ - \nu_{i,r}^- \in \mathbb{Z}$ and k_j is the reaction constant of the rth reaction [Bachar et al., 2013, Klipp, 2010].

Example 2.6. *For the two cell type example considered above (see Figure 2.1), the propensity vector is*

$$a(\boldsymbol{x}) = \begin{pmatrix} \alpha \cdot [A] \\ \beta \cdot [A] \\ \gamma_A \cdot [A] \\ \gamma_B \cdot [B] \end{pmatrix}$$

and the stoichiometric matrices are given by

$$\nu = (\nu^+ - \nu^-) = \begin{array}{c} \\ A \\ B \end{array}\!\!\begin{array}{c} \begin{array}{cccc} R_1 & R_2 & R_3 & R_4 \end{array} \\ \left[\begin{array}{cccc} 0 & 2 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{array}\right] \end{array} - \begin{array}{c} \\ A \\ B \end{array}\!\!\begin{array}{c} \begin{array}{cccc} R_1 & R_2 & R_3 & R_4 \end{array} \\ \left[\begin{array}{cccc} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right] \end{array} = \begin{array}{c} \\ A \\ B \end{array}\!\!\begin{array}{c} \begin{array}{cccc} R_1 & R_2 & R_3 & R_4 \end{array} \\ \left[\begin{array}{cccc} -1 & 1 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{array}\right] \end{array}$$

which results in the CME

$$\begin{aligned} \frac{dP(\boldsymbol{x}|t)}{dt} &= \alpha \cdot [A+1] \cdot P(x_1 = [A+1], x_2 = [B-1]|t) - \alpha \cdot [A] \cdot P(x_1 = [A], x_2 = [B]|t) \\ &+\beta \cdot [A-1] \cdot P(x_1 = [A-1], x_2 = [B]|t) - \beta \cdot [A] \cdot P(x_1 = [A], x_2 = [B]|t) \\ &+\gamma_A \cdot [A+1] \cdot P(x_1 = [A+1], x_2 = [B]|t) - \gamma_A \cdot [A] \cdot P(x_1 = [A], x_2 = [B]|t) \\ &+\gamma_B \cdot [B+1] \cdot P(x_1 = [A], x_2 = [B+1]|t) - \gamma_B \cdot [B] \cdot P(x_1 = [A], x_2 = [B]|t). \end{aligned}$$

For some models, the CME can be solved analytically, see for instance Jahnke and Huisinga [2007]. Within the scope of a Bachelor's thesis [Rösch, 2018], we investigated four rather simple stochastic cell cycle models and could derive the analytic solution of the CME by using the probability generating function representation, which was however not trivial for some of the considered models. In general, the solution of the CME is analytically and numerically intractable [Resat et al., 2009], due to the large or infinite number of possible states $\boldsymbol{x}$. This often restricts the choice of the inference approach that is used to estimate model parameters $\boldsymbol{\theta}$ (see section 2.2.3), as likelihood-based parameter inference (see section 2.2.1) requires a solution of the CME. If the solution of the CME is intractable, one can use a likelihood-free inference approach (see section 2.2.2) and simply forward simulate the system for several values of $\boldsymbol{\theta}$ to compare the model output to the data set. Alternatively, one can transform the CME into a deterministic model described by a closed-form ODE system (see section 2.1.2.3), that can easily be solved and used for model simulations and maximum likelihood estimation. The trick in the latter approach is to calculate the moment equations from the CME, which l captures the stochasticity of the system if not only the first but also higher order moments are considered.

2.1.2.3 Approximation by moment equations

The moment equations can be directly calculated from the CME. The first and second order moments (mean $\mu_\cdot$, variance and covariance $C_{\cdot,\cdot}$) are defined by

$$\begin{aligned} \mu_i(t) &:= E[X_i(t)] = \sum_{x_i} x_i P(\boldsymbol{x}|t) \\ C_{i,j}(t) &:= Cov[X_i(t), X_j(t)] = \sum_{x_i,x_j} (x_i - \mu_i(t))(x_j - \mu_j(t))^T P(\boldsymbol{x}|t), \end{aligned} \tag{2.5}$$

with $i, j = 1, 2, ..., n_s$ denoting the cell state index [Engblom, 2006]. One can calculate the derivatives to get the evolution equations for the first and second order moments:

$$\begin{aligned}
\frac{d\mu_i(t)}{dt} &= \sum_{r=1}^{n_r} \nu_{(i,r)} \left(a_{(r)}(\mu(t), \boldsymbol{\theta}) + \frac{1}{2} \sum_{l_1,l_2} \frac{\partial^2 a_{(r)}(\mu(t), \boldsymbol{\theta})}{\partial x_{l_1} \partial x_{l_2}} C_{l_1,l_2}(t) \right) \\
\frac{dC_{i,j}(t)}{dt} &= \sum_{r=1}^{n_r} \left(\nu_{(i,r)} \sum_{l_1} \frac{\partial a_{(r)}(\mu(t), \boldsymbol{\theta})}{\partial x_{l_1}} C_{l_1,j}(t) + \nu_{(j,r)} \sum_{l_2} \frac{\partial a_{(r)}(\mu(t), \boldsymbol{\theta})}{\partial x_{l_2}} C_{i,l_2}(t) \right) \\
&\quad + \sum_{r=1}^{n_r} \nu_{(i,r)} \nu_{(j,r)} \left(a_{(r)}(\mu(t), \boldsymbol{\theta}) + \frac{1}{2} \sum_{l_1,l_2} \frac{\partial^2 a_{(r)}(\mu(t), \boldsymbol{\theta})}{\partial x_{l_1} \partial x_{l_2}} C_{l_1,l_2}(t) \right),
\end{aligned} \tag{2.6}$$

with initial conditions $\mu_i(0)$, $C_{i,j}(0), i, j = 1, 2, ..., n_s$. For any parameter vector $\boldsymbol{\theta}$, the first and second order moment can be generated by solving the first and second order moment equations (see equation 2.6). As the ith moment depends on the $(i + 1)$th moment, the ODE system of moment equations is in general not closed but rather an infinite set of coupled differential equations. To solve this issue, one can apply moment closure as an approximation method [Kuehn, 2016]. Note, that if the reaction propensities are linear in cell states, the resulting ODE system is a linear system and can be solved exactly, without applying moment closure.

Example 2.7. *The simple two cell-type model 2.1 has reaction propensities linear in cell states and its moment equations are given by the linear ODE system*

$$\begin{pmatrix} \frac{d\mu_1(t)}{dt} \\ \frac{d\mu_2(t)}{dt} \\ \frac{dC_{1,1}(t)}{dt} \\ \frac{dC_{1,2}(t)}{dt} \\ \frac{dC_{2,2}(t)}{dt} \end{pmatrix} = \begin{pmatrix} -\alpha + \beta - \gamma_A & 0 & 0 & 0 & 0 \\ \alpha & -\gamma_B & 0 & 0 & 0 \\ \alpha + \beta + \gamma_A & 0 & 2(-\alpha + \beta - \gamma_A) & 0 & 0 \\ -\alpha & -\gamma_B & \alpha & -\alpha + \beta - \gamma_A - \gamma_B & 0 \\ \alpha & -\gamma_B & 0 & 2\alpha & -2\gamma_B \end{pmatrix} \begin{pmatrix} \mu_1(t) \\ \mu_2(t) \\ C_{1,1}(t) \\ C_{1,2}(t) \\ C_{2,2}(t) \end{pmatrix}.$$

2.1.3 Deterministic compartmental models

For large cell populations, the stochastic effects resulting from the behaviour of single cells can cancel each other and are less prominent in the observed mean population behaviour. Therefore one can neglect the stochasticity of the system and model its dynamics with a deterministic approach and a set of differential equations [Klipp, 2010, Kremling, 2012]. In case the set of differential equations depends on only one variable, ordinary differential equations (ODE) are used to model the biological system.

Definition 2.8. *An ODE of* n*th order is given by* $x^{(n)} = f(t, x, x^{(1)}, ..., x^{(n-1)})$*, where* n *is the highest derivative of* x*. A number of coupled ODEs describes an ODE system. An ODE system is linear if the right-hand side can be written as linear combination of the derivatives of* x *[Robers, 2018]. In case the right-hand side of the ODE system does not depend on time it is said to be autonomous.*

To model cell division and differentiation, one can consider a CRN of cell states and mass action kinetics.

Definition 2.9. *A mass-action CRN is a quadrupel $(\mathcal{S}, \mathcal{C}, \mathcal{R}, k)$, where $(\mathcal{S}, \mathcal{C}, \mathcal{R})$ is a CRN, $k \in \mathbb{R}_+^{\mathcal{R}}$ is the vector of reaction rate constants and the respective ODE system describes the evolution of the state vector $\boldsymbol{x} = [x_{S_1}, ..., x_{S_{ns}}]$ defined by*

$$\dot{\boldsymbol{x}} = \sum_{r=1}^{n_r} (\nu^+_{(\cdot,r)} - \nu^-_{(\cdot,r)}) \cdot k_r \cdot \prod_{i=1}^{n_s} x_{S_i}^{\nu^-_{(i,r)}}, \tag{2.7}$$

where n_r is the number of reactions and n_s is the number of species considered in the network [Craciun and Pantea, 2008].

Example 2.10. *For the simple two cell type model 2.1, the change in the number of A and B cells over time is described by the following ODE system:*

$$\begin{aligned}\dot{\boldsymbol{x}} &= \begin{pmatrix} \dot{x_A} \\ \dot{x_B} \end{pmatrix} \\ &= \begin{pmatrix} -1 \\ 1 \end{pmatrix} \cdot \alpha \cdot x_A^1 \cdot x_B^0 + \begin{pmatrix} 1 \\ 0 \end{pmatrix} \cdot \beta \cdot x_A^1 \cdot x_B^0 + \begin{pmatrix} -1 \\ 0 \end{pmatrix} \\ &\cdot \gamma_A \cdot x_A^1 \cdot x_B^0 + \begin{pmatrix} 0 \\ -1 \end{pmatrix} \cdot \gamma_B \cdot x_A^0 \cdot x_B^1 \\ &= \begin{pmatrix} -(\alpha - \beta + \gamma_A) & 0 \\ \alpha & -\gamma_B \end{pmatrix} \cdot \begin{pmatrix} x_A(t) \\ x_B(t) \end{pmatrix}.\end{aligned}$$

Note, that this ODE system corresponds to the equations for the mean in the moment equations (see equation 2.7).

Each cell division and differentiation process considered within the scope of this dissertation describes the transition from one cell state to another state (cell differentiation and death), or to the same state (proliferation). From this follows that $\nu^-_{(i,r)} \in \{0, 1\} \forall i$. Such systems are always described by a 1st order non-autonomous homogeneous linear ODE system with constant coefficients $\{a_{i,j}\}_{i,j=1,...,n_s}$ which can be written as

$$\begin{aligned}\dot{x}_i &= \frac{d}{dt} x_i \\ &= a_{i1}(\boldsymbol{\theta}, \boldsymbol{k}, t) \cdot x_1(t) + a_{i2}(\boldsymbol{\theta}, \boldsymbol{k}, t) \cdot x_2(t) + \dots a_{in}(\boldsymbol{\theta}, \boldsymbol{k}, t) \cdot x_n(t) \\ &= f_i(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{k}, t), i = 1, ..., n_s\end{aligned} \tag{2.8}$$

and in matrix notation

$$\dot{\boldsymbol{x}} = A(\boldsymbol{\theta}, \boldsymbol{k}, t) \cdot \boldsymbol{x} \tag{2.9}$$

[Robers, 2018].

2.1.4 Existence and uniqueness of the solution of first-order ODE systems

While studying differentiation processes the solution of the corresponding ODE system is of interest [Klipp, 2010] as it is either partly or completely defining the observables and thus required for parameter inference, model validation and model predictions.
For 1st order homogeneous linear ODE systems, the following theorem about the existence of a solution holds [Robers, 2018]:

Theorem 2.11. *If $A(\boldsymbol{\theta}, \boldsymbol{k}, t)$ is continuous on some interval $\mathcal{I} = (b_1, b_2)$ - that is, if $a_{i,j}(\boldsymbol{\theta}, \boldsymbol{k}, t)$ is a continuous function on $\mathcal{I}$ $\forall i, j = 1, 2, ..., n_s$, then there exist n_s linearly dependent solutions of the homogeneous linear system $\dot{\boldsymbol{x}} = A(\boldsymbol{\theta}, \boldsymbol{k}, t) \cdot \boldsymbol{x}$ on the interval $\mathcal{I}$.*

The general solution of a 1st order homogeneous linear ODE system is defined as follows [Robers, 2018]:

Definition 2.12. *Let $x_1, x_2, ..., x_{n_s}$ be linearly independent solutions of the homogeneous linear system $\dot{\boldsymbol{x}} = A(\boldsymbol{\theta}, \boldsymbol{k}, t) \cdot \boldsymbol{x}$ on the interval $\mathcal{I}$ and $c_1, ..., c_{n_s}$ arbitrary scalar constants. The linear combination $x(t) = c_1 \cdot x_1(t) + c_2 \cdot x_2(t) + ... + c_{n_s} \cdot x_{n_s}(t)$ is called the general solution of the homogeneous linear system on $\mathcal{I}$.*

The general solution can be easily calculated using the following theorem [Klipp, 2010, Robers, 2018]:

Theorem 2.13. *If $\lambda_1, ..., \lambda_{n_s}$ are the eigenvalues (not necessarily distinct) of a constant matrix $A \in \mathbb{R}^n$ and if $v_1, ..., v_n$ are associated linearly independent Eigenvectors, then the general solution of the homogeneous linear system $\dot{\boldsymbol{x}} = A \cdot \boldsymbol{x}$ is*

$$\dot{\boldsymbol{x}}(t) = c_1 \cdot v_1 \cdot e^{\lambda_1 t} + \cdots + c_{n_s} \cdot v_{n_s} \cdot e^{\lambda_{n_s} t}, \tag{2.10}$$

where $c_1, ..., c_{n_s}$ are arbitrary constants.

Example 2.14. *For the simple two cell type model, the general solution can be calculated by first calculating the eigenvalues and eigenvectors of matrix A.*

$$A \cdot v_i \overset{!}{=} \lambda_i v_i \quad \forall i = 1, ..., n_s \iff det(A - I\lambda) \overset{!}{=} 0$$

$$det \begin{pmatrix} -(\alpha - \beta + \gamma_A) - \lambda & 0 \\ \alpha & -\gamma_B - \lambda \end{pmatrix} \overset{!}{=} 0$$

leads to solutions $\lambda_1 = \gamma_B$ and $\lambda_2 = \beta - \alpha - \gamma_A$ for eigenvalues and $v_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ and $v_2 =$

$\begin{pmatrix} \frac{-(\alpha-\beta+\gamma_A-\gamma_B)}{\alpha} \\ 1 \end{pmatrix}$ *for the corresponding eigenvectors. The general solution is:*

$$\begin{pmatrix} x_A(t) \\ x_B(t) \end{pmatrix} = \begin{pmatrix} -c_2 \cdot \frac{\alpha-\beta+\gamma_A-\gamma_B}{\alpha} \cdot e^{(\beta-\alpha-\gamma_A)t} \\ c_1 \cdot e^{\gamma_B t} + c_2 \cdot e^{(\beta-\alpha-\gamma_A)t} \end{pmatrix} \tag{2.11}$$

Definition 2.15. *A linear system initial value problem (IVP) consists of solving the system of n_s first-order equations*

$$\begin{aligned} \dot{x}_1(t) &= a_{11}(\boldsymbol{\theta}, \boldsymbol{k}, t) \cdot x_1(t) + a_{12}(\boldsymbol{\theta}, \boldsymbol{k}, t) \cdot x_2(t) + \ldots a_{1n_s}(\boldsymbol{\theta}, \boldsymbol{k}, t) \cdot x_{n_s}(t) \\ &\vdots \\ \dot{x}_{n_s}(t) &= a_{n_s1}(\boldsymbol{\theta}, \boldsymbol{k}, t) \cdot x_1(t) + a_{n_s2}(\boldsymbol{\theta}, \boldsymbol{k}, t) \cdot x_2(t) + \ldots a_{n_sn_s}(\boldsymbol{\theta}, \boldsymbol{k}, t) \cdot x_{n_s}(t) \end{aligned}$$

subject to n_s initial conditions $x_1(c) = d_1, ..., x_{n_s}(c) = d_{n_s}$, $c \in \mathbb{R}$

Theorem 2.16. *If the functions $a_{ij}(t)$ $i, j = 1, ..., n_s$ are all defined and continuous on the interval $\mathcal{I} = (b_1, b_2)$ and if $c \in \mathbb{I}$, then there exists a unique solution to the linear system initial value problem on the interval $\mathcal{I}$.*

For first-order non-autonomous non-linear ODE systems, the above stated definitions and theorems can be formulated more generally [Robers, 2018].

Definition 2.17. *A system of n_s first-order differential equations*

$$\dot{x}_i = f_i(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{k}, t), \;\; i = 1, ..., n_s$$

has a solution on the interval $\mathcal{I} = (b_1, b_2)$, if there exists a set of n_s functions $\{x_i(t), ..., x_{n_s}(t)\}$, which all have continuous first derivatives on the interval $\mathcal{I}$. The set of functions $\{x_i(t), ..., x_{n_s}(t)\}$ is then called solution of the system on the interval $\mathcal{I}$.

Definition 2.18. *An IVP for a system of first-order differential equations consists of solving a system of equations of the form*

$$\dot{x}_i = f_i(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{k}, t), \;\; i = 1, ..., n_s$$

subject to a set of constraints, called initial conditions of the form

$$x_1(c) = d_1, ..., x_{n_s}(c) = d_{n_s}, c \in \mathbb{R}.$$

For such IVPs, the solution is unique under certain circumstances, as the following theorem states:

Theorem 2.19. *Let $R = (t, x_1, ..., x_{n_s})| m < t < n$ and $a_i < y_i < b_i, i = 1, ..., n_s$ where $m, n, a_i.b_i$ are all finite real constants. If*

(i) each of the n_s functions $f_i(t, x_1, ..., x_{n_s})$, $i = 1, ..., n_s$ is a continuous function of $t, x_1, x_2, ...,$ and x_{n_s} in R,

(ii) each of the n_s^2 partial derivatives $\frac{\partial f_i}{\partial y_j}$, $i, j = 1, ..., n_s$ is a continuous function of $t, x_1, x_2, ...,$ and x_{n_s} in R, and

(iii) $(c, d_1, ..., d_{n_s}) \in R$,
then there exists a unique solution to the linear system initial value problem

$$\dot{x}_i = f_i(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{k}, t), \quad x_1(c) = d_1, ..., x_{n_s}(c) = d_{n_s}$$

on some interval $\mathcal{I} = (c - h, c + h)$ where $\mathcal{I}$ is a subinterval of (m, n).

Finding the solution of an IVP for a system of first-order differential equations on a specified interval requires numerical methods [Robers, 2018].

2.1.5 Numerical integration methods and stiff ODE systems

Solving an ODE system requires the numerical integration of the respective ODEs. Given the IVP

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{\theta}, \boldsymbol{k}, t), \quad \boldsymbol{x}(t_0) = \boldsymbol{x_0},$$

one aims to find $\boldsymbol{x} : [t_0, T] \rightarrow \mathbb{R}^n$.

This is achieved by stepwise approximating the solution $x(t)$ with

$$\begin{aligned} \boldsymbol{x}(t_{i+1}) &= \boldsymbol{x}(t_i) + \int_{t_i}^{t_{i+1}} \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{\theta}, \boldsymbol{k}, t) dx \\ &= \boldsymbol{x}(t_i) + I_i. \end{aligned} \tag{2.12}$$

Therefore, one can apply explicit or implicit integration methods [Butcher, 2016].

Definition 2.20. *A numerical integration method is explicit if $\boldsymbol{x}$ can directly be calculated by computation of known quantities, i.e. by evaluating $\boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{\theta}, \boldsymbol{k}, t)$, and is implicit otherwise [Butcher, 2016].*

Some numerical integration methods are listed in table 2.1, where $h_i := t_{i+1} - t_i$ defines the step size [Butcher, 2016].

	method name	I_i	$\boldsymbol{x}_{i+1}$
explicit	forward Euler	$\approx h_i\boldsymbol{f}(\boldsymbol{x}_i(t_i),\boldsymbol{\theta},\boldsymbol{k},t_i)$	$= \boldsymbol{x}_i + h_i\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{k},t_i)$
	modified forward Euler	$\approx h_i\boldsymbol{f}(\boldsymbol{x}_i(\frac{t_i+t_{i+1}}{2}),\boldsymbol{\theta},\boldsymbol{k},\frac{t_i+t_{i+1}}{2})$	$= \boldsymbol{x}_i$ $+h_i\boldsymbol{f}(\boldsymbol{x}_i + \frac{h_i}{2}\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{k},t_i),\boldsymbol{\theta},\boldsymbol{k},t_i + \frac{h_i}{2})$
implicit	backward Euler	$\approx h_i\boldsymbol{f}(\boldsymbol{x}(t_{i+1}),\boldsymbol{\theta},\boldsymbol{k},t_{i+1})$	$= \boldsymbol{x}_i + h_i\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{k},t_{i+1})$
	modified backward Euler		$= \boldsymbol{x}_i$ $+h_i\boldsymbol{f}(\frac{\boldsymbol{x}_i+\boldsymbol{x}_{i+1}}{2},\boldsymbol{\theta},\boldsymbol{k},t_i + \frac{h_i}{2}))$
	Crank-Nicolson	$\approx h_i\frac{1}{2}(\boldsymbol{f}(\boldsymbol{x}(t_i),\boldsymbol{\theta},\boldsymbol{k},t_i)$ $+\boldsymbol{f}(\boldsymbol{x}(t_{i+1}),\boldsymbol{\theta},\boldsymbol{k},t_{i+1}))$	$= x_i + h_i\frac{1}{2}(\boldsymbol{f}(\boldsymbol{x}_i,\boldsymbol{\theta},\boldsymbol{k},t_i)$ $+\boldsymbol{f}(\boldsymbol{x}_{i+1},\boldsymbol{\theta},\boldsymbol{k},t_{i+1}))$
	Heun	$\approx h_i\frac{1}{2}(\boldsymbol{f}(\boldsymbol{x}(t_i),\boldsymbol{\theta},\boldsymbol{k},t_i)$ $+\boldsymbol{f}(\boldsymbol{x}(t_{i+1}),\boldsymbol{\theta},\boldsymbol{k},t_{i+1}))$	$= x_i + h_i\frac{1}{2}(\boldsymbol{f}(\boldsymbol{x}_i,\boldsymbol{\theta},\boldsymbol{k},t_i)$ $+\boldsymbol{f}(\boldsymbol{x}_i + h_i\boldsymbol{f}(\boldsymbol{x}_i,\boldsymbol{\theta},\boldsymbol{k},t_i),\boldsymbol{\theta},\boldsymbol{k},t_{i+1}))$

Table 2.1: Explicit and implicit numerical integration methods.

The above mentioned methods can be generalized as implicit and explicit Runge-Kutta methods [Butcher, 2016, Griffiths and Higham, 2010]. Besides, there exist some multi-step procedures, which attempt to increase efficiency by exploiting information from previous steps. Multi-step methods are for instance Adams-Bashforth methods (explicit), Adams-Moulton method (implicit), or the backward differentiation formula (implicit), which approximate either $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{k},t)$ or $\boldsymbol{x}(t)$ by a lagrange polynomial to approximate the integral I [Griffiths and Higham, 2010]. These methods can also be combined [Yang et al., 2005]. The Adams-Bashforth-Moulton method for instance, approximates $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{k},t)$ by a lagrange polynomial, calculates a predicted estimate of x_{i+1} and then repeats this procedure with using the predicted estimate for evaluating f to calculate the corrected estimate x_{i+1} [Yang et al., 2005].

Although implicit methods require more computational effort, they are useful for solving stiff problems. A differential equation is said to be stiff if the magnitudes of $\dot{x}_1(t),...,\dot{x}_{n_s}(t)$ are significantly different [Yang et al., 2005] and are thus hard to solve numerically. For chemical reaction networks, the resulting IVPs often result in stiff problems if the reactions obtain both very high, and very low rates and are thus on various time scales [Hairer and Wanner, 1996]. Although there is no clear distinction between a stiff and a non-stiff differential equation, one can determine its degree of stiffness if it is possible to transform the ODE system into a linear time-invariant state equation of the form

$$\dot{\boldsymbol{x}} = A\boldsymbol{x}(t) + Bu(t), \quad \boldsymbol{x}(0) = \boldsymbol{x_0},$$

where $u(t)$ is the input, and A and B are matrices of coefficients. The stiffness measure is then calculated from the negative real parts of the eigenvalues λ_i of matrix A:

$$\eta(A) = \frac{max\ |Re(\lambda_i)|}{min\ |Re(\lambda_i)| \neq 0}.$$

The higher the value for $\eta(A)$, the higher the unbalance between fast and slow mode and the more stiff the equation.

By increasing the step size, one can observe the numerical stability of the solution for a particular method. For stiff equations the step-size should be chosen carefully in order to avoid numerical instability [Yang et al., 2005]. In general, the stability properties of implicit methods allow for the accurate determination of the solution with quite large time-step sizes, even if the problem is stiff. In contrast, if the usually less computationally demanding explicit methods are applied to stiff problems, they are forced to apply a small step size due to their constrained stability area and require a much longer computation time in this case [Hairer and Wanner, 1996, Hirschfelder et al., 1954].

Fortunately, there are strategies to efficiently deal with stiff equations by adaptively determining the step size, as implemented in MATLAB for several ode solver build-in routines, which saves computation time [Yang et al., 2005].

2.1.6 Transition time distributions for compartmental models

For stochastic and deterministic compartmental models, the transition times from one compartment to another are per definition exponentially distributed [Matis and Wehrly, 1990]. As has been observed experimentally (e.g. [Filipczyk et al., 2015]), cell processes such as proliferation, differentiation and cell death require a certain time and are rather Erlang than exponentially distributed. Thus, the distribution of transit times as it is assumed in compartmental models might not be accurate enough to model cell proliferation, differentiation and cell death. To describe these cell processes more accurately, one can introduce intermediate compartments for each cell type compartment while specifying the model. This model adaptation results in Erlang distributed transition times between the cell type compartments, as the following theorem states.

Theorem 2.21. *Let T be an exponentially distributed random variable $T \sim exp(\lambda)$ representing the transition time for the next reaction to occur. The sum of k exponentially distributed transition times $Y = \sum_{i=1}^{k} T_i, k \in \mathbb{N}$ is then $Erlang(k, \lambda)$ distributed.*

Proof. Let $T_i \sim exp(\lambda), i = 1, ..., k$ and $Y = \sum_{i=1}^{k} T_i \sim Erlang(k, \lambda), \lambda \in \mathbb{R}$ be random variables with probability density functions $f_T(x, \lambda)$ and $f_Y(x, k, \lambda)$, respectively. As the moment generating function of a random variable X with probability density function $f(x, \boldsymbol{\theta})$ is defined as

$$\begin{aligned} \phi_X(s) &= E(e^{sX}) \\ &= \int_{-\infty}^{\infty} e^{sx} f(x, \boldsymbol{\theta}) dx, s \in \mathbb{R} \end{aligned} \tag{2.13}$$

[Bachar et al., 2013, van Kampen, 2007] and as for the generating function of the sum of independent random variables X_i

$$Z = \sum_{i=1}^{k} X_i,$$

it holds that

$$\phi_Z(s) = E(e^{s(\sum_{i=1}^k X_i})= E(\prod_{i=1}^k e^{sX_i}) \overset{independentX_i}{=} \prod_{i=1}^k E(e^{sX_i}) = \prod_{i=1}^k \phi_{X_i}(s), \tag{2.14}$$

we can show that the above stated theorem holds by deriving the moment generating functions of random variables T and Y. From

$$\begin{aligned}\phi_T(s) &= \int_{-\infty}^{\infty} e^{st} f(t,\lambda)dt = \int_0^{\infty} e^{st}\lambda e^{-\lambda t}dt = \left[\frac{\lambda}{s-\lambda}e^{(s-\lambda)t}\right]_0^{\infty} \\ &= \lim_{t\to\infty}\frac{\lambda}{s-\lambda}e^{(s-\lambda)t} - \frac{\lambda}{s-\lambda} = \frac{\lambda}{\lambda-s} = \frac{1}{1-\frac{s}{\lambda}} = (1-s\lambda^{-1})^{-1}, s<\lambda\end{aligned}$$

and

$$\begin{aligned}\phi_Y(s) &= \int_{-\infty}^{\infty} e^{sy} f(y,k,\lambda)dy, s\in\mathbb{R} = \int_0^{\infty} e^{sy}\frac{1}{\Gamma(k)}\lambda^k y^{k-1}e^{-\lambda y}dy = \frac{\lambda^k}{\Gamma(k)}\int_0^{\infty} e^{(s-\lambda)y}y^{k-1}dy \\ &= \frac{\lambda^k}{\Gamma(k)}\left(\left[\frac{1}{s-\lambda}e^{(s-\lambda)y}y^{k-1}\right]_0^{\infty} - \int_0^{\infty}(k-1)y^{k-2}\frac{1}{(s-\lambda)^2}e^{(s-\lambda)y}dy\right) \\ &= \frac{\lambda^k}{\Gamma(k)}\left(\left[\frac{1}{(s-\lambda)^2}e^{(s-\lambda)y}(k-1)y^{k-2}\right]_0^{\infty} - \int_0^{\infty}(k-2)(k-1)y^{k-3}\frac{1}{(s-\lambda)^2}e^{(s-\lambda)y}dy\right) \\ &= \frac{\lambda^k}{\Gamma(k)}\int_0^{\infty}(k-1)!\frac{1}{(s-\lambda)^{k-1}}e^{(s-\lambda)y}dy = \frac{\lambda^k}{(k-1)!}\left[\frac{(k-1)!}{(s-\lambda)^{k-1}}\frac{1}{s-\lambda}e^{(s-\lambda)y}\right]_0^{\infty} \\ &= \left(\frac{\lambda}{\lambda-s}\right)^k = (1-s\lambda^{-1})^{-k}\end{aligned}$$

it follows that

$$\phi_Y(s) = (1-s\lambda^{-1})^{-k} = \prod_{i=1}^k (1-s\lambda^{-1})^{-1} = \prod_{i=1}^k \phi_{T_i}(s)$$

□

Alternatively one can directly integrate any time distribution such as Erlang or lag-exponential, by deriving a Volterra integral equation representation of the generating function, which is at least feasible for very simple stochastic models that can be analytically solved with the probability generating function representation [Rösch, 2018]. For more complicated stochastic models, one can use a sampling based inference method such as approximate Bayesian computation (see section 2.2.2) for fitting the model to measurements, which allows to specify any time distribution while simulating from the model (see section 2.1.7 and 2.2.2).

2.1.7 Model simulation

In order to perform numerical parameter inference, a fast and efficient simulation of the model observables is required which can then be compared to the experimental observations during optimization. In addition, model validation, i.e. comparison of the model to independent data or observations not used for inference, and model predictions require simulations of the model for the optimal parameter vector.

In case of deterministic models, for which the observables are defined as the solution of the ODE system, the model can simply be forward simulated by evaluating the solution at the respective time points for the parameters of interest.

For stochastic CRN models, one can simulate individual realizations of the underlying random processes with the stochastic simulation algorithm (SSA), which is also known as Gillespie algorithm [Gillespie, 1976]. The idea is to simulate every reaction explicitly such that each realization is a sample of the solution of the CME. Based on an initial model state, transition times are drawn from the underlying reaction time distributions for all possible reactions that could occur in the next step. The lowest transition time determines which reaction takes place, leading to the next model state. This is repeated until a stopping criteria is fulfilled, e.g. that the number of reactants is equal to zero, the maximum simulation time has been exceeded, or the maximum number of simulation steps is reached. Formally, the Gillespie algorithm can be formulated as follows:

Gillespie algorithm

Step 0 Initialize the sampling time $t_0 = 0$ and the iteration counter $i = 0$, and specify a stopping time t_{stop}. Determine the initial state $\boldsymbol{x}_0$ and calculate the reaction probability density function for this initial state according to

$$P(\tau, r) = a_r(\boldsymbol{x}_0) \exp\left(\sum_{k=1}^{M} a_k(\boldsymbol{x}_0)\tau\right).$$

Step 1 Generate a random pair (τ, r) according to the joint probability density function $P(\tau, r)$, i.e. with the direct method by

(i) drawing $r_1 \sim U(0,1)$ and calculating $\tau = \frac{1}{\sum_{k=1}^{M} a_k(\boldsymbol{x}_i)} \ln\left(\frac{1}{r_1}\right)$, and

(ii) drawing $r_2 \sim U(0,1)$ and finding reaction r that fulfills

$$\sum_{k=1}^{r-1} a_k(\boldsymbol{x}_i) < r_2 \sum_{k=1}^{M} a(\boldsymbol{x}_i) \leq \sum_{k=1}^{r} a_k(\boldsymbol{x}_i).$$

Step 2 Update $t_{(i+1)} = t_i + \tau$, $\boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \nu_{(.,r)}$ and $a(\boldsymbol{x}_{i+1}) = a(\boldsymbol{x}_i + \nu_{(.,r)})$ according to reaction R_r. Set $i = i + 1$.

Step 3 If $t < t_{stop}$ or $\boldsymbol{x}_i = \boldsymbol{0}$ terminate the calculation, otherwise return to **Step 1**.

Simulating realizations of the solution of the CME often requires a high numerical effort [Gillespie, 2001, Klipp, 2010]. To improve the efficiency and speed of the algorithm, several attempts have been made in previous years. A commonly used approach is τ-leaping, which updates the propensity vector not after each reaction but only after an interval of length τ [Cao et al., 2006]. Moreover, R-leaping has been suggested which predefines the number of reaction firings [Auger et al., 2006]. Both methods can be combined to S-leaping [Lipková et al., 2018]. Furthermore, it has been shown that the recycling of random numbers for instance could reduce the simulation time by 25% without significantly reducing the accuracy [Yates and Klingbeil, 2013].

2.2 Parameter inference

Parameter inference approaches aim to identify the values of the parameter vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_{n_\theta})$ for which the model best describes the data and can thereby be interpreted as model calibration to the experimental measurements. In this section I will introduce likelihood-based and likelihood-free optimization approaches and methods to analyse parameter identifiability and model uncertainty.

2.2.1 A likelihood-based approach: maximum likelihood estimation

If a closed-form solution exists for the model observables $y^{\mathcal{M}}(\boldsymbol{\theta}, t_k)$ the likelihood $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})$ of a particular model $\mathcal{M}(\boldsymbol{\theta})$ can be calculated in order to assess how well $\mathcal{M}(\boldsymbol{\theta})$ explains the experimental data $y^{\mathcal{D}}(t_k)$ for a certain set of parameters $\boldsymbol{\theta}$.

Definition 2.22. *The Likelihood function $\mathcal{L}$ corresponds to the probability of observing the data $y^{\mathcal{D}}(t_k)$ given the model $\mathcal{M}(\boldsymbol{\theta})$ and is defined by*

$$\begin{aligned} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = P(\mathcal{D}|\boldsymbol{\theta}) \quad &= \quad \prod_{k=1}^{n_t} P(y^{\mathcal{D}}(t_k)|y^{\mathcal{M}}(\boldsymbol{\theta}, t_k), \omega) \\ &\overset{independent\ y_j}{=} \quad \prod_{k=1}^{n_t}\prod_{j=1}^{n_y} P(y_j^{\mathcal{D}}(t_k)|y_j^{\mathcal{M}}(\boldsymbol{\theta}, t_k), \omega), \end{aligned} \tag{2.15}$$

where n_t is the number of time points t_k and n_y the number of observables $y_j^{\mathcal{M}}(\boldsymbol{\theta}, t_k)$ and data points $y_j^{\mathcal{D}}(t_k)$ at time t_k.
Accordingly, the log*-likelihood function is defined by*

$$\ell_D(\boldsymbol{\theta}) \quad = \quad \sum_{k=1}^{n_t}\sum_{j=1}^{n_y} \log P(y_j^{\mathcal{D}}(t_k)|y_j^{\mathcal{M}}(t_k), \omega), \tag{2.16}$$

where ω describes parameters of the underlying noise distribution [Fröhlich et al., 2018].

Definition 2.23. *The maximum likelihood estimate of the unknown parameter vector $\boldsymbol{\theta}$ is defined as the solution of the following optimization problem:*

$$\hat{\boldsymbol{\theta}}^{ML} = \underset{\substack{\boldsymbol{\theta} \\ subject\ to\ \mathcal{M}}}{\arg\max}\ \ell_D(\boldsymbol{\theta}). \tag{2.17}$$

Maximum likelihood estimation (MLE) thereby identifies the parameter values $\boldsymbol{\theta}^{ML}$ for which the conditional probability of observing the data given the model $\mathcal{M}(\boldsymbol{\theta})$ is maximized. The resulting estimate obtains as statistical properties consistency, asymptotic normality and efficiency [Bachar et al., 2013]. For specification of the likelihood function, a certain noise model needs to be assumed (see section 2.2.1.1). Moreover, to solve the optimization problem (see equation 2.17) an appropriate optimization algorithm needs to be specified, which depends on the properties of the specified optimization problem (see section 2.2.1.2).

2.2.1.1 Noise models

The difference between measured and predicted output is defined as residual vector

$$\boldsymbol{r}(t,\boldsymbol{\theta}) = \overline{y}^{\mathcal{D}}(t) - y^{\mathcal{M}}(t,\boldsymbol{\theta}). \tag{2.18}$$

If the model would describe the experimental system accurately and precisely, the model solutions for the parameters would fit the data exactly and the residuals would be equal to zero [Droste, 1998]. In practice, this is usually not the case as both, experimental data and model observables underlie uncertainties. The observed data is noisy due to human or technical measurement errors, and inherent biological variation. Model uncertainty is further explained and discussed in section 2.2.3.

While solving the optimization problem introduced in 2.2.1, one needs to specify a noise distribution in order to account for the uncertainty in the data and accurately estimating the parameters. The residual distribution assumption should ideally agree with its true underlying assumption. Table 2.2 lists a couple of noise models which can be applied for modelling biological processes. Additive normally distributed noise (see table 2.2) is widely used for modelling continuous observables of biological processes and can therefore be used if cell concentrations are modelled. If the moments of the cell concentrations or counts are modelled, the assumption of additive normally distributed noise is also valid as it can be shown that the moments are asymptotically normal [Moore, 1986]. In case the data observables are likely to be outlier corrupted, it is recommended to use a more heavy-tailed noise distribution than the normal distribution, such as the Laplace distribution (see table 2.2) or Student's t distribution [Maier et al., 2017]. As all these distributions are symmetric, and

Error model	Likelihood function $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})$
Additive normal noise $y_j^{\mathcal{D}}(t_k) = y_j^{\mathcal{M}}(t_k) + \epsilon_{j,k}$ with $\epsilon_{j,k} \sim \mathcal{N}(0, \sigma_{j,k}^2)$	$\prod_{k=1}^{n_t}\prod_{j=1}^{n_y} \mathcal{N}(y_j^{\mathcal{D}} \vert y_j^{\mathcal{M}}, \sigma_{j,k}^2)$ $= \prod_{k=1}^{n_t}\prod_{j=1}^{n_y} \frac{1}{\sqrt{2\pi}\sigma_{j,k}} \exp\left\{-\frac{(y_j^{\mathcal{D}}(t_k) - y_j^{\mathcal{M}}(t_k,\boldsymbol{\theta}))^2}{2\sigma_{j,k}}\right\}$
Multiplicative log-normal noise $y_j^{\mathcal{D}}(t_k) = y_j^{\mathcal{M}}(t_k) \cdot \nu_{j,k}$ with $\nu_{j,k} \sim log\mathcal{N}(0, \sigma_{j,k}^2)$	$\prod_{k=1}^{n_t}\prod_{j=1}^{n_y} log\mathcal{N}(y_j^{\mathcal{D}} \vert y_j^{\mathcal{M}}, \sigma_{j,k}^2)$ $= \prod_{k=1}^{n_t}\prod_{j=1}^{n_y} \frac{1}{\sqrt{2\pi}\sigma_{j,k}} \exp\left\{-\frac{(log(y_j^{\mathcal{D}}(t_k)) - log(y_j^{\mathcal{M}}(t_k,\boldsymbol{\theta})))^2}{2\sigma_{j,k}}\right\}$
Additive laplace noise $y_j^{\mathcal{D}}(t_k) = y_j^{\mathcal{M}}(t_k) + \epsilon_{j,k}$ with $\epsilon_{j,k} \sim \mathcal{L}\text{aplace}(0, b)$	$\prod_{k=1}^{n_t}\prod_{j=1}^{n_y} \mathcal{L}\text{aplace}(y_j^{\mathcal{D}} \vert y_j^{\mathcal{M}}, b)$ $= \prod_{k=1}^{n_t}\prod_{j=1}^{n_y} \frac{1}{2b} \exp\left\{-\frac{\vert y_j^{\mathcal{D}}(t_k) - y_j^{\mathcal{M}}(t_k,\boldsymbol{\theta})\vert}{b}\right\}$

Table 2.2: Selection of noise distributions useful for maximum likelihood estimation (MLE).

can observe negative values. For continuous model observables that take exclusively positive values (e.g. concentrations), one should rather model the noise with a positively skewed distribution, such as the log-normal distribution (see table 2.2). For modelling noise of discrete count data describing the independent events occurring with a time-invariant rate, the Poisson distribution might be a good choice if the event of interest is rather rare. However, for events that occur more often, the inherent Poisson assumption of the distribution variance equal to the distribution mean usually does not hold.

Whether the noise distribution assumption agrees with its true underlying distribution can be investigated while optimizing the log-likelihood function by comparing calculated and theoretical quantiles. In a Q-Q-plot of the respective calculated and theoretic quantiles, the values should roughly lie on a line to conclude that the noise distribution assumption is fulfilled.

For a certain optimization problem, one might be able to directly calculate the noise parameters from the data. If the data set is rather small one should consider to apply bootstrapping and repeatedly estimate the noise parameter based on subsets of the data. If this is not possible and the noise parameters have to be estimated while performing MLE, the unknown noise parameters can either be instances of the parameter vector $\boldsymbol{\theta}$ and estimated together with reaction rates and scaling constants, or they can be analytically approximated during optimization in each evaluation of the log-likelihood function by using hierarchical optimization [Loos et al., 2018]. The idea in hierarchical optimization is to use the sufficient condition for an optimum

$$\nabla_n \ell_D(\boldsymbol{\theta}, n)|_{\hat{n}=0} \tag{2.19}$$

in order to derive the analytic approximation of the noise parameter n. The resulting approximations of n for all noise models considered in table 2.2 are shown in table 2.3.

Error model	Approximation of noise parameter in hierarchical optimization
Additive normal noise $y_j^{\mathcal{D}}(t_k) = y_j^{\mathcal{M}}(t_k) + \epsilon_{j,k}$ with $\epsilon_{j,k} \sim \mathcal{N}(0, \sigma_{j,k}^2)$	$\prod_{k=1}^{n_t} \prod_{j=1}^{n_y} \mathcal{N}(y_j^{\mathcal{D}} \| y_j^{\mathcal{M}}, \sigma_{j,k}^2)$ $\hat{\sigma}_j = \frac{1}{n_t} \sum_{k=1}^{n_t} (y_j^{\mathcal{D}}(t_k) - y_j^{\mathcal{M}}(t_k, \boldsymbol{\theta}))^2, j = 1, ..., n_t$
Multiplicative log-normal noise $y_j^{\mathcal{D}}(t_k) = y_j^{\mathcal{M}}(t_k) \cdot \nu_{j,k}$ with $\nu_{j,k} \sim \log\mathcal{N}(0, \sigma_{j,k}^2)$	$\prod_{k=1}^{n_t} \prod_{j=1}^{n_y} \log\mathcal{N}(y_j^{\mathcal{D}} \| y_j^{\mathcal{M}}, \sigma_{j,k}^2)$ $\hat{\sigma}_j = \frac{1}{n_t} \sum_{k=1}^{n_t} (log(y_j^{\mathcal{D}}(t_k)) - log(y_j^{\mathcal{M}}(t_k, \boldsymbol{\theta})))^2, j = 1, ..., n_t$
Additive laplace noise $y_j^{\mathcal{D}}(t_k) = y_j^{\mathcal{M}}(t_k) + \epsilon_{j,k}$ with $\epsilon_{j,k} \sim \mathcal{L}\text{aplace}(0, b)$	$\prod_{k=1}^{n_t} \prod_{j=1}^{n_y} \mathcal{L}\text{aplace}(y_j^{\mathcal{D}} \| y_j^{\mathcal{M}}, b)$ $\hat{b}_j = \frac{1}{n_t} \sum_{k=1}^{n_t} \|y_j^{\mathcal{D}}(t_k) - y_j^{\mathcal{M}}(t_k, \boldsymbol{\theta})\|, j = 1, ..., n_t$

Table 2.3: Analytic approximation for noise parameters for selection of noise distributions used in maximum likelihood estimation (MLE).

2.2.1.2 Optimization procedures

Once the likelihood function is defined, the aim is to infer the unknown parameters $\boldsymbol{\theta}$ by solving the optimization problem introduced in equation 2.17.

Definition 2.24. *An optimization problem is given by:*

$$\underset{\substack{\text{subject to}\\ f_i(\boldsymbol{\theta})\leq b_i, i=1,...,m}}{\text{minimize}} \quad f_0(\boldsymbol{\theta}), \tag{2.20}$$

where the vector $\boldsymbol{\theta} = \theta_1, ..., \theta_{n_\theta}$ *is the optimization variable, the function* $f_0 : \mathbb{R}^{n_\theta} \rightarrow \mathbb{R}$ *is the objective function, the functions* $f_i : \mathbb{R}^{n_\theta} \rightarrow \mathbb{R}$ *are the (in)equality constraint functions and the constants* b_i *are the limits for the constraints. A vector* $\boldsymbol{\theta}$ *is called optimal, or a solution of the optimization problem, if it has the smallest objective function value among all vectors that satisfy the constraints [Boyd and Vandenberghe, 2004].*

To solve an optimization problem, one can use a broad range of specified gradient-based algorithms, such as gradient or steepest decent methods, several Newton methods, interior point methods or trust-region-reflective methods [Boyd and Vandenberghe, 2004] with largely varying performance and applicability. In general one aims to reliably, i.e. accurately and precisely infer parameters with minimal computation time. The effectiveness of an optimization algorithm very much depends on the structure of the problem. Therefore it is helpful to define classes of optimization problems. One distinguishes between linear, convex, and non-linear optimization problems according to the following definitions.

Definition 2.25. *An optimization problem is called*

(a) linear if the objective and constraint functions are linear, i.e. satisfy
$f_i(\alpha\boldsymbol{x} + \beta\boldsymbol{y}) = \alpha f_i(\boldsymbol{x}) + \beta f_i(\boldsymbol{y}),\ i = 1, ..., m,\ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{n_\theta}$ *and* $\forall \alpha, \beta \in \mathbb{R}$.

(b) convex if the objective and constraint functions are convex, i.e. satisfy the inequality
$f_i(\alpha\boldsymbol{x} + \beta\boldsymbol{y}) \leq \alpha f_i(\boldsymbol{x}) + \beta f_i(\boldsymbol{y})\ , i = 1, ..., m,\ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{n_\theta}$ *and* $\forall \alpha, \beta \in \mathbb{R}_{\geq 0}$
with $\alpha + \beta = 1$.

(c) non-linear if the objective and constraint functions are not linear and not known to be convex [Boyd and Vandenberghe, 2004].

Additional important properties for optimization algorithm effectiveness are the number of variables and constraints, and the sparsity of the problem, i.e. the constraint function dependency on only a small number of variables [Boyd and Vandenberghe, 2004].
According to the definition above any linear optimization problem belongs to the class of convex optimization problems. Convex problems such as least squares optimization problems can in general be solved reliably and effectively, e.g. with interior point methods even for hundreds of variables and thousands of constraints in at most a few tens of seconds [Boyd and Vandenberghe, 2004].

However, this does not hold for non-linear optimization problems as even simple looking problems with only a few variables can be extremely challenging and more complex problems with hundreds of variables can even be intractable [Boyd and Vandenberghe, 2004]. In contrast to least-squares problems, MLE optimization problems are often non-convex. This of course depends on the noise model and the model observables' structure.

To solve even non-convex optimization problems efficiently and effectively, one could use algorithms which solve the non-linear optimization problem not globally, but only locally. Such algorithms can be fast and even handle large-scale non-linear problems since they only require differentiability of the functions $f_i, i = 0, 1, ..., m$, but they require a good initial guess for the parameter vector, i.e. a starting value and are sensitive to this starting value and algorithm parameter settings. Specifying the algorithm and its settings often requires experimenting [Boyd and Vandenberghe, 2004].

Local optimization algorithms can be combined with multi start procedures which sample a number of starting values from the whole parameter space and solve (simultaneously) several local optimization problems to potentially obtain the global optimum. Such approaches require a certain number of starting values and the implementation of an appropriate sampling method (e.g. latin hypercube sampling [Eliáš and Vořechovský, 2016]) for obtaining uniformly distributed starting values in order to efficiently capture the whole parameter space. To ensure that the optimization procedure converged to the global optimum, one should check if the optimal value and maximum objective function value is observed several times for different starting values. [Boyd and Vandenberghe, 2004] suggests techniques to formulate an optimization problem as a convex optimization problem. Moreover, an optimizer algorithms performance can be remarkably improved by log-transformation of the parameters, which has been suggested by [Raue et al., 2013], [Villaverde et al., 2019b] and [Kreutz, 2016] and further investigated in [Hass et al., 2019], as it often results in (more) convex optimization problems. Importantly, avoiding the estimation of noise parameters remarkably improves the convergence during optimization and simplifies the problem one aims to solve as it reduces the number of parameters [Loos et al., 2018].

2.2.2 A likelihood-free approach: approximate Bayesian computation

In case the likelihood does not exist in a closed form or is too costly to evaluate, approximate Bayesian computation (ABC) algorithms can be used to approximate the posterior distribution $P(\boldsymbol{\theta}|y^{\mathcal{D}})$ of the parameters $\boldsymbol{\theta}$ given the data $y^{\mathcal{D}}$ in order to estimate model parameters [Toni et al., 2009].

The main idea in ABC is to repeatedly generate Monte Carlo samples from the uncertain parameter distribution which can then be used to simulate from the specified model $\mathcal{M}(\boldsymbol{\theta})$ and subsequently assess the model agreement with experimental data to re-specify the parameter distribution $P(\boldsymbol{\theta}|y^{\mathcal{D}})$[Prescott and Baker, 2018]. The aim of an ABC algorithm is to converge to an accurate approximation of the posterior distribution $P(\boldsymbol{\theta}|y^{\mathcal{D}})$. As in general ABC requires a large amount of simulations, it is computationally expensive [Sisson, 2018] and several approaches have been suggested to reduce the computational costs and improve the performance by using parallelisation, low-fidelity models, abortion of unpromising simulations, and by exploring the parameter space more efficiently by using markov chain monte carlo (MCMC) or sequential monte carlo (SMC)

[Del Moral et al., 2006, Prescott and Baker, 2018, Sisson, 2018]. One of the first algorithms that has been suggested is the ABC rejection sampler.

The ABC rejection sampler first samples the unknown parameter vectors from a prior distribution $P(\boldsymbol{\theta})$ and subsequently calculates the model observables $y^{\mathcal{M}(\boldsymbol{\theta})}$ for each parameter vector according to the specified simulation model $\mathcal{M}(\boldsymbol{\theta})$. The model observables are then compared to the observed values by evaluating a distance function $d(y^{\mathcal{D}}, y^{\mathcal{M}})$ and if the distance function value is lower than a predefined threshold, the respective parameter vector gets accepted [Toni et al., 2009].

ABC rejection algorithm

```
Step 0 Sample $\boldsymbol{\theta}^*$ from a prior distribution $P(\boldsymbol{\theta})$.
Step 1 Simulate $y^{\mathcal{M}(\boldsymbol{\theta}^*)}$ from $f(y|\boldsymbol{\theta}^*)$.
Step 2 If $d(y^{\mathcal{D}}, y^{\mathcal{M}(\boldsymbol{\theta}^*)}) \leq \epsilon$, accept $\boldsymbol{\theta}^*$, otherwise reject $\boldsymbol{\theta}^*$. Continue with Step 0.
```

If the prior distribution is very different from the unknown posterior distribution, the acceptance rate is very low and the computational cost very high. This is avoided by ABC MCMC, which generates the distribution $P(\boldsymbol{\theta}|d(y^{\mathcal{D}}, y^{\mathcal{M}}) \leq \epsilon)$ by a Markov chain [Toni et al., 2009].

ABC MCMC algorithm

```
Step 0 Set $i = 0$ and initialize $\boldsymbol{\theta}_i$ with values from the parameter space.
Step 1 Propose $\boldsymbol{\theta}^*$ according to a proposal distribution $q(\boldsymbol{\theta}|\boldsymbol{\theta}_i)$.
Step 2 Simulate $y^{\mathcal{M}(\boldsymbol{\theta}^*)}$ from $f(y|\boldsymbol{\theta}^*)$.
Step 3 If $d(y^{\mathcal{D}}, y^{\mathcal{M}(\boldsymbol{\theta}^*)}) \leq \epsilon$, continue with Step 4, otherwise set $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i$ and continue with Step
       5.
Step 4 Set $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}^*$ with probability

           $\alpha = \min\left(1, \frac{P(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}_i|\boldsymbol{\theta}^*)}{P(\boldsymbol{\theta}_i)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_i)}\right)$

       and $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i$ with probability $1 - \alpha$.
Step 5 Set $i = i + 1$ and continue with Step 1.
```

The Markov chain may get stuck in regions of low probability due to low acceptance probability and correlated samples [Toni et al., 2009]. The problems of ABC rejection sampler and the ABC MCMC approach can partly be overcome by using ABC SMC algorithm for which several versions exist [Del Moral et al., 2006, Sisson, 2018].

In approximate Bayesian computation sequential monte carlo (ABC SMC), the posterior distribution is determined sequentially. In each step, an intermediate distribution of parameter values $\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(N)}$, which are called particles, is calculated. The parameter values for the current intermediate distribution are drawn with weights from the previous distribution and are perturbed with a kernel function. In each round, the weights for the next round are calculated and a gradually stricter threshold is applied for the distance acceptance, such that after a few steps this intermediate distributions of parameter values converges against the posterior distribution [Filippi et al., 2013, Toni et al., 2009].

ABC SMC algorithm

Step 0 Initialize thresholds $\epsilon_1, ..., \epsilon_T$ as descending sequence and set the population indicator $t = 0$.

Step 1 Set the particle indicator $i = 1$.

Step 2 If $t = 0$, sample $\boldsymbol{\theta}^{**}$ independently from $P(\boldsymbol{\theta})$, otherwise sample $\boldsymbol{\theta}^{*}$ from the previous population $\boldsymbol{\theta}_{t-1}^{(i)}$ with weights w_{t-1} and perturb the particle to obtain $\boldsymbol{\theta}^{**} \sim K_t(\boldsymbol{\theta}|\boldsymbol{\theta}^{*})$, where K_t is a perturbation kernel. If $P(\boldsymbol{\theta}^{**}) = 0$ repeat Step 2, otherwise simulate $y^{\mathcal{M}(\boldsymbol{\theta}^{*})}$ from $f(y|\boldsymbol{\theta}^{**})$ and if $d(y^{\mathcal{D}}, y^{\mathcal{M}(\boldsymbol{\theta}^{*})}) \geq \epsilon_t$ repeat Step 2.

Step 3 Set $\boldsymbol{\theta}_t^{(i)} = \boldsymbol{\theta}^{**}$ and calculate the weight for particle $\boldsymbol{\theta}_t^{(i)}$:

$$w_t^{(i)} = \begin{cases} 1, & \text{if } t = 0 \\ \frac{P(\boldsymbol{\theta}_t^{(i)})}{\sum_{j=1}^{N} w_{t-1}^{(j)} K_t(\boldsymbol{\theta}_{t-1}^{(j)}, \boldsymbol{\theta}_t^{(i)})}, & \text{if } t > 0. \end{cases} \tag{2.21}$$

If $i < N$, set $i = i + 1$ and go to Step 2.

Step 4 Normalize the weights $w_t^{(i)} = \frac{w_t^{(i)}}{\sum_{i=1}^{N} w_t^{(i)}}$. If $t < T$ set $t = t + 1$ and continue with Step 1.

From the last population of particles, i.e. the approximation of the posterior distribution, one can determine the posterior mean, posterior median and/ or maximum a posteriori (MAP) estimate

$$\hat{\boldsymbol{\theta}}^{MAP} = \underset{\boldsymbol{\theta}}{\text{argmax}} P(\boldsymbol{\theta}|y^{\mathcal{D}}) = \underset{\boldsymbol{\theta}}{\text{argmax}} \frac{P(y^{\mathcal{D}}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(y^{\mathcal{D}})}. \tag{2.22}$$

as point estimates and quantiles of the posterior distribution as $(1-\alpha)\%$ credibility intervals. The MAP estimator is consistent, which means it converges to the true value for large sample sizes. Moreover it is an asymptotically normal estimator but not an efficient estimator as it is biased towards the prior distribution $P(\boldsymbol{\theta})$ [Bachar et al., 2013, Schwartz, 1965].

2.2.3 Identifiability and uncertainty analysis

Every mathematical model $\mathcal{M}(\boldsymbol{\theta})$ underlies uncertainty due to several reasons. Although the parameters $\boldsymbol{\theta}$ are assumed to obtain fixed optimal values, different parameter values, different combinations of parameter values, or different model structures can lead to approximately the same values in model observables and therefore also to the same objective function values during optimization. In addition, uncertainty can result from the experimental data, as every observation is noisy due to measurement errors. Thus, not every model parameter might be uniquely identifiable while performing parameter inference. To be aware which parameters are unidentifiable and which model states are unobservable from the experimental data it is necessary to perform an identifiability and uncertainty analysis. An unidentifiable model can lead to wrong parameter estimates and subsequently to meaningless or misleading predictions [Villaverde and Banga, 2017] even in

an ideal scenario of nearly noise-free experimental data with many observations. Thus, the model should always be designed in such a way that its parameters $\boldsymbol{\theta} \in \mathcal{P}$ can in principle be inferred from the experimental data.
To tell apart uncertainties resulting from model structure and the ones resulting from both, the model specification and experimental data, one distinguishes between structural and practical parameter identifiability.

2.2.3.1 Structural identifiability

A model is structurally identifiable if it is possible to determine parameter values from measurements of the model output [Villaverde et al., 2016]. Structural identifiability analysis assumes ideal data (noise-free with a large sample size) and can be performed while specifying the model before the experiment is conducted and is therefore also termed a priori identifiability analysis. It can even be used to design the experiment [Walter, 1987] as it is independent of the exact parameter values but rather related to the model structure and type of input [Chis et al., 2011].

Definition 2.26. *A parameter θ is structurally*

(i) uniquely (globally) identifiable from a set of observables $y^{\mathcal{M}} = h(\boldsymbol{x}, \theta, k)$ if it can be evaluated uniquely from the equations for $y^{\mathcal{M}}$ plus any other information about θ_i, such that for any $\theta^* \in \mathcal{P}$ it holds

$$\mathcal{M}(\theta) = \mathcal{M}(\theta^*) \Rightarrow \theta = \theta^*.$$

(ii) locally identifiable from $y^{\mathcal{M}}$ if it has a countable number $n \geq 1$ of solutions and non-uniquely identifiable if $n > 1$. Local identifiability is fulfilled if for almost any $\theta^* \in \mathcal{P}$, there exists a neighbourhood $\mathcal{V}(\theta)$ of θ such that

$$\theta^* \in \mathcal{V}(\theta) \text{ and } \mathcal{M}(\theta) = \mathcal{M}(\theta^*) \Rightarrow \theta = \theta^*.$$

(iii) unidentifiable from $y^{\mathcal{M}}$ if it has an infinite number of solutions, that is for almost any $\theta^* \in \mathcal{P}$, there exists no neighbourhood $\mathcal{V}(\theta)$ of θ such that

$$\theta^* \in \mathcal{V}(\theta) \text{ and } \mathcal{M}(\theta) = \mathcal{M}(\theta^*) \Rightarrow \theta = \theta^*.$$

[Chis et al., 2011, Davidescu and Jørgensen, 2008, Walter, 1987, Walter and Pronzato, 1997]

If all model parameters $\theta_i, i = 1, ..., n_{n_\theta}$ are identifiable, the model $\mathcal{M}(\boldsymbol{\theta})$ is said to be identifiable. Sometimes only combinations of subsets of the parameter vector $\boldsymbol{\theta}$ (i.e. the product of two parameters) are identifiable but not its single components [Walter, 1987].
To analyse a models' structural identifiability various methods have been suggested in previous

years, i.e. power series expansion of the solution [Pohjanpalo, 1978] or differential algebra approaches [Ljung and Glad, 1994], which lead to global results but are often only applicable to linear or simple non-linear systems [Chis et al., 2011].

Local identifiability approaches are the Taylor series expansion of the sensitivity matrix [Srinath and Gunawan, 2010, Yao et al., 2003] and the differential geometry approach that uses the generating series expansion of the observables based on Lie derivatives [Brendel et al., 2006, Villaverde and Banga, 2017, Walter and Pronzato, 1996].

Others [Brendel et al., 2006, Craciun and Pantea, 2008, Davidescu and Jørgensen, 2008] investigated structural identifiability analysis of dynamic reaction networks. Craciun and Pantea [2008] and Brendel, Bonvin, and Marquardt [2006] defined a reaction rate constant k_r as identifiable if the respective column of the stoichiometric matrix $\nu_{(\cdot,r)}$ is linear independent of the remaining columns of the stoichiometric matrix. If the reaction rate constant however depends on several parameters θ_i, one still cannot conclude that all θ_i are identifiable. Davidescu and Jørgensen [2008] combine this approach with the Lie derivative approach, which allows the structural identifiability analysis of all parameters that define the reaction rate. In general, the algebraic manipulations involved in identifiability assessment can become tedious or even impossible, especially for larger systems [Chis et al., 2011].

To assess the structural identifiability of cell differentiation processes, we focus on a method introduced by Villaverde and Banga [2017] which is a differential geometry approach and can be used for the local identifiability analysis of state space ODE models by considering structural identifiability as generalization of observability. A model is observable if it is possible to reconstruct the states $\boldsymbol{x}$ of the model from its observables $y^{\mathcal{M}} = h(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{k})$. In this case two different states would lead to two different outputs and thus observability can be assessed by analysing if the mapping from $y^{\mathcal{M}} = h(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{k})$ to $\boldsymbol{x}$ is locally unique. This is done by determining the rank of the generalized observability-identifiability matrix $\mathcal{O}(\tilde{\boldsymbol{x}})$ which is defined based on Lie derivatives and given by

$$\mathcal{O}(\tilde{\boldsymbol{x}}) = \begin{pmatrix} \frac{\partial}{\partial \tilde{\boldsymbol{x}}} h(\tilde{\boldsymbol{x}}) \\ \frac{\partial}{\partial \tilde{\boldsymbol{x}}} (L_f h(\tilde{\boldsymbol{x}})) \\ \frac{\partial}{\partial \tilde{\boldsymbol{x}}} (L_f^2 h(\tilde{\boldsymbol{x}})) \\ \vdots \\ \frac{\partial}{\partial \tilde{\boldsymbol{x}}} (L_f^{n_s+n_\theta-1} h(\tilde{\boldsymbol{x}})) \end{pmatrix} \tag{2.23}$$

where $\tilde{\boldsymbol{x}} = [\boldsymbol{x}, \boldsymbol{\theta}]$, the Lie derivatives are defined by

$$\begin{aligned} L_f h(x) \quad &= \tfrac{\partial h(x)}{\partial (x)} f(x,u) + \textstyle\sum_{j=0}^{\infty} \tfrac{\partial h(x)}{\partial u^{(j+1)}} \\ \text{and} \quad & \\ L_f^i h(x) &= \tfrac{\partial L_f^{i-1} h(x)}{\partial (x)} f(x,u) + \textstyle\sum_{j=0}^{\infty} \tfrac{\partial L_f^{i-1} h(x)}{\partial u^{(j)}} u^{(j+1)}, \end{aligned} \tag{2.24}$$

and $u^{(j)}$ denotes the jth derivative of the input [Villaverde and Banga, 2017]. The Lie derivatives thereby evaluate the change of the observables $h(x)$ along the model dynamics $f(x,u)$. The model contains unidentifiable parameters or unobservable states if

$$\mathrm{rank}(\mathcal{O}(\tilde{\boldsymbol{x}})) < n_s + n_\theta.$$

Which parameters are non-identifiable and which states are unobservable can be identified by eliminating the jth column of the generalized observability-identifiability matrix and observing the ranks of the full $\mathcal{O}(\tilde{\boldsymbol{x}})$ and the reduced matrix $\mathcal{O}(\tilde{\boldsymbol{x}})_{\cdot,-j}$. The j-th parameter is non-identifiable if

$$\mathrm{rank}(\mathcal{O}(\tilde{\boldsymbol{x}})) - \mathrm{rank}(\mathcal{O}(\tilde{\boldsymbol{x}})_{\cdot,-j}) = 1.$$

2.2.3.2 Practical identifiability

If a parameter cannot be uniquely identified from the available experimental data, it is said to be practically unidentifiable [Gábor et al., 2017]. This unidentifiability can occur due to poor data quality, low sample size, lack of influence of the observables or an interdependence among parameters. Practical identifiability, which is also termed a posteriori identifiability, describes a parameter estimation accuracy and precision problem under the assumption that all model parameters are structurally identifiable [Walter, 1987]. One aims to quantify how accurate the parameter estimates are, given a certain model and data set.

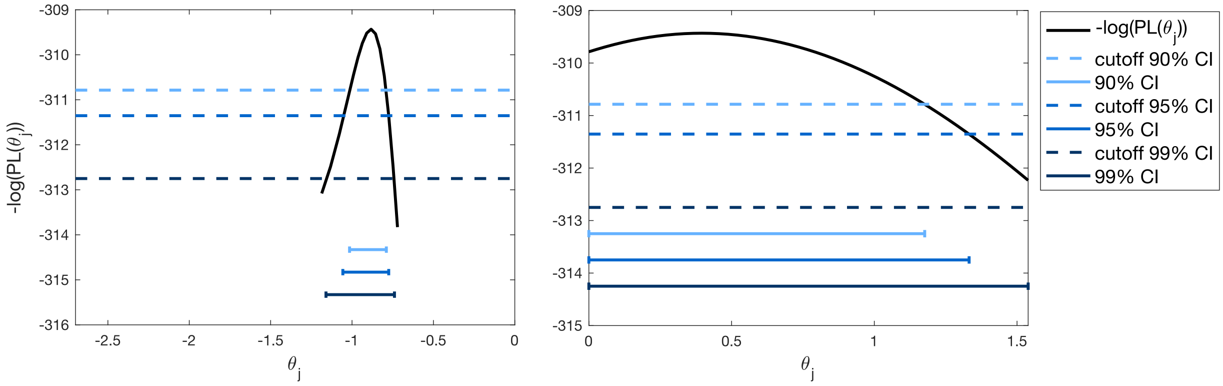


Figure 2.2: Profile likelihood (PL) for an identifiable (left) and an unidentifiable (right) parameter compared to the threshold values defining the 90%, 95% and 99% confidence interval assuming a χ^2_1 distribution.

By solving the optimization problem with the frequentist's approach, i.e. by using maximum likelihood estimation (see section 2.2.1), one accounts for data uncertainty during optimization by specifying a noise model. Thus, by re-optimizing the likelihood function at θ_j^{ML} for all remaining parameter components θ_i, $i \neq j$, the profile likelihood (PL) and parameter confidence intervals can be computed, as the following definition states.

Definition 2.27. *The PL of parameter component j is defined by*

$$PL(\theta_j) = \max_{\theta_{j \neq i}} \mathcal{L}_D(\boldsymbol{\theta}).$$

The PL-based confidence interval of parameter component j at confidence level α is then given by

$$CI_{j,\alpha} = \left\{ \theta_j \left| \frac{PL(\theta_j)}{\mathcal{L}_D(\hat{\boldsymbol{\theta}}^{ML})} > exp(-\frac{\Delta_\alpha}{2}) \right. \right\}, \tag{2.25}$$

where the threshold Δ_α is given by the α-quantiles of the inverse cumulative density function of a χ^2_1 distribution for a sufficiently large number of observations [Kreutz et al., 2013].

In case the negative PL exceeds the threshold in at least one direction, it is practically non-identifiable [Kreutz et al., 2013]. Figure 2.2 illustrates the negative log-transformed PL function of an identifiable (left) and an unindentifiable (right) parameter and the respective 90%, 95% and 99% confidence intervals assuming a χ^2_1 distribution. For the unidentifiable parameter (bottom) the lower bound of the confidence intervals are smaller than the lower parameter bound.
Several other approaches [Gábor et al., 2017, Gunawan et al., 2006, Yao et al., 2003, Zak, 2003] use the properties of the sensitivity matrix which can be derived from the output function of the model $\mathcal{M}(\boldsymbol{\theta})$ and is defined as follows.

Definition 2.28. *The sensitivities s of the output vector $(y_1, ..., y_{n_o})$ is given by the first order derivative of the model output with respect to the parameters and defined by*

$$s_{lj} = \frac{\partial y_j(t)}{\partial \theta_l}.$$

The sensitivities reflect how much the output will be affected by changes in the respective parameter value and reveal which group of parameters will cause a proportional change to the output [Srinath and Gunawan, 2010].

A particular practical identifiability assessment approach, which is based on sensitivity analysis was suggested by Gábor, Villaverde, and Banga [2017]. While calculating the root mean squared sensitivity for each parameter and comparing it to a threshold, parameters are considered to be either influential or non-influential to the output.
Let

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{N_e} \sum_{j=1}^{N_{i,y}} \sum_{k=1}^{N_{i,j,t}} w_{ikj}(y^{\mathcal{M}}_{ijk}(\boldsymbol{\theta}, t_k) - y^{\mathcal{D}}_{ijk})^2$$

be the cost function used while optimizing parameters $\boldsymbol{\theta}$, where N_e is the number of experiments, $N_{i,y}$ the number of observed quantities in the i-th experiment, $N_{i,j,t}$ the number of measured time points of the j-th observed quantity in the i-th experiment and w_{ijk} denote the weights. The scaled sensitivities for an output j and a parameter l are defined by $\tilde{s}_{lj} = \sqrt{w_j} \frac{\partial y_j}{\partial \theta_l}$. The root mean squared sensitivity is then defined as

$$\tilde{s}_i^{msqr} = \sqrt{\frac{1}{N_D} \sum_{j=1}^{N_D} \tilde{s}^2_{lj}}, \; l = 1, ..., n_\theta, \; N_D = \sum_{i=1}^{N_e} \sum_{j=1}^{N_{i,y}} \sum_{k=1}^{N_{i,j,t}} 1.$$

Their method also investigates the interplay among influential parameters by calculating a collinearity index which is defined by

$$CI_K = \frac{1}{min_{||\alpha||=1}||\tilde{s}_K \alpha||}.$$

A value of 20 means that 95% of the variation in the model output caused by changing one of the parameter in the subset can be compensated by changing the other parameter in the set. The collinearity of parametric sensitivities can result in an unidentifiable model. Subsequently they define the largest identifiable set which is the subset of parameters whose elements are influential and whose sensitivity vectors are not collinear ($CI_K < 20$).

The sensitivity analysis outcome does not only depend on parameters and model structure, but also the choice of initial conditions and external stimuli [Gábor et al., 2017, Villaverde et al., 2018]. Poor experimental data can corrupt parameter identifiability.

2.3 Model selection

While studying biological systems with mathematical models, one aims to find a model that details the most important biological properties of the process, such that it can explain the experimental data well enough, but that is also reduced to the key principles, such that it leads to a good prediction [Klipp, 2010]. Based on the considered process, one has to decide which prior knowledge to incorporate, which model structure and level of complexity to consider, and which parameter boundaries and constraints to assume while performing parameter inference. A model with low complexity will obtain a low variance but a high bias and under-fit the data as it can describe the data precisely but not accurately. Any additional model parameter will result in a more complex model that will be able to more accurately describe the data and thereby decrease the bias, but will also increase the variance and thereby the chance to over-fit the data. This should be prevented as model predictions could be misleading and wrong. In general, one aims to find the model complexity that minimizes the variance and the bias and thereby balance goodness of fit and parsimony, which is known as bias-variance trade-off (see Figure 2.3).

In order to find the necessary model complexity to explain the data well enough but at the same time avoid over-fitting, one can consider several models with varying complexities and perform model selection [Guthery, 2008]. Applying model selection can also be useful to test and rank several pre-existing biological theories in order to investigate which one is most likely given the data set. In order to define the set of competing models one could alternatively (or additionally) use a computational strategy such as step wise model selection procedures (see section 2.3.1).

This section introduces various possibilities to assess model performance on the given data set and thereby computationally compare and rank competing models (see sections 2.3.2 and 2.3.4). The result will of course depend on the predefined set of candidate models, which are abstract descriptions of the underlying process. No matter how complex a model is, it will never be an exact copy of the underlying biological process [Klipp, 2010]. Thus, model selection will in general not lead to identifying the 'true' model [Burnham and Anderson, 2004], but dependent on the set of candidate models it could help to identify a (set of) model(s) which provide(s) a sufficiently good approximation of the reality [Burnham and Anderson, 2003] and can be useful to answer the respective biological questions (see section 2.3.3). Subsequent model validation, i.e. comparison to data not used for the parameter inference and model selection, is always recommended.

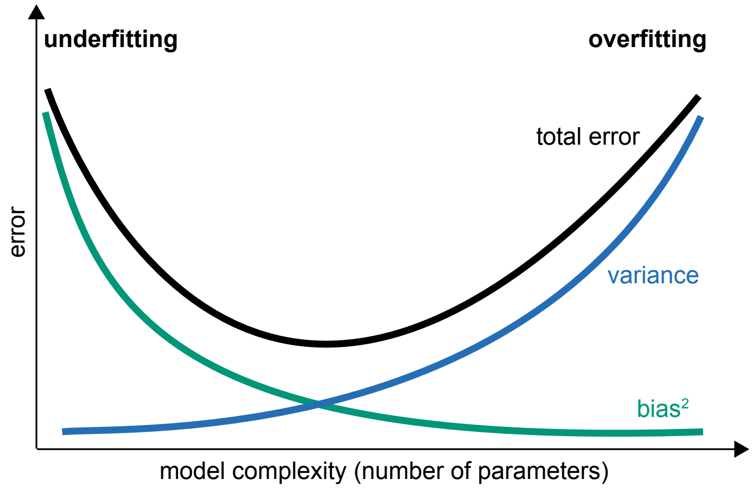


Figure 2.3: Bias-variance trade-off showing variance (blue line) increases and bias^2 (green line) declines with increasing model complexity resulting in a total error (black line) which is minimal for medium model complexity. Model is underfitted for too few parameters and overfitted for too many parameters.

2.3.1 Stepwise model selection procedures

Step wise model selection procedures are a useful tool to define the set of plausible models and are either implemented as forward or backward parameter selection procedures. Forward selection begins with a very simple model with low complexity, which is termed null model, and subsequently adds the parameter(s) that improve(s) the model fit the most. Backward selection begins with the most complex model, which is termed the full model, and step wise eliminates the parameter(s) with the lowest impact on the model performance.

The step wise procedure can be computationally demanding if one tests all possible combinations of the parameter which is included or removed in each step. The procedure of identifying the order in which parameters should be added or removed from the model can be optimized by applying regularization techniques such as ridge, lasso or elastic net, which have first been introduced for linear regression problems [Hoerl and Kennard, 1970, Tibshirani, 1996, Zou and Hastie, 2005]. The idea is to add a parameter dependent penalization term which is proportional to some penalty strength λ to the objective function (see equation 2.15) while solving the least squares optimization problem, such that for a given penalty strength the penalization term is smaller the closer the parameter values are to zero. By choosing different values of λ one can force more or less parameters to be close to zero. In forward selection, λ is reduced successively to force less parameters to zero and in backward selection λ is increased successively to force more parameters to zero.

In maximum likelihood estimation, one can adopt this idea and integrate various penalty terms or even a combination of penalty terms [Chamroukhi and Huynh, 2018] while defining a penalized log-likelihood function.

Definition 2.29. *Let $\ell_D(\boldsymbol{\theta})$ be the log-likelihood function which one aims to maximize in order to* *solve the optimization problem*

$$\hat{\boldsymbol{\theta}}^{ML} = \underset{\substack{\boldsymbol{\theta} \\ \textit{subject to } \mathcal{M}}}{\arg\max}\ \ell_D(\boldsymbol{\theta}).$$

The penalized log-likelihood function is then defined as

$$\tilde{\ell}_D(\boldsymbol{\theta}) = \sum_{k=1}^{n_t}\sum_{j=1}^{n_y} \log P(y_j^{\mathcal{D}}(t_k)|y_j^{\mathcal{M}}(t_k), \omega) - p(\lambda, \theta_i), \tag{2.26}$$

where $p(\lambda, \theta_i)$ is the penalty function (see table 2.4) and $\mathcal{T}$ is the index set containing all indices of parameters which could be excluded or included in the model.

Regularization technique	Penalty function $p(\lambda, \theta_i)$	Properties
Lasso/ L_1 regularization	$\lambda \cdot \sum_{i\in\mathcal{T}} \lvert\theta_i\rvert$	Promotes sparsity of coefficients (forces θ_i to zero)
Ridge/ L_2 regularization	$\lambda \cdot \sum_{i\in\mathcal{T}} \lvert\theta_i\rvert^2$	Promotes smaller coefficients (supports finding of reasonable θ_i)
Elastic net/ combined L_1 and L_2 regularization	$\lambda_1 \cdot \sum_{i\in\mathcal{T}} \lvert\theta_i\rvert$ $+\lambda_2 \cdot \sum_{i\in\mathcal{T}} \lvert\theta_i\rvert^2$	Combines properties of L_1 and L_2 regularization

Table 2.4: Regularization techniques which can be used for variable selection in MLE.

It can be shown that L1 regularization can be probabilistically interpreted as optimizing the likelihood function with a Laplacian prior [Park and Casella, 2008]. However, to compare the set of models which one obtained by regularized MLE, one can either use scores (introduced in section 2.3.2), or conduct likelihood ratio tests (see section 2.3.4).

2.3.2 Model selection scores

A score which is nowadays often used in model selection has been derived by Akaike [1973] in the 1970s. The Akaike information criterion (AIC) is defined as follows:

Definition 2.30. *Let n_θ be the number of parameters of model $\mathcal{M}(\boldsymbol{\theta})$. The AIC of $\mathcal{M}(\boldsymbol{\theta})$ is then given by*

$$AIC \ := \ -2\ell_{\mathcal{D}}(\boldsymbol{\theta}^{\hat{ML}}) + 2n_\theta \tag{2.27}$$

Minimization of the AIC value aims to optimize the bias-variance trade off, as the first term $-2\ell_{\mathcal{D}}(\boldsymbol{\theta}^{\hat{ML}})$ potentially decreases the more parameters are used, while the second term $2n_\theta$ increases with additional parameters.

For a large number of parameters in relation to a small sample size, a corrected version of the AIC [Akaike, 1973] has been introduced.

Definition 2.31. *Let n_θ be the number of parameters of model $\mathcal{M}(\boldsymbol{\theta})$ and n_{obs} be the number of observations used for parameter inference. The corrected AIC of $\mathcal{M}(\boldsymbol{\theta})$ is then given by*

$$AIC_c \quad := \quad -2\ell_{\mathcal{D}}(\hat{\boldsymbol{\theta}}^{ML}) + 2n_\theta + \frac{2n_\theta(n_\theta+1)}{n_{obs}-n_\theta-1}. \tag{2.28}$$

Note that for increasing n_{obs} AIC$_c$ converges to AIC. As a rule of thumb AIC$_c$ should be used instead of AIC when the ratio $\frac{n_{obs}}{n_\theta} \geq 40$ [Burnham and Anderson, 2003]. The AIC score definitions correspond to the frequentist probability interpretation which defines probability as the limit of the relative frequency of a particular event.

An alternative interpretation is the Bayesian or subjective probability, which also considers degrees of belief by including prior knowledge. From a Bayesian perspective, an exact selection method would be to perform model selection based on Bayes factors [Burnham and Anderson, 2004, Lewis and Raftery, 1997]. This however requires sampling from the posterior distribution of the respective model, which is depending on the parameter inference approach not always available (see sections 2.2.2 and 2.2.1). An approximation of Bayes factors which does not require knowledge of the posterior distribution is given by the Bayesian information criterion (BIC).

Definition 2.32. *Let n_θ describe the number of model parameters and n_{obs} the number of observations used for model fitting. The BIC [Bhat and Kumar, 2010, Schwarz, 1978] is then defined as*

$$BIC \quad := \quad -2\ell_{\mathcal{D}}(\hat{\boldsymbol{\theta}}^{ML}) + n_\theta \cdot \log(n_{obs}). \tag{2.29}$$

Calculation of these scores provides a ranking of all considered models $S_j,\ j \in J$, in which the best performing model is the one with the lowest score

$$\mathcal{M}_{\text{rank }1} = \mathcal{M}_i, \text{ if } S_i = \underset{j\in J}{min}(S_j). \tag{2.30}$$

To derive the set of plausible and implausible models usually the differences

$$\Delta_i^S \quad := \quad S_i - S_{min}, \tag{2.31}$$

with S being the AIC, AIC$_c$ or BIC score and i the index of the respective model, are calculated and model i can be rejected if $\Delta_i^S > 10$ [Burnham and Anderson, 2003].

As an alternative to the above introduced scores, one could perform Bayesian model selection using Bayes factors. These require the evaluation of marginal likelihoods of the competing models, which can for instance be realized using population annealing [Murakami, 2014], thermodynamic integration [Calderhead and Girolami, 2009, Friel and Pettitt, 2008, Hug et al., 2016, Lartillot and Philippe, 2006], or an adaptive scheduling scheme using Simpson's rule [Hug et al., 2015].

2.3.3 Model averaging

Structurally diverse models can lead to very similar values in model observables (see section 2.2.3). With model selection scores (introduced in 2.3.2) it is possible to rank these models, but sometimes one does not find a single best model but rather a set of plausible almost equally well performing models. Model averaging is a technique that allows to incorporate the inference result of the whole set of candidate models instead of identifying a single best performing model based on a score [Link and Barker, 2006, Posada and Buckley, 2004]. The idea is to calculate a weight based on the score for each model and to use this to calculate a weighted average of parameter values resulting from all considered n_m models.

As the likelihood of model $\mathcal{M}_k(\boldsymbol{\theta})$ given the data $\ell_{\mathcal{D}}(\hat{\boldsymbol{\theta}})$ is proportional to $\exp(-\frac{1}{2}\Delta_k^{\mathrm{AIC}})$ [Burnham and Anderson, 2003], the Akaike weights are defined as

$$w_k^{\mathrm{AIC}} = \frac{\exp(-\frac{1}{2}\Delta_k^{\mathrm{AIC}})}{\sum_j^{n_m}\exp(-\frac{1}{2}\Delta_j^{\mathrm{AIC}})} \tag{2.32}$$

$$= \frac{\exp(-\frac{1}{2}\mathrm{AIC}_k)}{\sum_j^{n_m}\exp(-\frac{1}{2}\mathrm{AIC}_j)}. \tag{2.33}$$

An approximation of the posterior model probability assuming identical prior probabilities for all models leads to the analogous formulation for BIC weigths [Link and Barker, 2006]:

$$P(\mathcal{M}_k|\mathcal{D}) \approx \frac{\exp(-\frac{1}{2}\mathrm{BIC}_k)}{\sum_j \exp(-\frac{1}{2}\mathrm{BIC}_j)} =: w_k^{\mathrm{BIC}}. \tag{2.34}$$

The AIC or BIC weights w^k can be used to calculate the weighted mean of parameters:

$$\overline{\hat{\boldsymbol{\theta}}_w^{ML}} = \sum_{k=1}^{n_m} \hat{\boldsymbol{\theta}}_k^{ML} \cdot w_k, \tag{2.35}$$

and the standard error of the weighted mean of parameters [Cochran, 1977, Gatz and Smith, 1995]:

$$\begin{aligned} SE_{\overline{\hat{\boldsymbol{\theta}}_w^{ML}}} = \frac{n_m}{(n_m-1)(\sum_{k=1}^{n_m} w_k)} \Bigg(&\sum_{k=1}^{n_m}(\theta_k^{ML} \cdot w_k - \overline{w} \cdot \overline{\boldsymbol{\theta}_w^{ML}})^2 \\ &- 2\overline{\hat{\boldsymbol{\theta}}_w^{ML}} \sum_{k=1}^{n_m}(w_k - \overline{w})(\hat{\boldsymbol{\theta}}_k^{ML} \cdot w_k - \overline{w} \cdot \overline{\theta_w^{ML}}) \\ &+ \overline{\hat{\boldsymbol{\theta}}_w^{ML}}^2 \sum_{k=1}^{n_m}(w_k - \overline{w})^2 \Bigg). \end{aligned} \tag{2.36}$$

2.3.4 Model selection with statistical tests

For the comparison of two nested models $\mathcal{M}_1$ and $\mathcal{M}_2$ the likelihood ratio test (LRT) is widely used [Klipp, 2010]. The test statistic of the LRT is given by

$$T(\mathcal{M}_1, \mathcal{M}_2) = 2\log\left(\frac{\mathcal{L}_{\mathcal{D}}^{\mathcal{M}_2}(\hat{\theta})^{ML}}{\mathcal{L}_{\mathcal{D}}^{\mathcal{M}_1}(\theta)}\right) = 2(\ell_{\mathcal{D}}^{\mathcal{M}_2}(\theta) - \ell_{\mathcal{D}}^{\mathcal{M}_1}(\theta)). \tag{2.37}$$

The null hypothesis $\mathcal{H}_0$: "Model $\mathcal{M}_1$ (which is nested within model $\mathcal{M}_2$) is true." can be rejected if the test statistic exceeds the 95% quantile of the χ^2 distribution with $(n_\theta^{\mathcal{M}_2} - n_\theta^{\mathcal{M}_1})$ degrees of freedom [Lewis et al., 2010], i.e. if

$$T(\mathcal{M}_1, \mathcal{M}_2) > q^{\chi^2}_{0.95,\ (n_\theta^{\mathcal{M}_2} - n_\theta^{\mathcal{M}_1})}.$$

Note that the difference in AIC, corrected AIC and BIC values of two models $\mathcal{M}_1$ and $\mathcal{M}_2$ is equal to

$$\begin{aligned}
\Delta^{\text{AIC}}_{\mathcal{M}_1,\mathcal{M}_2} &= \text{AIC}_{\mathcal{M}_1} - \text{AIC}_{\mathcal{M}_2} = T(\mathcal{M}_1, \mathcal{M}_2) - 2(n_\theta^{\mathcal{M}_2} - n_\theta^{\mathcal{M}_1}) \\
\Delta^{\text{AIC}_c}_{\mathcal{M}_1,\mathcal{M}_2} &= T(\mathcal{M}_1, \mathcal{M}_2) - 2\left(n_\theta^{\mathcal{M}_2} + \frac{n_\theta^{\mathcal{M}_2}(n_\theta^{\mathcal{M}_2}+1)}{n_{obs} - n_\theta^{\mathcal{M}_2} - 1} - n_\theta^{\mathcal{M}_1} - \frac{n_\theta^{\mathcal{M}_1}(n_\theta^{\mathcal{M}_1}+1)}{n_{obs} - n_\theta^{\mathcal{M}_1} - 1}\right) \\
\Delta^{\text{BIC}}_{\mathcal{M}_1,\mathcal{M}_2} &= T(\mathcal{M}_1, \mathcal{M}_2) - log(n_{obs})(n_\theta^{\mathcal{M}_2} - n_\theta^{\mathcal{M}_1}).
\end{aligned} \tag{2.38}$$

By applying a threshold of 10 to the score difference in order to reject $\mathcal{M}_1$, $T(\mathcal{M}_1, \mathcal{M}_2)$ would have to exceed critical values of

$$\begin{aligned}
v^{\text{AIC}}_{crit} &= 10 + 2(n_\theta^{\mathcal{M}_2} - n_\theta^{\mathcal{M}_1}), \\
v^{\text{AIC}_c}_{crit} &= 10 + 2\left(n_\theta^{\mathcal{M}_2} + \frac{n_\theta^{\mathcal{M}_2}(n_\theta^{\mathcal{M}_2}+1)}{n_{obs} - n_\theta^{\mathcal{M}_2} - 1} - n_\theta^{\mathcal{M}_1} - \frac{n_\theta^{\mathcal{M}_1}(n_\theta^{\mathcal{M}_1}+1)}{n_{obs} - n_\theta^{\mathcal{M}_1} - 1}\right) \\
v^{\text{BIC}}_{crit} &= 10 + \log(n_{obs})(n_\theta^{\mathcal{M}_2} - n_\theta^{\mathcal{M}_1}).
\end{aligned} \tag{2.39}$$

Figure 2.4 shows critical values of the three model selection criteria to reject the null hypothesis $\mathcal{H}_0$ for increasing difference in the number of parameters $n_\theta^{\mathcal{M}_2} - n_\theta^{\mathcal{M}_1}$ and varying significance levels α in case of LRT and varying number of observations n_{obs} in case of BIC. According to the critical values, BIC is more conservative compared to AIC and LRT for the recommended threshold of a score difference equal to 10. Only the corrected AIC score can be a more conservative criterion than the BIC, but only if the ratio $\frac{n_\theta^{\mathcal{M}_1}}{n_{obs}}$ is large enough (such that $v^{\text{BIC}}_{crit} < v^{\text{AIC}_c}_{crit}$ is fulfilled). The larger the data set, the closer $v^{\text{AIC}_c}_{crit}$ will be to v^{AIC}_{crit}. LRT is the least conservative method, even for a significance level equal to 0.01.

AIC and BIC can be easily used to compare a set of models which are not nested. The LRT however would require some adaptation to be applicable to compare non-nested models, as the distribution of the test statistic under the null hypothesis is not χ^2 distributed in this case. A possible solution to this would be to simulate the distribution of the test statistic under $\mathcal{H}_0$ and calculate the quantile of the simulated distribution to decide if the null hypothesis can be rejected, as described in detail by [Lewis et al., 2010]. As it is also not clear which model one should specify as the null model, it is recommended to conduct two tests to investigate if none, both or one of the models can be rejected [Williams, 1970]. For the comparison of a large set of models, this simulation based LRT approach can however get computationally very demanding.

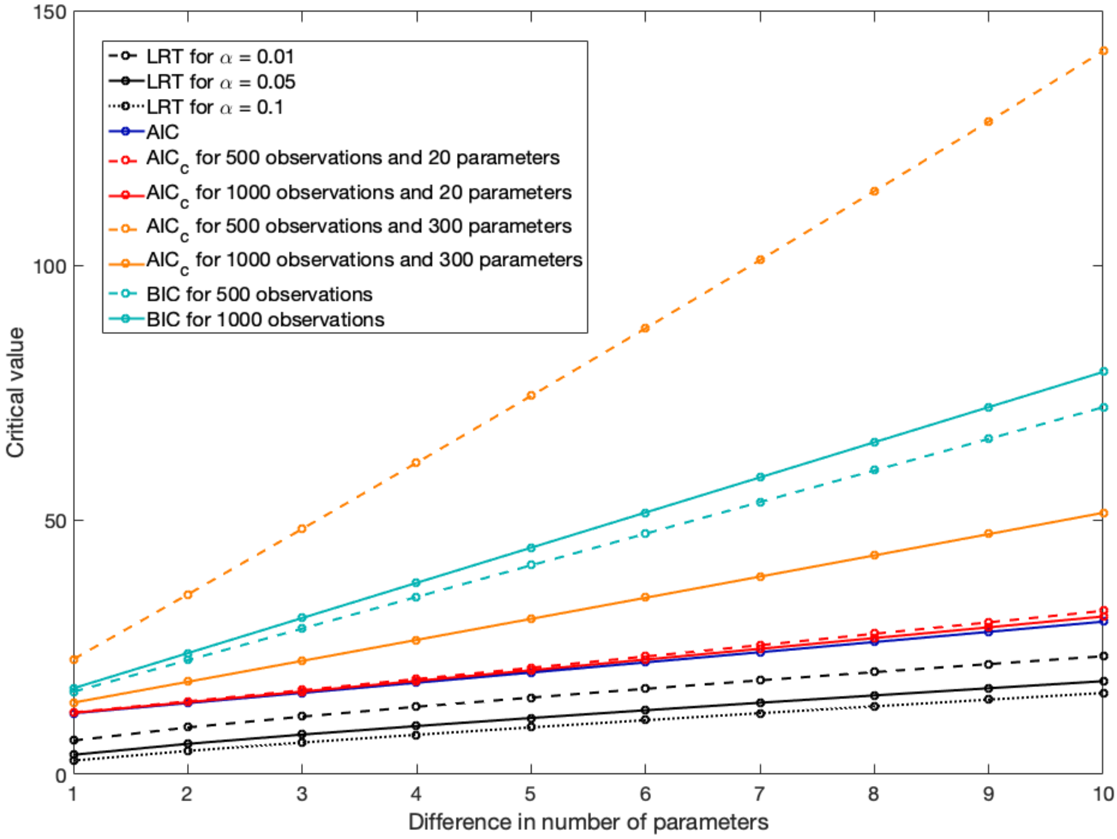


Figure 2.4: Critical values of different model selection criteria (LRT, AIC, corrected AIC, and BIC) for increasing difference in the number of parameters in models $\mathcal{M}_1$ and $\mathcal{M}_2$.

3 Application I: Modeling adult neurogenesis to infer age-related changes and division modes

A first example for a computational cell division and differentiation model is explained in detail in this chapter. The content of this chapter has been published in Cell Reports [Bast et al., 2018] and was restructured and slightly modified for my dissertation. It stems from a collaborative project with the biologists Dr. Filippo Calzolari and Prof. Jovica Ninkovic, and the computational biologists Dr. Carsten Marr, Dr. Michael Strasser, Prof. Jan Hasenauer and Prof. Fabian Theis. The data result from experiments that have been performed by Filippo Calzolari. My contribution was

(i) literature search about adult neurogeneisis to define biologically meaningful model structure, assumptions, parameter boundaries and constraints, and
(ii) the computational analysis, including implementation of model specification, parameter inference, model selection and prediction.

In this chapter I will derive stochastic models describing the mechanisms of adult neurogenesis on the macroscopic cellular level (see section 3.2) and use them to infer age-reated changes in model parameters in a systematic way and by using experimental data. As the respective experiment (see section 3.1) was performed with young and aged adult mice, I will model both groups separately and subsequently perform parameter inference (see section 3.3) to determine the difference in model parameters. In order to identify the most likely division mode of proliferating cell states, I will compare a set of 64 competing models for each group with the Bayesian Information Criterion (see section 3.5.2). The results will then be validated with dependent and independent experimental data (see section 3.5), which allows us to then use model simulations as predictions (see section 3.6) to draw further conclusions.
The code accompanying the analysis described in this chapter is publicly available at
https://github.com/marrlab/NeurogenesisAnalysis.

3.1 Biological background and experimental data

Neurogenesis describes the process of generating neurons from neural stem cells via some progenitor cell types. In the mammalian brain this process not only occurs during development, but continues throughout life in certain regions [Kriegstein and Alvarez-Buylla, 2009]. In mice, one of these brain regions that harbors adult neural stem cells is the subependymal zone (SEZ) [Ninkovic and Götz, 2013]. Upon differentiation, the cells migrate along the rostal migatory stream till they finally arrive at the olfactory bulb where they differentiate to neurons [Ninkovic and Götz, 2013] or die

[Platel et al., 2010]. The neural stem cells located at the SEZ have only a limited capacity to self-renew and generate declining numbers of olfactory bulb neurons with age [Bouab et al., 2011]. It was experimentally proven that there exists a pool of stem cells which does not actively divide [Daynac et al., 2016, Shook et al., 2012]. Which alterations with age result in this declining but retaining neurogenesis has not been identified so far.

To understand the process of adult neurogenesis in this particular region, an *in vivo* lineage tracing experiment was performed in which adult neural stem cells of the SEZ were clonally labelled using confetti reporters in young (3 months old) and middle-aged (1 year old) mice. Thus cells observed with the same color label are assumed to belong to the same clone, i.e. result from the same stem cell. To ensure that the probability to label more than one stem cell is negligibly small, the dosage of the label-inducing drug tamoxifen was estimated prior to the experiment [Calzolari et al., 2015]. Per mouse and brain hemisphere up to three clones could be observed. The animals were sacrificed at 7, 21, 35 and 56 days post-labelling, their brain hemispheres were cut at a thickness of $80\mu m$ at the microtome. The resulting sections were stained and analysed with a confocal laser scanning microscope. Based on the cell morphology, the location of the cells and the immunohisto fluorescent staining, the respective cell type (neural stem cell, transitamplifying progenitor, neuroblast or neuron) was assigned to each cell, see Figure 3.1.

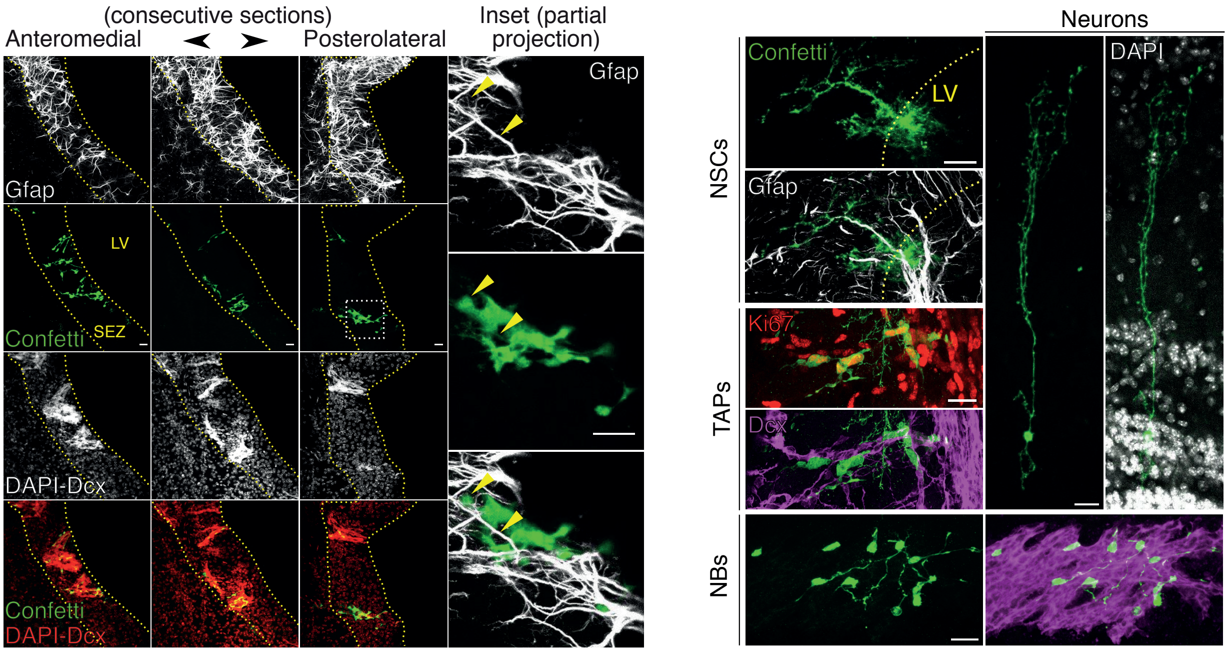


Figure 3.1: Progeny of a clone distributed across three consecutive SEZ sections in a 1-year old brain (left). Progeny classification into four cell types: NSC, transitamplifying progenitor (TAP), neuroblast (NB) and neuron (N) via marker expression (right). Dashed curves indicate borders between SEZ and lateral ventricle, dashed box highlights the inset. Scale bars correspond to $20\mu m$. Yellow arrowheads point to GFAP signal in the soma and radial process. Graphic is taken from Bast et al. [2018].

In detail, GFAP positive cells are classified as NSCs and Dcx positive cells as NBs. TAPs and Ns were defined by a combination of lack of marker expression, localization and morphology. The proliferation marker Ki67 is shown to confirm the TAP identity of SEZ-localized Dcx-negative cells, but was not regularly used to identify cells.

The experimental data contain the number of cells per cell type observed for each clone, the number of days post-labelling and the age of the mice, see Figure 3.2 and table 3.1. Using this data set, the goal is to infer parameters of a model describing the molecular mechanisms of adult neurogenesis in order to identify the changes between the two groups (young and aged mice).

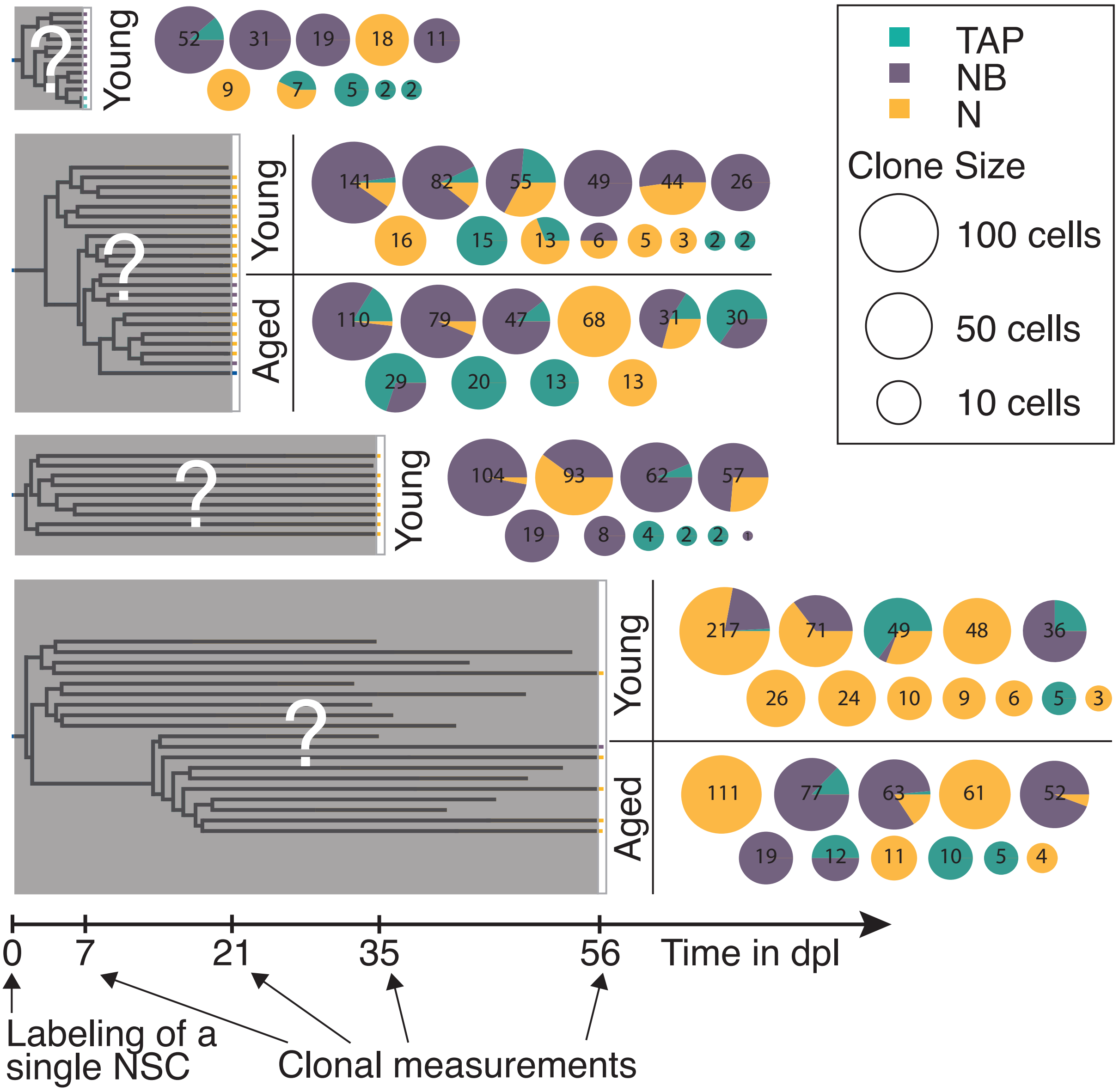


Figure 3.2: Experimental design showing the clonal progeny of a single labelled NSC is observed at one of four different time points 7, 21, 35 and 56 days post labelling (dpl) in young and two different time points 21 and 56 dpl in aged mice. The resulting data is shown as pie charts indicating the number and cell type (TAP, NB and N) composition of clones observed at each time point. Graphic is taken from Bast et al. [2018].

Time in days post labelling (dpl)	young			aged		
	TAP	NB	N	TAP	NB	N
7	0	11	0			
	0	19	0			
	6	46	0			
	2	0	0			
	0	0	18			
	0	31	0			
	2	0	0			
	5	0	0			
	3	0	4			
	0	0	9			
21	15	0	0	5	17	9
	0	49	0	0	0	13
	6	0	0	5	42	0
	3	124	14	13	0	0
	2	0	0	21	9	0
	0	26	0	0	74	5
	0	23	21	19	10	0
	0	3	3	18	90	2
	0	0	3	20	0	0
	13	24	18	0	0	68
	0	0	16			
	0	0	5			
	4	0	9			
35	0	42	15			
	2	0	0			
	0	19	0			
	0	8	0			
	0	101	3			
	4	58	0			
	0	37	56			
	0	1	0			
	4	0	0			
	2	0	0			
56	0	0	26	1	52	10
	0	25	46	0	0	11
	9	27	0	5	0	0
	2	46	169	0	0	111
	0	0	24	0	0	61
	32	2	15	0	49	3
	0	0	9	0	19	0
	0	0	10	0	0	4
	5	0	0	10	0	0
	0	0	48	10	67	0
	0	0	6	6	6	0
	0	0	3			

Table 3.1: Number of counted TAPs, NBs and Ns, in young and aged adult mice at different days post labelling in clonal lineage tracing experiments.

3.2 Specification of a set of candidate models

The experimental observations describe cell type and number of the progeny, which is assumed to result from a single labelled neural stem cell. They exhibit a strong heterogeneity in clone size and clonal composition (see Figure 3.2). This suggests that the underlying process is stochastic and cannot be described fully with a deterministic model. As it has been shown that tissue homoeostasis on the cellular level can be accurately described by stochastic models [Gardiner, 2009, Klein and Simons, 2011], the underlying dynamics of adult neurogenesis are modeled as a Markov jump process [Fröhlich et al., 2016, Resat et al., 2009] using the CME (see section 2.1.2.2). As solving the CME is due to its infinite dimension in general not possible [Resat et al., 2009], the link between CME and moment equations [Sotiropoulos and Kaznessis, 2011] is used and the first and second order moment equations [Fröhlich et al., 2016, Resat et al., 2009] are solved instead.

Model scheme

According to experimental observations, the process of adult neurogenesis can be described as follows: starting from the pool of dormant stem cells (dS), which is depleted over time, stem cells can then be activated and inactivated by switching between the quiescent (qS) and active (aS) state [Basak et al., 2012, Costa et al., 2011, Daynac et al., 2016, Shook et al., 2012, Urbán et al., 2016]. Active stem cells (aS), transitamplifying progenitors (TAPs) and neuroblasts of type I (NB I) proliferate [Ponti et al., 2013]. In contrast, Neuroblasts of type II (NB II) do not divide, but migrate along the SEZ to the olfactory bulb [Petreanu and Alvarez-Buylla, 2002] where they eventually become neuroblasts of type III, that are either depleted via cell death, or become neurons (N) [Platel et al., 2010]. All possible transitions between the cell states are schematically depicted in Figure 3.3.

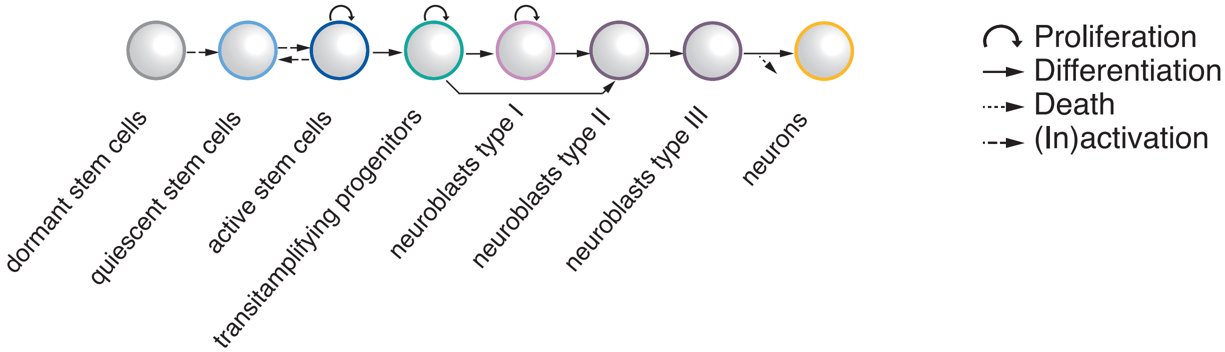


Figure 3.3: Model scheme for adult neurogenesis describing stem cell activation and inactivation, proliferation and differentiation of stem and progenitor cells, migration of type II neuroblasts, and neuroblast cell death. Graphic is taken from Bast et al. [2018].

For the three proliferating cell states (aS, TAP, NB I) I introduced four different division modes (see Figure 3.4): asymmetric (A), symmetric (S), constrained (C), where the proportion of symmetric and asymmetric divisions is regulated by a single parameter p_d, and unconstrained (U), where any combination of asymmetric division, self-renewal and symmetric differentiation probabilities is allowed. An asymmetric division is defined as a cell division followed by the transition of only one daughter cell to the next possible cell type in the model before it possibly divides again (see section 1.1). The other daughter cell initially persists in the "parental" cell state, to later divide again or undergo transitions as allowed by the model (e.g. in the case of an aS, to return to a quiescent qS state).

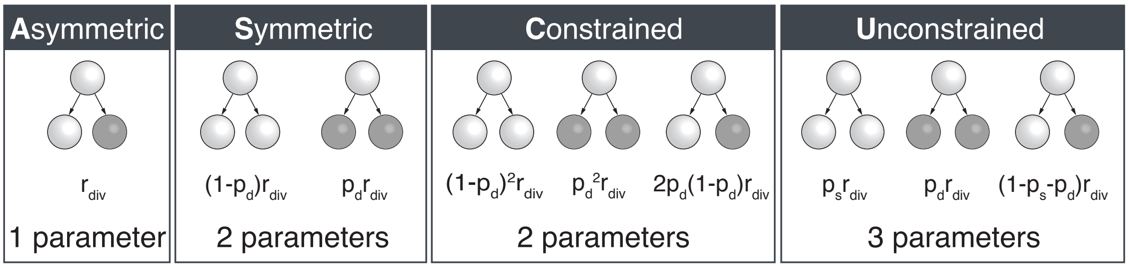


Figure 3.4: Division modes for dividing cell types: Asymmetric divisions (A) give rise to a daughter cell of the same type and a daughter cell of the subsequent type, symmetric divisions (S) produce two daughters of the same cell type, constrained divisions (C) assume independent differentiation between sister cells, while the unconstrained division (U) is the most flexible mode. The number of model parameters increases from left to right with equal model complexity for modes S and C. Graphic is taken from Bast et al. [2018].

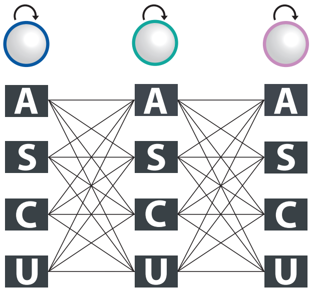


Figure 3.5: Combinations of the four division modes for dividing aS, TAP, and NB I lead to 64 different models. Graphic is taken from Bast et al. [2018].

It should be noted that while the model directly couples certain cell state transitions to cell division, some of these cell fate choices may in reality also happen some time after the cell has divided. My approach allows exploration of a very diverse set of proliferative behaviours. Combining the four different division modes across three proliferative compartments results in a set of $4^3 = 64$ candidate models (see Figure 3.5) with a varying number of parameters and complexity. According to the model scheme (see Figure 3.3), every cell state is modelled as a compartment and according to the transitions between the compartments, all occurring reactions and the respective Chemical Master Equation (see section 2.1.2.2) can be derived.

Cell states and model parameters

Let $S = (dS, qS, aS, T, B_1, B_2, B_3, N) \in \mathcal{S}$ denote the vector of cell state variables (dormant, quiescent, active neural stem cells, transitamplifying progenitors, neuroblasts type I, neuroblasts type II, neuroblasts type III and neurons) and

$$\theta = \begin{pmatrix} r_{act1} \\ r_{act2} \\ r_{inact} \\ r_{div} \\ p_{S\to SS} \\ p_{S\to TT} \\ p_{T\to TT} \\ p_{T\to BB} \\ p_{B_1} \\ p_{B_1\to B_1B_1} \\ p_{B_1\to B_2B_2} \\ r_{mig} \\ p_N \\ p_{dS0} \\ p_{qS0} \end{pmatrix}$$

the vector of model parameters, consisting of (in)activation ($r_{act1}, r_{act2}, r_{inact}$), division ($r_{div}$) and neuroblast migration (r_{mig}) rates, and probabilities of self-renewal ($p_{S\to SS}, p_{T\to TT}, p_{B_1\to B_1B_1}$), of differentiation ($p_{S\to TT}, p_{T\to BB}, p_{B_1\to B_2B_2}$), of a TAP differentiating to neuroblast type I (p_{B_1}), of a neuroblast type III to become a neuron (p_N) and of a dormant and quiescent stem cell getting initially labelled (p_{dS0}, p_{qS0}). Note that the remaining probabilities are not part of the parameter vector as they can be defined as complementary probabilities

$$\begin{aligned} p_{S\to ST} &= 1 - p_{S\to SS} - p_{S\to TT}, \\ p_{T\to TB} &= 1 - p_{T\to TT} - p_{T\to BB}, \\ p_{B_1\to B_1B_2} &= 1 - p_{B_1\to B_1B_1} - p_{B_1\to B_2B_2}, \\ p_{aS0} &= 1 - p_{dS0} - p_{qS0}, \text{ and} \\ p_{B_2} &= 1 - p_{B_1}, \\ p_{B_3} &= 1 - p_N, \end{aligned} \tag{3.1}$$

where p_{B_2} is the probability of a TAP to differentiate to a neuroblast type II and p_{B_3} the probability of a neuroblast type III to die.

Reactions

The reactions of a particular model depend on the division mode of the proliferating cell states (see Figure 3.4). For the most general model, in which all proliferating cell states divide according to the unconstrained (U) division mode, the $n_r = 17$ reactions are:

$$
\begin{aligned}
R_1 :&\quad \mathrm{dS} \xrightarrow{r_{act1}} \mathrm{qS} \\[2em]
R_2 :&\quad \mathrm{qS} \xrightarrow{r_{act2}} \mathrm{aS} \\[2em]
R_3 :&\quad \mathrm{aS} \xrightarrow{r_{inact}} \mathrm{qS} \\
R_4 :&\quad \mathrm{aS} \xrightarrow{r_{div} \cdot p_{S \to SS}} 2\,\mathrm{aS} \\
R_5 :&\quad \mathrm{aS} \xrightarrow{r_{div} \cdot p_{S \to TT}} 2\,\mathrm{T} \\
R_6 :&\quad \mathrm{aS} \xrightarrow{r_{div} \cdot (1 - p_{S \to SS} - p_{S \to TT})} \mathrm{T} + \mathrm{aS} \\[2em]
R_7 :&\quad \mathrm{T} \xrightarrow{r_{div} \cdot p_{T \to TT}} 2\,\mathrm{T} \\
R_8 :&\quad \mathrm{T} \xrightarrow{r_{div} \cdot p_{T \to BB} \cdot p_{B_1}^2} 2\,\mathrm{B}_1 \\
R_9 :&\quad \mathrm{T} \xrightarrow{r_{div} \cdot p_{T \to BB} \cdot (1 - p_{B_1})^2} 2\,\mathrm{B}_2 \\
R_{10} :&\quad \mathrm{T} \xrightarrow{r_{div} \cdot p_{T \to BB} \cdot 2 \cdot p_{B_1} (1 - p_{B_1})} \mathrm{B}_1 + \mathrm{B}_2 \\
R_{11} :&\quad \mathrm{T} \xrightarrow{r_{div} \cdot (1 - p_{T \to TT} - p_{T \to BB}) \cdot p_{B_1}} \mathrm{B}_1 + \mathrm{T}
\end{aligned}
\tag{3.2}
$$

$$
\begin{aligned}
R_{12} :&\quad \mathrm{T} \xrightarrow{r_{div} \cdot (1 - p_{T \to TT} - p_{T \to BB}) \cdot (1 - p_{B_1})} \mathrm{B}_2 + \mathrm{T} \\[2em]
R_{13} :&\quad \mathrm{B}_1 \xrightarrow{r_{div} \cdot p_{B_1 \to B_1 B_1}} 2\,\mathrm{B}_1 \\
R_{14} :&\quad \mathrm{B}_1 \xrightarrow{r_{div} \cdot p_{B_1 \to B_2 B_2}} 2\,\mathrm{B}_2 \\
R_{15} :&\quad \mathrm{B}_1 \xrightarrow{r_{div} \cdot (1 - p_{B_1 \to B_1 B_1} - p_{B_1 \to B_2 B_2})} \mathrm{B}_1 + \mathrm{B}_2 \\[2em]
R_{16} :&\quad \mathrm{B}_2 \xrightarrow{r_{mig}} \mathrm{B}_3 \\[2em]
R_{17} :&\quad \mathrm{B}_3 \xrightarrow{1000 p_N} \mathrm{N} \\
R_{18} :&\quad \mathrm{B}_3 \xrightarrow{1000 p_{B3}} \emptyset\,\cdot
\end{aligned}
$$

Chemical master equation (CME)

Let $P(\mathbf{x}|t)$ be the probability to be in a certain state $\mathbf{x}$, that is to observe a certain number of cells in states $dS, qS, aS, T, B_1, B_2, B_3$ and N, at time t. The CME describes the change of P over time (see section 2.1.2.2). For the reactions of the most general model (see equation 4.1), the CME is defined by

$$\frac{dP(\mathbf{x}|t)}{dt} = \sum_{r=1}^{18} a_{(r)}(\mathbf{x} - \nu_{(\cdot,r)})P(\mathbf{x} - \nu_{(\cdot,r)}|t) - a_{(r)}(\mathbf{x})P(\mathbf{x}|t), \tag{3.3}$$

where $\boldsymbol{x}$ is the state vector $\boldsymbol{x} = (x_{dS}, x_{qS}, x_{aS}, x_T, x_{B_1}, x_{B_2}, x_{B_3}, x_N)$, $\nu_{(\cdot,r)}$ indicates the rth column of the stoichiometric matrix

$$\nu = \begin{array}{c} \\ dS \\ qS \\ aS \\ T \\ B_1 \\ B_2 \\ B_3 \\ N \end{array} \begin{array}{c} \begin{array}{cccccccccccccccccc} R_1 & R_2 & R_3 & R_4 & R_5 & R_6 & R_7 & R_8 & R_9 & R_{10} & R_{11} & R_{12} & R_{13} & R_{14} & R_{15} & R_{16} & R_{17} & R_{18} \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 1 & 1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 1 & 1 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 & 0 & 1 & 0 & 2 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & +1 & 0 \end{bmatrix} \end{array}$$

and $a_{(r)}$ the rth entry of the propensity vector

$$a(\mathbf{x}) = \begin{pmatrix} r_{act1} \cdot [dS] \\ r_{act2} \cdot [qS] \\ r_{inact} \cdot [aS] \\ r_{div} \cdot p_{S\to SS} \cdot [aS] \\ r_{div} \cdot p_{S\to TT} \cdot [aS] \\ r_{div} \cdot (1 - p_{S\to SS} - p_{S\to TT}) \cdot [aS] \\ r_{div} \cdot p_{T\to TT} \cdot [T] \\ r_{div} \cdot p_{T\to BB} \cdot p_{B_1}^2 \cdot [T] \\ r_{div} \cdot p_{T\to BB} \cdot (1 - p_{B_1})^2 \cdot [T] \\ r_{div} \cdot p_{T\to BB} \cdot 2 \cdot p_{B_1}(1 - p_{B_1}) \cdot [T] \\ r_{div} \cdot (1 - p_{T\to TT} - p_{T\to BB}) \cdot p_{B_1} \cdot [T] \\ r_{div} \cdot (1 - p_{T\to TT} - p_{T\to BB}) \cdot (1 - p_{B_1}) \cdot [T] \\ r_{div} \cdot p_{B_1\to B_1B_1} \cdot [B_1] \\ r_{div} \cdot p_{B_1\to B_2B_2} \cdot [B_1] \\ r_{div} \cdot (1 - p_{B_1\to B_1B_1} - p_{B_1\to B_2B_2}) \cdot [B_1] \\ r_{mig} \cdot [B_2] \\ 1000 p_N \cdot [B_3] \\ 1000 p_{B3} \cdot [B_3] \end{pmatrix}.$$

Whenever a reaction R_r (for instance R_2: activation of qS) occurs, the system jumps from a particular state $\mathbf{x}$, ($\mathbf{x} = ([dS] = 0, [qS] = 1, [aS] = 0, [T] = 0, [NB]_1 = 0, [NB_2] = 0, [NB_3] = 0, [N] = 0)$) to state $\mathbf{x} + \nu_{(\cdot,r)}$ (in this case to state $([dS] = 0, [qS] = 0, [aS] = 1, [T] = 0, [NB]_1 = 0, [NB]_2 = 0, [NB]_3 = 0, [N] = 0)$). The CME therefore describes the stochastic evolution of the state vector $\mathbf{x}$.

Moment equations

As the solution of the CME is analytically and numerically intractable [Resat et al., 2009], the first and second order moment equations are calculated instead (see equation 3.5). This is done by using a definition of mean (μ_i), variance and covariance ($C_{i,j}$) (that is the first and second order moments) of cell state abundances based on the solution of the CME [Engblom, 2006].

$$\begin{aligned} \mu_i(t) &:= E[X_i(t)] = \sum_{x_i} x_i P(\mathbf{x}|t) \\ C_{i,j}(t) &:= Cov[X_i(t), X_j(t)] = \sum_{x_i,x_j} (x_i - \mu_i(t))(x_j - \mu_j(t))^T P(\mathbf{x}|t), \end{aligned} \tag{3.4}$$

with $i, j = 1, 2, ..., 7$ denoting the cell state index. I calculated the derivatives to get the evolution equations for the first and second order moment equations:

$$\begin{aligned} \frac{d\mu_i(t)}{dt} &= \sum_{r=1}^{n_r} \nu_{(i,r)} \left(a_{(r)}(\mu(t), \theta) + \frac{1}{2} \sum_{l_1,l_2} \frac{\partial^2 a_{(r)}(\mu(t), \theta)}{\partial x_{l_1} \partial x_{l_2}} C_{l_1,l_2}(t) \right) \\ \frac{dC_{i,j}(t)}{dt} &= \sum_{r=1}^{n_r} \left(\nu_{(i,r)} \sum_{l_1} \frac{\partial a_{(r)}(\mu(t), \theta)}{\partial x_{l_1}} C_{l_1,j}(t) + \nu_{(j,r)} \sum_{l_2} \frac{\partial a_{(r)}(\mu(t), \theta)}{\partial x_{l_2}} C_{i,l_2}(t) \right) \\ &+ \sum_{r=1}^{n_r} \nu_{(i,r)} \nu_{(j,r)} \left(a_{(r)}(\mu(t), \theta) + \frac{1}{2} \sum_{l_1,l_2} \frac{\partial^2 a_{(r)}(\mu(t), \theta)}{\partial x_{l_1} \partial x_{l_2}} C_{l_1,l_2}(t) \right). \end{aligned} \tag{3.5}$$

For any parameter vector θ, the first and second order moments can be generated by solving the first and second order moment equations (see equation 3.5). Note that as the reaction propensities are linear in cell states, the moments are closed and application of moment closure is not required.

Model assumptions

To solve the ODE system, initial conditions have to be specified according to the experimental setting. In previous work [Calzolari et al., 2015], the probability to label more than one stem cell was calculated to be $0.0024, 0.0115, 0.0696$ and 0.2227 at times $t = 3, 7, 21$ and 56 days post labelling for a tamoxifen dose of $10\frac{\mu g}{g}$. Accordingly, I assumed for the models that exactly one neural stem cell is labelled at $t_0 = 0$, which can be dormant, quiescent or active according to probabilities p_{dS0}, p_{qS0} and $p_{aS0} = 1 - p_{dS0} - p_{qS0}$. Considering the labelling of a certain neural stem cell state as binomial $B_{1,p.}(k), k \in 0, 1$ distributed, I derived the initial conditions for the ODE system describing the change in first and second order moments over time (see equation 3.5):

$$\begin{aligned} \mu_1(0) &= p_{dS0} & C_{1,1}(0) &= p_{dS0}(1 - p_{dS0}) \\ \mu_2(0) &= p_{qS0} & C_{2,2}(0) &= p_{qS0}(1 - p_{qS0}) \\ \mu_3(0) &= 1 - p_{dS0} - p_{qS0} & C_{3,3}(0) &= (1 - p_{dS0} - p_{qS0})(p_{dS0} + p_{qS0}) \\ \mu_k(0) &= 0, \quad k = 4, ..., 8 & C_{k,k}(0) &= 0, \quad k = 4, ..., 8 \\ & & C_{1,2}(0) &= C_{2,1}(0) = -p_{dS0} \cdot p_{qS0} \\ & & C_{1,3}(0) &= C_{3,1}(0) = -p_{dS0} \cdot (1 - p_{dS0} - p_{qS0}) \\ & & C_{2,3}(0) &= C_{3,2}(0) = -p_{qS0} \cdot (1 - p_{dS0} - p_{qS0}) \\ & & C_{k,l}(0) &= 0, \quad k, l = 4, ..., 8. \end{aligned}$$

In order to infer model parameters, the optimization problem is specified, which requires the definition of biologically meaningful parameter boundaries (see table 3.2).

parameter	fixed value	estimated		
(reference)		lower	upper	inequality
		boundary	boundary	constraint
r_{act1} [Daynac et al., 2016] [Shook et al., 2012]	0.0002	-	-	-
r_{act2}, r_{inact} [Daynac et al., 2016] [Shook et al., 2012]	-	$\frac{1}{1000}h^{-1}$	$1h^{-1}$	$-0.3 \leq r_{inact} - r_{act2} \leq 0.4$
r_{div} [Ponti et al., 2013]	-	$\frac{1}{25}h^{-1}$	$\frac{1}{15}h^{-1}$	-
$p_s(i) = p_{x_i \rightarrow x_i x_i}$, $p_d(i) = p_{x_i \rightarrow x_{i+1} x_{i+1}}$, $i = 3, 4, 5$	-	0	1	$p_s + p_d \leq 1$
p_{B_1} [Ponti et al., 2013]	0.55	-	-	-
r_{mig} [Petreanu and Alvarez-Buylla, 2002]	-	$\frac{1}{1000}h^{-1}$	$\frac{1}{10}h^{-1}d^{-1}$	-
p_N [Platel et al., 2010]	-	0.65	0.85	-
p_{dS0}, p_{qS0}	-	0	1	$p_{dS0} + p_{qS0} \leq 1$

Table 3.2: Boundaries, constraints or values assumed for model parameters

The parameter boundaries are based on findings of [Ponti et al., 2013] for the cell division rate (r_{div}) and [Petreanu and Alvarez-Buylla, 2002] for the migration rate (r_{mig}). According to the analysis of Ponti et al. [2013] which showed that only 55% of neuroblasts divide, I introduced another neuroblast state and assumed the probability for a TAP to differentiate into the proliferating neuroblast state (NB I) to be $p_{B_1} = 0.55$. Platel et al. [2010] experimentally determined the percentage of neuroblasts differentiating to neurons to be 78% in the SEZ of P20-P30 mice, I therefore estimate this percentage for the three-months- (young) and one-year-old (aged) mice by assuming a range of $[65, 85]$% in the model.

In Addition, I fitted a stem cell compartment ODE model (see Figure 3.6 and Equation 3.6) to cell counts of Shook et al. [2012] and Daynac et al. [2016].

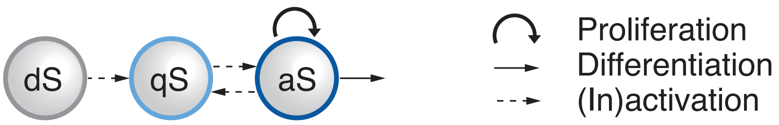
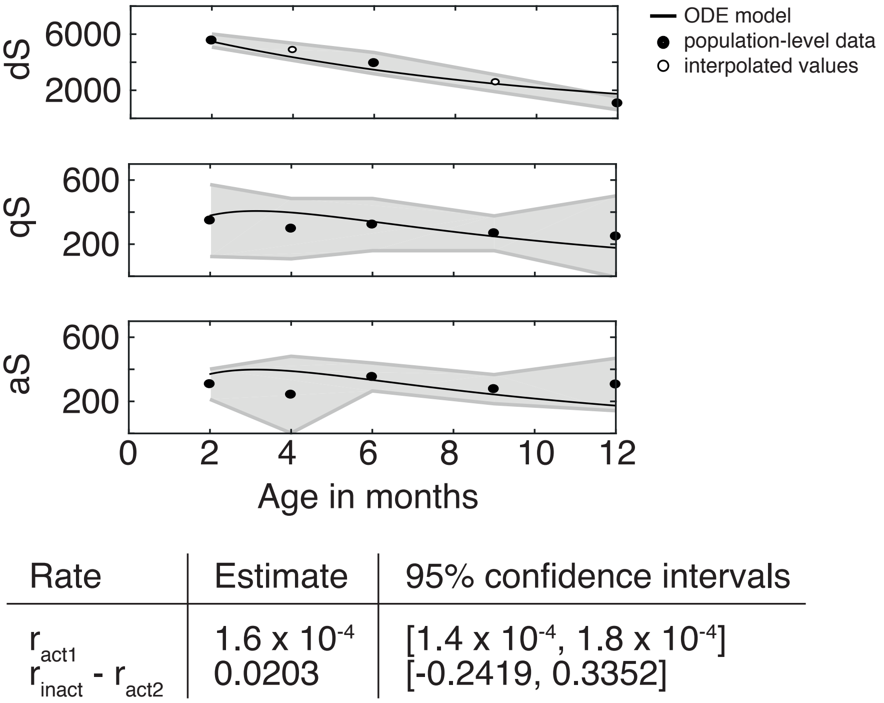


Figure 3.6: Model scheme as depicted in Figure 3.3 reduced to transitions of stem cell compartments. Graphic is taken from Bast et al. [2018].

$$\begin{aligned}
\frac{d[dS](t)}{dt} &= -r_{act1}[dS](t) &&, \frac{d[dS](0)}{dt} = dS_0 \\
\frac{d[qS](t)}{dt} &= r_{act1}[dS](t) + r_{inact}[aS](t) - r_{act2}[qS] &&, \frac{d[qS](0)}{dt} = qS_0 \\
&= r_{act1}[dS](t) + (r_{act2} + (r_{inact} - r_{act2}))[aS](t) - r_{act2}[qS] && \\
\frac{d[aS](t)}{dt} &= (-r_{diff_S} - r_{inact})[aS](t) + r_{act2}[qS] &&, \frac{d[aS](0)}{dt} = aS_0 \\
&= (-r_{diff_S} - (r_{act2} + (r_{inact} - r_{act2})))[aS](t) + r_{act2}[qS]. &&
\end{aligned} \tag{3.6}$$

The model fit and the inferred (identifiable) parameters can be seen in Figure 3.7.



Rate	Estimate	95% confidence intervals
r_{act1}	1.6×10^{-4}	$[1.4 \times 10^{-4}, 1.8 \times 10^{-4}]$
$r_{inact} - r_{act2}$	0.0203	[-0.2419, 0.3352]

Figure 3.7: Resulting model fit to population level data and inferred identifiable parameters with 95% Pl-based confidence intervals (see equation 2.27). Graphic is taken from Bast et al. [2018].

I performed this analysis using MATLAB toolboxes AMICI [Fröhlich et al., 2016] for model definition and PESTO [Stapor et al., 2017] for parameter estimation with interior point optimization algorithm. Figure 3.8 shows that the multi start optimization procedure converged and likely found the global optimum within the specified boundaries.
This pre-analysis led to two constraints for the set of considered models
(i) the dS activation rate was fixed to $r_{act1} = 0.000171$,
(ii) the difference between qS activation and aS inactivation rates was constrained to
$-0.3 \leq r_{inact} - r_{act2} \leq 0.4$.
Another inequality constraint was introduced for division mode U, in which the sum of probabilities for symmetric self-renewal and symmetric differentiation must be lower or equal to 1. The same holds for the sum of probabilities to initially label a dormant stem cell and to initially label a

quiescent stem cell.

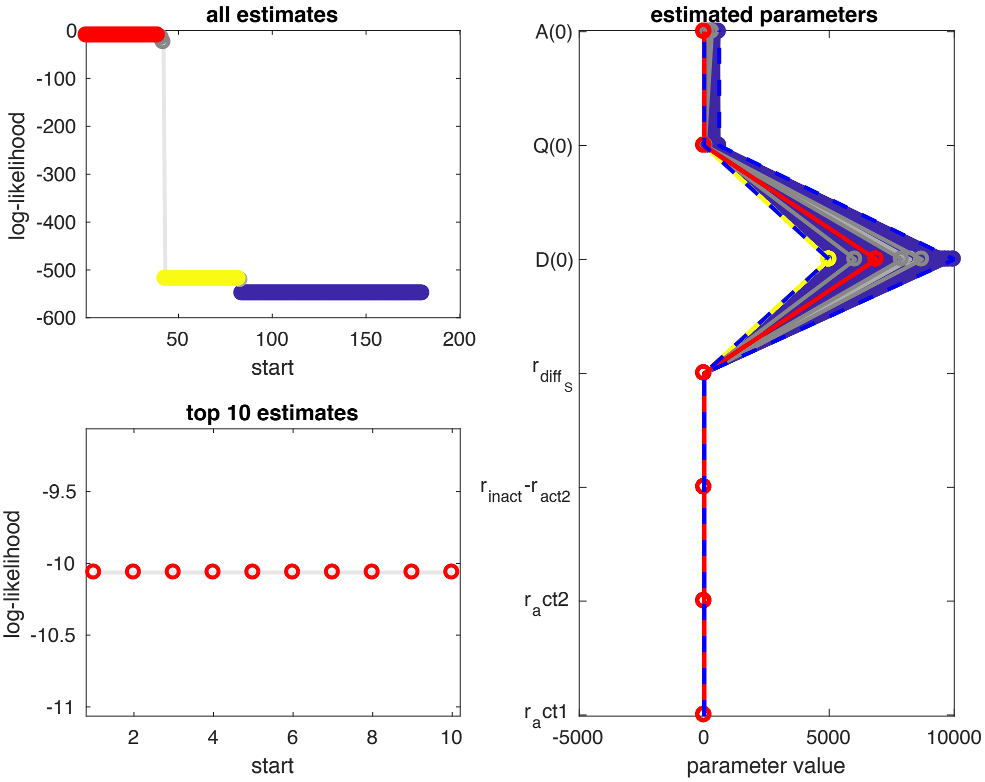


Figure 3.8: While optimizing parameters of ODE stem cell compartment model (see equation 3.6) with multi start approach the highest log likelihood value was found several times.

3.3 Parameter inference

The derived models contain between 7 and 13 unknown parameters. These parameters were estimated with maximum likelihood estimation by minimizing the discrepancy between observed and modelled first and second order moments. Moreover, I analysed the identifiability of parameters.

3.3.1 Maximum likelihood estimation

Let $\boldsymbol{x} = (x_{dS}, x_{qS}, x_{aS}, x_T, x_{B_1}, x_{B_2}, x_{B_3}, x_N)$ be the state vector and $\mathcal{M}(\theta)$ be a particular model I consider consisting of dynamics $\dot{\mathbf{x}} = f(\mathbf{x}, \theta)$ and model observables $y^{\mathcal{M}} = h(x, \theta)$, which is given by

$$\mathcal{M}(\theta) : \left\{ \begin{array}{lll} \dot{\mathbf{x}} & = f(\mathbf{x}, \theta) = & \left\{ \frac{d\mu_i(t)}{dt}, \frac{dC_{i,j}(t)}{dt} \right\}_{i,j=1,\dots,8}, \quad x_0(\theta) = x_0 \\ y^{\mathcal{M}} & = h(x, \theta) = & \{\mu_4(t), \mu_5(t) + \mu_6(t) + \mu_7(t), \mu_8(t), \\ & & C_{4,4}(t), C_{4,5}(t) + C_{4,6}(t) + C_{4,7}(t), C_{4,8}(t), \\ & & C_{5,5}(t) + 2C_{5,6}(t) + 2C_{5,7}(t) + C_{6,6}(t) + 2C_{6,7}(t) + C_{7,7}(t), \\ & & C_{5,8}(t) + C_{6,8}(t) + C_{7,8}(t), C_{8,8}(t). \end{array} \right\} \tag{3.7}$$

Furthermore, let $y^{\mathcal{D}}(t_k)$ denote the observed moments at time t_k which were directly calculated from clonal observations (see Figure 3.2, Table 3.1) and $\mathcal{D} = \left\{t_k, y^{\mathcal{D}}(t_k)\right\}_{k=1}^{n_t}$ be the data I want to fit with the model. Due to false cell type assignment or counting errors in the clonal data, I assume the observed moments $y^{\mathcal{D}}(t_k)$ underlie additive normally distributed measurement noise [Raue et al., 2013] (see section 2.2.1.1)

$$y_l^{\mathcal{D}}(t_k) = y_l^{\mathcal{M}}(t_k, \theta) + \epsilon, \text{ with } \epsilon \sim \mathcal{N}(0, \sigma_{l,k}^2), \qquad k = 1, ..., n_t, l = 1, ..., n_y. \tag{3.8}$$

The variation in experimentally observed moments ($\sigma_{l,k}^2$) was estimated by drawing 1000 bootstraps from the clonal data.

In order to assess how well a particular model fits the experimental data for a certain set of parameters θ, the log-likelihood function $\ell_D(\theta)$ (see equation 2.16) was calculated. Under the assumption of additive normally distributed measurement noise $\ell_D(\theta)$ is given by

$$\ell_D(\theta) = -\frac{1}{2} \sum_{k=1}^{n_t} \sum_{l=1}^{n_y} \log(2\pi\sigma_{l,k}^2) + \left(\frac{\left(y_l^{\mathcal{D}}(t_k) - y_l^{\mathcal{M}}(t_k, \theta)\right)^2}{\sigma_{l,k}^2} \right), \tag{3.9}$$

where n_t is the number of time points and $n_y = 9$ is the number of considered moment equations.

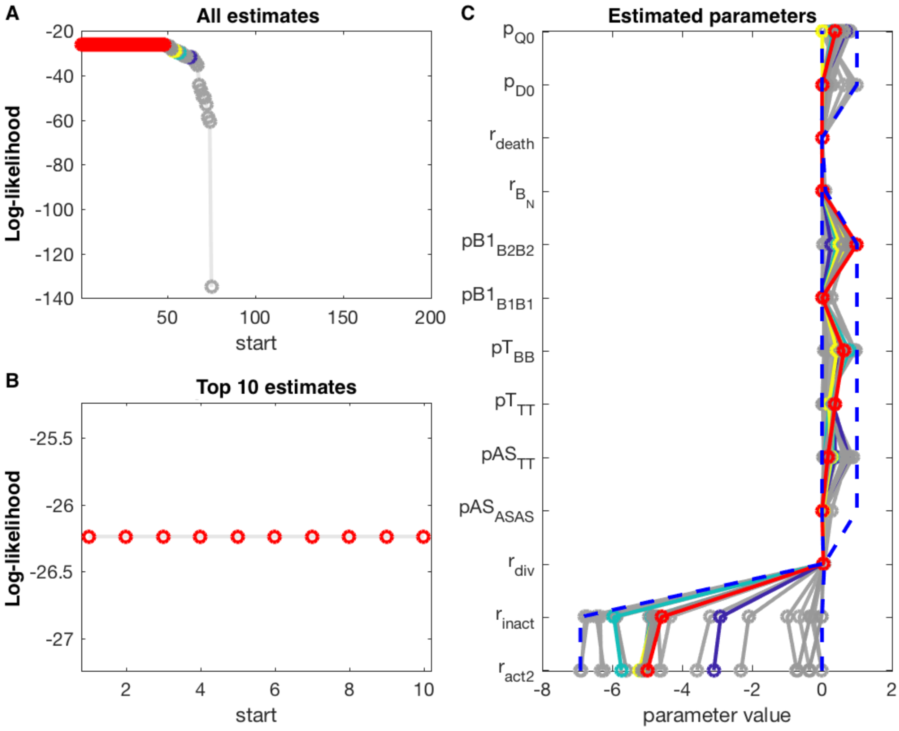


Figure 3.9: Result of multi-start optimization procedure shows observed optimized log-likelihood values (A), the 10 highest log-likelihood values (B) and the corresponding optimal parameter vectors (C) for 200 multi-starts. The plateau in (A) indicates convergence of the optimization algorithm.

In order to estimate the unknown parameter vector θ, the optimization problem

$$\theta^{ML} = \underset{\substack{\theta \\ \text{subject to } \mathcal{M}, \\ A\theta \leq b}}{\text{argmax}} \ \ell_D(\theta), \tag{3.10}$$

was solved using multi-start local optimization with interior point algorithm [Boyd and Vandenberghe, 2004] (see section 2.2.1.2). The starting values $(\theta_i^{start})_{i=1,\dots,200}$ (initial parameter vectors) were determined according to latin hypercube sampling [Eliáš and Vořechovský, 2016]. $A \in \mathbb{R}^{pxn}$ and $b \in \mathbb{R}^p$ define the inequality constraints for θ, which are introduced in 3.2.
The resulting optimal parameter set is observed at the highest ℓ_D value. To ensure that the optimization procedure converged, I investigated if this best log-likelihood value is observed several times for different starting values. Figure 3.9 illustrates the multi-start local optimization result for a randomly picked model. One can observe a plateau of the highest log-likelihood value (marked in red), indicating the implemented optimization procedure converged. The parameter estimation was performed individually for all $4^3 = 64$ models.

3.3.2 Identifiability analysis

Using the MATLAB toolbox STRIKE-GOLDD, a structural identifiability analysis was performed by calculating the generalized observability-identifiability matrix (see equation 2.23 and section 2.2.3). The analysis of the most general model $\mathcal{M}(\theta)$ (see equation 3.7) with
$\theta = (r_{act2}, r_{inact}, r_{div}, p_{S\to SS}, p_{S\to TT}, p_{T\to TT}, p_{T\to BB}, p_{B_1\to B_1B_1}, p_{B_1\to B_2B_2}, r_{mig}, p_n, p_{dS0}, p_{qS0})^t$ revealed all parameters except initial condition parameters p_{dS0} and p_{qS0} are structurally identifiable and all model states are observable.

3.4 Model comparison and averaging

The 64 models result from all possible combinations of the four division strategies (see Figure 3.4) for each of the three proliferating cell states (aS, T, B_1, see Figure 3.3). I compared and ranked the 64 different models based on their BIC value (see equation 2.29). Since I could not identify a single best performing model, I applied model averaging, see section 2.3.3.

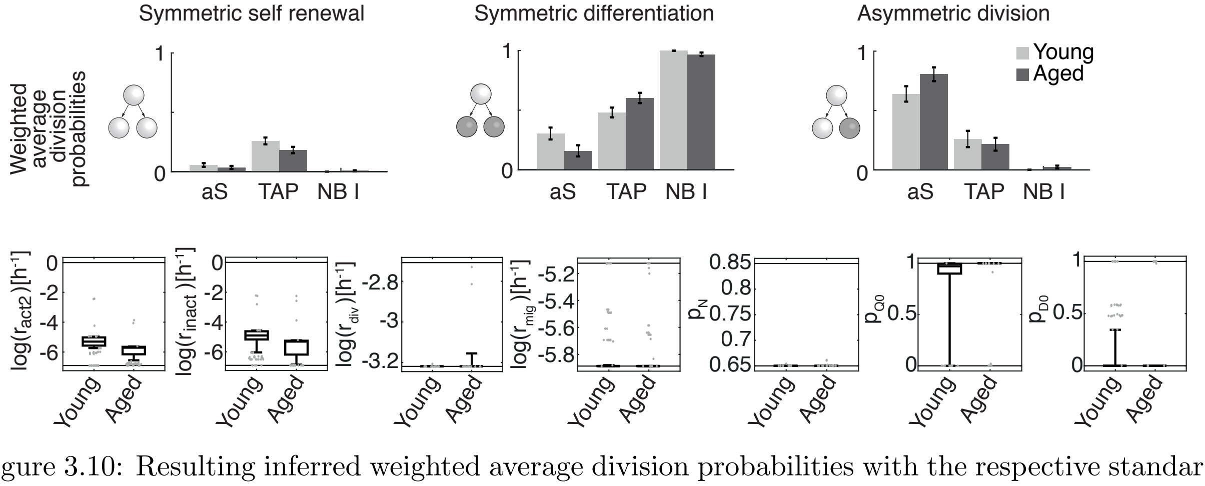


Figure 3.10: Resulting inferred weighted average division probabilities with the respective standard error of the weighted mean for all three proliferating cell types in young (light grey bars) and aged (dark grey bars) mice (upper row). 64 resulting parameter estimates (grey dots) for rates and initial condition probabilities with the corresponding weighted box plots for young and aged mice. Boxes depict the 1st, 2nd and 3rd quartiles and horizontal lines at top and bottom represent parameter boundaries (bottom row). Graphic is taken from Bast et al. [2018].

Assuming identical prior probabilities for all models, the posterior model probability was approximated to calculate the BIC weights (w_k^{BIC}), which allows the calculation of the weighted mean of parameters $\overline{\theta_w^{ML}}$ (see equation 2.35) and the standard error of the weighted mean of parameters $SE_{\overline{\theta_w^{ML}}}$ (see equation 2.36), which are together with the resulting parameter estimates visualized in Figure 3.10 for all 64 models and both data sets (young and aged adult).

As can be seen in Figure 3.10, this analysis reveals an increase in the probability of asymmetric neural stem cell divisions at the expense of the corresponding symmetric differentiation probability, and longer quiescence in aged mice.

The 10 best performing models, the BIC score difference to the best model and the BIC weights resulting from this analysis are shown in Figure 3.11. The BIC values for all (64 young and 64 aged) models and the respective resulting division probabilities can be found in Tables 3.3 and 3.4.

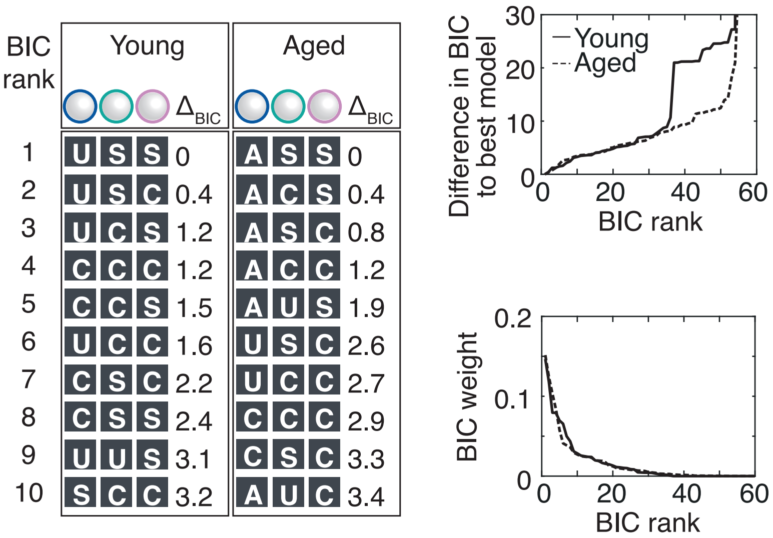


Figure 3.11: Best performing models with their BIC score difference to the rank 1 model for young and aged adult (left table), The difference in BIC score to the rank 1 model as function of the rank (right upper graphic) and BIC weight as function of the rank (right bottom graphic). Graphic is taken from Bast et al. [2018].

model			division probabilities (young)									BIC^y
mode			aS			T			B_1			
aS	T	B_1	p_s	p_d	$1-p_s-p_d$	p_s	p_d	$1-p_s-p_d$	p_s	p_d	$1-p_s-p_d$	
U	U	U	0.001	0.1296	0.8694	0.3706	0.6294	0.00	0.00	0.999	0	100
U	U	S	0,001	0,1282	0,8708	0,3716	0,6284	0,00	0,001	0,999	0	96,19
U	U	C	0,001	0,1296	0,8694	0,3706	0,6294	0,00	0	0,998	0,002	96,55
U	U	A	0,001	0,999	0	0,001	0,999	0,00	0	0	1	120,29
U	S	U	0,001	0,1266	0,8724	0,369	0,631	0,00	0,001	0,999	0	97,09
U	S	S	0,001	0,127	0,872	0,3692	0,6308	0,00	0,001	0,999	0	93,08
U	S	C	0,001	0,1266	0,8724	0,369	0,631	0,00	0	0,998	0,002	93,51
U	S	A	0,001	0,999	0	0,001	0,999	0,00	0	0	1	117,64
U	C	U	0,001	0,158	0,841	0,1558	0,3664	0,48	0,001	0,999	0	98,27
U	C	S	0,001	0,1583	0,8407	0,1559	0,3662	0,48	0,001	0,999	0	94,3
U	C	C	0,001	0,158	0,841	0,1558	0,3664	0,48	0	0,998	0,002	94,69
U	C	A	0,001	0,999	0	0	0,998	0,00	0	0	1	116,71
U	A	U	0,001	0,999	0	0	0	1,00	0,001	0,999	0	120,29
U	A	S	0,001	0,999	0	0	0	1,00	0,001	0,999	0	116,431
U	A	C	0,001	0,999	0	0	0	1,00	0	0,998	0,002	116,71
U	A	A	0,001	0,999	0	0	0	1,00	0	0	1	768,71

model			division probabilities (young)									BIC^y
mode			aS			T			B_1			
aS	T	B_1	p_s	p_d	$1-p_s-p_d$	p_s	p_d	$1-p_s-p_d$	p_s	p_d	$1-p_s-p_d$	
S	U	U	0,1605	0,8395	0	0,0434	0,1115	0,85	0,001	0,999	0	101,26
S	U	S	0,154	0,846	0	0,0393	0,1053	0,86	0,001	0,999	0	97,59
S	U	C	0,1605	0,8395	0	0,0434	0,1115	0,85	0	0,998	0,002	97,68
S	U	A	0,001	0,999	0	0,001	0,999	0,00	0	0	1	117,86
S	S	U	0,2822	0,7178	0	0,4216	0,5784	0,00	0,001	0,999	0	101,98
S	S	S	0,2831	0,7169	0	0,4218	0,5782	0,00	0,001	0,999	0	98,74
S	S	C	0,2822	0,7178	0	0,4216	0,5784	0,00	0	0,998	0,002	98,39
S	S	A	0,001	0,999	0	0,001	0,999	0,00	0	0	1	114,36
S	C	U	0,2149	0,7851	0	0,1989	0,307	0,49	0,001	0,999	0	99,93
S	C	S	0,218	0,782	0	0,1988	0,307	0,49	0,001	0,999	0	96,7
S	C	C	0,2149	0,7851	0	0,1989	0,307	0,49	0	0,998	0,002	96,36
S	C	A	0,001	0,999	0	0	0,998	0,00	0	0	1	114,27
S	A	U	0,001	0,999	0	0	0	1,00	0,001	0,999	0	117,66
S	A	S	0,001	0,999	0	0	0	1,00	0,001	0,999	0	114,27
S	A	C	0,001	0,999	0	0	0	1,00	0	0,998	0,002	114,08
S	A	A	0,001	0,999	0	0	0	1,00	0	0	1	759,55
C	U	U	0,0654	0,5538	0,3807	0,0581	0,145	0,80	0,001	0,999	0	100,27
C	U	S	0,0622	0,5635	0,3743	0,0508	0,1344	0,81	0,001	0,999	0	96,62
C	U	C	0,0654	0,5538	0,3807	0,0581	0,145	0,80	0	0,998	0,002	96,69
C	U	A	0	0,998	0,002	0,001	0,999	0,00	0	0	1	117,85
C	S	U	0,1335	0,4027	0,4638	0,4008	0,5992	0,00	0,001	0,999	0	98,81
C	S	S	0,1334	0,4029	0,4637	0,4015	0,5985	0,00	0,001	0,999	0	95,5
C	S	C	0,1335	0,4027	0,4638	0,4008	0,5992	0,00	0	0,998	0,002	95,23
C	S	A	0	0,998	0,002	0,001	0,999	0,00	0	0	1	114,36
C	C	U	0,0936	0,4817	0,4247	0,1898	0,3185	0,49	0,001	0,999	0	97,89
C	C	S	0,0948	0,479	0,4262	0,1898	0,3185	0,49	0,001	0,999	0	94,57
C	C	C	0,0936	0,4817	0,4247	0,1898	0,3185	0,49	0	0,998	0,002	94,31
C	C	A	0	0,998	0,002	0	0,998	0,00	0	0	1	114,27
C	A	U	0	0,998	0,002	0	0	1,00	0,001	0,999	0	117,65
C	A	S	0	0,998	0,002	0	0	1,00	0,001	0,999	0	114,26
C	A	C	0	0,998	0,002	0	0	1,00	0	0,998	0,002	114,06
C	A	A	0	0,998	0,002	0	0	1,00	0	0	1	765,59
A	U	U	0	0	1	0,214	0,786	0,00	0,001	0,999	0	103,74
A	U	S	0	0	1	0,2148	0,7852	0,00	0,001	0,999	0	100,04
A	U	C	0	0	1	0,214	0,786	0,00	0	0,998	0,002	100,16
A	U	A	0	0	1	0,001	0,999	0,00	0	0	1	330,78
A	S	U	0	0	1	0,216	0,784	0,00	0,001	0,999	0	100,76
A	S	S	0	0	1	0,2144	0,7856	0,00	0,001	0,999	0	97,45
A	S	C	0	0	1	0,216	0,784	0,00	0	0,998	0,002	97,17
A	S	A	0	0	1	0,001	0,999	0,00	0	0	1	332,01
A	C	U	0	0	1	0,0402	0,6393	0,32	0,001	0,999	0	101,79
A	C	S	0	0	1	0,0398	0,6407	0,32	0,001	0,999	0	98,46
A	C	C	0	0	1	0,0402	0,6393	0,32	0	0,998	0,002	98,2
A	C	A	0	0	1	0	0,998	0,00	0	0	1	327,19
A	A	U	0	0	1	0	0	1,00	0,001	0,999	0	1031,62
A	A	S	0	0	1	0	0	1,00	0,001	0,999	0	1010,85
A	A	C	0	0	1	0	0	1,00	0	0,998	0,002	1028,04
A	A	A	0	0	1	0	0	1,00	0	0	1	333.466

Table 3.3: Resulting division probabilities and BIC values for 64 models fitted to dataset of young mice

The model which assumes symmetric division strategies for all three cell types is highlighted in red in tables 3.3 and 3.4 as we compared the probabilities with work from Obernier et al. [2018], see 3.5.

model			division probabilities (aged)									BIC^a
mode			aS			T			B_1			
aS	T	B_1	p_s	p_d	$1-p_s-p_d$	p_s	p_d	$1-p_s-p_d$	p_s	p_d	$1-p_s-p_d$	
U	U	U	0.001	0.116	0.883	0.3858	0.6142	0	0.001	0.999	0	116
U	U	S	0,001	0,1086	0,8904	0,3786	0,6214	0	0,001	0,999	0	109,97
U	U	C	0,001	0,116	0,883	0,3858	0,6142	0	0	0,998	0,002	113,32
U	U	A	0,001	0,999	0	0,001	0,999	0	0	0	1	124,71
U	S	U	0,001	0,1097	0,8893	0,3811	0,6189	0	0,001	0,999	0	109,24
U	S	S	0,001	0,109	0,89	0,3799	0,6201	0	0,001	0,999	0	107,34
U	S	C	0,001	0,1097	0,8893	0,3811	0,6189	0	0	0,998	0,002	106,35
U	S	A	0,001	0,999	0	0,001	0,999	0	0	0	1	117,69
U	C	U	0,001	0,1302	0,8688	0,1609	0,3587	0,4804	0,001	0,999	0	109,37
U	C	S	0,001	0,13	0,869	0,1602	0,3596	0,4801	0,001	0,999	0	107,43
U	C	C	0,001	0,1302	0,8688	0,1609	0,3587	0,4804	0	0,998	0,002	106,48
U	C	A	0,001	0,999	0	0	0,998	0,002	0	0	1	121,82
U	A	U	0,001	0,999	0	0	0	1	0,001	0,999	0	117,97
U	A	S	0,001	0,999	0	0	0	1	0,001	0,999	0	113,41
U	A	C	0,001	0,999	0	0	0	1	0	0,998	0,002	115,08
U	A	A	0,001	0,999	0	0	0	1	0	0	1	583,38
S	U	U	0,0031	0,9969	0	0,001	0,0921	0,9069	0,001	0,2366	0,7624	111,58
S	U	S	0,0886	0,9114	0	0,001	0,0369	0,9621	0,001	0,999	0	110,54
S	U	C	0,0295	0,9705	0	0,001	0,0889	0,9101	0,1352	0,3999	0,465	108,71
S	U	A	0,001	0,999	0	0,001	0,999	0	0	0	1	115,93
S	S	U	0,3289	0,6711	0	0,4164	0,5836	0	0,001	0,999	0	111,66
S	S	S	0,3294	0,6706	0	0,4171	0,5829	0	0,001	0,999	0	109,67
S	S	C	0,3289	0,6711	0	0,4164	0,5836	0	0	0,998	0,002	108,76
S	S	A	0,001	0,999	0	0,001	0,999	0	0	0	1	115,26
S	C	U	0,2321	0,7679	0	0,1649	0,3527	0,4824	0,001	0,2696	0,7294	110,96
S	C	S	0,2632	0,7368	0	0,1737	0,3401	0,4862	0,3022	0,6978	0	109,13
S	C	C	0,2519	0,7481	0	0,1692	0,3465	0,4843	0,1115	0,4436	0,4449	108,11
S	C	A	0,001	0,999	0	0	0,998	0,002	0	0	1	113,04
S	A	U	0,001	0,999	0	0	0	1	0,001	0,999	0	115,53
S	A	S	0,001	0,999	0	0	0	1	0,001	0,999	0	111,84
S	A	C	0,001	0,999	0	0	0	1	0	0,998	0,002	112,64
S	A	A	0,001	0,999	0	0	0	1	0	0	1	613,73
C	U	U	0,0963	0,4757	0,428	0,001	0,0869	0,9121	0,001	0,999	0	110,75
C	U	S	0,0917	0,486	0,4222	0,001	0,0834	0,9156	0,001	0,999	0	109,66
C	U	C	0,0963	0,4757	0,428	0,001	0,0869	0,9121	0	0,998	0,002	107,86
C	U	A	0	0,998	0,002	0,001	0,999	0	0	0	1	115,92
C	S	U	0,1351	0,3999	0,4649	0,4112	0,5888	0	0,001	0,999	0	109,91
C	S	S	0,1355	0,3993	0,4652	0,4118	0,5882	0	0,001	0,999	0	107,9
C	S	C	0,1351	0,3999	0,4649	0,4112	0,5888	0	0	0,998	0,002	107,02
C	S	A	0	0,998	0,002	0,001	0,999	0	0	0	1	115,26
C	C	U	0,1262	0,4158	0,4581	0,1821	0,3286	0,4893	0,001	0,999	0	109,5
C	C	S	0,1264	0,4153	0,4583	0,1826	0,3279	0,4894	0,001	0,999	0	107,5
C	C	C	0,1262	0,4158	0,4581	0,1821	0,3286	0,4893	0	0,998	0,002	106,61
C	C	A	0	0,998	0,002	0	0,998	0,002	0	0	1	113,03
C	A	U	0	0,998	0,002	0	0	1	0,001	0,999	0	115,52
C	A	S	0	0,998	0,002	0	0	1	0,001	0,999	0	111,83
C	A	C	0	0,998	0,002	0	0	1	0	0,998	0,002	112,63
C	A	A	0	0,998	0,002	0	0	1	0	0	1	580,12

model			division probabilities (aged)									BIC^a
mode			aS			T			B_1			
aS	T	B_1	p_s	p_d	$1 - p_s - p_d$	p_s	p_d	$1 - p_s - p_d$	p_s	p_d	$1 - p_s - p_d$	
A	U	U	0	0	1	0,2179	0,7821	0	0,001	0,999	0	110,07
A	U	S	0	0	1	0,2152	0,7848	0	0,001	0,999	0	105,66
A	U	C	0	0	1	0,2179	0,7821	0	0	0,998	0,002	107,18
A	U	A	0	0	1	0,001	0,999	0	0	0	1	293,08
A	S	U	0	0	1	0,2141	0,7859	0	0,001	0,999	0	107,39
A	S	S	0	0	1	0,2112	0,7888	0	0,001	0,999	0	103,74
A	S	C	0	0	1	0,2141	0,7859	0	0	0,998	0,002	104,5
A	S	A	0	0	1	0,001	0,999	0	0	0	1	291,36
A	C	U	0	0	1	0,0418	0,6329	0,3253	0,001	0,999	0	107,82
A	C	S	0	0	1	0,0406	0,6377	0,3218	0,001	0,999	0	104,13
A	C	C	0	0	1	0,0418	0,6329	0,3253	0	0,998	0,002	104,93
A	C	A	0	0	1	0,0001	0,9794	0,0205	0	0	1	147,29
A	A	U	0	0	1	0	0	1	0,001	0,999	0	913,95
A	A	S	0	0	1	0	0	1	0,001	0,999	0	887,48
A	A	C	0	0	1	0	0	1	0,0001	0,9813	0,0186	141,15
A	A	A	0	0	1	0	0	1	0	0	1	41974,59

Table 3.4: Resulting division probabilities and BIC values for 64 models fitted to dataset of aged mice

3.5 Validation of modeling results

In order to test the reliability of the modelling results, I performed a robustness test and investigated if simulations of the average models agree with features of the clonal and independent data.

3.5.1 Robustness test

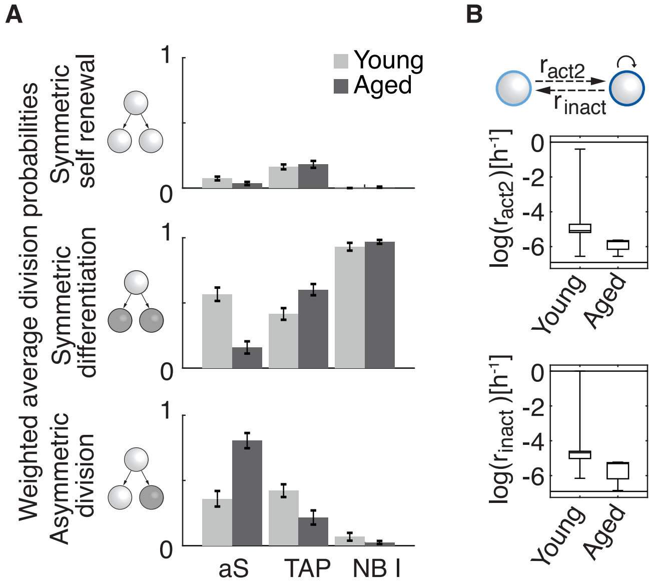


Figure 3.12: Form the robustness test resulting inferred weighted average division probabilities with the respective standard error of the weighted mean for all three proliferating cell types in young (light grey bars) and aged (dark grey bars) mice (A). 64 resulting parameter estimates for (in)activation rates shown as weighted box plots for young and aged mice. Boxes depict the 1st, 2nd and 3rd quartiles and horizontal lines at top and bottom represent parameter boundaries (B). Graphic is taken from Bast et al. [2018].

In order to test the robustness of the estimated weighted mean parameter differences between the two groups (young and aged), I repeated the analysis using only the measurements observed at days 21 and 56 for parameter estimation. Resulting probabilities for division strategies can be seen in Figure 3.12. Weighted parameters differ only slightly from parameters inferred from the complete data set (see Figure 3.10). Thus, the conclusions of longer quiescence in aged mice, and an increase in the probability of asymmetric neural stem cell divisions at the expense of symmetric differentiation could be confirmed, even for fitting only parts of the data set.

3.5.2 Comparison of model to dependent and independent data

As stated in section 2.1, I used recently published data [Daynac et al., 2016, Shook et al., 2012] to constrain the stem cell population in- and activation rates in the model (see table 3.2). From the same analysis, cell counts of subsequent cell-states are available, which were not included in the parameter estimation procedure. To evaluate if the model is able to describe the cell count dynamics per cell-state on the population level, I calculated the first order moment (mean) for each cell-state. We set the initial values of the model to the earliest observed measurements. This analysis was performed based on the average young model (age-independent model) and on an age-dependent model, in which the parameters change with ageing from the weighted mean parameter in group young ($\overline{\theta_{w,y}^{ML}}$) to the weighted mean parameter in group aged ($\overline{\theta_{w,a}^{ML}}$). This change in parameters was modelled with Hill functions (see Figure 3.13).

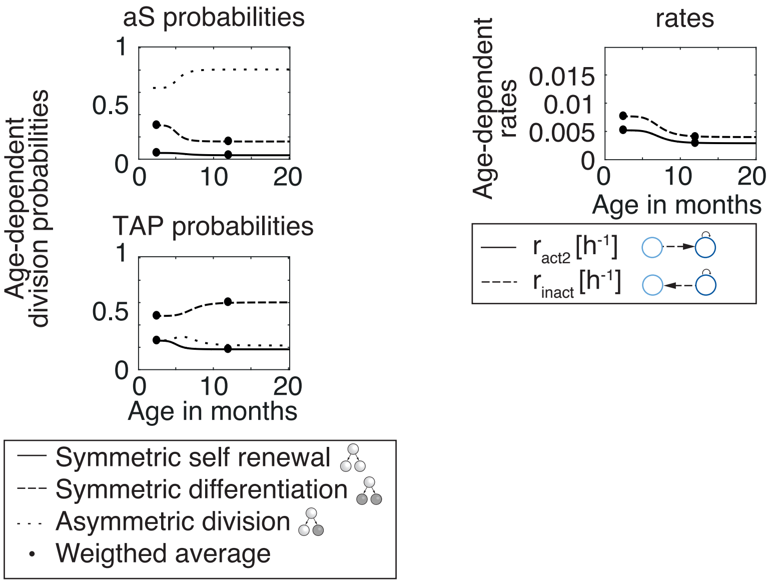


Figure 3.13: Hill function fits to model age-dependent stem cell and TAP division probabilities and (in)activation rates. Graphic is taken from Bast et al. [2018].

The Hill function is defined as follows:

$$H(a, s, n, y_{min}, y_{max}) := \frac{y_{max} - y_{min}}{(as)^n + 1} + y_{min}, \tag{3.11}$$

where a denotes the age of mice, n is the Hill coefficient, $\frac{1}{s}$ is the age at which the saddle point of the Hill function is observed and y_{min}, y_{max} describe the minimum and maximum values of the

Hill function [Gesztelyi et al., 2012]. y_{min}, and y_{max}, have been set to

$$y_{min} = min(\overline{\theta_{w,y}^{ML}}, \overline{\theta_{w,a}^{ML}})$$
$$y_{max} = max(\overline{\theta_{w,y}^{ML}}, \overline{\theta_{w,a}^{ML}})$$

and optimized for s and n, assuming boundaries

$$s \in \left[\frac{1}{t_{aged}}, \frac{1}{t_{young}}\right]$$

and

$$n \in \begin{cases} [1, 10] & \text{, if } \overline{\theta_{w,y}^{ML}} \leq \overline{\theta_{w,a}^{ML}} \\ [-10, -1] & \text{, otherwise.} \end{cases}$$

The models include halfway migration of neuroblasts to be consistent with the population study data, which are the number of cells obtained after dissociating the lateral wall of the lateral ventricles, thus accessing only a portion of the whole NB population.

One can observe that the model behaviour over time agrees with the experimental observations made by [Daynac et al., 2016] (see A in Figure 3.14).

In addition, I calculated the clone size and cell fractions from 500 SSA simulations [Gillespie, 2001] to test if the model accurately predicts these statistics. The model variability which can be calculated from 500 simulations is shown as grey shaded area and includes most of the data points of the data set (see B and C in Figure 3.14).

Another model validation step was the comparison of the percentage of clones consisting exclusively of neurons (neuron-only clones) at day 56. I simulated 1000 times the number of lineage trees observed at 56 days post labelling (12 in young and 11 in old adult) and calculated for every of the 1000 runs the percentage of neuron-only clones to observe a distribution of neuron-only clones. For the simulation of trees, exponentially distributed (in)activation times and Erlang distributed division and migration times were assumed (see Figure 3.15). Panel D in Figure 3.14 shows that the average model (grey distribution plot) correctly predicts the decline of neuron-only clones with age, which was observed from the data set (black dots, see table 3.1).

A last validation step was to compare the division probabilities resulting from the model which assumes exclusively symmetric division modes to the probabilities inferred by Obernier et al. [2018]. In their *in vivo* clonal lineage tracing study of active NSCs, assumed exclusively symmetric divisions and reported probabilities of $0.2-0.34$ and $0.7-0.8$ for symmetric self-renewal and symmetric differentiation of NSCs respectively. As can be seen in tables 3.3 and 3.4 the corresponding probabilities I inferred are 0.2831 ans 0.77169 in young and 0.3294 and 0.6706 in aged mice.

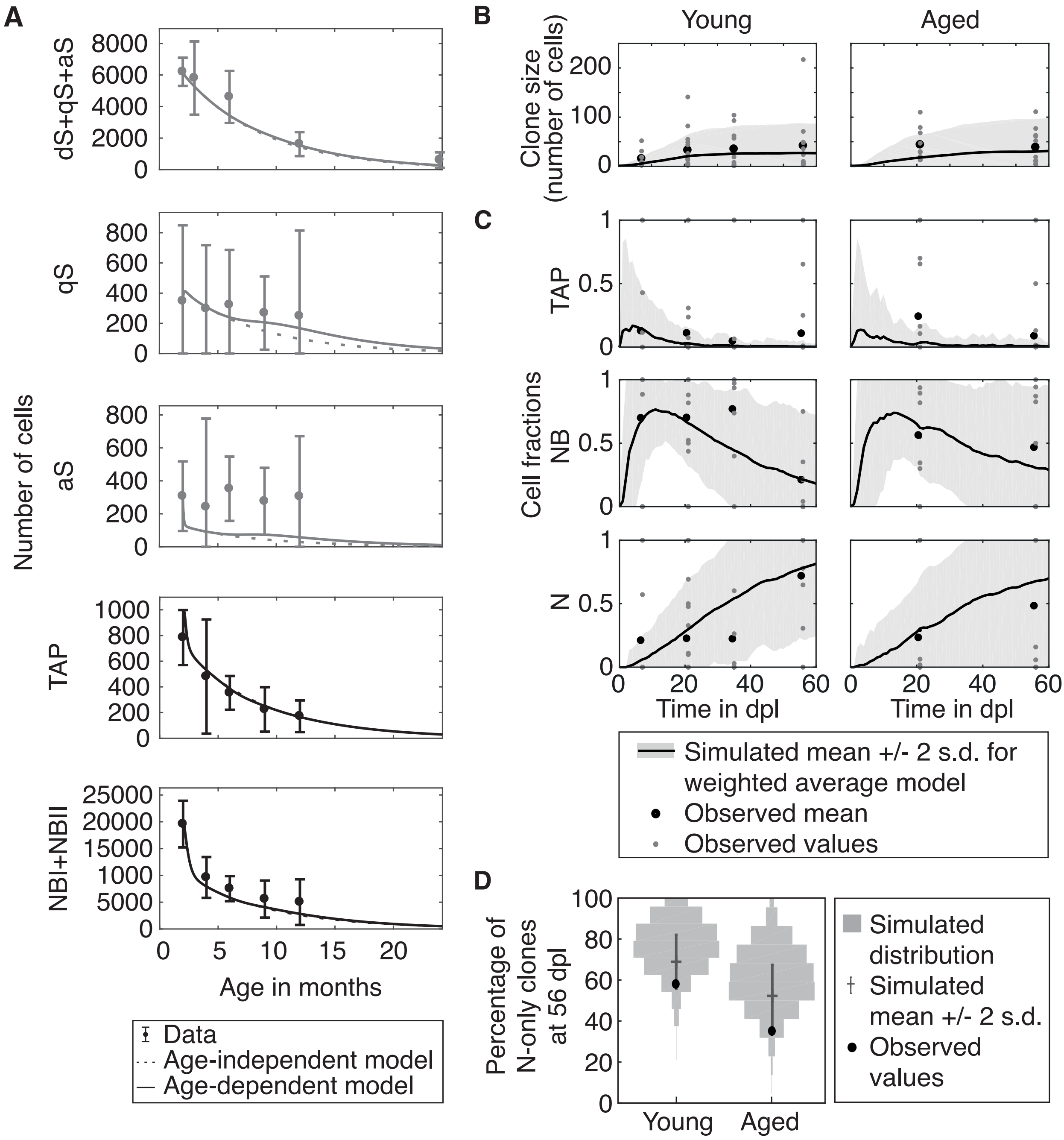


Figure 3.14: Comparison of age-dependent and age-independent (solid and dashed lines) weighted average models to dependent (three upper panels, grey) and independent (two lower panels, black) population data from Daynac et al. [2016] (mean ± 2 s.d.)(A). Clone Size (B) and cell fractions (C) predicted by average models (solid lines) compared to observed cell numbers (small grey dots) and their mean (large black dots). Neuron-only clones predicted by Gillespie simulations of average model (grey distribution plot) compared to observed values (black dots) (D). Graphic is taken from Bast et al. [2018].

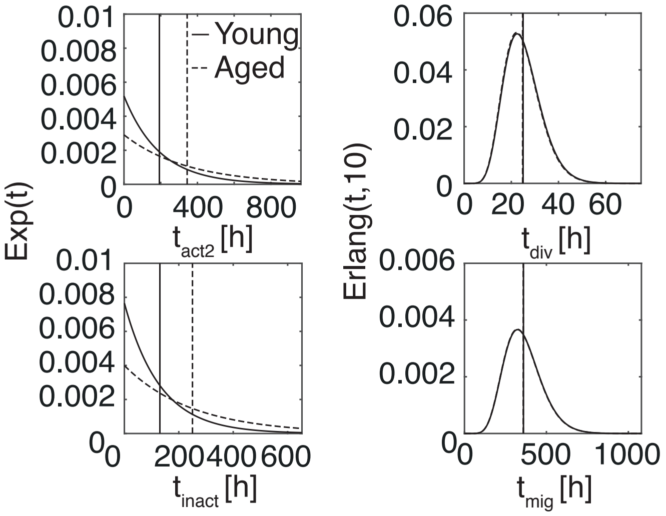


Figure 3.15: Biologically plausible distributions for (in)activation, division and migration times, which are used for tree simulations. Horizontal lines show the mean of distribution (weighted average rates). Graphic is taken from Bast et al. [2018].

3.6 Prediction from average models

Validation of the resulting average models allows to make predictions. I simulated individual trees, which were used to calculate self-defined genealogical metrics (see Figure 3.18). The genealogical metrics estimation was performed by assuming exponentially distributed (in)activation times and Erlang distributed division and migration times. Resulting mean and median of genealogical metrics are shown in table 3.5 for both groups.

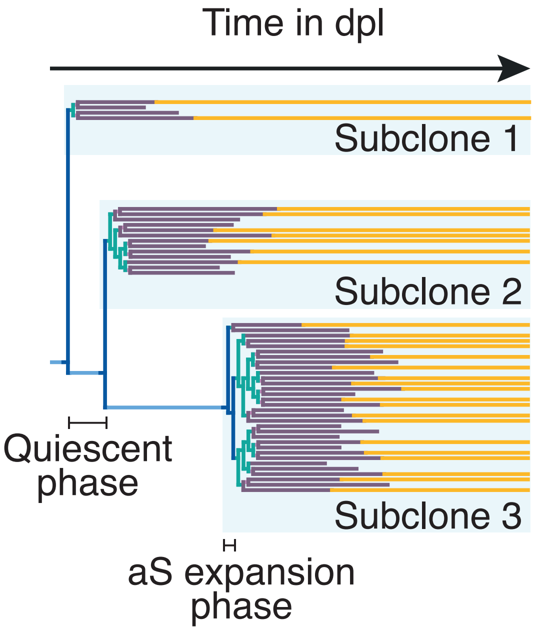


Figure 3.16: Definition of genealogical metrics to characterize simulated lineage trees. Graphic is taken from Bast et al. [2018].

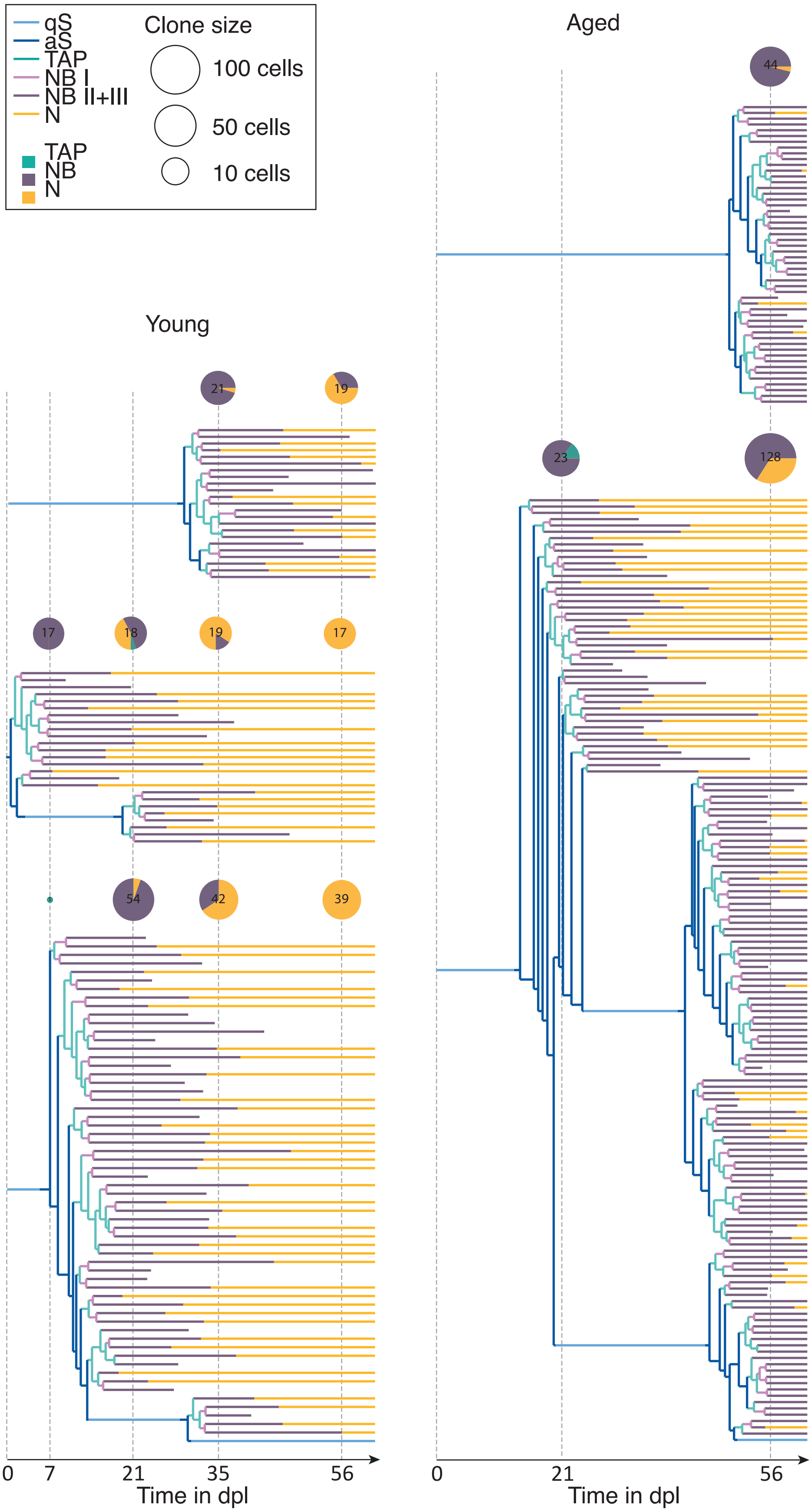


Figure 3.17: Five exemplary simulated lineage trees from the young (left) and aged (right) model. Pie charts indicate clone size and composition (number of TAPs, NBs and Ns) at the experimental time points to allow a comparison with the experimental data (see Figure 3.2). Graphic is taken from Bast et al. [2018].

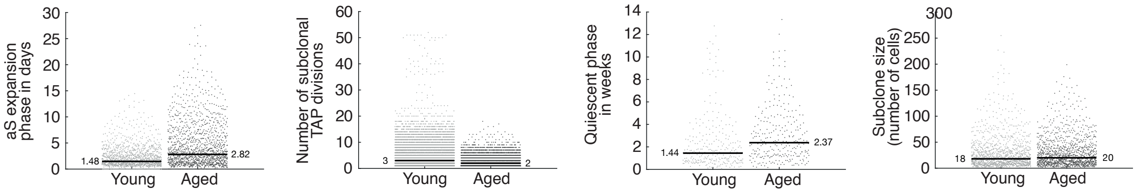


Figure 3.18: Predicted distributions for four genealogical metrics (active stem cell expansion phase, number of subclonal TAP divisions, quiescent phase, subclone size) calculated from 1000 simulated lineage trees from the average young and aged model. Medians are shown as a black line. Graphic is taken from Bast et al. [2018].

Genealogical metrics	young mice		aged mice	
	mean	median	mean	median
number of subclones	1.48	1	1.49	2
number of active subclones	1.23	1	1.34	1
number of inactive subclones	0.25	0	0.14	0
subclone size (number of cells)	28.16	18	29.28	20
aS expansion phase duration [d]	2.12	1.48	4.27	2.83
quiescent phase [weeks]	2.27	1.44	2.97	2.36
subclonal activity duration [d]	8.35	7.10	8.67	7.33
mean subclonal branch length (mean number of generations)	7.83	7.19	7.26	6.80
subclonal lifespan [weeks]	10.44	10.37	6.53	6.49
clonal lifespan [weeks]	6.24	5.91	6.57	6.18
number of subclonal TAP divisions	4.78	3	2.35	2
number of subclonal NB I divisions	1.00	1	1.04	1

Table 3.5: Genealogical metrics observed from 1000 simulated trees.

In addition, I performed the genealogical metrics estimation assuming exclusively exponentially distributed rates. This led to very similar results and all metrics showed the same qualitative behavior.

4 Application II: Modeling hematopoiesis to infer lineage hierarchies and uncover changes with age and upon disease

A second example for a computational cell division and differentiation model is explained in detail in this chapter. The content of this chapter resulted in two drafted manuscripts, of which one is published in iScience Cell Press [Bast et al., 2021] and the other one is drafted and will be submitted soon. The content of these two manuscripts was restructured and slightly modified for my dissertation. It stems from collaborative projects with the clinicians Michèle Buck, Prof. Katharina Götze, Prof. Robert Oostendorp, Judith Hecker, Prof. Florian Bassermann, and the computational biologists Dr. Carsten Marr and Prof. Fabian Theis. The data result from experiments that have been performed by Michèle Buck. My contribution was

(i) literature search about hematopoiesis to define biologically meaningful model assumptions, parameter boundaries, and lineage hierarchies, and

(ii) the computational analysis, including model specification, parameter inference, identifiability analysis, and model selection.

Firstly, I will introduce the most important findings about human hematopoiesis and relevant hematopoietic disorders, and the performed experiment together with the measured data and show data analysis results (see section 4.1). Secondly, I will derive deterministic models describing the mechanisms of adult hematopoiesis in young, aged and diseased individuals on the cell population level (see section 4.2). By fitting this set of considered models to experimental time-resolved cell count data (see section 4.1), I will introduce a maximum likelihood estimation approach to infer unknown model parameters and asses structural and practical parameter identifiability (see section 4.3). To answer the question which lineage hierarchies are plausible to describe hematopoiesis and which ones can be rejected based on our experimental data, I will perform model selection (see section 4.4). Using the best performing model to fit experimental data of healthy and dieseased donors, I will compare resulting parameter values to analyse which rates change during age in healthy individuals and with disease in comparison to healthy donors and combine the inferred rates with patient and donor information (see section 4.5). Lastly, I will validate the results based on *in silico* data and unseen measurements (see section 4.6).

The code accompanying the analysis described in this chapter can be found at
https://github.com/marrlab/HematopoiesisModelComparison and
https://github.com/LisaBast/HematopoieticDisorderAnalysis

4.1 Biological background and experimental data

Hematopoiesis describes the process by which blood cellular components are produced. As humans produce roughly 500 billion mature blood cells per day, the hematopoietic cells form one of the most generative tissues in the human body [Fliedner et al., 2002]. Hematopoiesis is assumed to be a continuous process which occurs through differentiation of hematopoietc stem cells (HSCs) into mature blood cells via several hierarchically defined progenitor compartments of decreasing potency and increasing maturity [Jagannathan-Bogdan and Zon, 2013]. While hematopoietic stem and progenitor cells (HSPCs) are located in the bone marrow, mature cells are mainly found in the peripheral blood [Walenda et al., 2014]. This differentiation process is known to be regulated cell intrinsically by regulator genes, but also cell-extrinsically by signals from a cells' environment, which is referred to as the bone marrow niche [Pinho and Frenette, 2019]. In the past, murine and human lineage commitment was studied *in vivo* and *in vitro* via different methods including label-retaining transplantation assays, colony forming capacity-assessment, fluorescence activated cell sorting (FACS) and transcriptome analysis [Adolfsson et al., 2005, Akashi et al., 2000, Doulatov et al., 2010, 2012, Forsberg et al., 2006, Giebel et al., 2006, Goardon et al., Haas et al., 2018, Hao et al., 2001, Laurenti et al., 2018, Månsson et al., 2007, Notta et al., 2016, Perié et al., 2015, Pronk et al., 2007, Reynaud et al., 2003, Rossi et al., 2008, Sanjuan-Pla et al., 2013, Takano et al., 2004, Velten et al., 2017], proposing multiple differentiation paths which describe how cells transit through defined cell type compartments. These differentiation paths include for instance the existence of lineage restricted progenitors which bypass multipotent progenitors by directly transiting to mature cell types and thereby challenge the classical model of hematopoiesis. So far the plausibility of these proposed differentiation transitions has not been analysed comprehensively and quanitatively by deriving and comparing different lineage hierarchies with a systems biology approach.

Furthermore, it has been observed that upon ageing less mature blood cells are produced [Chung and Park, 2017, Lee et al., 2019]. Based on clonal assays of the murine hematopoietic system, it has been suggested that the decline in hematopiesis is due to HSCs, which lose their self-renewal capacity [Dykstra et al., 2011]. Additionally, an increased frequency of HSCs and myeloid skewing, i.e. the more prominent production of myeloid cells compared to lymphoid cells, has been observed with age in mice [Dykstra et al., 2011, Sudo et al., 2000]. Similarly, studies of human hematopoiesis revealed preferred myeloid differentiation with age, decreased lymphoid progenitors, and increased HSC frequencies [Pang et al., 2011]. If these increased HSC frequencies result from an increase of the HSC population or a decrease in cell numbers of downstream compartments such as progenitors or mature cells, has so far not been investigated.

Additionally, upon ageing the incidence of clonal hematopoietic disorders increases [Lee et al., 2019], which are driven by the acquisition of mutations in HSPCs [Sperling et al., 2017]. Elderly individuals whose HSPCs exhibit the most common somatic mutations observed in hemtopoietic disorders, but which do not suffer from hematologic diseases, are termed clonal hematopoiesis of indeterminate potential (CHIP) or age-related clonal hematopoiesis individuals [Genovese et al., 2014, Jaiswal et al., 2014, Steensma et al., 2015]. It has been investigated that CHIP individuals have an increased risk of developing blood cancer and that this risk is proportional to the size of the somatic clone [Watson et al., 2015]. In contrast to CHIP, myelodysplastic syndrome (MDS) is a particular age-related hematopoietic disorder, which is characterized by ineffective hematopoiesis and peripheral cytopenias, i.e. the lack of cellular components in the blood (see Figure 4.1). MDS patients obtain an increased risk of transformation to acute myeloid leukaemia [Chung and Park, 2017]. Distinct acquired mutations, have been detected in MDS HSCs and are considered disease-initiating events, leading to a dominant MDS clone [Pang et al., 2013, Shastri et al., 2017], but so far it is unclear why and how some age-related mutations achieve clonal dominance and lead to MDS. Possible mechanisms are growth advantages of a specific clone and suppression of healthy hematopoisis and can be either induced by cell-intrinsic deregulation or by cell-extrinsic niche-mediated interference or by a combination of both. It has also not been investigated if proliferation, differentiation, or cell death, or combinations of these processes and which cell types are affected.

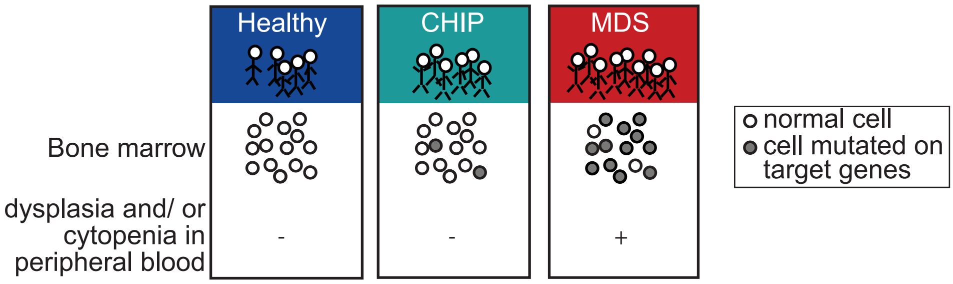

Healthy
CHIP
MDS
Bone marrow
dysplasia and/ or cytopenia in peripheral blood
-
-
+
○ normal cell
● cell mutated on target genes

Figure 4.1: Overview of cell abnormalities in bone marrow and peripheral blood in case of CHIP or MDS in comparison to healthy hematopoiesis. Graphic is taken from unpublished manuscript Buck et al.

To assess the plausibility of a set of previously suggested lineage hierarchies and to infer age- and disease-related cell-intrinsic hematopoietic changes that induce clonal dominance, an *in vitro* experiment has been performed. In this experiment cell intrinsic differentiation and proliferation potential of human HSCs and progenitors has been investigated by purifying bone marrow samples for HSCs, culturing the HSCs for up to 7 days in a medium supplemented with 8 growth factors and assessing their progeny by Fluorescent activated cell sorting (FACS) after different time points t_i in the interval $t_i \in [1, 7]$ days (see Figure 4.2).

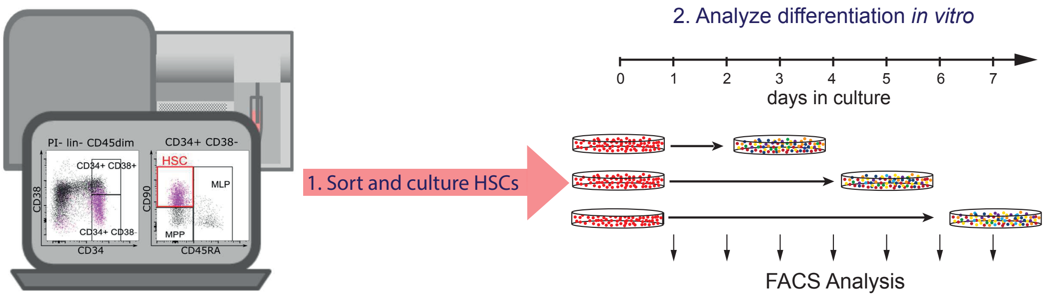


Figure 4.2: Design of time resolved *in vitro* HSC culture experiment, which was individually performed with each bone marrow sample. Graphic is taken from Bast et al. [2021].

The bone morrow samples used in this experiment were collected from donors courtesy of the Stiftung Aktion Knochenmarkspende Bayern or obtained from femoral heads of patients undergoing hip replacement surgery courtesy of Dr. Martin Nolde (SANA Klinik, München-Solln, Germany) for healthy individuals and were obtained from MDS patients undergoing routine clinical evaluation. At day 0 of the experiment, $CellTrace^{TM}$ Violet stain was added to the medium to track the number of divisions for each cell population later in the FACS analysis at day $t_i \in [1, 7]$. For the FACS analysis, fluorescence-coupled antibody staining with anti-CD34, anti-CD38, anti-CD90, anti-CD45RA and anti-CD123 was performed at the observed time point t_i to determine the number of cells in each compartment (see Figure 4.3).

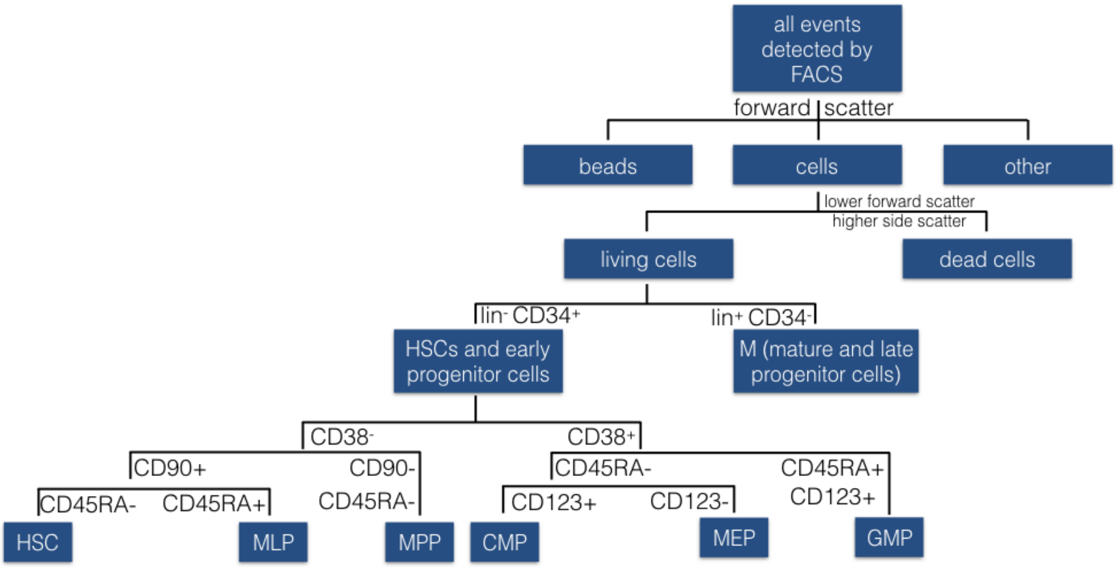


Figure 4.3: FACS gating scheme used to determine number of cells in each compartment.

To gain accurate absolute cell population counts, 50 μl of Flow-Count Fluorospheres ($B_{total} \approx 50000$ beads) were added to the cell culture at t_i prior to FACS analysis. The corrected number of cells in the ith compartment N_i^c was calculated according to:

$$N_i^c = N_i * \frac{B_{total}}{B_{counted}},$$

where B_{total} is the ground truth number of beads added to the sample before FACS analysis and $B_{counted}$ is the observed number of beads while performing FACS analysis. The experiment including FACS gating was performed by Michéle Buck using the FloJo V10 software. The experimental data describe cell counts and division distributions of HSCs, multipotent progenitor cells (MPPs), common myelocyte progenitors (CMPs), multipotent lymphocyte progenitors (MLPs), megakaryocyte erythrocyte progenitors (MEPs), granulocyte monocyte progenitors (GMPs) and CD34- cells which correspond to mature cells and late progenitors (M, see Figure 4.4). Upon availability of bone marrow samples, the experiment was repeated, which is termed repetition, or measured several twice at the same time points using cell subpopulations from the same HSC-sort, which is termed replicate in section 4.3.

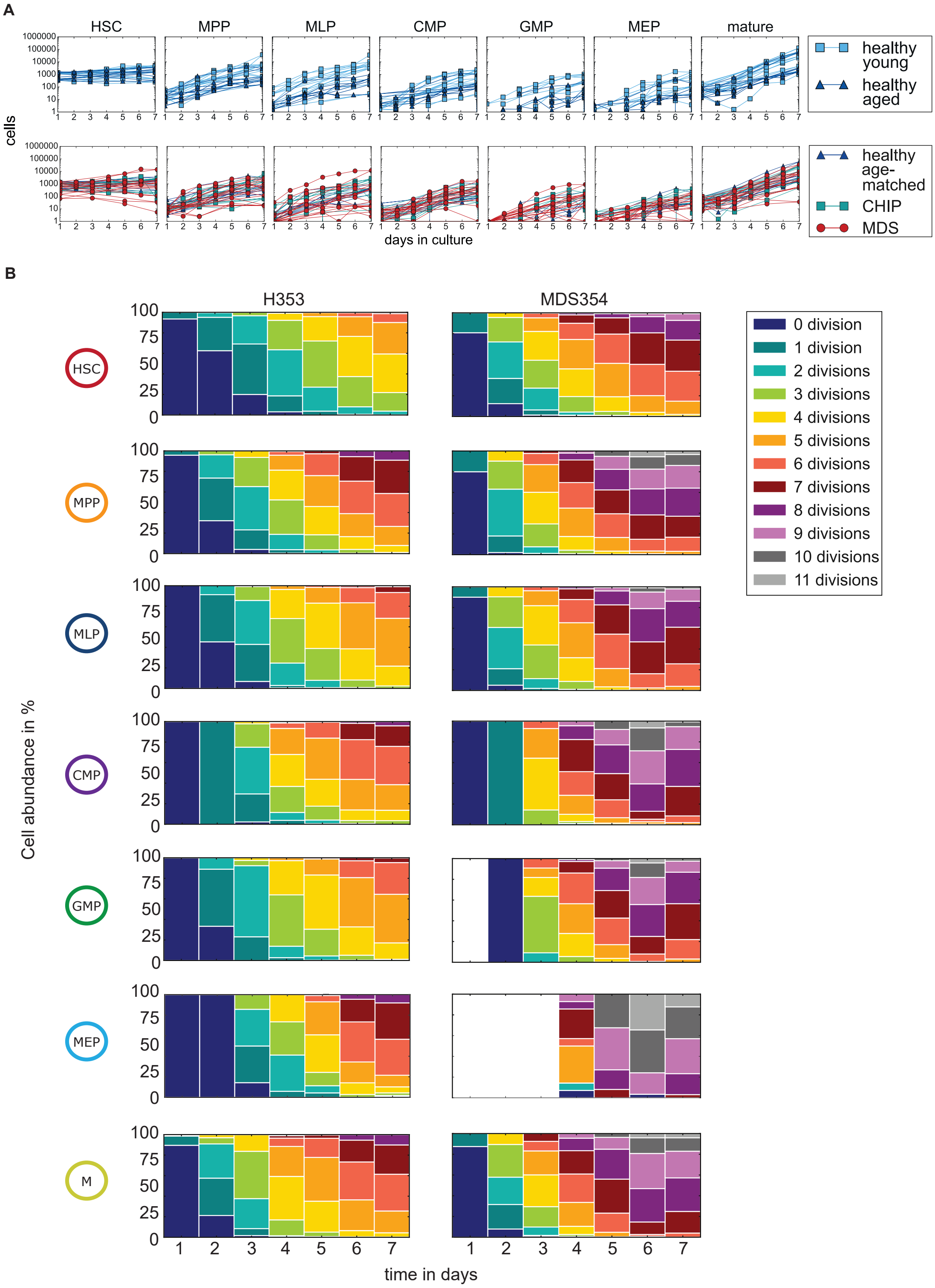


Figure 4.4: Measurements observed from *in vitro* HSC culture experiment describe cell counts of HSC, MPP, CMP, MLP, MEP, GMP and M compartments (A) at various time points $t_i \in [1, 7]d$ for bone marrow samples of young healthy (light blue), vs. aged healthy (dark blue) donors (upper row), and of donors with CHIP (green), vs. MDS patients (red), vs. healthy age-matched individuals (dark blue, bottom row). Division distributions (B) are exemplary shown for all cell type compartments of healthy individual H353 (left) and MDS patient MDS354 (right). Graphic is taken from unpublished manuscript Buck et al. and slightly modified.

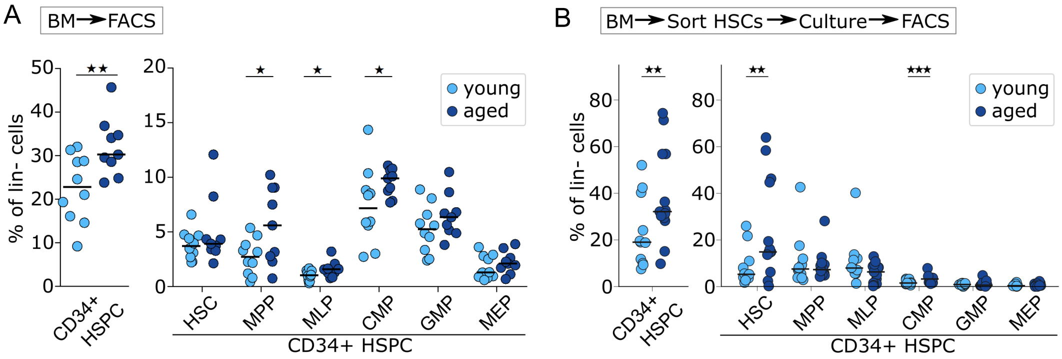


Figure 4.5: Snapshot analysis (A) of bone marrow cells shows a higher proportion of HSPCs (p=0.0086, Wilcoxon-Mann-Whitney test) in aged bone marrow (BM) due to a higher fraction of MPPs (p=0.0472), MLPs (p=0.0393), and CMPs (p=0.0325). *In vitro* culture of sorted HSCs for 7 days (B) shows a significantly (p = 0.024, Wilcoxon-Mann-Whitney test) higher proportion of CD34+ HSPC in aged individuals, resulting from higher HSC (p = 0.024) and CMP (p = 0.006) fractions. Progenitor compartments were normalized to viable, lin- cells. Black horizontal lines show median value. Graphic is taken from unpublished manuscript Buck et al.

Analysing the bone marrow cells of samples from healthy donors via FACS before and after culturing revealed a significantly increased proportion of HSPCs in the aged group in both cases (see Figure 4.5). The relative fraction F_i of a cell population i is calculated by dividing the number of cells N_i by the number of all lin- cells:

$$F_i = \frac{N_i}{[lin-]}.$$

Donors with age up to 60 years were defined as young and with age above 60 as aged. Interestingly, the increased proportion of HSCs which was observed in previous studies [Dykstra et al., 2011, Pang et al., 2011, Sudo et al., 2000], could only be confirmed after 7 days in culture (p = 0.024, Wilcoxon-Mann-Whitney test) but not for the snapshot analysis of bone marrow samples. The observed increased HSPC fraction found in BM of aged individuals can be caused by

(i) an absolute increase of HSPC number while the number of CD34- (mature) cells stays constant,

(ii) a decrease of CD34- (mature) cells while the number of HSPCs stays constant, or

(iii) through a decrease of HSPCs in combination with a strong decrease of CD34- (mature) cells

(see Figure 4.6).

To investigate which theory is most supported by the experimental data, one can calculate the yield, which is defined as

$$yield(t) = \frac{N_i(t)}{[HSC](t = t_0)}.$$

The yield at day $t_i = 7$ is reduced from a median of 11.2 per input HSC in young to 4.1 per input HSC in aged donors for HSPCs and from 33.7 per input HSC in young to 5.0 per input HSC in aged donors for CD34- cells. According to Wilcoxon-Mann-Whitney test this reduction in aged donors is significant with p-values of 0.022 for HSPCs and 0.019 for CD34- cells. At the same time, the HSC yield stays roughly constant, as can be seen in Figure 4.7, suggesting theory (iii) is the most plausible one.

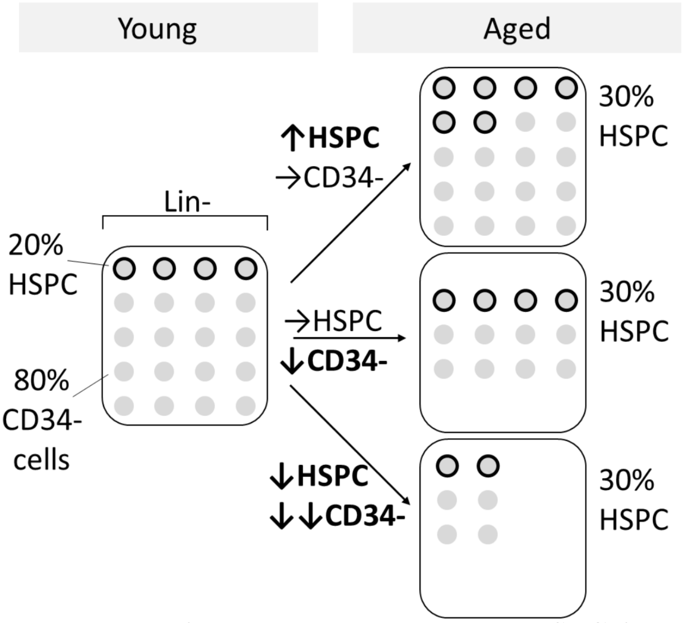


Figure 4.6: Three possible theories for the observed increased HSPC fraction in bone marrow of aged donors. Graphic is taken from unpublished manuscript Buck et al.

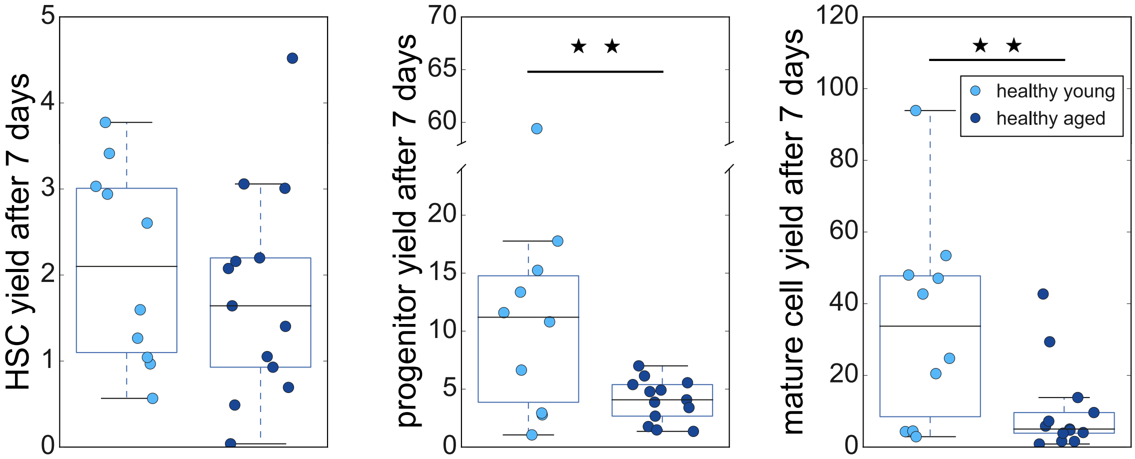


Figure 4.7: Yield at day 7 shows aged HSCs produce significantly less progenitor (p = 0.022, Wilcoxon-Mann-Whitney test) and mature (p = 0.019) cells but the same amount of HSCs in comparison to young HSCs. Graphic is taken from unpublished manuscript Buck et al. and was slightly modified.

This reduced blood cell production in aged individuals can originate from increased cell death, reduced differentiation, or reduced proliferation in several compartments of the hematopoietic hierarchy. Which changes arise with ageing can be revealed with the compartmental model introduced in section 4.2 by performing parameter inference (see section 4.3) on the experimental data from young and aged healthy donors (see section 4.5, Figure 4.8).

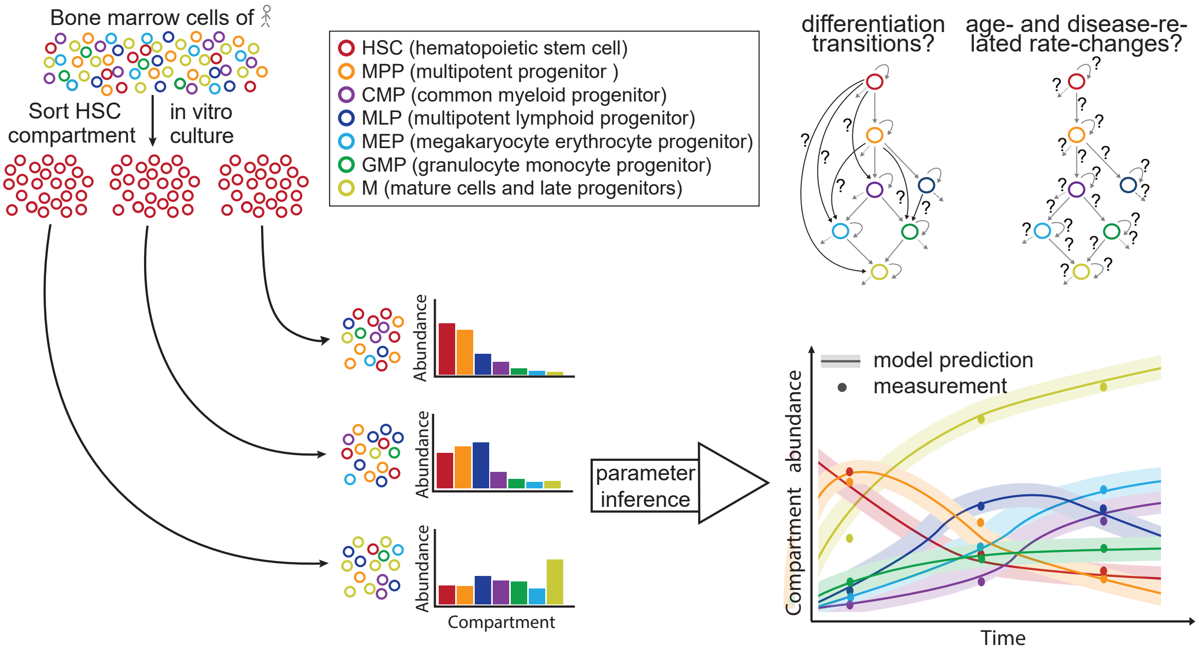

Bone marrow cells of
Sort HSC compartment
in vitro culture
HSC (hematopoietic stem cell)
MPP (multipotent progenitor)
CMP (common myeloid progenitor)
MLP (multipotent lymphoid progenitor)
MEP (megakaryocyte erythrocyte progenitor)
GMP (granulocyte monocyte progenitor)
M (mature cells and late progenitors)
differentiation transitions?
age- and disease-related rate-changes?
?
Abundance
Compartment
parameter inference
model prediction
measurement
Compartment abundance
Time

Figure 4.8: Outline of approach using experimental data of an arbitrary donor sample, whose HSCs were cultured *in vitro* and whose HSC progeny was observed at several time points. Subsequent modeling and parameter inference reveals unknown rates for proliferation, differentiation and cell death with cell type resolution. Approacg allows to determine the most plausible lineage hierarchies and to uncover changes with age and upon disease. Graphic is taken from unpublished manuscript Buck et al. and was slightly modified.

4.2 Specification of a set of candidate models

Cell states and model parameters

As we observed hundreds of hematopoietic cells in cell type and division specific compartments at time points $t_0, ..., t_{n_t}$ (see Figure 4.4 and Figure 4.8), a compartmental model using ordinary differential equations (ODEs) is suitable to describe the underlying dynamics (see section 2.1.3). Thus, healthy and dysfunctional hematopoiesis is modelled as a biochemical reaction network in which each cell compartment (species) is a reactant

$$\mathcal{S} = \{HSC, MPP, MLP, CMP, GMP, MEP, M\}$$

and each reaction corresponds to a differentiation, proliferation, or cell death transition between the compartments with transition rates $\alpha_{(\cdot)}, \beta_{(\cdot)}$, and $\gamma_{(\cdot)}$.

Derivation of a set of 10 comparable lineage hierarchies

Based on the classical model of hematopoiesis (model A, see Figure 4.9 A) and recently reported experimental evidence, we derived nine alternative models, likewise containing compartments HSC, MPP, CMPs, MLP, MEP, GMP, and M, but with different unique sets of direct differentiation transitions between them (models B-J, see Figure 4.9 B-J).

In detail, several studies in humans [Doulatov et al., 2010, 2012, Giebel et al., 2006, Goardon et al., Hao et al., 2001, Reynaud et al., 2003, Rossi et al., 2008] show that progenitor cells in the CD34+CD38 compartment which are CD90+ and CD45RA+, correspond to multipotent lymphoid progenitor cells (MLP), and have lymphoid, macrophage, and dendritic potential. As these results suggest that MLPs can also differentiate to GMPs, we have incorporated this transition in models B, C, E and I (see Figure 4.9 B,C,E).

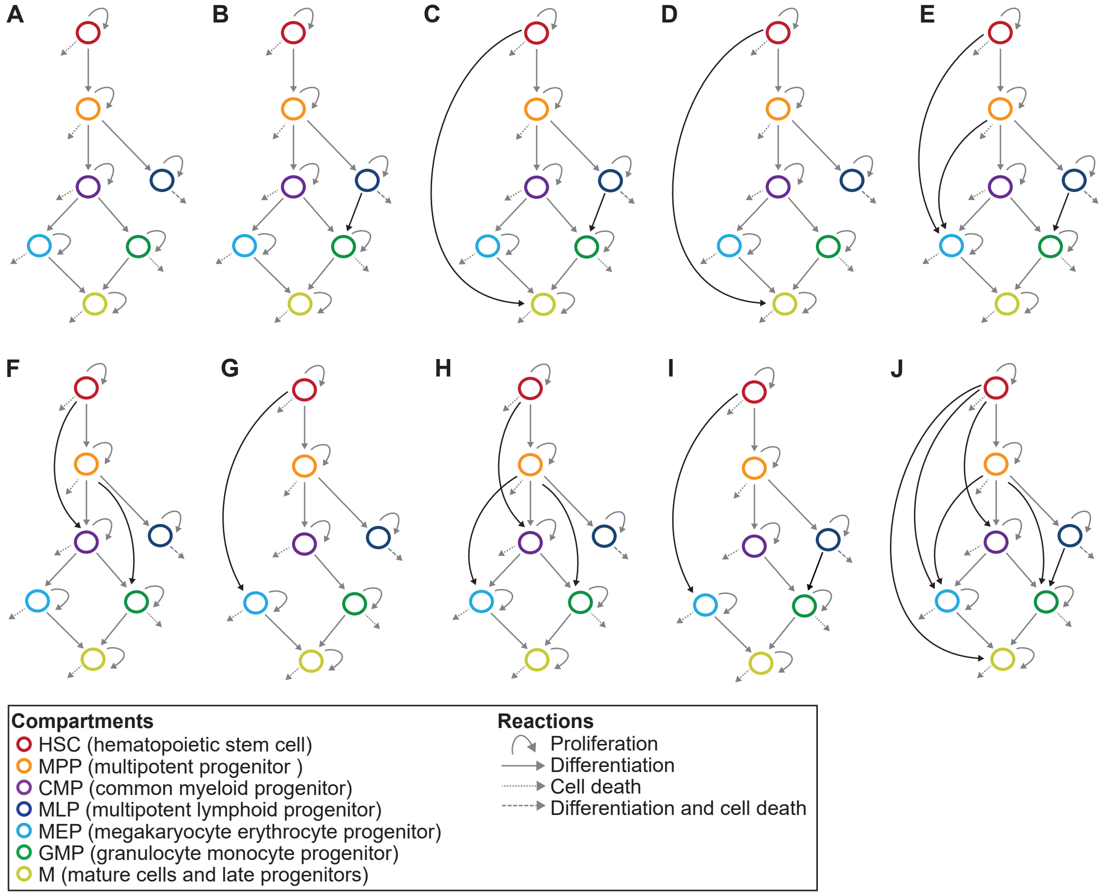

A
B
C
D
E
F
G
H
I
J
Compartments
○ HSC (hematopoietic stem cell)
○ MPP (multipotent progenitor)
○ CMP (common myeloid progenitor)
○ MLP (multipotent lymphoid progenitor)
○ MEP (megakaryocyte erythrocyte progenitor)
○ GMP (granulocyte monocyte progenitor)
○ M (mature cells and late progenitors)
Reactions
Proliferation
Differentiation
Cell death
Differentiation and cell death

Figure 4.9: Suggested lineage hierarchies describing healthy hematopoiesis consisting of HSCs, progenitors (MPPs, MLPs, CMPs, MEPs, GMPs), and a compartment of late progenitors and mature cells (M). Graphic is taken from Bast et al. [2021].

Moreover, in a study investigating adult blood lineage commitment in mice, Adolfsson et al. [2005] proposed a revised model of hematopoiesis. They identified a new cell type, the lymphoid-primed multipotent progenitors (LMPPs), which are FLT3+ Lin Sca-1+c-Kit+ cells (LSK Flt3+ cells), that possess B-cell, T-cell and granulocyte-monocyte (GM) potential but lack megakaryocyte-erythrocyte (MegE) potential. Mouse LSK cells include long-term HSCs, short-term HSCs, and MPPs. The existence of a distinct LSK subtype that does not have MegE potential may indicate that MEPs can directly arise from HSCs. Furthermore, loss of MegE potential in the newly defined LMPP compartment indicates a direct LMPP to GMP transition, without differentiation into CMPs first. Adolfsson et al. [2005] also suggested a hematopoietic model which extends the classical model of hematopoiesis by their findings. In this model, HSCs can generate LMPPs with lymphocyte and GM potential, and CMPs with MegE and GM potential. For the cell type compartments we considered, these findings led on the one hand to the possible direct transition from the HSC to CMP compartment and from MPPs to GMPs in models F and H (see Figure 4.9 F,H), and on the other hand to a transition between the HSC and MEP compartment in models E, G and I (see Figure 4.9 E,G,I). The direct differentiation path from HSCs to MEPs (see Figure 4.9 E,G,I) was also supported by in vitro studies of Takano et al. [2004], who investigated colony forming units of LSK daughter and granddaughter cells. However, a separate study from Forsberg et al. [2006] also investigated the lineage potential of FLT3+ LMPPs but found conflicting results, which instead support the classical model of hematopoiesis (see Figure 4.9 A).

In another mouse study, a fraction of phenotypically defined HSCs was shown to express von Willebrand factor (vWF), a protein mainly expressed by platelets and endothelium [Månsson et al., 2007]. The existence of a megakaryocyte-primed HSC subset was also experimentally investigated by Sanjuan-Pla et al. [2013] generating vWF-eGFP transgenic mice, isolating LSK CD150+CD48CD34 HSCs with a high eGFP expression and transplanting them into irradiated mice. They found that vWF-eGFP+ HSCs were platelet biased, additionally contributing to other myeloid lineages whereas their lymphoid contribution was very marginal.

Models E and H furthermore include the direct differentiation path from MPPs to MEPs, which was suggested by Pronk et al. [2007] (see Figure 4.9 E, H). By studying the phenotypic, functional and molecular characteristics of myeloerythroid precursors, they identified MPPs which give rise to erythroid and megakaryocytic progeny through various intermediate stages. This finding is supported by human studies, in which BAH1 and CD71 were identified as erythroid and megakaryocytic differentiation markers within the CD34+ CD38- MPP compartment [Notta et al., 2016].

Accordingly, the full set of reactions, which describe proliferation, differentiation and cell death transitions is given by

$$\begin{aligned}
R_1 :&\quad HSC \xrightarrow{\beta_{HSC}} 2\,HSC\\
R_2 :&\quad MPP \xrightarrow{\beta_{MPP}} 2\,MPP\\
R_3 :&\quad CMP \xrightarrow{\beta_{CMP}} 2\,CMP\\
R_4 :&\quad MLP \xrightarrow{\beta_{MLP}} 2\,MLP\\
R_5 :&\quad MEP \xrightarrow{\beta_{MEP}} 2\,MEP\\
R_6 :&\quad GMP \xrightarrow{\beta_{GMP}} 2\,GMP\\
R_7 :&\quad M \xrightarrow{\beta_{M}} 2\,M
\end{aligned}$$

$$
\begin{aligned}
R_8: &\quad \mathit{HSC} \xrightarrow{\alpha_{HSC\to MPP}} \mathit{MPP} \\
R_9: &\quad \mathit{HSC} \xrightarrow{\alpha_{HSC\to CMP}} \mathit{CMP} \\
R_{10}: &\quad \mathit{HSC} \xrightarrow{\alpha_{HSC\to MEP}} \mathit{MEP} \\
R_{11}: &\quad \mathit{HSC} \xrightarrow{\alpha_{HSC\to M}} M \\
\\
R_{12}: &\quad \mathit{MPP} \xrightarrow{\alpha_{MPP\to CMP}} \mathit{CMP} \\
R_{13}: &\quad \mathit{MPP} \xrightarrow{\alpha_{MPP\to MLP}} \mathit{MLP} \\
R_{14}: &\quad \mathit{MPP} \xrightarrow{\alpha_{MPP\to MEP}} \mathit{MEP} \\
R_{15}: &\quad \mathit{MPP} \xrightarrow{\alpha_{MPP\to GMP}} \mathit{GMP} \\
\\
R_{16}: &\quad \mathit{CMP} \xrightarrow{\alpha_{CMP\to MEP}} \mathit{MEP} \\
R_{17}: &\quad \mathit{CMP} \xrightarrow{\alpha_{CMP\to GMP}} \mathit{GMP} \\
\\
R_{18}: &\quad \mathit{MLP} \xrightarrow{\alpha_{MLP\to \cdots}} \emptyset \\
R_{19}: &\quad \mathit{MLP} \xrightarrow{\alpha_{MLP\to GMP}} \mathit{GMP} \\
\\
R_{20}: &\quad \mathit{MEP} \xrightarrow{\alpha_{MEP\to M}} M \\
\\
R_{21}: &\quad \mathit{GMP} \xrightarrow{\alpha_{GMP\to M}} M \\
\\
R_{22}: &\quad \mathit{HSC} \xrightarrow{\gamma_{HSC}} \emptyset \\
R_{23}: &\quad \mathit{MPP} \xrightarrow{\gamma_{MPP}} \emptyset \\
R_{24}: &\quad \mathit{CMP} \xrightarrow{\gamma_{CMP}} \emptyset \\
R_{25}: &\quad \mathit{MEP} \xrightarrow{\gamma_{MEP}} \emptyset \\
R_{26}: &\quad \mathit{GMP} \xrightarrow{\gamma_{GMP}} \emptyset \\
R_{27}: &\quad M \xrightarrow{\gamma_{M}} \emptyset
\end{aligned}
$$

where all proliferation and cell death reactions ($R_1, ..., R_7$ and $R_{22} - R_{27}$) are shared between the models, unlike some of the differentiation reactions leading to various model complexities (Table T 4.1).

The classical, widely used lineage hierarchy (model A) is together with model G the simplest model with the least parameters. In contrast, model J (see Figure 4.10) includes all in literature discussed differentiation paths resulting in the most complex model out of the 10 considered models.

To analyse which lineage hierarchies are plausible and which ones can be rejected based on our experimental data, we implemented and fitted a set of 10 biologically motivated lineage hierarchies.

Model	# Reaction rates	Reactions
model A	21	$R_1 - R_8, R_{12} - R_{13}, R_{16} - R_{18}, R_{20} - R_{27}$
model B	22	$R_1 - R_8, R_{12} - R_{13}, R_{16} - R_{27}$
model C	23	$R_1 - R_8, R_{11} - R_{13}, R_{16} - R_{27}$
model D	22	$R_1 - R_8, R_{11} - R_{13}, R_{16} - R_{18}, R_{20} - R_{27}$
model E	24	$R_1 - R_8, R_{10}, R_{12} - R_{14}, R_{16} - R_{27}$
model F	23	$R_1 - R_9, R_{13}, R_{16} - R_{18}, R_{20} - R_{27}$
model G	21	$R_1 - R_8, R_{10}, R_{12} - R_{13}, R_{17}, R_{18}, R_{20} - R_{27}$
model H	22	$R_1 - R_9, R_{13} - R_{14}, R_{16} - R_{18}, R_{20} - R_{27}$
model I	22	$R_1 - R_8, R_{10}, R_{12} - R_{13}, R_{17} - R_{27}$
model J	27	$R_1 - R_{27}$

Table 4.1: Complexity of models A-J

ODE system

As hundreds of cells are observed in bulk in the experiment, the cell division and differentiation process is described deterministically with ODEs (see 2.1). Under the assumption that the cell culture medium contains a sufficiently high concentration of the 8 growth factors, which is higher than the amount required for cell growth, one can model cell growth as an unlimited process with constant proliferation rates. Note, that for MLPs the out-flux reaction is defined as net differentiation and describes differentiation combined with cell death. As the MLP downstream compartment is not measured, the individual MLP rates can not be estimated accurately and combining them is the only possibility to ensure structural identifiability of model parameters and thus to obtain accurate parameter estimates.

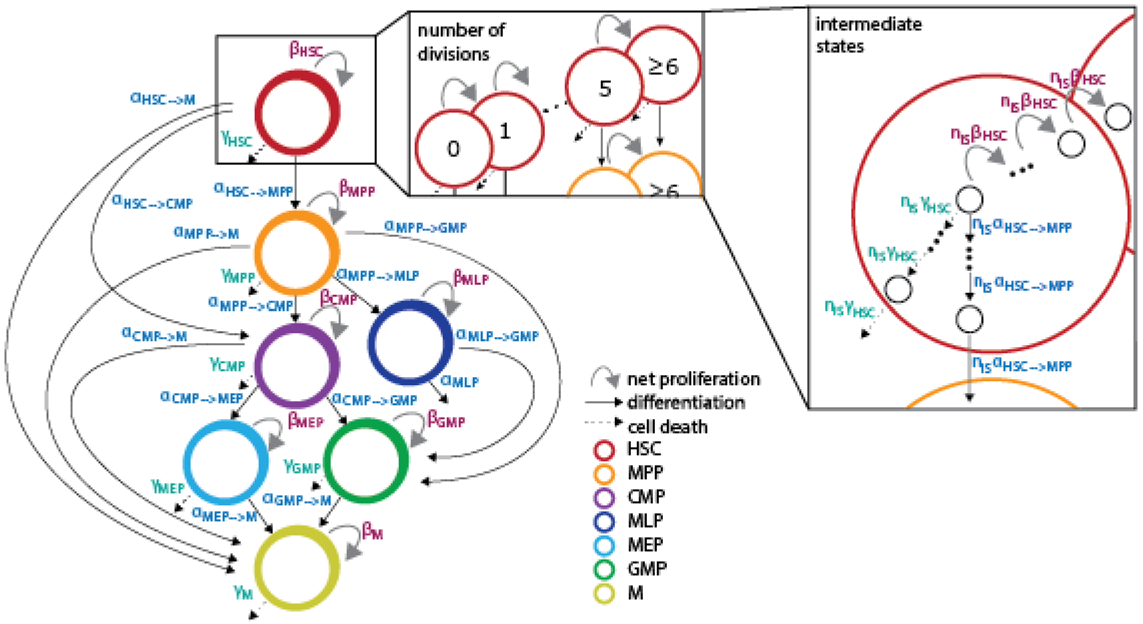


Figure 4.10: Compartmental model with model extensions to include the number of cell divisions and intermediate states for realistic rate distribution assumptions, exemplary shown for lineage hierarchy J. Unknown rates describe cell type specific proliferation, differentiation and death rates (parameter vector). Graphic is taken from Bast et al. [2021] and was slightly modified.

The ODE system describing the evolution of the cell concentrations over time for each compartment can be derived from the reactions above by applying the law of mass action. As all reaction rates are positive, differentiation and cell death reduces and proliferation increases the number of cells within the compartment proportionally to the cell concentration of the respective compartment at time t. The model equations are in the following derived in several steps, including a first extension to incorporate the number of cell divisions and a second extension to incorporate intermediate states, which ensure realistic time distributions for division, differentiation and death reactions (see Figure 4.10).

As an example, the ODE system for model A without intermediate states ($n_{IS} = 1$) and neglecting the number of undergone cell divisions is given by

$$\begin{aligned}
\dot{x}_1 &:= \frac{d[HSC]}{dt} = -(\alpha_{HSC\to MPP} - \beta_{HSC} + \gamma_{HSC})[HSC] \\
\dot{x}_2 &:= \frac{d[MPP]}{dt} = \alpha_{HSC\to MPP}[HSC] - (\alpha_{MPP\to CMP} + \alpha_{MPP\to MLP} - \beta_{MPP} + \gamma_{MPP})[MPP] \\
\dot{x}_3 &:= \frac{d[MLP]}{dt} = \alpha_{MPP\to MLP}[MPP] - (\alpha_{MLP\to \dots} - \beta_{MLP})[MLP] \\
\dot{x}_4 &:= \frac{d[CMP]}{dt} = \alpha_{MPP\to CMP}[MPP] - (\alpha_{CMP\to GMP} + \alpha_{CMP\to MEP} - \beta_{CMP} - \gamma_{CMP})[CMP] \\
\dot{x}_5 &:= \frac{d[GMP]}{dt} = \alpha_{CMP\to GMP}[CMP] - (\alpha_{GMP\to M} - \beta_{GMP} + \gamma_{GMP})[GMP] \\
\dot{x}_6 &:= \frac{d[MEP]}{dt} = \alpha_{CMP\to MEP}[CMP] - (\alpha_{MEP\to M} - \beta_{MEP} + \gamma_{MEP})[MEP] \\
\dot{x}_7 &:= \frac{d[M]}{dt} = \alpha_{GMP\to M}[GMP] + \alpha_{MEP\to M}[MEP] - (\beta_M + \gamma_{MEP})[M] \qquad (4.1)
\end{aligned}$$

with initial conditions $\mathbf{x}(0) = \mathbf{x}_0$.

For any model hierarchy (see Figure 4.9 A-J), the ODE system for $n_{IS} = 1$ can be formulated as

$$\dot{x}_j := \frac{dS_j}{dt} = \sum_{i\in I_j} \alpha_{i\to j} \cdot S_i(t) + \left(\beta_j - \gamma_j - \sum_{o\in O_j} \alpha_{j\to o}\right) \cdot S_j(t), \qquad (4.2)$$

$\forall j = 1, ..., |\mathcal{S}|$, where I_j is the set of influx compartments and O_j the set of outflux compartments of the respective species $S_j \in \mathcal{S}$ and the initial conditions are given by $\mathbf{x}(0) = \mathbf{x}_0$.

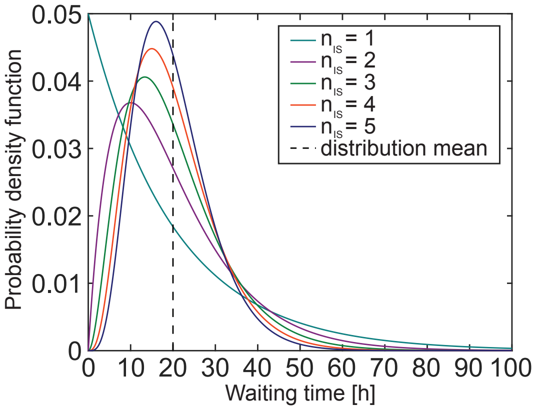


Figure 4.11: Waiting time distributions for 1-5 intermediate states assuming a mean reaction (proliferation, differentiation or death) rate of 1/20 [h-1]. Graphic is taken from Bast et al. [2021].

Incorporating the information of the number of cell divisions N_{div}, the ODE system is expanded by introducing additional states which indicate not only the cell type but also the number of divisions occurring within the time interval of interest $[t_0, t_i^{obs}]$ (see Figure 4.9). Hence, each ODE describes the evolution of cell abundances of species $S_{j,n_{div}} \in \mathcal{S}, j = 1, ...|\mathcal{S}|$ which divided $n_{div} \in \{0, ..., N_{div}\}$ times over time t, which is denoted as $\frac{d[S_{j,n_{div}}]}{dt}$. This leads to an ODE system of $N_{div} \cdot n_c$ equations:

$$\begin{aligned}
\frac{dS_{j,0}(t)}{dt} &= \sum_{i \in I_j} \alpha_{i \to j} \cdot S_{i,0}(t) + \left(\beta_j - \gamma_j - \sum_{o \in O_j} \alpha_{j \to o} \right) \cdot S_j(t) \\
&:= \dot{x}_{(j-1) \cdot N_{div} + 1} \\
\frac{dS_{j,n_{div}}(t)}{dt} &= 2\beta_j \cdot S_{j,n_{div}-1}(t) + \sum_{i \in I_j} \alpha_{i \to j} \cdot S_{i,n_{div}}(t) - \left(\beta_j + \gamma_j + \sum_{o \in O_j} \alpha_{j \to o} \right) \cdot S_{j,n_{div}}(t) \\
&:= \dot{x}_{(j-1) \cdot N_{div} + n_{div} + 1} \\
\frac{dS_{j,N_{div}}(t)}{dt} &= 2\beta_j \cdot S_{j,N_{div}-1}(t) + \sum_{i \in I_j} \alpha_{i \to j} \cdot S_{i,N_{div}}(t) + \left(\beta_j - \gamma_j - \sum_{o \in O_j} \alpha_{j \to o} \right) \cdot S_{j,N_{div}}(t) \\
&:= \dot{x}_{j \cdot N_{div}} \qquad (4.3)
\end{aligned}$$

where $n_{div} = 1, ..., N_{div} - 1$, $j = 1, ..., |\mathcal{S}|$ and initial conditions $\mathbf{x}(0) = \mathbf{x}_0$.
According to the law of mass action, the waiting time T for a reaction to occur is anti-proportional to its reaction rate r and follows an exponential distribution

$$T \sim \exp(r), \quad r \in \{\alpha_{(.)}, \beta_{(.)}, \gamma_{(.)}\}.$$

This is in contrast to the observation that the considered processes cell division, differentiation, and cell death require a minimum time to be completed. To more accurately describe transition times between cell states, we introduced intermediate states and further expanded the model (see Figures 4.10 and 4.14).
By introducing intermediate states, the waiting time to stay in a particular state corresponds to the sum of exponentially distributed waiting times of its n_{IS} intermediate states and is thereby per definition Erlang(n_{IS}, r) distributed [Matis and Wehrly, 1990]. This model extension results in an ODE system with

$$N_{eq} = \left\{ \begin{array}{ll} N_{div} \cdot \left(\left(n_{IS} \cdot \sum_{j=1}^{|\mathcal{S}|} (n_j^{out} + 2) \right) + 1 \right), & n_{IS} > 1 \\ N_{div} \cdot |\mathcal{S}| & , n_{IS} = 1 \end{array} \right\} \qquad (4.4)$$

equations, where n_j^{out} is the number of outfluxes of compartment S_j and n_{IS} is the number of intermediate states within each compartment S_j. Each ODE describes the time evolution of the number of cells of species $S_j \in \mathcal{S}$ that divided $n_{div} \in \{0, ..., N_{div}\}$ times and are in the j-th proliferation, the k-th differentiation and the m-th cell death intermediate state, which is denoted by $\frac{d[S_{j,n_{div},i_o}^{(k,l,m)}(t)]}{dt}$, $k, l, m = 1, ..., n_{IS}$, and $i_o = 1, ..., n_j^{out}$ and $n_{div} = 1, ..., N_{div}$. The ODE system is given by

$$
\begin{array}{rcl}
\dfrac{d[S^{(0,0,0)}_{j,n_{div},1}]}{dt} & := & \begin{cases}
n_{IS}\cdot \left(\sum\limits_{i\in I_j} \alpha_{i\to j}\cdot S^{(0,n_{IS},0)}_{i,n_{div},1}(t) \right. \\
\quad \left. -(\beta_j + \sum\limits_{o\in O_j} \alpha_{j\to o} + \gamma_j)\cdot S^{(0,0,0)}_{j,n_{div},1}(t) \right) & \text{, if } n_{div}=0 \\
n_{IS}\cdot \left(\sum\limits_{i\in I_j} \alpha_{i\to j}\cdot S^{(0,n_{IS},0)}_{i,n_{div},1}(t) \right. \\
\quad -(\beta_j + \sum\limits_{o\in O_j} \alpha_{j\to o} + \gamma_j)\cdot S^{(0,0,0)}_{j,n_{div},1}(t) \\
\quad \left. +\beta_j\cdot S^{(n_{IS},0,0)}_{j,n_{div}-1,1}(t) \right) & \text{, if } n_{div}\in\{1,...,N_{div}-1\} \\
n_{IS}\cdot \left(\sum\limits_{i\in I_j} \alpha_{i\to j}\cdot S^{(0,n_{IS},0)}_{i,n_{div},1}(t) \right. \\
\quad -(\beta_j + \sum\limits_{o\in O_j} \alpha_{j\to o} + \gamma_j)\cdot S^{(0,0,0)}_{j,n_{div},1}(t) \\
\quad \left. +\beta_j\cdot S^{(n_{IS},0,0)}_{j,n_{div}-1,1}(t) + 2\cdot\beta_j\cdot S^{(n_{IS},0,0)}_{j,n_{div},1}(t) \right) & \text{, if } n_{div}=N_{div}
\end{cases} \\
 & = & \dot{x}_{(j-1)\cdot(N_{div}\cdot(2\cdot n_{IS}+1)+\sum_{c=1}^{j-1} n_c^{out}\cdot n_{IS}\cdot N_{div}+(n_{div}+1)} \\
\dfrac{d[S^{(k,0,0)}_{j,n_{div},1}(t)]}{dt} & := & n_{IS}\cdot\beta_j\left(S^{(k-1,0,0)}_{j,n_{div},1}(t) - S^{(k,0,0)}_{j,n_{div},1}(t)\right) \\
 & = & \dot{x}_{(j-1)\cdot(N_{div}\cdot(2\cdot n_{IS}+1)+\sum_{c=1}^{j-1} n_c^{out}\cdot n_{IS}\cdot N_{div}+(n_{div}+k+1)} \\
\dfrac{d[S^{(0,0,m)}_{j,n_{div},1}]}{dt} & := & n_{IS}\cdot\gamma_j\left(S^{(0,0,m-1)}_{j,n_{div},1}(t) - S^{(0,0,m)}_{j,n_{div},1}(t)\right) \\
 & = & \dot{x}_{(j-1)\cdot(N_{div}\cdot(2\cdot n_{IS}+1)+\sum_{c=1}^{j-1} n_c^{out}\cdot n_{IS}\cdot N_{div}+(n_{div}+m+n_{IS}+1)} \\
\dfrac{d[S^{(0,l,0)}_{j,n_{div},i_o}]}{dt} & := & n_{IS}\cdot\alpha_{j\to i_o}\left(S^{(0,l-1,0)}_{j,n_{div},i_o}(t) - S^{(0,l,0)}_{j,n_{div},i_o}(t)\right) \\
 & = & \dot{x}_{(j-1)\cdot(N_{div}\cdot(2\cdot n_{IS}+1)+\sum_{c=1}^{j-1} n_c^{out}\cdot n_{IS}\cdot N_{div}+(n_{div}+l+i_o+2\cdot n_{IS}+1)}
\end{array}
\tag{4.5}
$$

where $n_{div} = 0, ..., N_{div}$, $i_o = 1, ..., n_j^{out}$, $k, l, m = 1, ..., n_{IS}$, $j = 1, ..., |\mathcal{S}|$ and initial conditions $\mathbf{x}(0) = \mathbf{x}_0$. The model allows to describe up to N_{div} division compartments per cell type compartment and if cells divided more often (more than N_{div} times), they accumulate in the N_{div}-compartment of the respective species S_j.

Note that for both model extensions, the number of states increases, but the number of parameters stays constant.

Model assumptions

In order to perform MLE (see section 4.3), biologically meaningful boundaries for parameter values $\boldsymbol{\theta}$ need to be specified (see table T 4.2). I assumed a minimum mean transition time of 1 and a maximum mean transition time of 500 hours for proliferation, differentiation and cell death times. For initial conditions of the first HSC compartment ($n_{div} = 0$), I assumed the lower parameter boundary to be equal to the lowest observed number of HSCs which divided once at the first observed time point t_1. and an upper boundary equal to the HSC starting population N_{input}. For other compartments the lower boundary for the initial condition is 0 cells and the upper boundary is the maximum plus 10% of the observed number of Cells in the respective compartment at the first observed time point t_1.

parameter	boundaries
$\beta_{S_j}, \gamma_{S_j}, \alpha_{S_{j1} \to S_{j2}}$, where $S_j, S_{j1}, S_{j2} \in \mathcal{S}$	$\left[\frac{1}{500}, 1\right] h^{-1}$
$x_0^{(w)}$	$\left[\underset{r}{min}\left\{y_{1,0}^{\mathcal{D}(w,r)}(t_1)\right\}, N_{input}\right]$, if $j = 1$ (for HSCs) $\left[0, \underset{r}{max}\left\{y_{j,0}^{\mathcal{D}(w,r)}(t_1)\right\} + 0.1 * \underset{r}{max}\left\{y_{j,0}^{\mathcal{D}(w,r)}(t_1)\right\}\right]$, otherwise

Table 4.2: Parameter boundaries used for fitting model to experimental data

4.3 Parameter inference

In order to assess how well $\mathcal{M}(\boldsymbol{\theta})$ fits the experimental data for a certain set of parameters θ, the log-likelihood $\ell_D(\boldsymbol{\theta})$ is calculated and maximized as introduced in section 2.2.1. The 10 considered models contain between 29 and 42 unknown parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_{n_\theta})$, which are the reaction rates $\beta_{S_j}, \gamma_{S_j}, \alpha_{S_{j1} \to S_{j2}}$, where $S_j, S_{j1}, S_{j2} \in \mathcal{S}$ and the initial conditions of repetition $w \in \{1, 2\}$ is given by $\mathbf{x}^{(w)}(0) = \mathbf{x}_0^w(\theta)$, where

$$\mathbf{x}_0^{(w)}(\theta) = S_{j,n_{div},i_o}^{(k,l,m)(w)}(0) = \begin{cases} \theta_j, & \text{if } n_{div} = k = l = m = 0 \text{ and } i_o = 1 \\ 0, & \text{otherwise} \end{cases}$$

for $j = 1, ..., |\mathcal{S}|$, repetition $w \in \{1, 2\}$ and $n_{div} = 1, ..., 6$. These parameters are estimated by minimizing the weighted difference between observed and modelled cell counts by applying maximum likelihood estimation.

Let $\mathcal{M}(\theta)$ be a particular model consisting of dynamics $\dot{\mathbf{x}} = f(\mathbf{x}, \theta)$ and model observables $y^{\mathcal{M}} = h(x, \theta)$:

$$\mathcal{M}(\theta) : \left\{ \begin{array}{l} \dot{\mathbf{x}} = f(\mathbf{x}, \theta) = \left\{ \frac{d[S_{j,n_{div},i_o}^{(k,l,m)}](t)}{dt} \right\}, \ \mathbf{x}^{(w)}(0) = \mathbf{x}_0^w(\theta), \\ \mathbf{y}^{\mathcal{M}} = h(x, \theta) = \left\{ \sum_{k,l,m=0}^{n_{IS}} \sum_{i_o=1}^{n_{out}^j} [S_{j,n_{div},i_o}^{(k,l,m)}](t) \right\} \end{array} \right\}, \tag{4.6}$$

where $j = 1, ..., |\mathcal{S}|$, $k, l, m = 1, ..., n_{IS}$, $i_o = 1, ..., n_{out}^j$, $n_{div} = 1, ..., N_{div}$ and $w \in \{1, 2\}$ and let

$$\mathcal{D} = \left\{t_s, y_{j,n_{div}}^{\mathcal{D}}(t_s)\right\}_{s=1,..,n_t,\ j=1,...,|\mathcal{S}|,\ n_{div}=1,...,N_{div}} \tag{4.7}$$

be the data (see Figure 4.4). Here $\mathbf{y}^{\mathcal{D}}(t_s)$ denotes the vector of observed cell counts of species $j = 1, ..., |\mathcal{S}|$ that divided $n_{div} = 1, ..., N_{div}$ times at time t_s of a particular individual. For parameter estimation, I assumed the observations $y_{j,n_{div}}^{\mathcal{D}}(t_s)$ are subject to multiplicative log-normally distributed measurement noise

$$\begin{aligned} y_{j,n_{div}}^{\mathcal{D}}(t_s) &= y_{j,n_{div}}^{\mathcal{M}}(t_s) \cdot \nu, \text{ with } \nu \sim log\mathcal{N}(0, \sigma_{j,n_{div}}^2) \\ \log(y_{j,n_{div}}^{\mathcal{D}}(t_s)) &= \log(y_{j,n_{div}}^{\mathcal{M}}(t_s)) + \epsilon, \text{ with } \epsilon \sim \mathcal{N}(0, \sigma_{j,n_{div}}^2) \end{aligned} \tag{4.8}$$

due to counting errors (i.e. technical error of the FACS machine) or false cell type assignment while processing raw FACS data by gating (see section 2.2.1.1).

4.3.1 Maximum likelihood estimation

According to the assumed multiplicative log-normally distributed measurement noise the log-likelihood function $\ell_D(\theta)$ is defined as

$$\ell_D(\theta) = -\frac{1}{2}\sum_{w=1}^{n_w}\sum_{r=1}^{n_r}\sum_{s=1}^{n_t}\sum_{n_{div}=1}^{N_{div}}\sum_{j=1}^{|\mathcal{S}|}\log(2\pi\sigma^2_{j,n_{div}}) + \left(\frac{\left(\log(\mathbf{y}^{\mathcal{D}(w,r)}_{j,n_{div}}(t_s)+1) - \log(\mathbf{y}^{\mathcal{M}(w,r)}_{j,n_{div}}(t_s,\theta)+1)\right)^2}{\sigma^2_{j,n_{div}}}\right). \tag{4.9}$$

In order to estimate the unknown parameter vector θ, the optimization problem

$$\theta^{ML} = \underset{\substack{\theta \\ \text{subject to } \mathcal{M}}}{\text{argmax}}\ \ell_D(\theta), \tag{4.10}$$

is solved using local hierarchical optimization [Loos et al., 2018] with trust-region-reflective algorithm [Moré and Sorensen, 1983] and $n_{MS} = 1000$ multi starts (see section 2.2.1.2). With the hierarchical optimization approach $\sigma^2_{j,n_{div}}, j = 1, ..., |\mathcal{S}|, n_{div} = 1, ..., n_{div}$ is analytically calculated each time the log-likelihood function is evaluated. The noise parameter is therefore not part of the parameter vector θ. The starting values $(\theta_i^{start})_{i=1,...,n_{MS}}$ (initial parameter vectors) are determined according to latin hypercube sampling [Eliáš and Vořechovskỳ, 2016]. The resulting optimal parameter is observed at the highest ℓ_D value. To ensure that the optimization procedure converged, I checked if the highest log-likelihood value is observed several times for different starting values.

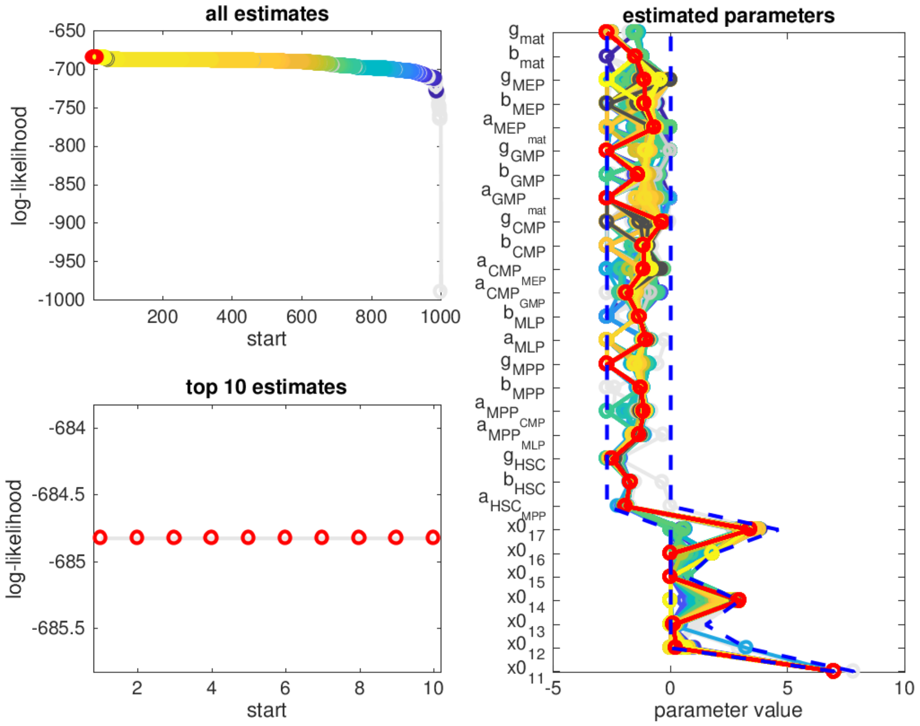


Figure 4.12: While optimizing parameters of healthy donor (H311) with model A for $n_{MS} = 1000$ different starting values the highest log likelihood value (right upper panel) and the corresponding optimal parameter values (left) was found more than 10 times (right bottom panel).

As can be seen in Figure 4.12, the highest log-likelihood value ($\ell_D(\theta^{ML}) = -244$) is observed several times (more than 10 times), suggesting the algorithm converged to the local optimum within the defined parameter boundaries (see Table 4.2). I performed this analysis for all measurements observed at $t_i \in [1, 7]$ of each sample individually and with every model hierarchy A-J using MATLAB toolboxes AMICI [Fröhlich et al., 2016] for model definition and PESTO [Stapor et al., 2017] for parameter estimation.

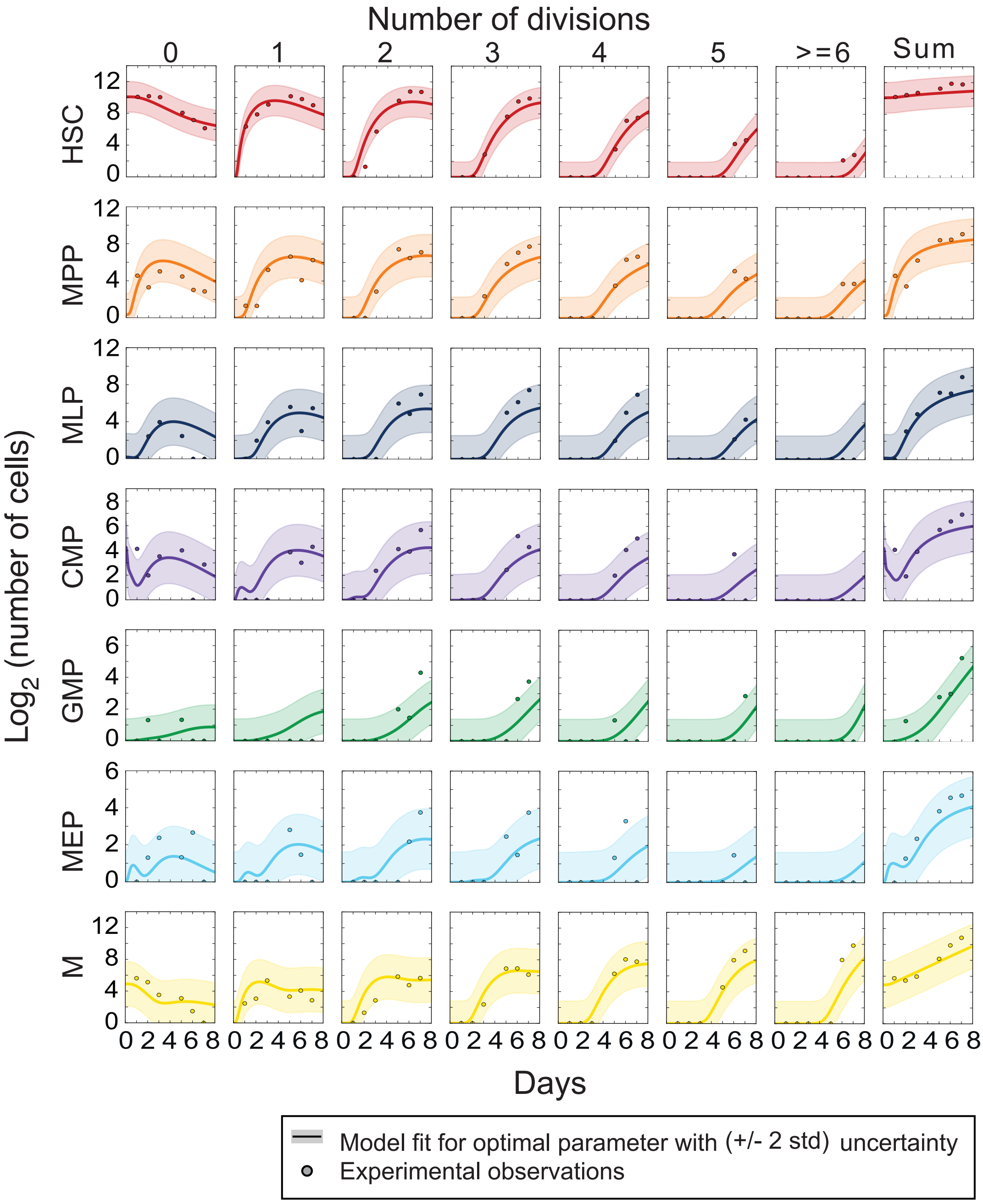


Figure 4.13: Model A fitted (line) to experimental data obtained from bone marrow sample of a healthy donor (H311, dots). The error band corresponds to the model fit for the optimal parameter $\pm 2\hat{\sigma}$ and depicts the model uncertainty. Rows indicate cell type compartments and the first seven columns refer to division compartments. The last column shows the sum over division compartments for every cell type compartment and was not explicitly fitted. Graphic is taken from Bast et al. [2021].

An exemplary model fit can be seen in Figure 4.13. The ODE model for hierarchy A is able to explain the experimentally observed cell counts of the various compartments for sample H311 with the introduced parameter inference approach. This has likewise been observed for almost every other donor sample regardless of disease status.

For very few donor samples either the number of cells, or the number of time points analysed, or both were not sufficient, which led to inaccurate measurements and poor model fits. These samples have been discarded from the analyses and are not included in the results depicted in following sections.

4.3.2 Identifiability analysis

To analyse parameter identifiability, I performed a structural and a practical identifiability analysis (see section 2.2.3). A structural identifiability analysis was performed using a method introduced by [Villaverde and Banga, 2017] and the MATLAB toolbox STRIKE-GOLDD [Villaverde et al., 2019a]. This analysis revealed which parameters are unidentifiable for the different hierarchies if one assumes ideal noise-free data with a large sample size. The results for models A-J under the assumption of no and three intermediate states can be found in Table T4.3. Interestingly, increasing the state space by introducing 3 intermediate states improves the structural identifiability of parameters, as for $n_{IS} = 3$ only the initial conditions are non-identifiable.

Model	Unidentifiable parameters	
	$n_{IS} = 1$	$n_{IS} = 3$
model A	$\mathbf{x}_0(\theta)$	$\mathbf{x}_0(\theta)$
model B	$\mathbf{x}_0(\theta)$	$\mathbf{x}_0(\theta)$
model C	$a_{GMP\to M}, g_{GMP}, \mathbf{x}_0(\theta)$	$\mathbf{x}_0(\theta)$
model D	$a_{GMP\to M}, g_{GMP}, \mathbf{x}_0(\theta)$	$\mathbf{x}_0(\theta)$
model E	$a_{CMP\to MEP}, g_{CMP}, \mathbf{x}_0(\theta)$	$\mathbf{x}_0(\theta)$
model F	$\mathbf{x}_0(\theta)$	$\mathbf{x}_0(\theta)$
model G	$\mathbf{x}_0(\theta)$	$\mathbf{x}_0(\theta)$
model H	$\mathbf{x}_0(\theta)$	$\mathbf{x}_0(\theta)$
model I	$\mathbf{x}_0(\theta)$	$\mathbf{x}_0(\theta)$
model J	$a_{MLP\to GMP}, a_{MLP}, \mathbf{x}_0(\theta)$	$\mathbf{x}_0(\theta)$

Table 4.3: Unidentifiable parameters for models A-J

Practical identifiability as introduced in section 2.2.3.2 was analysed by calculating the Profile-Likelihood-based confidence intervals for all maximum likelihood estimates, which can be seen in Figures 4.15 (see section 4.4), 4.18, and 4.22 (see section 4.5). Depending on the sample and the respective rate, parameters obtained very narrow confidence intervals, or rather broad confidence intervals, but were in most cases practically identifiable. The confidence intervals of some cell death and differentiation rates of downstream compartments as for instance MEP and GMP reached the upper or lower parameter boundary and were thus not practically identifiable.

4.4 Comparison of lineage hierarchies

To computationally test the plausibility of the 10 lineage hierarchies (see Figure 4.9), I performed parameter estimation with all models $\mathcal{M}_q \in \{A, B, C, D, E, F, G, H, I, J\}$ for a subset of 10 healthy individuals (H439, H547, H482, H522, H370, H353, H311, H380, H312, H559) of varying ages.

Determination of Hyperparameter

First, I optimized the number of intermediate states n_{IS} that are required to accurately and efficiently estimate and practically identify model parameters. To optimize this hyperparameter at affordable computational cost, I randomly picked a subset of 4 samples (H353, H559, H482, H522) and fitted them with every model $\mathcal{M}_q$ for a range of $n_{IS} \in \{1, 2, 3, 4, 5\}$ intermediate states. One can observe that for all models A-J the log-likelihood increases with n_{IS} (see Figure 4.14 A). However, the mean percentage increase in log-likelihood per additional intermediate state is below 3% if the model contains more than 3 intermediate states. As the percentage of practically identifiable parameters plateaus at 3 intermediate states (see Figure4.14 B) and the computation time per sample increases exponentially with n_{IS} (see Figure4.14 C), I fixed the hyperparameter n_{IS} and performed the remaining analysis with 3 intermediate states.

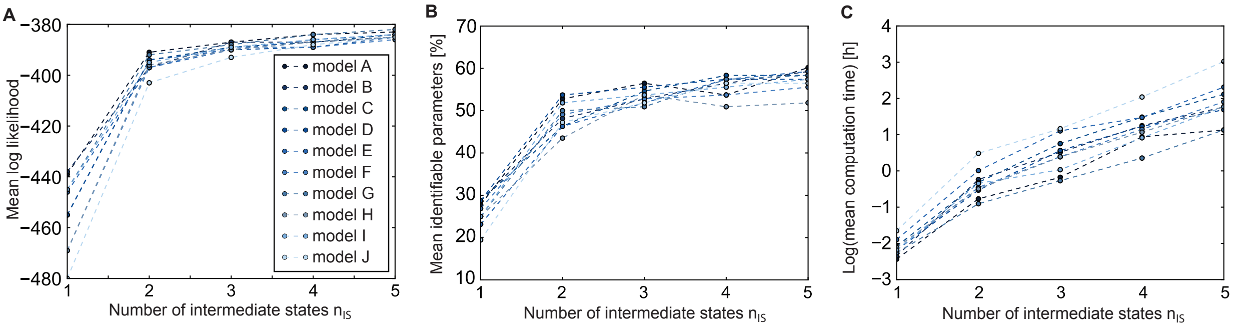


Figure 4.14: Mean log likelihood for fitting models A-J (10 lines) to 4 individual samples increases with any additionally introduced intermediate state n_{IS} (A). Mean percentage of practically identifiable parameters sorted by models A-J (10 lines) based on 4 individual samples (B). Log mean computation time in hours for fitting models A-J (10 lines) to the 4 individual samples increases with the number of intermediate states irrespective of the model hierarchy. For every sample and lineage hierarchy, optimization of 1000 multi starts was run in parallel on 24 workers (C). Graphic is taken from Bast et al. [2021].

Performance assessment of the 10 lineage hierarchies

After fixing the hyperparameter to $n_{IS} = 3$ intermediate states, I fitted every model A-J to all 10 samples from healthy donors (see Figure 4.15 for parameter estimates and their 95% confidence intervals resulting from fitting with model A) and ranked the different models based on their BIC value (see equation 2.29 in section 2.3.2) to compare them.

The BIC values for models $\mathcal{M}_q \in \{A, B, C, D, E, F, G, H, I, J\}$ with $n_{IS} = 3$ intermediate states for every individual separately and in total are visualized in Figure 4.16 A. Additionally, we assessed how often a respective model was the best performing one (lowest score amongst the considered models), amongst the plausible models (BIC difference in score to best model < 10), or how often it was rejected (BIC difference to best model ≥ 10) and summarized this result in a stacked barplot (see Figure 4.16 B). Based on our analysis, there is no evidence in our data for models E, H, and J, which were rejected for all donor samples according to BIC. Models C, F and I also performed poorly, as they were rejected for 80-90% of samples.

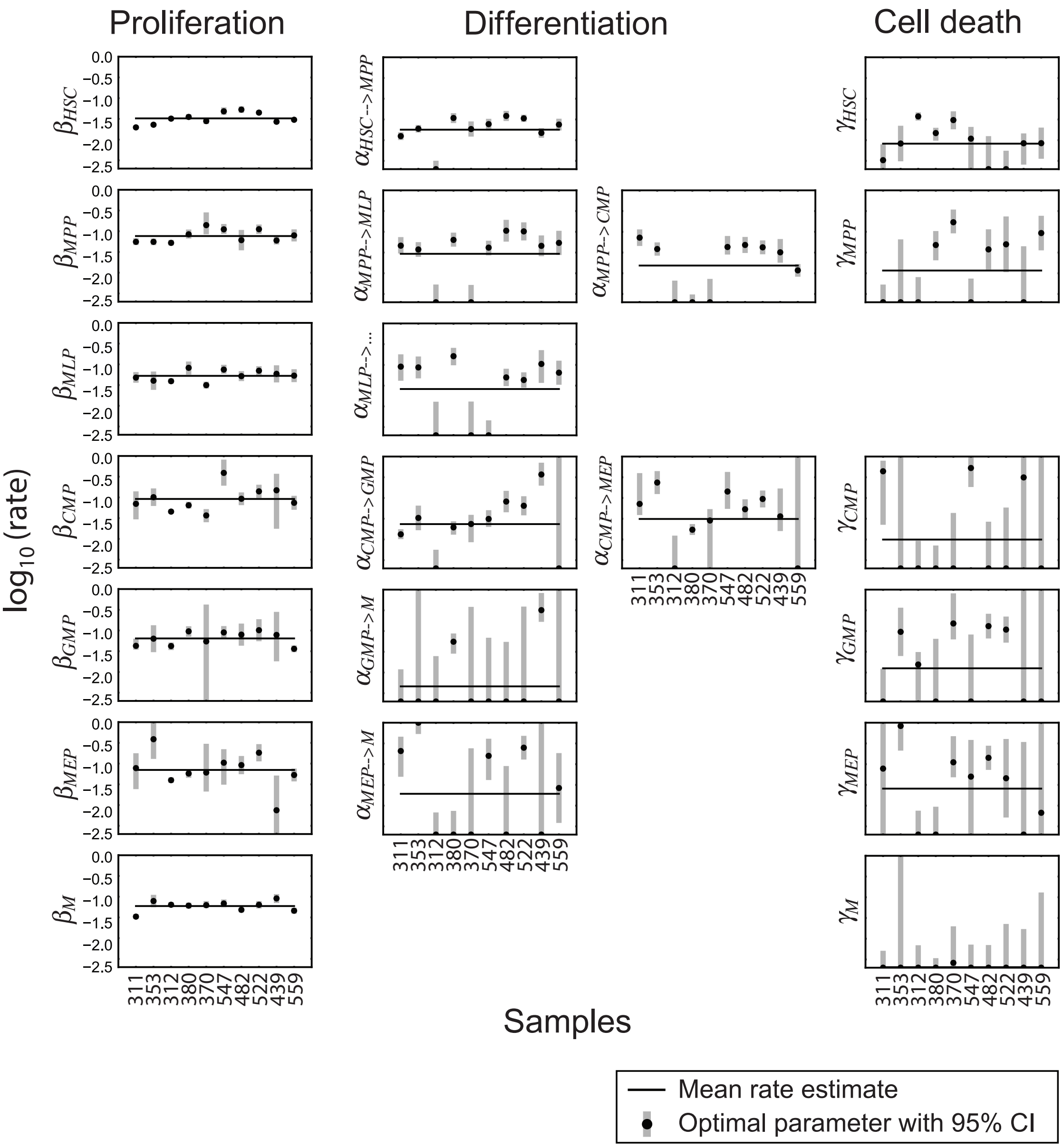
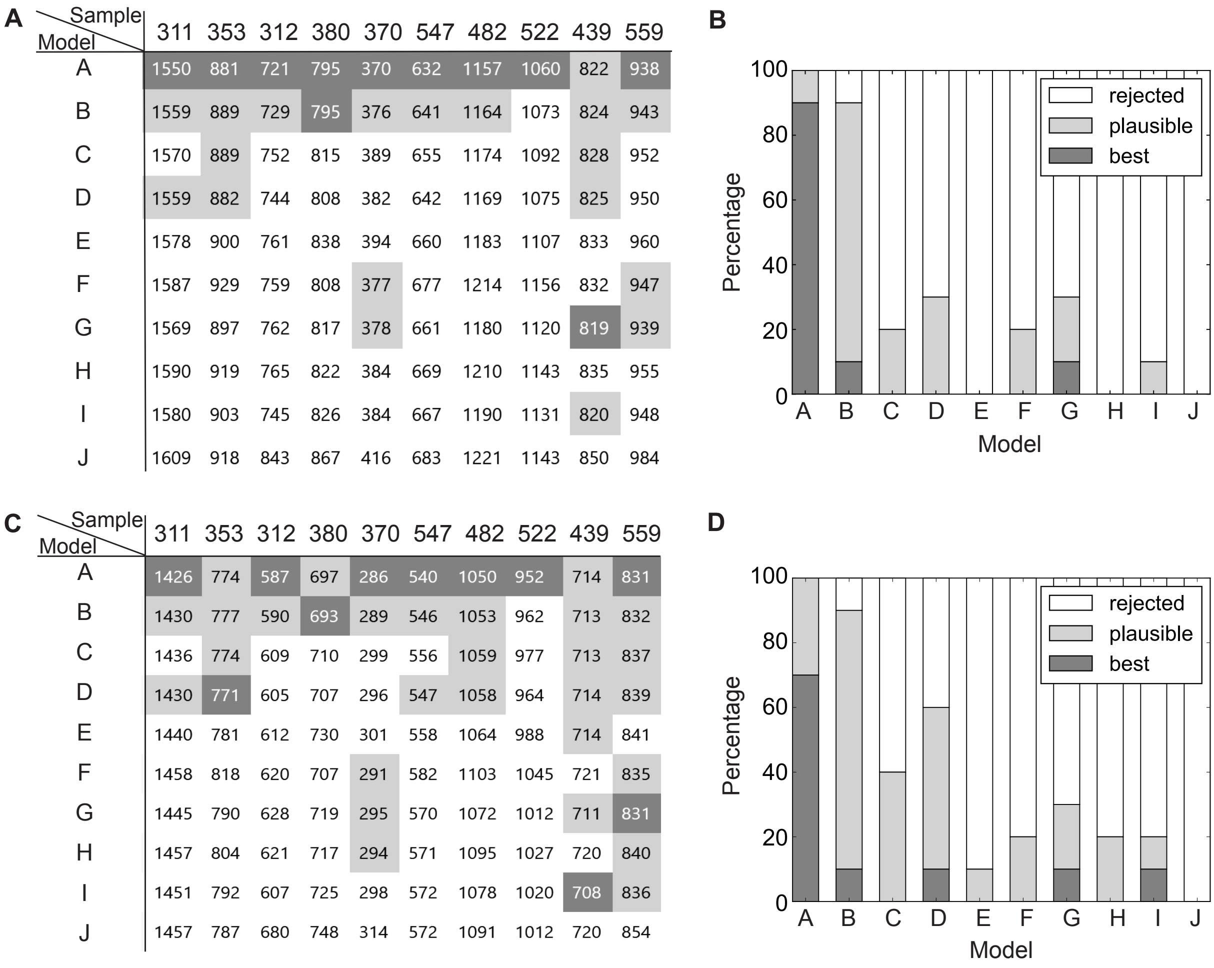


Figure 4.15: Optimal parameter values (dots) with 95% confidence intervals (boxes) resulting from fitting model A to all 10 samples. Graphic is taken from Bast et al. [2021].

The only model which was not rejected by a single sample is the classical lineage hierarchy A. It was selected as the best performing model in 90% of the samples based on BIC. Interestingly, model B has been considered as plausible for 90% of the samples according to BIC. There is also some evidence for models D and G, which are plausible for 30% of the samples.

To investigate if the outstanding performance of model A mainly stems from its low complexity, we additionally calculated the Akaike Information criterion (see equation 2.27 in section 2.3.2, Figure 4.16 C,D). As this selection criteria penalizes the number of parameters in the model differently than BIC, it can potentially select and reject different models. As can be seen in Figure 4.17, AIC is a less conservative scoring method compared to BIC for the number of datapoints $n_{obs} \in [49 \cdot 3, 49 \cdot 7]$ that were used for the fit and the absolute differences in the number of parameters $|n_\theta^{\mathcal{M}_i} - n_\theta^{\mathcal{M}_j}| \in [0, 6]\,,\ \ i, j = 1, ..., 10$. For AIC we again found support for models A and B whereas models E, F, H, I and J performed as poorly as in the BIC ranking (Figure 4.16 C,D). According to AIC there is more evidence for model C (rejected in only 60% of samples as compared to 80% with BIC) and model D (rejected in only 40% of the samples respectively compared to 70% with BIC), but not for model G (rejected in 70% of the samples for both criteria). With a rejection percentage of at least 70% of samples according to both criteria (AIC and BIC), models E-J can be overall rejected.



A

Model \ Sample	311	353	312	380	370	547	482	522	439	559
A	1550	881	721	795	370	632	1157	1060	822	938
B	1559	889	729	795	376	641	1164	1073	824	943
C	1570	889	752	815	389	655	1174	1092	828	952
D	1559	882	744	808	382	642	1169	1075	825	950
E	1578	900	761	838	394	660	1183	1107	833	960
F	1587	929	759	808	377	677	1214	1156	832	947
G	1569	897	762	817	378	661	1180	1120	819	939
H	1590	919	765	822	384	669	1210	1143	835	955
I	1580	903	745	826	384	667	1190	1131	820	948
J	1609	918	843	867	416	683	1221	1143	850	984

C

Model \ Sample	311	353	312	380	370	547	482	522	439	559
A	1426	774	587	697	286	540	1050	952	714	831
B	1430	777	590	693	289	546	1053	962	713	832
C	1436	774	609	710	299	556	1059	977	713	837
D	1430	771	605	707	296	547	1058	964	714	839
E	1440	781	612	730	301	558	1064	988	714	841
F	1458	818	620	707	291	582	1103	1045	721	835
G	1445	790	628	719	295	570	1072	1012	711	831
H	1457	804	621	717	294	571	1095	1027	720	840
I	1451	792	607	725	298	572	1078	1020	708	836
J	1457	787	680	748	314	572	1091	1012	720	854

Figure 4.16: BIC (A) and AIC (C) values resulting from fitting all considered lineage hierarchies A-J to samples of 10 healthy individuals. Three categories were defined based on BIC scoring: models were categorized into best (for the lowest score, dark gray), plausible (a difference to the lowest score of less than or equal to 10, light gray) and rejected (a difference to the lowest score of more than 10, white) models for each donor sample. The stacked barplots of BIC (B) and AIC (D) values details the relative frequency in % of a model to belong to one of the three categories based on the 10 donor samples. Graphic is taken from Bast et al. [2021] and was slightly modified.

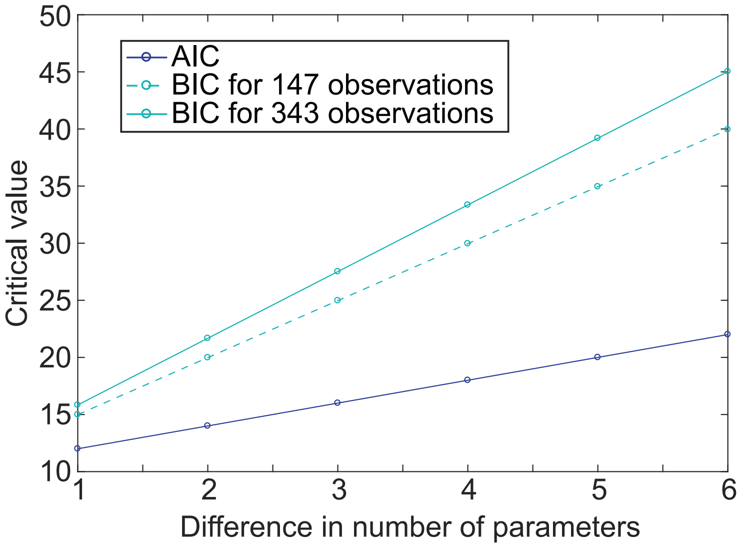


Figure 4.17: Critical values for rejecting a model over another less complex model according to BIC and AIC for a difference in number of parameters of 1 to 6 are shown for the upper and lower boundary of observations. BIC shows larger critical values and is thus always a more conservative criterion compared to AIC. Graphic is taken from Bast et al. [2021].

4.5 Comparison of healthy and perturbed hematopoiesis

Fitting the best performing model A (see section 4.4) to all available donor samples allowed me to resolve parameter changes that arise with age in healthy individuals (see Figure 4.18), and with disease in comparison to healthy age-matched controls (see Figure 4.22).

Age-related changes in human hematopoiesis

To statistically test if donor age has a significant influence on parameters, I performed a linear regression analysis for each rate (target variable) with donor age as covariate. I found that most rates do not significantly change, only HSC and CMP proliferation rates and the CMP death rate are significantly decreasing ($p = 0.015$, $p = 0.026$ and $p = 0.029$ respectively) with age on a significance level of $\alpha = 0.05$ (see Figure 4.19). After correction for multiple testing with Bonferroni [Miller, 1966], none of these p-values (0.315, $0,546$ and 0.609, respectively) is significant. Moreover, this linear regression analysis does not take practical parameter identifiability into account and CMP death rates were not practically identifiable for most individuals (see Figure 4.18), such that one can only conclude that decreased HSC and CMP proliferation rates mainly contribute to an age-related decrease in hematopoiesis, as they were overall practically identifiable.

Interestingly, Busch et al. [2015] found with their computational approach based on *in vivo* mouse lineage tracing experimental data, that the differentiation from MPPs to CMPs stays constant with age, while differentiation towards the lymphoid lineage declines with age. Based on my analysis on human *in vitro* data, both differentiation rates ($\alpha_{MPP \to MLP}$ and $\alpha_{MPP \to CMP}$) decline slightly, but not significantly with age (see Figure 4.19).

Importantly, the here uncovered decrease in HSC and CMP proliferation rates with age can explain why a larger fraction of HSCs is observed with age (see section 4.1 and Figure 4.5): decreasing HSC and CMP proliferation leads to a reduction in newly produced HSCs and HSPCs, and subsequently to a drastic reduction in newly produced mature cells, which can be seen in the yield distributions for young and aged individuals (see Figure 4.7). Thus, the decrease in the two proliferation rates also supports theory (iii), as introduced in Figure 4.6.

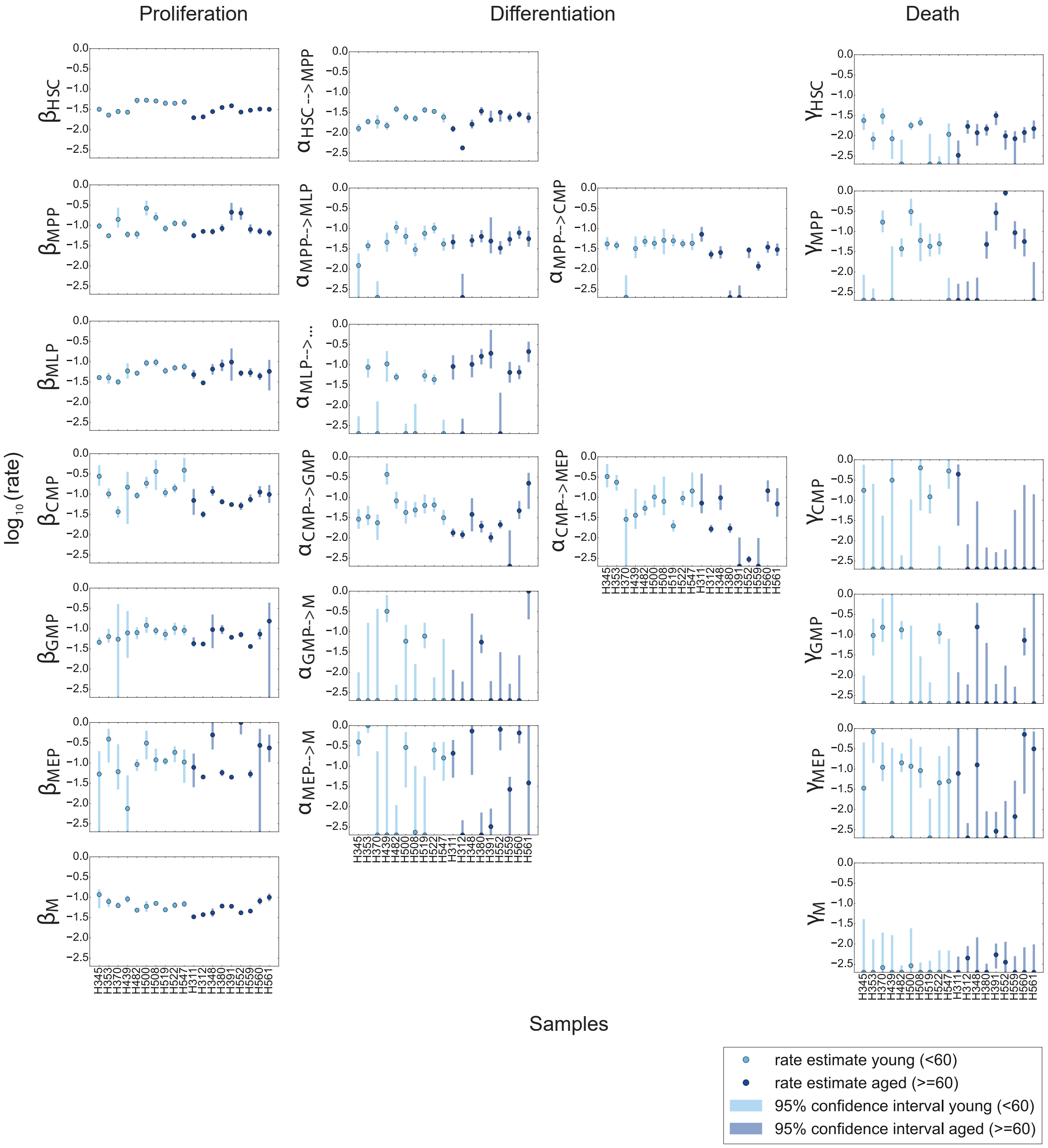


Figure 4.18: Optimal parameter values and their 95% confidence interval resulting from fitting healthy young (light blue) and aged (dark blue) donor samples with model A. Graphic is taken from unpublished manuscript Buck et al.

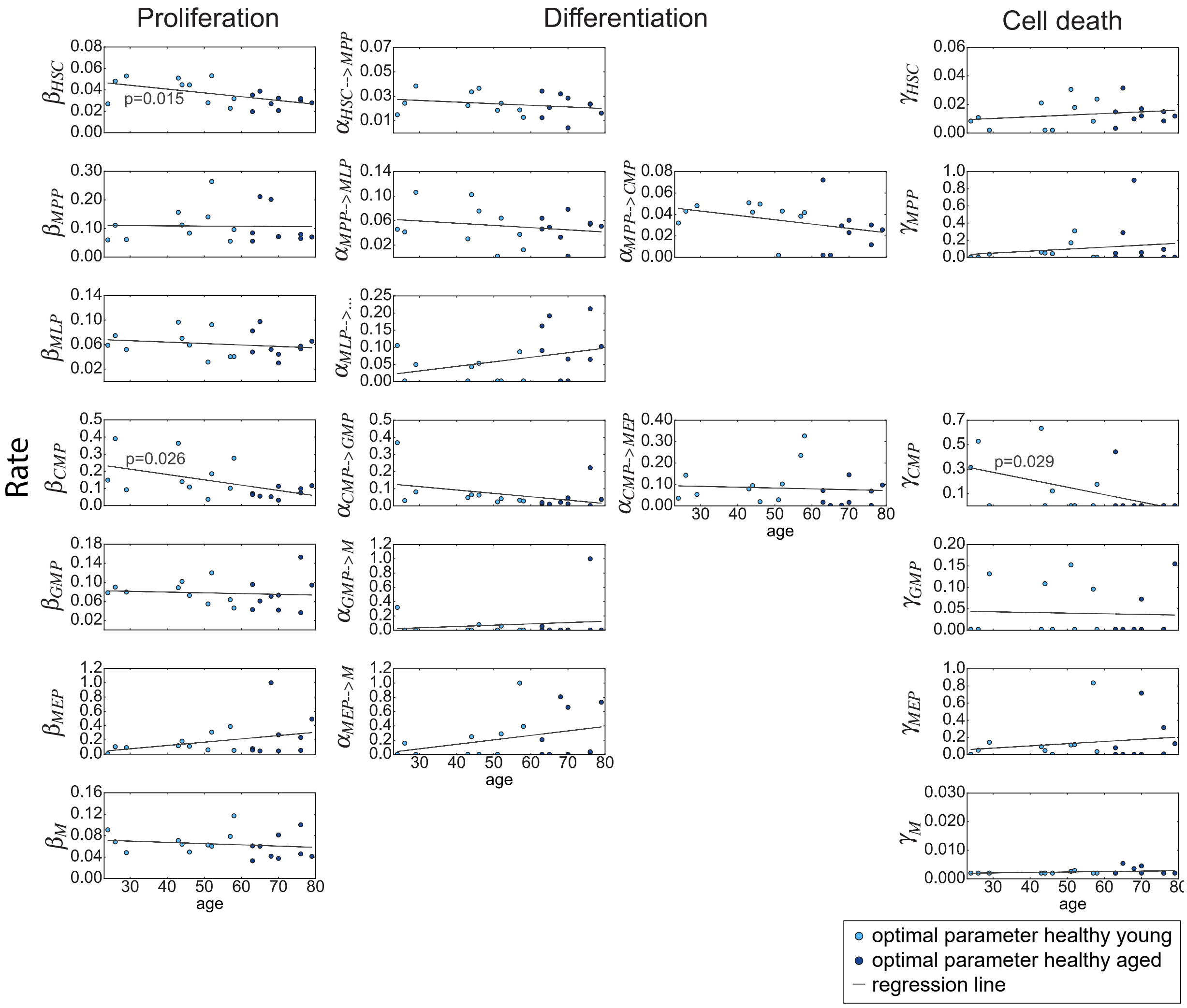


Figure 4.19: Proliferation, differentiation, and cell death rates (columns) of all cell type compartments (rows) vs. age of healthy young and aged donors. Lines show linear regression line for the influence of donor age on the respective rate and p-values are shown if donor age has a significant ($p < 0.005$) influence on rate. Graphic is taken from unpublished manuscript Buck et al.

Hematopoietic changes in CHIP and MDS in comparison to age-matched healthy controls

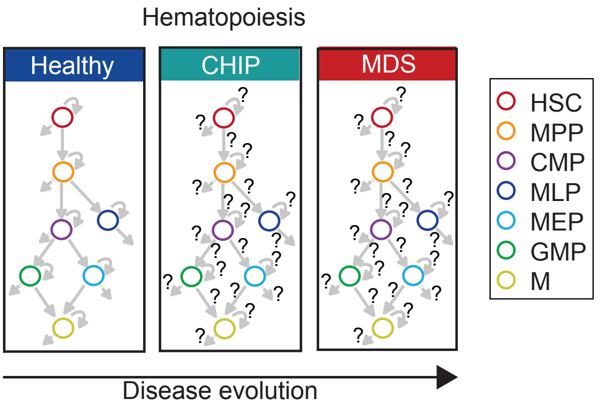
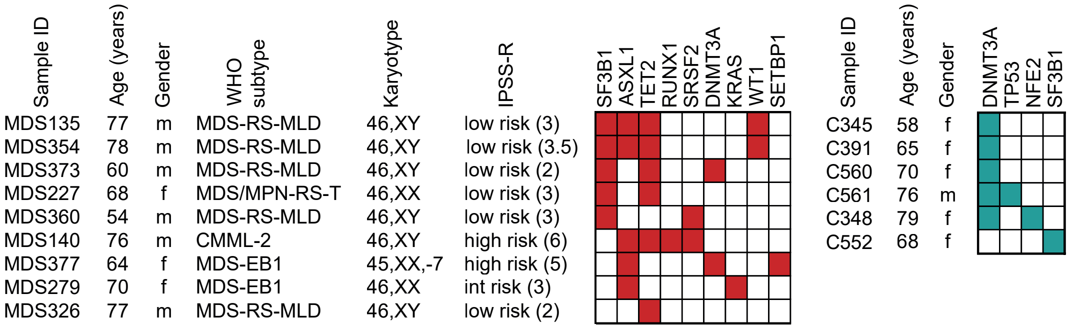


Figure 4.20: Currently unsolved is which rates are altered in CHIP or MDS in comparison to healthy hematopoiesis. Graphic is taken from unpublished manuscript Buck et al.

To resolve kinetic differences at the cell type level that occur with acquired mutations (see Figure 4.20), I fitted data from CHIP and MDS donors (donor information including age, gender and mutations detected in donor stem cells is listed in Figure 4.21) with model A and compared inferred rates to estimates from age-matched healthy samples (see Figure 4.22).



Sample ID	Age (years)	Gender	WHO subtype	Karyotype	IPSS-R	SF3B1	ASXL1	TET2	RUNX1	SRSF2	DNMT3A	KRAS	WT1	SETBP1
MDS135	77	m	MDS-RS-MLD	46,XY	low risk (3)	🟥	🟥	🟥					🟥	
MDS354	78	m	MDS-RS-MLD	46,XY	low risk (3.5)	🟥	🟥	🟥					🟥	
MDS373	60	m	MDS-RS-MLD	46,XY	low risk (2)	🟥		🟥			🟥			
MDS227	68	f	MDS/MPN-RS-T	46,XX	low risk (3)	🟥		🟥						
MDS360	54	m	MDS-RS-MLD	46,XY	low risk (3)	🟥				🟥				
MDS140	76	m	CMML-2	46,XY	high risk (6)		🟥	🟥	🟥	🟥				
MDS377	64	f	MDS-EB1	45,XX,-7	high risk (5)		🟥				🟥			🟥
MDS279	70	f	MDS-EB1	46,XX	int risk (3)		🟥					🟥		
MDS326	77	m	MDS-RS-MLD	46,XY	low risk (2)			🟥						

Sample ID	Age (years)	Gender	DNMT3A	TP53	NFE2	SF3B1
C345	58	f	🟩			
C391	65	f	🟩			
C560	70	f	🟩			
C561	76	m	🟩	🟩		
C348	79	f	🟩		🟩	
C552	68	f				🟩

Figure 4.21: MDS (left) and CHIP (right) donor information including age, gender, and presence/ absence of target mutations in stem cells for every individual analysed sample is shown. For MDS donors additional information about WHO subtype, Karyotype and IPSS-R score indicating severity of disease is listed. Graphic is taken from unpublished manuscript Buck et al.

To identify considerably increased or decreased rates in MDS patients or CHIP individuals (see Figure 4.23), I calculated a common 95% confidence interval for rate estimates of healthy donors without mutations and used its uper and lower boundary as cut-off values for dysregulated rates in CHIP and MDS. As the optimization approach does not reveal the full posterior distribution of each individual rate estimate, I approximated the 95% confidence interval by calculating the percentiles of a sampled common rate distribution of all healthy donors (see dark blue error band in Figure 4.22). Therefore I assumed that the posterior distribution of each individual rate corresponds to a log-normal distribution. This assumption is reasonable as

(i) rate estimates are per definition always positive and

(ii) the lower bound of the individual profile likelihood-based confidence intervals are for all rates closer to the optimal rate value than the upper bound,

two features which can adequately be described by a log-normal distribution. In order to determine a common distribution of all healthy donors, I sampled $n = 1000$ random numbers from 5 log-normal distributions, each belonging to a healthy donor and each obtaining different shape parameters based on the respective individual rate estimate and the corresponding profile likelihood-based 95% confidence interval $\left[CI^{l}_{(i,j)}, CI^{u}_{(i,j)}\right]$.

Let $r^{H}_{(i,j)}, i = 1, ..., 21, j = 1, ..., 5$ be the estimates of the reaction rates $\beta_{S_j}, \gamma_{S_j}, \alpha_{S_{j1} \to S_{j2}}$, where $S_j, S_{j1}, S_{j2} \in \mathcal{S}$, each belonging to one out of 5 healthy donors without mutations. Let X be a log-normally distributed random variable

$$X \sim log\mathcal{N}(\mu, \sigma).$$

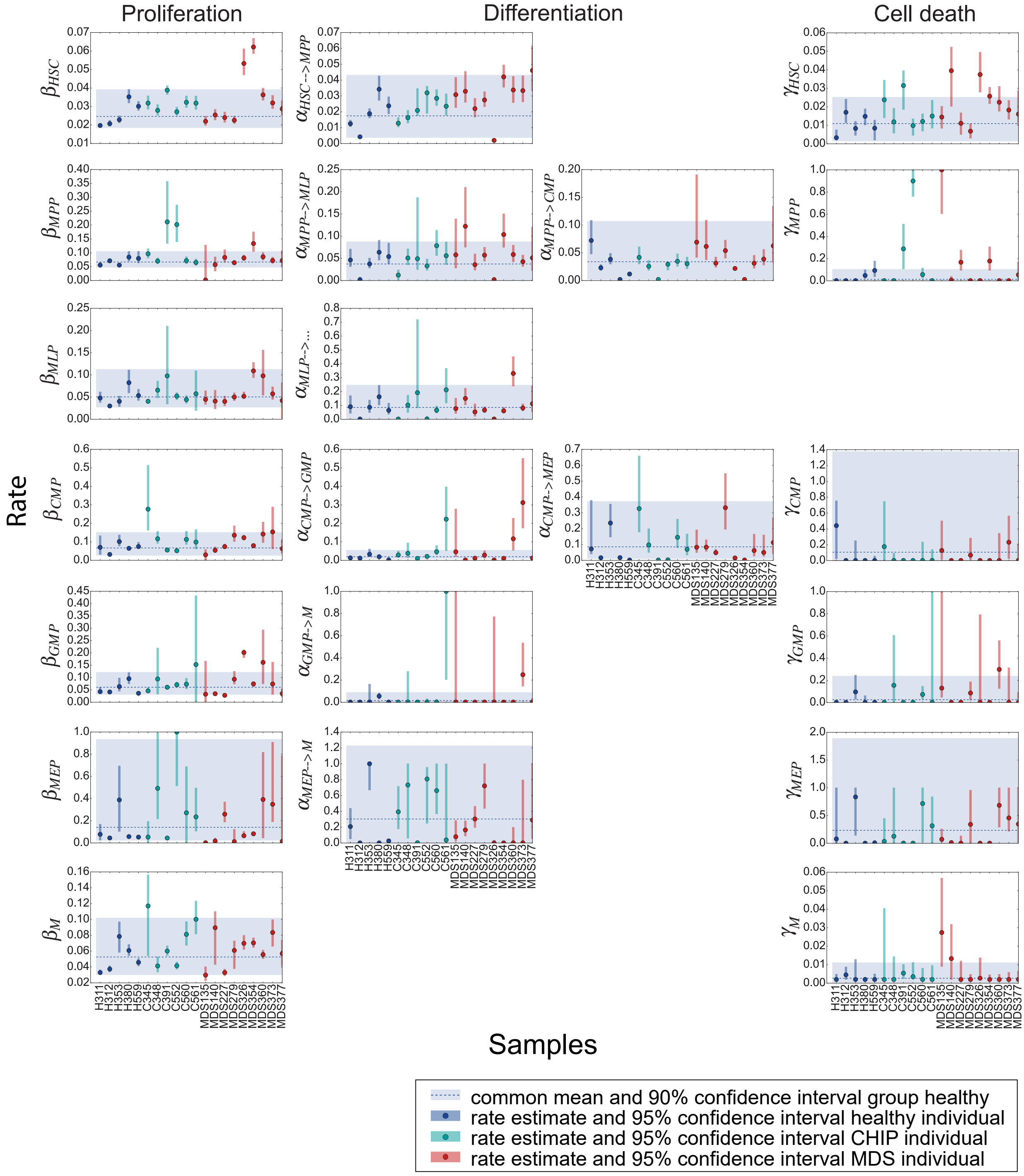


Figure 4.22: Optimal parameter values (dots) and their 95% confidence intervals (bars) resulting from fitting healthy age-matched (blue), CHIP (green), and MDS (red) donor samples with model A. Blue band shows the common 90% confidence interval for rate estimates of healthy donors without mutations, which was used as cutoff for dysregulated rates in MDS and CHIP (dots outside the blue band). Graphic is taken from unpublished manuscript Buck et al.

Then its mean $E(X)$ and α-quantiles $Q_\alpha(X)$ are given by

$$\begin{aligned} E(X) &= e^{\mu+\frac{\sigma^2}{2}} \\ Q_\alpha(X) &= e^{\mu+q_\Phi(\alpha)\cdot\sigma}, \end{aligned} \tag{4.11}$$

where $q_\Phi(\alpha)$ is the $\alpha-$quantile of the standard normal distribution $\mathcal{N}(0,1)$.
Thus, by assuming the $r^H_{(i,j)}$ follow a log-normal distribution with distribution parameters $\mu_{(i,j)}$ and $\sigma_{(i,j)}$ $\forall i$ and j and then solving equations

$$\begin{aligned} r^H_{(i,j)} &\overset{!}{=} e^{\mu_{(i,j)}+\frac{\sigma^2_{(i,j)}}{2}} \\ CI^l_{(i,j)} &\overset{!}{=} e^{\mu_{(i,j)}+q_\Phi(0.025)\cdot\sigma_{(i,j)}} \\ CI^u_{(i,j)} &\overset{!}{=} e^{\mu_{(i,j)}+q_\Phi(0.975)\cdot\sigma_{(i,j)}} \end{aligned} \tag{4.12}$$

for μ and σ, one can approximate the log-normal distribution parameters which are given by:

$$\hat{\mu}_{(i,j)} = \ln\left(r^H_{(i,j)}\right) - \frac{\hat{\sigma}^2_{(i,j)}}{2} \tag{4.13}$$

$$\hat{\sigma}_{(i,j)} = \frac{q_\Phi(0.975) + \sqrt{q_\Phi(0.975)^2 - 2\ln\left(\frac{CI^u_{(i,j)}}{r^H_{(i,j)}}\right)} + q_\Phi(0.025) - \sqrt{q_\Phi(0.025)^2 - 2\ln\left(\frac{CI^l_{(i,j)}}{r^H_{(i,j)}}\right)}}{2}.$$

By sampling $n = 1000$ random numbers from each individual $log\mathcal{N}(\hat{\mu}_{(i,j)}, \hat{\sigma}_{(i,j)})$ distribution, $j = 1, ..., 5$, $i = 1, ..., 21$ and calculating the 0.05 and 0.95 percentiles of the 5000 pooled random numbers for each $i = 1, ..., 21$, I approximated the 90% confidence interval for all $i = 1, ..., 21$ rates based on their estimates and their individual profile likelihood-based confidence intervals of all age-matched healthy donors. A rate of a CHIP or MDS sample is then defined as strongly increased, if their rates estimate is above the upper common 90% confidence bound and as strongly decreased if it is below the lower common 90% confidence bound of the common confidence interval of healthy age-matched donors (see Figure 4.22).

In CHIP samples, accelerated rates were found in both early and later progenitor stages, but decelerated rates were only observed at later progenitor stages (see Figure 4.23 bottom row). While HSC and MPP proliferation was higher in C391, where a DNMT3A mutation was detected with a variant allele frequency (VAF) of 12% (see Figure 4.21), I also found deregulation of later progenitor stages in a sample with a DNMT3A mutation (C345, VAF = 4% and C561, VAF = 1%) 4.23). Surprisingly, in two CHIP samples (C560, VAF=5% and C348, VAF=25%), no rates were identified to be strongly dysregulated.

In MDS samples, HSC proliferation was clearly outside the healthy range for MDS326 and MDS354 (see Figure 4.22) with a proliferation rate of 0.053 h^{-1} (95% CI [0.047, 0.061]h^{-1}) and 0.062h^{-1} (95% CI [0.058, 0.067]h^{-1}), as shown in Figure 4.22). For all 9 MDS cases we found at least one rate that was outside the 90% CI of the healthy age-matched group. While HSC rates were considerably deregulated for MDS140, MDS326, MDS354, MDS377, deregulated rates also appeared at later progenitor stages in other MDS samples (MDS279, MDS135, MDS227, MDS373, MDS360) where MPP, MEP, GMP, MEP or CMP proliferation was considerably altered (see Figure 4.23). Thus, changes in rates were not restricted to the HSC compartment, but were heterogeneously observed throughout the hierarchy in the MDS samples we studied. In both CHIP and MDS, deregulation affected proliferation, differentiation and cell death rates not only in the HSC compartment, but also further downstream in the hematopoietic hierarchy.

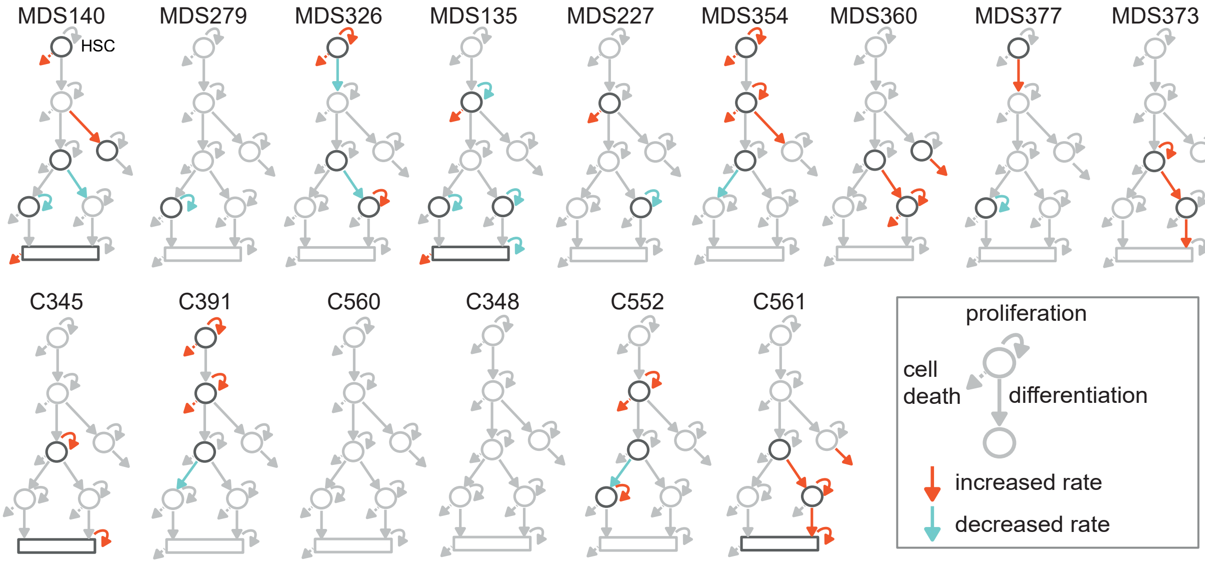


Figure 4.23: Increased (red) and decreased (green) rates for every donor sample with MDS (upper row) or CHIP (bottom row). Graphic is taken from unpublished manuscript Buck et al.

To analyse whether the heterogeneity of deregulated rates in MDS segregated into discernable subgroups, I performed a weighted principal component analysis (PCA) on the estimated parameters (see Figure 4.24). PCA is a linear transformation method which calculates the directions that maximize the variance of the data and are called principal components. Thus, a PCA of all rate estimates, which are stored in matrix $X \in \mathbb{R}^{n_{rates} \times n_{patients}}$, allows us to compress the information of the rate estimates by projecting the n_{rates} dimensional information of each patient to a lower dimensional space. This lower dimensional space is spanned by the principal components that explain most of the variance between the individuals.

In order to perform a weighted PCA, I defined a weight $w_{(i,j)}$ for each rate i and sample j. The weights were calculated by taking into account the parameter uncertainty, represented by the profile likelihood-based 95% confidence intervals $\left[CI^l_{(i,j)}, CI^u_{(i,j)}\right]$ of the rate estimates $\theta_{(i,j)}$, and the number of data points $n^{\mathcal{D}}_{obs,j}$ used for the parameter inference of sample j. I thereby weight parameter estimates higher are if more data points were observed and if their confidence intervals are relatively small as this results in more accurate and reliable parameter estimates. The weights

$$w^r_{(i,j)} := \frac{r_{max_i} - r_{min_i}}{CI^u_{(i,j)} - CI^l_{(i,j)}} \in [0,1], \quad i = 1, ..., n_{rates}, j = 1, ..., n_{samples}$$

are then higher for low parameter uncertainty, which corresponds to a small confidence interval and the weights

$$w^{\mathcal{D}}_{(i,j)} = \frac{n^{\mathcal{D}}_{obs,j}}{\sum_{j=1}^{n_{patients}} n^{\mathcal{D}}_{obs,j}} \in [0,1], \quad i = 1, ..., n_{rates}, j = 1, ..., n_{samples}$$

are higher if many data points were observed for patient j and used to fit the respective rates. Upon normalization, the overall weights in weight matrix W are then given by

$$w_{(i,j)} = \frac{\frac{w^r_{(i,j)}}{\sum_{i=1}^{n_{rates}} \sum_{j=1}^{n_{patients}} w^r_{(i,j)}} + w^{\mathcal{D}}_{(i,j)}}{2} \in [0,1]. \tag{4.14}$$

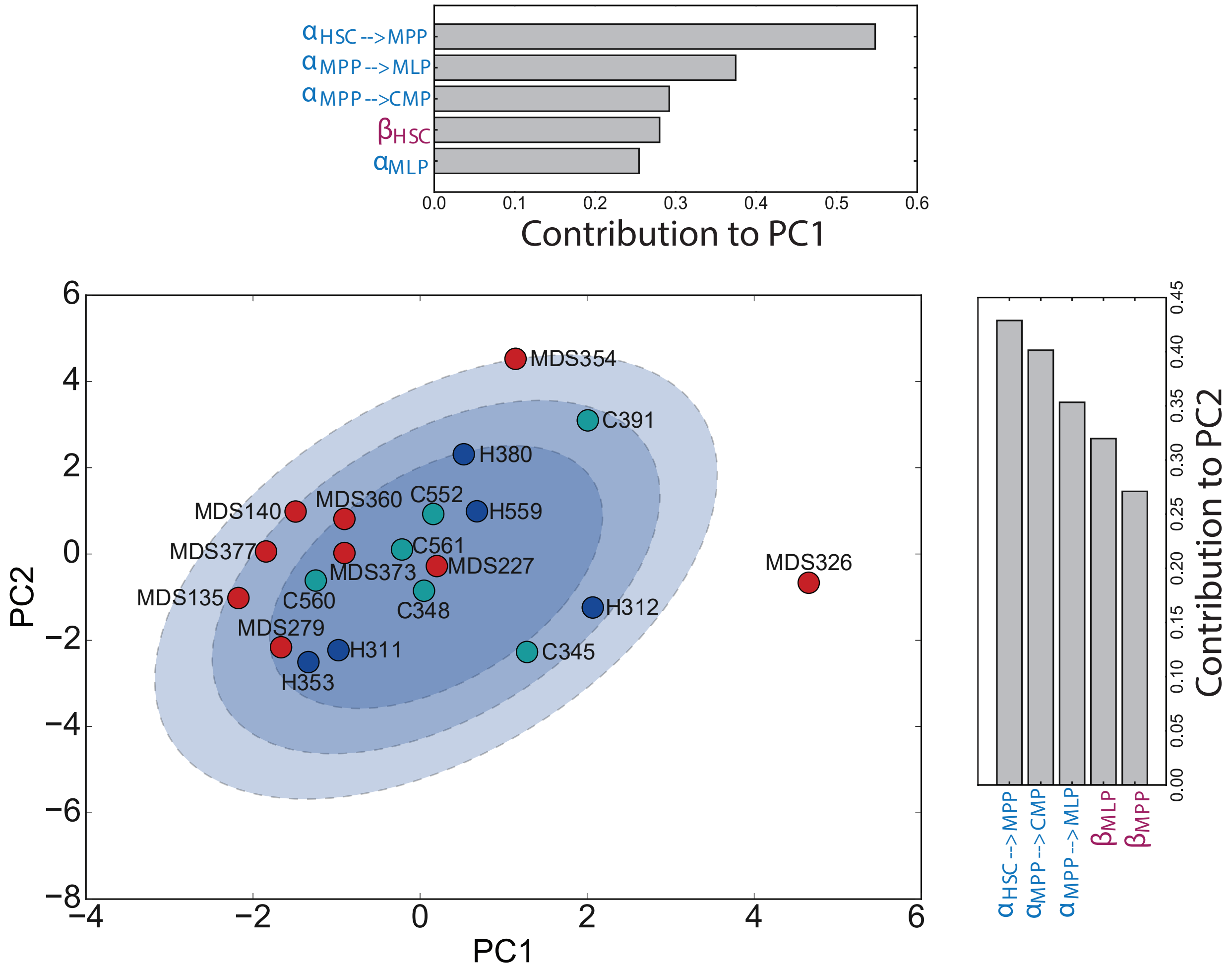


Figure 4.24: Weighted principal component analysis (PCA) performed on rate estimates of MDS patients. Rate estimates of healthy individuals with and without CHIP were projected onto the two-dimensional subspace spanned by PC1 and PC2 . Blue ellipses illustrate 70% (small), 85% (middle) and 95% (large) confidence areas for PC1 and PC2 values of healthy cohort. Greater heterogeneity of MDS patients (red dots) compared to age-matched healthy controls with CHIP (green dots) and without CHIP (blue dots) can be seen in values of PC1. Top 5 rate contributions to the first two principal components shows PC1 and PC2 are mainly spanned by differentiation and proliferation rates of early stem and progenitor compartments. Graphic is taken from unpublished manuscript Buck et al.

Taking into account the weight $w_{(i,j)}$ of each parameter i and sample j, the weighted principal components can be calculated by singular value decomposition of the weighted covariance matrix of the data matrix X, which is given by

$$\sigma_w^2 = \frac{(X \odot W)(X \odot W)^T}{WW^T},$$

where $W \in \mathbb{R}^{n_{rates} \times n_{patients}}$ is the matrix of weights for each observation in X and $\odot$ denotes the element-wise product.

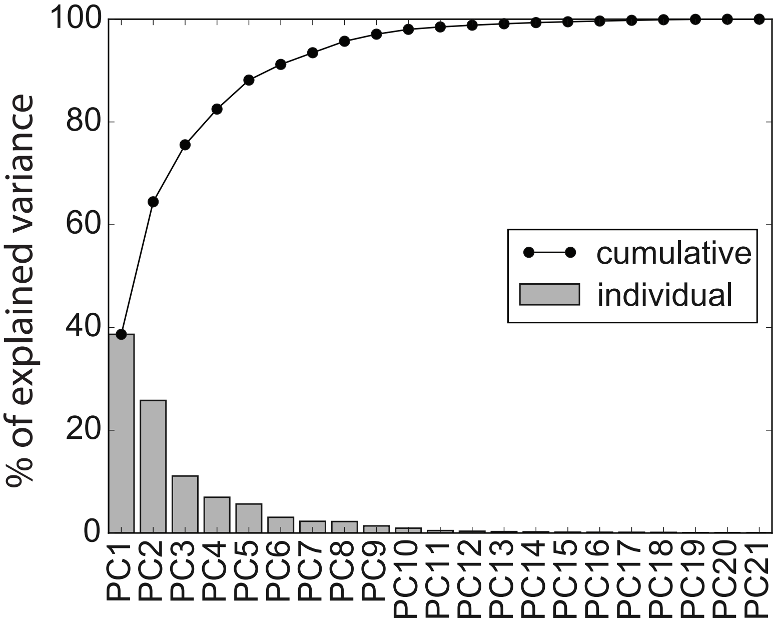


Figure 4.25: Percentage of total (bars) and cumulative (line) explained variance for all 21 principal components of PCA shows more than 60% of the variance is explained by the first two principal components. Graphic is taken from unpublished manuscript Buck et al.

Each Eigenvector of the weighted covariance matrix can be interpreted as the rate contribution to the respective principal component (see Figure 4.24) and the absolute value of each eigenvalue normalized to the total sum of absolute eigenvalues defines how much of the variance in the data is explained by the corresponding principal component (see Figure 4.25). By defining matrix $U \in \mathbb{R}^{n_{patients} \times n_{PCs}}$, which contains the n_{PCs} Eigenvectors with the highest absolute eigenvalues as rows, we can transform the data to the principal component space by calculating

$$Y = X \cdot U,$$

as it has been done for rate estimates of healthy donors with and without CHIP (see Figure 4.24). Interestingly, the heterogeneity between MDS samples can be mainly explained by differentiation and proliferation rates of early HSPC compartments. This can be seen in Figures 4.24 and 4.25, which show the explained variance per principal component and the rates of MDS patients that mainly contributed to PC1 and PC2. In the PCA plot (Figure 4.24), CHIP samples fell between healthy controls and MDS samples, in accordance with previous observations of co-existing mutated and unmutated HSCs in CHIP and almost exclusively mutated HSCs in MDS ([Pang et al., 2013]).

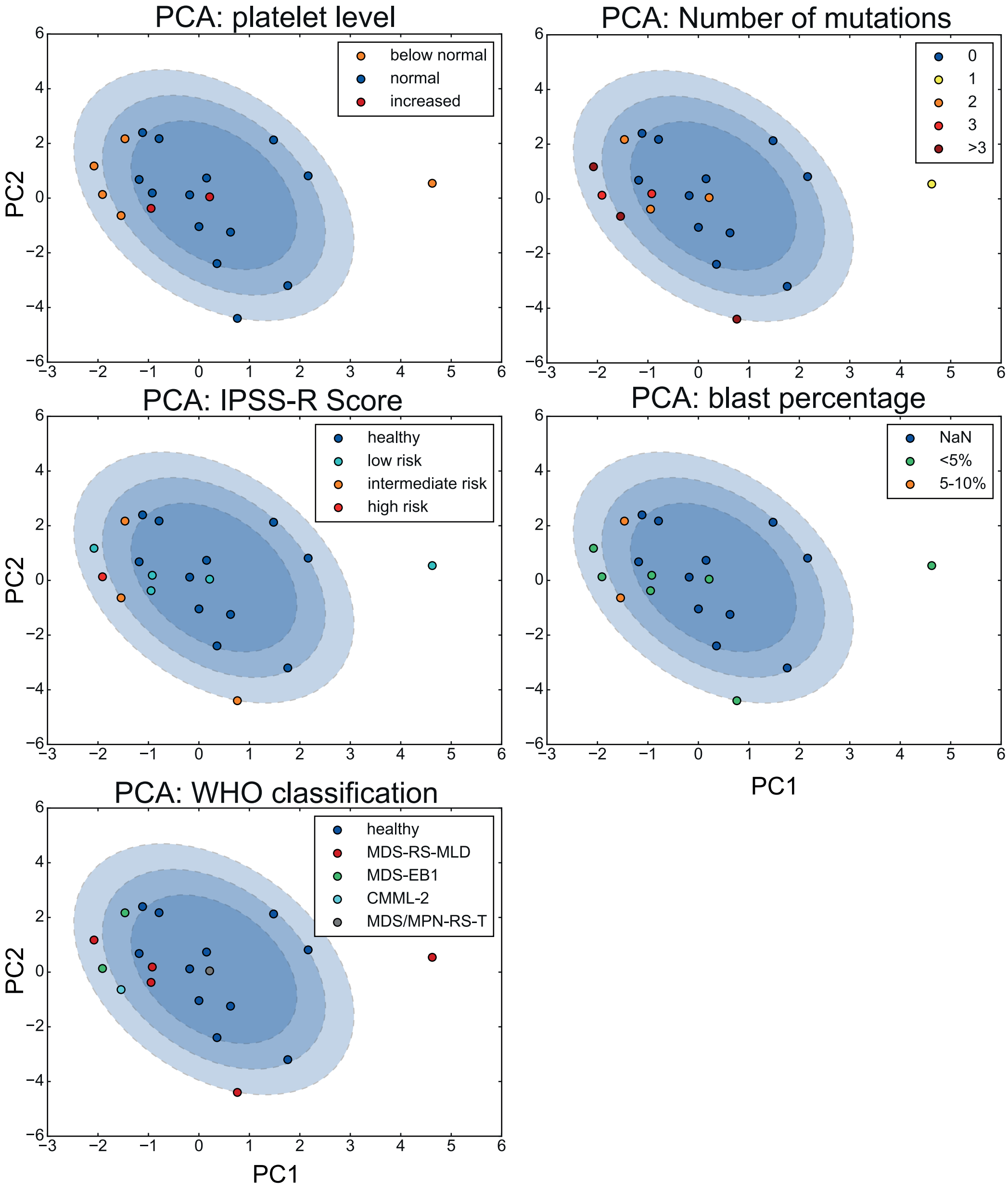


Figure 4.26: Projection of MDS patients and healthy donors to two-dimensional space spanned by PC1 and PC2 with donor information platelet level (top left), IPSSR-Score (middle left), World Health Organization (WHO) classification (bottom left), number of mutations (top right), and blast percentage (middle right) categories as color code. Graphic is taken from unpublished manuscript Buck et al.

Interestingly, some MDS cases were at the border or clearly outside the confidence areas defined by healthy age-matched samples with and without CHIP, while others showed kinetics that were consistent with the age-matched controls (see Figure 4.24). Mapping patient features on the PCA, we found no obvious correlation of HSPC kinetics with IPSS-R Score, WHO classification, number of MDS-related mutations, or blast percentage (see Figure 4.26). Interestingly, we found no obvious correlation of rates with either ASXL1 or SF3B1 mutations: both mutational subtypes showed deviations from the healthy age-matched controls (see Figure 4.26). However, MDS patients with similar kinetics to healthy age-matched samples showed increased or normal platelet levels whereas

MDS patients that observably differed from healthy donors had decreased platelet levels (see Figure 4.26). However, to draw general conclusions, a larger sample size will be required to investigate patient feature correlation with estimated kinetic rates.

4.6 Validation of modelling results

To test the reliability of the modelling and model selection results, I performed an *in silico* analysis to test if all parameters can be accurately and precisely be estimated for model A for a known realistic test parameter vector. To also validate the approach for the comparsion of lineage hierarchies, I expanded this initial *in silico* analysis and subsequently performed a comprehensive robustness test on *in silico* data with varying noise levels and considering all model hierarchies. As a second validation step for model A, I compared the model simulation to experimental data observed from the same sample at later time points which have not been used for the model fit. In addition, I compared parameter trends to independent data from another study in mice, in which hematopoiesis has been modelled and rates have been estimated with a similar approach.

4.6.1 *In silico* analysis

***In silico* analysis for model A**

To analyse how accurate and precise parameters can be estimated with model A an *in silico* analysis was performed. Assuming a realistic test parameter vector, I generated data by simulating with the test parameter from model A and perturbing data points with multiplicative log-normal noise of four different noise levels (see dots in Figure 4.27 A-B): weak noise ($\sigma_{j,n_{div}} = 0.4 \ \forall j, n_{div}$), middle ($\sigma_{j,n_{div}} = 0.8 \ \forall j, n_{div}$), strong ($\sigma_{j,n_{div}} = 1.2 \ \forall j, n_{div}$) and realistic ($\sigma_{j,n_{div}} \in [0.6, 1.1] \ \forall j, n_{div}$). The realistic noise level is equal to the mean of the estimated noise parameters calculated from fitting the samples of all individuals with model A. For weak noise, the measurement points deviate only lightly from the simulated values, whereas for strong noise the perturbed measurements scatter strongly (see Figure 4.27 B). The test parameter was calculated by averaging the optimal parameters observed based on samples from healthy individuals (without CHIP).

I fitted model A to the simulated data and could observe that for the assumed realistic noise level, true and fitted model agree well for almost any cell type and division compartment. Comparison of test parameter values to optimal values and their 95% PL confidence intervals (see red and black dots with grey bars in Figure 4.27 C) allowed me to assess if the test parameter value lays within the 95% PL confidence interval (grey bar) and how much true and optimal parameter deviate from each other. I could observe that all parameters are practically identifiable in the simulated setting under the assumption of weak noise. For middle, strong and realistic noise levels most parameters are still identifiable, only (some) progenitor death rates and differentiation from GMPs to mature cells ($\alpha_{GMP \to M}$) are practically unidentifiable as they observe confidence intervals that include the lower boundary of the parameter search interval. Most importantly, the true parameter (red dot in Figure 4.27 C) is contained in the 95% PL confidence interval for all parameters and noise levels except $\alpha_{MPP \to MLP}$ for strong and realistic noise. Confidence intervals of some parameters (i.e.

γ_{HSC} or γ_{CMP}) are larger for stronger noise levels, whereas for other parameters (i.e. proliferation rates $\beta_{(.)}$) they obtain roughly the same size for all considered noise levels. Moreover, the optimal parameter differs only slightly from the true test parameter for most parameters and noise levels. Dependent on the noise level, the absolute difference between test and optimal parameter values covers at most only 2.9% (weak), 5.7% (middle), 5.8% (realistic), and 5.6% (strong) of the search interval length, which underpins the high accuracy and precision of the approach, even for high noise levels.

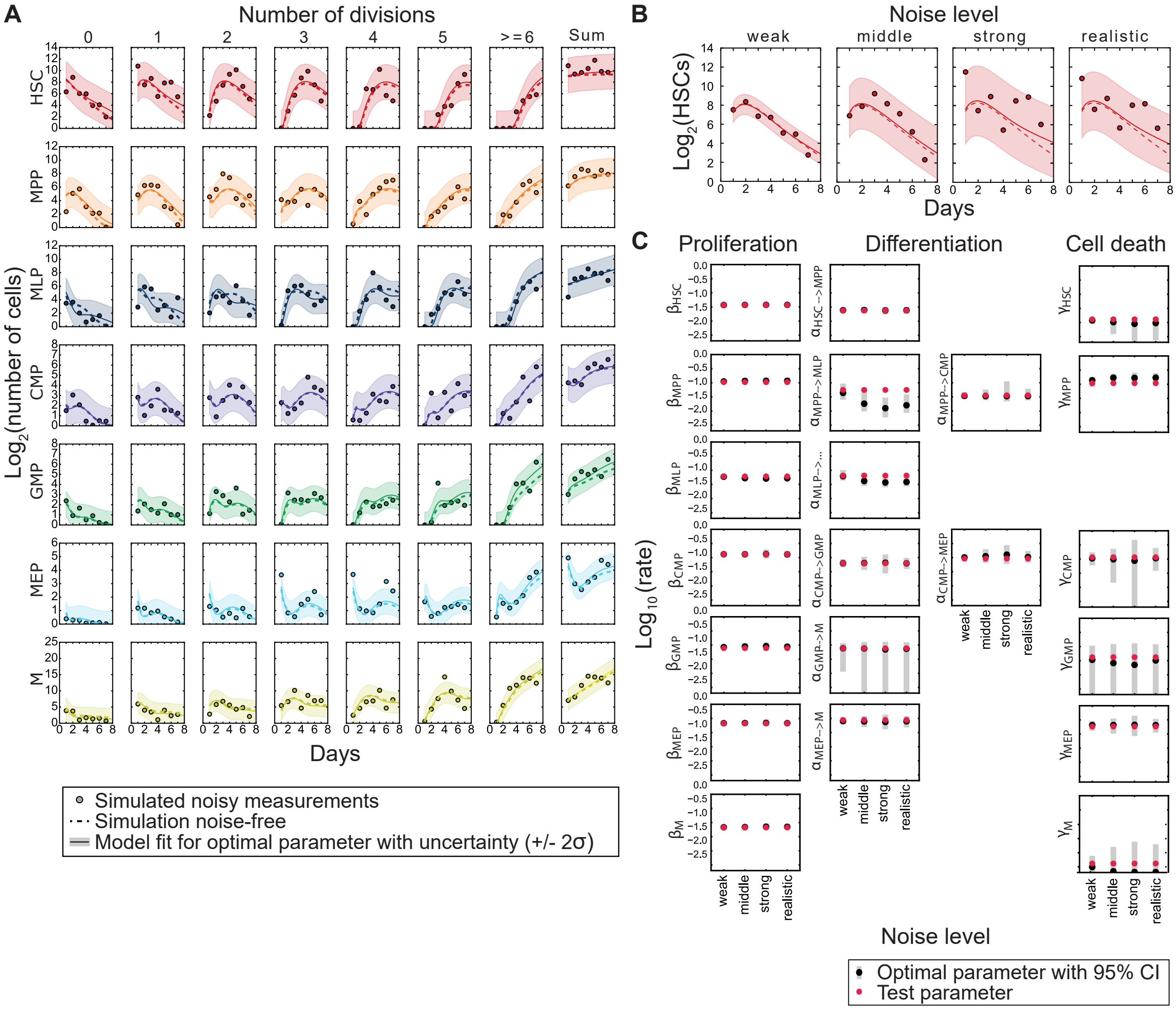


Figure 4.27: Model A was used to generate perturbed *in silico* data samples for a realistic test parameter (dots) and by assuming cell compartment specific log-normal noise (realistic noise level). True underlying (unperturbed) model observables (dashed line) and the model observables for the optimal parameter (solid line) deviate only slightly for the assumed noise level (A). Simulation and fit of data exemplarily shown for HSCs that divided once and which were simulated and fitted with model A. Simulated data were perturbed with a noise function assuming weak, middle, strong, and realistic noise (B). Test parameter values (red dots) for which perturbed samples were simulated by using model A are for most rates contained in 95% confidence interval (grey boxes) of optimal parameter values (black dots) (C). Graphic is partly taken from Bast et al. [2021].

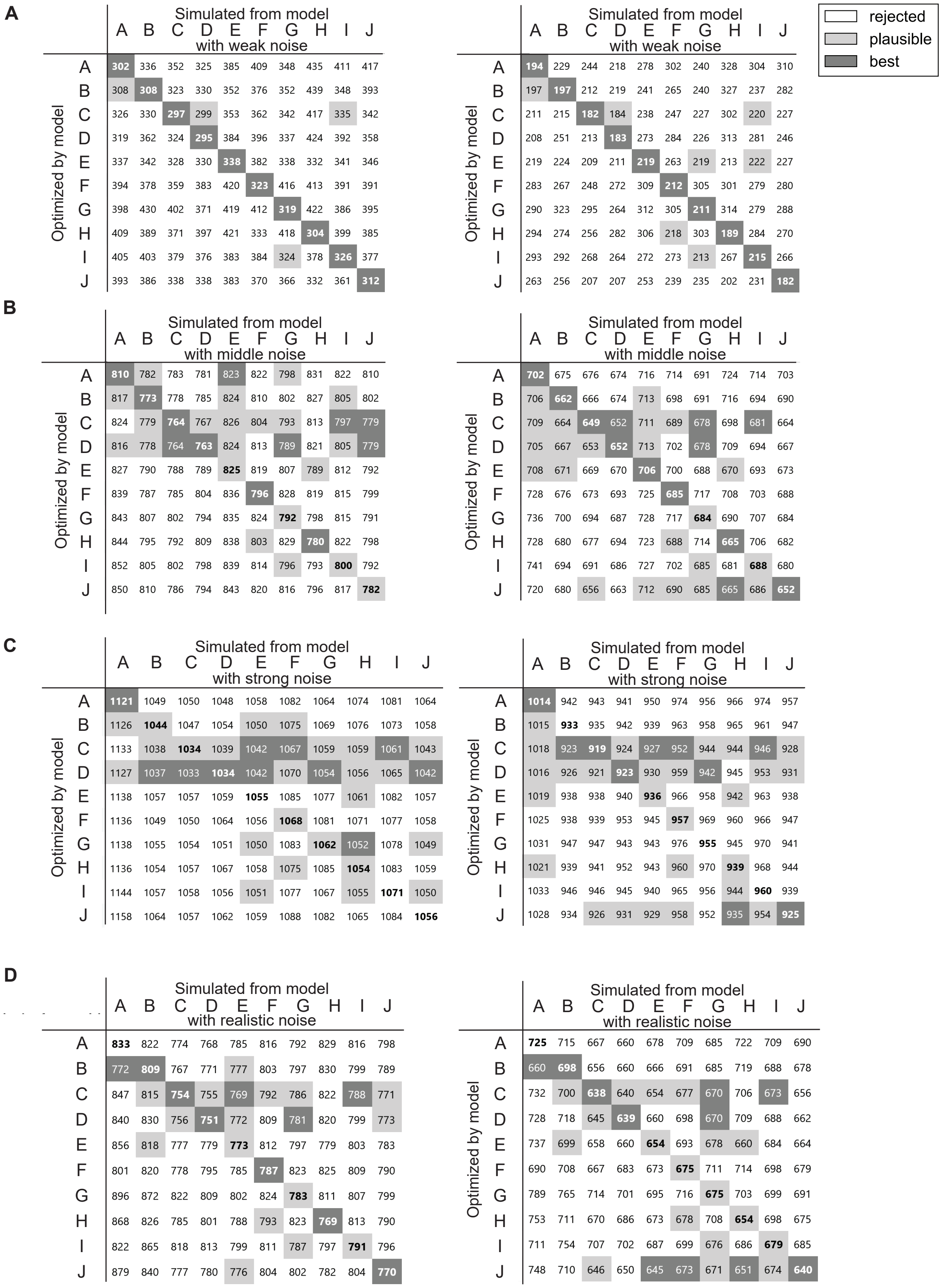
In silico analysis for model selection approach

To test the implementation, robustness and accuracy of my model selection approach the above introduced *in silico* analysis was performed for each considered lineage hierarchy model $\mathcal{M}_q \in$

$\{A, B, C, D, E, F, G, H, I, J\}$ for $n_{IS} = 3$ intermediate states. I set the test parameter to the mean over all maximum likelihood estimates observed from fitting the samples of all healthy individuals (without CHIP) with the respective model. I again applied different noise levels, which were defined as weak ($\sigma_{j,n_{div}} = 0.4\ \forall j, n_{div}$), middle ($\sigma_{j,n_{div}} = 0.8\ \forall j, n_{div}$), strong ($\sigma_{j,n_{div}} = 1.2\ \forall j, n_{div}$), and realistic ($\sigma_{j,n_{div}} \in [0.6, 1.1]\ \ \forall j, n_{div}$). The realistic noise level is equal to the mean of the estimated noise parameters calculated from fitting the samples of all individuals with the respective model. I performed MLE (see section 2.2.1) with all 10 models $\mathcal{M}_q \in \{A, B, C, D, E, F, G, H, I, J\}, q = 1, ..., 10$ on all $10 \cdot 4$ *in silico* generated samples to check the accuracy and precision of the analysis pipeline for all hierarchies and for the different noise levels. Comparing the BIC scores of each model for each *in silico* data sample (see Figure 4.28), additionally allows me to investigate if the ground truth model can successfully be identified as the best performing model. Especially I was interested in investigating if it is in principle possible to select the true model and not necessarily the least complex model.

For weak noise ($\sigma_{j,n_{div}} = 0.4\ \forall j, n_{div}$), the true model performed best and almost all (8 or 9 out of 9 other models) were rejected according to BIC (see Figure 4.28 A left). For middle noise level ($\sigma_{j,n_{div}} = 0.8\ \forall j, n_{div}$), the true underlying model was always at least amongst the plausible models and identified as the best model for models A, B, C, D, F, and H (see Figure 4.28 B left). Assuming a strong noise level ($\sigma_{j,n_{div}} = 1.2\ \forall j, n_{div}$), we found that the true model is only accurately identified for lineage hierarchies A and D (see Figure 4.28 C left). Model D is however also at least plausible if the data were simulated from any other model. For model I and the two most complex models E and J, other models are favoured and the true model is rejected. For realistic noise level ($\sigma_{j,n_{div}} \in [0.6, 1.1]\ \ \forall j, n_{div}$), 60% of the models are correctly identified as best performing model, similar to middle noise and the true model was at least amongst the plausible models for 9 out of 10 models (see Figure 4.28 D left).

This analysis shows how crucial the underlying noise is for the robust identification of the true model. However, model A was only once the best performing model when the in silico data was generated from another lineage hierarchy (data generated from model E with middle noise) and only twice considered as plausible (data generated from model B and G with middle noise), suggesting its low complexity is not the predominant feature of its outstanding performance. For models E and J on the contrary, the model complexity might have been a barrier in correctly identifying them: both were never identified as true model when the data were generated from another model and both were identified at most once as a plausible model. While simulating from model I, model C was selected as best model for noise levels middle, strong and realistic, despite its higher complexity (23 vs. 22 parameters), while model I itself was only plausible (middle and realistic noise) or even rejected (strong noise). For completion, the AIC scores, which represent a less conservative scoring method, are reported respectively (see Figure 4.28 right) and show a very similar pattern than the BIC values (see Figure 4.28 left). Based on this analysis I concluded that the parameter inference approach allows for the robust identification of lineage hierarchies in the presence of noise, if it is not over prominent. Most importantly, it is unlikely that model A was only selected due to its low complexity.


rejected
plausible
best

A

Simulated from model with weak noise (rows: Optimized by model)

	A	B	C	D	E	F	G	H	I	J
A	**302**	336	352	325	385	409	348	435	411	417
B	308	**308**	323	330	352	376	352	439	348	393
C	326	330	**297**	299	353	362	342	417	335	342
D	319	362	324	**295**	384	396	337	424	392	358
E	337	342	328	330	**338**	382	338	332	341	346
F	394	378	359	383	420	**323**	416	413	391	391
G	398	430	402	371	419	412	**319**	422	386	395
H	409	389	371	397	421	333	418	**304**	399	385
I	405	403	379	376	383	384	324	378	**326**	377
J	393	386	338	338	383	370	366	332	361	**312**

Simulated from model with weak noise (rows: Optimized by model)

	A	B	C	D	E	F	G	H	I	J
A	**194**	229	244	218	278	302	240	328	304	310
B	197	**197**	212	219	241	265	240	327	237	282
C	211	215	**182**	184	238	247	227	302	220	227
D	208	251	213	**183**	273	284	226	313	281	246
E	219	224	209	211	**219**	263	219	213	222	227
F	283	267	248	272	309	**212**	305	301	279	280
G	290	323	295	264	312	305	**211**	314	279	288
H	294	274	256	282	306	218	303	**189**	284	270
I	293	292	268	264	272	273	213	267	**215**	266
J	263	256	207	207	253	239	235	202	231	**182**

B

Simulated from model with middle noise (rows: Optimized by model)

	A	B	C	D	E	F	G	H	I	J
A	**810**	782	783	781	823	822	798	831	822	810
B	817	**773**	778	785	824	810	802	827	805	802
C	824	779	**764**	767	826	804	793	813	797	779
D	816	778	764	**763**	824	813	789	821	805	779
E	827	790	788	789	**825**	819	807	789	812	792
F	839	787	785	804	836	**796**	828	819	815	799
G	843	807	802	794	835	824	**792**	798	815	791
H	844	795	792	809	838	803	829	**780**	822	798
I	852	805	802	798	839	814	796	793	**800**	792
J	850	810	786	794	843	820	816	796	817	**782**

Simulated from model with middle noise (rows: Optimized by model)

	A	B	C	D	E	F	G	H	I	J
A	**702**	675	676	674	716	714	691	724	714	703
B	706	**662**	666	674	713	698	691	716	694	690
C	709	664	**649**	652	711	689	678	698	681	664
D	705	667	653	**652**	713	702	678	709	694	667
E	708	671	669	670	**706**	700	688	670	693	673
F	728	676	673	693	725	**685**	717	708	703	688
G	736	700	694	687	728	717	**684**	690	707	684
H	728	680	677	694	723	688	714	**665**	706	682
I	741	694	691	686	727	702	685	681	**688**	680
J	720	680	656	663	712	690	685	665	686	**652**

C

Simulated from model with strong noise (rows: Optimized by model)

	A	B	C	D	E	F	G	H	I	J
A	**1121**	1049	1050	1048	1058	1082	1064	1074	1081	1064
B	1126	**1044**	1047	1054	1050	1075	1069	1076	1073	1058
C	1133	1038	**1034**	1039	1042	1067	1059	1059	1061	1043
D	1127	1037	1033	**1034**	1042	1070	1054	1056	1065	1042
E	1138	1057	1057	1059	**1055**	1085	1077	1061	1082	1057
F	1136	1049	1050	1064	1056	**1068**	1081	1071	1077	1058
G	1138	1055	1054	1051	1050	1083	**1062**	1052	1078	1049
H	1136	1054	1057	1067	1058	1075	1085	**1054**	1083	1059
I	1144	1057	1058	1056	1051	1077	1067	1055	**1071**	1050
J	1158	1064	1057	1062	1059	1088	1082	1065	1084	**1056**

Simulated from model with strong noise (rows: Optimized by model)

	A	B	C	D	E	F	G	H	I	J
A	**1014**	942	943	941	950	974	956	966	974	957
B	1015	**933**	935	942	939	963	958	965	961	947
C	1018	923	**919**	924	927	952	944	944	946	928
D	1016	926	921	**923**	930	959	942	945	953	931
E	1019	938	938	940	**936**	966	958	942	963	938
F	1025	938	939	953	945	**957**	969	960	966	947
G	1031	947	947	943	943	976	**955**	945	970	941
H	1021	939	941	952	943	960	970	**939**	968	944
I	1033	946	946	945	940	965	956	944	**960**	939
J	1028	934	926	931	929	958	952	935	954	**925**

D

Simulated from model with realistic noise

	A	B	C	D	E	F	G	H	I	J
A	**833**	822	774	768	785	816	792	829	816	798
B	772	**809**	767	771	777	803	797	830	799	789
C	847	815	**754**	755	769	792	786	822	788	771
D	840	830	756	**751**	772	809	781	820	799	773
E	856	818	777	779	**773**	812	797	779	803	783
F	801	820	778	795	785	**787**	823	825	809	790
G	896	872	822	809	802	824	**783**	811	807	799
H	868	826	785	801	788	793	823	**769**	813	790
I	822	865	818	813	799	811	787	797	**791**	796
J	879	840	777	780	776	804	802	782	804	**770**

Simulated from model with realistic noise (rows: Optimized by model)

	A	B	C	D	E	F	G	H	I	J
A	**725**	715	667	660	678	709	685	722	709	690
B	660	**698**	656	660	666	691	685	719	688	678
C	732	700	**638**	640	654	677	670	706	673	656
D	728	718	645	**639**	660	698	670	709	688	662
E	737	699	658	660	**654**	693	678	660	684	664
F	690	708	667	683	673	**675**	711	714	698	679
G	789	765	714	701	695	716	**675**	703	699	691
H	753	711	670	686	673	678	708	**654**	698	675
I	711	754	707	702	687	699	676	686	**679**	685
J	748	710	646	650	645	673	671	651	674	**640**

Figure 4.28: BIC (left) and AIC (right) values from fitting models A-J (rows) to data simulated from models A-J (columns) and perturbing them with weak (A), middle (B), strong (C), and realistic (D) noise levels. Graphic is partly taken from Bast et al. [2021].

4.6.2 Comparison of model to dependent and independent data

Model prediction for unseen measurements observed at later time points

To validate the resulting model based on dependent data, I used measurements of sample MDS279 which were not used to estimate θ^{ML}.

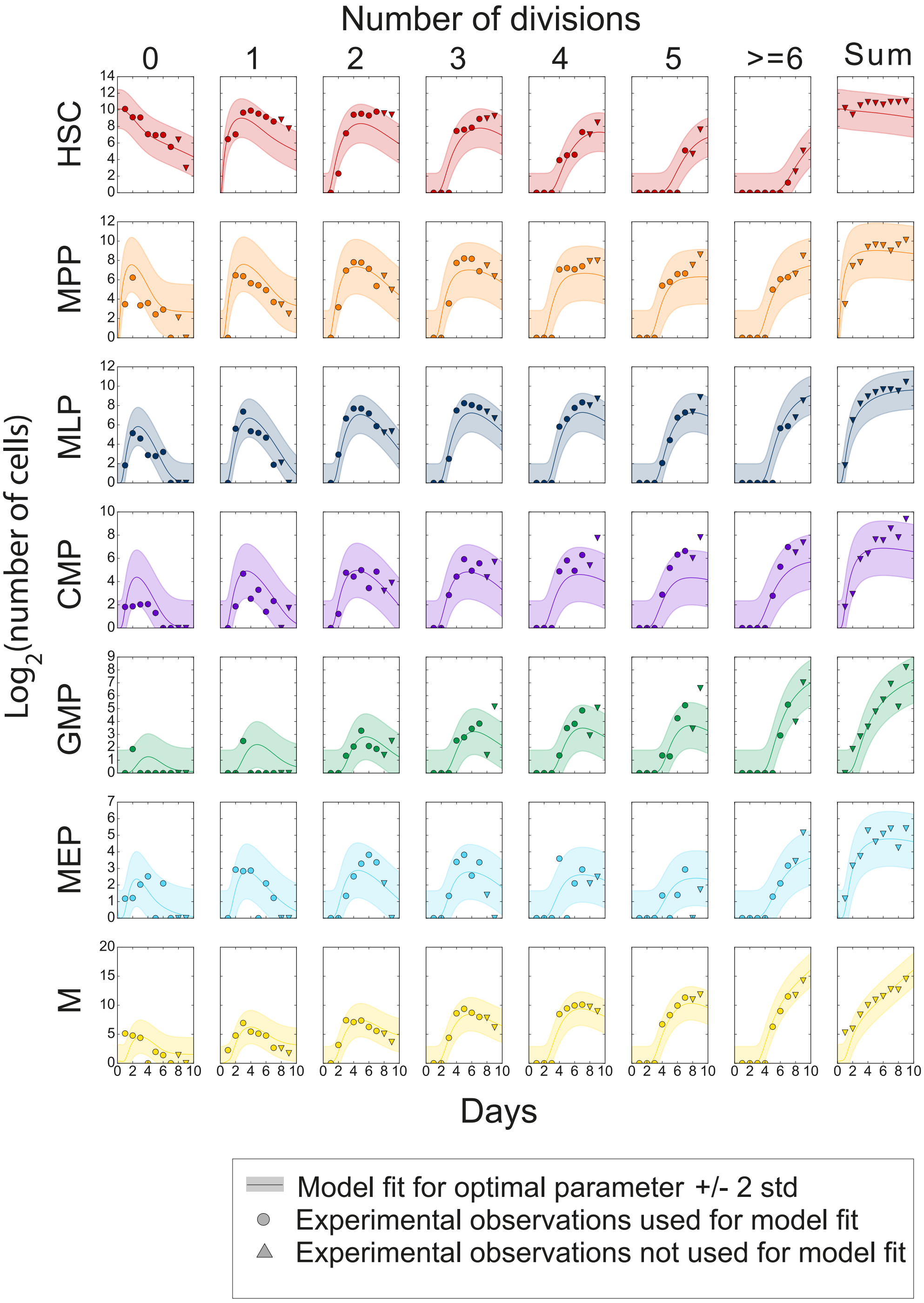


Figure 4.29: Model A (solid line) fitted to measurements of sample MDS279 for time points $t_i \leq 7$ days (dots) and model prediction compared to measurements of later time points $t_i > 7$ days (triangles). Sum over cell division compartments (last column) was not considered in log-likelihood function. Graphic is taken from unpublished manuscript Buck et al.

As can be seen in Figure 4.29, the model is able to predict later time points not used for the model fit (triangles) accurately and precisely as measurements lie within error band for almost any celltype- and division-compartment.

Comparison to independent data based on parameter-based metrics

To further verify the parameter estimation result, I compared parameter results to the results of another hematopoiesis study. Busch et al. [2015] inferred net proliferation and differentiation rates from an *in vivo* mouse labelling experiment. Similar to our approach, they compared young and aged hematopoiesis and used a compartmental model to infer rates. In contrast to my modelling approach, they considered early hematopoietic stem and progenitor cells, namely long- and short-term HSCs, MPPs, Common lymphoid progenitors and CMPs as compartments and did not distinguish between proliferation and cell death rates. To compare the two analyses, I calculated parameter metrics, which were introduced by Busch et al. [2015]:

Definition 4.1. *The cell type specific* ***net proliferation*** *is defined as difference between proliferation and cell death rate:*

$$\beta_j^{net} = \beta_j - \gamma_j \quad \forall j = 1, ..., |\mathcal{S}|.$$

A value of $\beta_j^{net} > 0$ means more cells in compartment j proliferate than die, a value of $\beta_j^{net} < 0$ means more cells the compartment j undergo cell death than proliferate. The closer the value to 0, the more balanced are proliferation and cell death in compartment j.

Definition 4.2. *The cell type specific* ***cellular exit time*** *describes the time cells spend in compartment j on average before they undergo cell death or differentiate and is given by*

$$T_j^{exit} = \frac{1}{\sum_{o \in O_j} \alpha_{j \to o} + \gamma_j} \quad \forall j = 1, ..., |\mathcal{S}|.$$

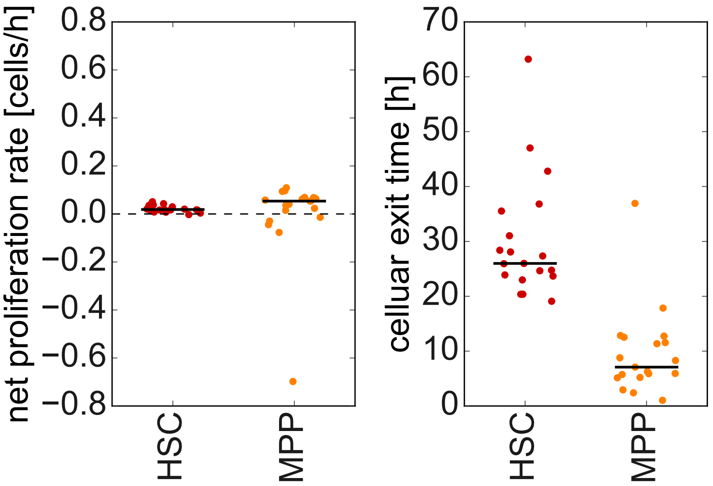


Figure 4.30: Net proliferation rates (left) and cellular exit times (right) of HSCs and MPPs for samples of healthy individuals. Graphic is taken from unpublished manuscript Buck et al.

As mouse and human are not one-by-one comparable and different compartments have been considered in this study, I compared only parameter trends. Moreover I only took samples from healthy individuals (without CHIP) into consideration for this analysis as Busch et al. [2015] performed their analysis in healthy mice. Interestingly, the metrics revealed stronger MPP net proliferation compared to HSC net proliferation (median of 0.05 vs. $0.02\,[cells/h]$) and a shorter cellular exit time in the MPP vs. HSC compartment (median of 7 vs. $26\,[h]$), which is the same trend that has been observed by Busch et al. [2015] (see Figure 4.30).

5 Discussion and outlook

5.1 Discussion

The data-driven modelling and model selection approach developed in this thesis has proven to be a powerful tool to gain mechanistic insights into cell division and differentiation processes based on quasi time-resolved cell count data. My approach is able to infer cell division, differentiation, and cell death rates, together with their uncertainties and to quantitatively compare different compartmental models based on snapshot data of the underlying dynamical process. The insights my analysis revealed either could not have been observed experimentally or would have required many resources in order to test each and every hypothesis experimentally.

Using the examples of adult neurogenesis (see chapter 3) and adult hematopoiesis (see chapter 4), my modelling and inference approach shed light on two tissue homoeostasis processes, which were not in detail understood before or frequently debated.

In the first example I modelled a comprehensive description of adult neurogenesis in the SEZ of the murine brain on the cell type level. By integrating the most important experimental findings from previous studies, I introduced a compartmental model consisting of three stem cell states: dormant, quiescent, and active, and additionally states for downstream cell types. The approach allowed the inference of cell type specific frequencies of division modes in young and aged. I could reveal that according to the experimental data, the most prominent age-related changes occur on the stem cell level. In detail I could show that active stem cells mostly divide asymmetrically but also undergo symmetric differentiation divisions in young adult mice. This symmetric differentiation capacity of neural stem cells diminishes with age until they almost exclusively divide asymmetrically. In addition, stem cells stay longer quiescent in aged individuals. These findings contributed to the understanding why and how neurogenesis gradually declines during ageing. By decreasing the symmetric differentiation divisions and switching less often to quiescence, active stem cells compensate the weakened influx of the almost emptied dormant stem cell pool. These subtle changes in stem cell dynamics are able to explain why less neurons are observed in aged mice. The findings of adult neurogenesis on the clonal level do not only help to understand cell population dynamics, they can also help to understand and reveal changes in gene regulation that are linked to a specific function such as increased or decreased proliferation or differentiation [Poiana et al., 2020]. Moreover, the findings and the introduced model are a solid starting point for further experiments and systems biology approaches to analyse dysfunctions in adult neurogenesis in neurodegenerative diseases [Cheyuo et al., 2019].

In the second example, my modelling and inference approach could reveal age- and disease-related changes in cell division, differentiation, and cell death rates, and in addition plausible and implausible lineage hierarchies for human hematopoiesis. In recent years, the hematopoietic lineage has often been debated and experimentalists suggested a range of differentiation possibilities for the set of defined hematopoietic progenitors. My computational data-driven modelling and model selection approach allowed to investigate the plausibility of these competing lineage hierarchies based on data observed from an *in vitro* bulk experiment. Based on this quasi time-resolved cell count data set, I could quantitatively reject several of the suggested lineage hierarchies and found most evidence in the data for the classical hematopoiesis lineage hierarchy. In addition, my approach could contribute to the understanding of how hematopoiesis is declining but maintained during ageing. As stem cell counts are comparably high in young and aged individuals, the observation that hematopoiesis declines with age is surprising but could also be confirmed by the experimental data I used for my analysis. Based on the *in vitro* data set I used, my modelling and inference approach could uncover that the number of newly produced mature blood cells is reduced in aged individuals due to decreasing proliferation of hematopoietic stem cells and common myeloid progenitors. By experimentally decoupling cell intrinsic from cell extrinsic effects, this work could thus contribute to the better understanding of cell intrinsic behaviour of hematopoietic cells. Moreover, my approach allows to identify which rates are disturbed in homeostasis-related diseases. Focusing on myelodysplastic syndromes (MDS), I could reveal for every patient which cell types are affected and if proliferation, differentiation or cell death, or a combination of them is disturbed. I found a large heterogeneity in the cohort of MDS patients. For every patient sample different rates were changed compared to age-matched healthy controls, potenitally affecting almost any cell type that has been investigated. This finding is important in two ways. First, it explains why there are so many subtypes of MDS, as reported by the WHO. Second, as it explains the heterogeneity of the disease on the patient level in more detail, it emphasizes the need for personalized medicine in MDS treatment. Of course it will be challenging to develop these targeted treatment strategies and many more analysis, especially on the molecular level, and based on a larger cohort are required to reach this goal.

In general, it is important to mention that the performance of my approach depends on the richness and quality of the data set used for parameter inference. Thus, it is not guaranteed that exactly the same conclusions would have been drawn based on another data set, resulting from the same or another experiment. By performing model selection I however ensured that the model is only as complex as required to explain the data and thereby avoided over-fitting. Additionally it also allows conclusions abut how much better or worse different models perform in comparison to any other considered model. Moreover, my modelling and inference approach can only reveal the biological signal that is present in the data and that is not overlaid by other biological or technical noise. In addition, the performance depends on the quality and correctness of incorporated prior knowledge. If the prior knowledge is wrong it will introduce a bias that can lead to wrong conclusions and predictions. It is always important to include only prior knowledge for which strong evidence could be found. For the analysis of cell division and differentiation processes on the molecular level, the compartmental modelling and model selection approach might not be suitable and other modelling approaches need to be specified, which could in principle be integrated into my analysis pipeline.

For stochastic models, parameter inference with ABC has been introduced in this thesis (see section 2.2.2). Performing parameter inference with ABC however turned out to be impractical for several reasons. The usage of the ABC SMC algorithm requires the specification of a prior distribution, a kernel function, the sequence of thresholds and a distance function. In practice often several combinations of these algorithm values and functions are tested till one finds a set-up which performs well for the considered optimization problem. One the one hand, this gives the analyst more flexibility in designing the optimization strategy, but on the other hand also limits the practicability. In addition, adaptation of the model specification, i.e. due to a change in model assumptions or integration of additional biological knowledge from other sources than the experimental data can get long-winded. Even if the likelihood-free parameter inference method works well for a specific set-up and model, it can get computationally demanding and even infeasible for a slightly more complex model, which makes the approach unsuitable for model selection of a broader set of models with varying complexity as it has been performed within the scope of this dissertation (see chapter 3). For rather small sets of considered models, one can use the suggested ABC SMC model selection algorithm, which approximates the full posterior distribution on the joint model and parameter space (see section 2.2.2). The probability of a specific model to be the true one can then be approximated by marginalization of this distribution [Toni and Stumpf, 2009] and Bayes factors can be used to perform model selection. Alternatively, one could improve ABC efficiency and practicability by combining ABC with machine learning methods. An example is the Bayesian inference approach published by Lueckmann et al. [2019], which does not rely on defined rejection thresholds or distance functions and allows incorporation of arbitrary waiting time distributions with the generating function approach, thus efficiently computes the posterior distribution. It would be interesting to implement a similar approach into my analysis pipeline and test it on the adult neurogenesis data set (see figure 3.2).

5.2 Outlook

Regarding the adult neurogenesis project (see chapter 3), one could also study neurological disorders such as schizophrenia in the future. So far, the role of adult neurogenesis in the SEZ and other brain regions is unexplored for schizophrenia [Weissleder et al., 2019]. One could perform additional experiments on disease mouse models, e.g. mouse with a homologous copy number variation mutation that increases the risk for the respective disease in humans, in order to compare healthy and diseased homoeostasis and identify which rates are likely to be disturbed in the respective neurological disorders. This could help to understand the underlying disease mechanisms and also which cell types are affected.

In the adult hematopoiesis project (see chapter 4), it would be interesting to analyse more samples to be able to more comprehensively investigate which de-regulated rate is correlated with specific patient information (see figure 4.26.) In order to more accurately model the clonal competition in CHIP individuals, one could introduce feedback terms into the model as has been suggested by Park et al. [2019]. In addition, it would be useful to also analyse bone marrow samples of acute myeloid leukaemia patients and subsequently analyse the differences between the four groups healthy, CHIP, MDS and acute myeloid leukaemia. To Further optimize the analysis pipeline, one could perform the FACS gating automatically with clustering algorithms that simultaneously perform the gating for all .fcs files. This would first of all standardize the data preprocessing and thus hopefully minimize technical error and additionally save the experimentalists plenty of work. Moreover, one could adapt the modelling approach by directly modelling the $CellTrace^{TM}$ Violet stained distribution instead of performing the gating, similar to the approach introduced by [Hross and Hasenauer, 2016]. To not only explore cell intrinsic effects but also niche effects that are thought to play a role in hematopoietic differentiation [Silberstein et al., 2016], one could establish an *in vitro* experiment that allows to culture stroma cells together with the currently cultured hematopoietic stem cells and thereby incorporate the niche cell influence on hematopoiesis. To model the dynamics accurately, one may have to include feedback terms in the model, i.e. limited growth of HSCs dependent on number of mature cells as has been suggested by Walenda et al. [2014], Stiehl et al. [2015] or [Klose et al., 2019] and could again derive and compare a set of plausible models.

In the past decades, advances in technology such as electron microscopy, immunohistochemistry, fluorescence activated cell sorting, fluorescence *in situ* hybridization and Next-Generation Sequencing allow for a more and more systematic cell type identification [Regev et al., 2017]. For instance, improvements in measuring transcript abundance on the single cell level (single cell RNA-sequencing) and the development of algorithms to analyse RNA-sequencing data have enabled the transcriptomic cell type identification [Hwang et al., 2018]. This cell type definition builds on the biological finding that while all cells within an organism share a common genotype, not all cells express the same genes as some are up- and some are down-regulated. Upon analysis of the transcriptome on the single cell level, which contains the information which genes are highly expressed in a particular cell, the changes between cells upon development are detectable with higher resolution compared to conventional methods which use cell morphology, location and a very limited set of cell type markers. My framework could in principle also be used to compare models that were trained on data which describe homoeostatic processes on the molecular level of gene regulation. The work of Fischer et al. [2019] for instance provides a framework to model developmental trajectories on the transcriptomic level with partial differential equations taking transcriptomic state space and time as dependent variables. With this model it is possible to estimate rates that describe diffusion, drift and net population growth functions. In order to analyse these mechanisms on the cell type level, one can discretize the state space by performing a clustering analysis and it would in principle also be possible to comprehensively study the mechanisms of a whole tissue in healthy and diseased individuals which could be done similarly to my analysis approach. How to discretize the state space is however not trivial as the identification of transcriptomic cell types or cell identities partly depends on the measurement technique, and preprocessing and clustering algorithms which were used to identify distinct groups of cells [Luecken and Theis, 2019]. Thus, different RNA-seq analysis pipelines can lead to different results in terms of cell type identity [Menon, 2017] which raises the philosophical question 'What exactly is a cell type?'. In the Human Cell Atlas project, researchers from all over the world address this question and collectively follow the mission to create reference maps of all human cells based on their distinct molecular profiles [Regev et al., 2017]. A similar program is The Human Protein Atlas, which is Swedish based and was already initiated in 2003 [Uhlen et al., 2010]. The establishment of a detailed human cellular network architecture will have a high impact on the understanding of tissue homoeostasis in healthy individuals and help to study dysfunctions in diseased individuals in the coming years.

In addition to the transcriptiomic gene regulation level, tissue homoeostasis can also be studied on the epigenetic level. The research field of epigenetics, which also entered the stage of computational biology recently due to technological developments, describes the study of gene expression changes that do not involve changes in the DNA sequence. These changes can be DNA methylation, histone modification and non-coding RNA-associated gene silencing [Beerman and Rossi, 2015, Waddington, 1942]. Several research groups analyse these epigenetic changes to assess their impact on cell development and dysfunctions. Bonev and Cavalli [2016] for instance study how chromatin structure is established, reset and maintained to shed light on cell regulation and cell fate decisions. Their work focuses on the identification of cell type specific transcription factors and enhancers to explain why some genes are more expressed in some cell types and less in others and to understand how this is associated with 3D chromatin structure. Because changes in the 3D chromatin structure conformation have to occur prior to gene activation, chromatin accessibility is thought to also determine cell fate [Andrey and Mundlos, 2017, Bonev and Cavalli, 2016]. In this context, epigenetic memory seems to be an important feature for cell fate choice determination. For some cell systems, it has already been investigated that cell fate choices require the initiation of heritable gene expression programs, that are at least partly achieved through changes in chromatin structure and DNA methylation [Wilson et al., 2002]. Mechanistic mathematical models are more and more used to distangle epigenetic mechanisms leading to developmental changes or diseases and even models, that take genetic and epigenetic stem cell regulation into account have been developed [Lei et al., 2014].

Moreover, models describing the spatial arrangement of cells within the tissue and their influence on division patterns, as e.g. introduced by Lupperger et al. [2017] could be included as additional information in the analysis framework for *in vivo* studies.

In the near future, the integration of various data sets into a common modelling and analysis framework will be the key to even more effectively infer parameters and more comprehensively validate models. This combination of different experimental data sets resulting from several analyses, ideally performed with material from the same individuals or even the same cells, will allow to study the system of interest at various scales and will allow to rule out even more hypotheses. The respective data analysis integration task one has to solve might be challenging but experimental techniques will also further improve and enable more comprehensive and high quality measurements and thus the advancement of both will be the key to gain a bigger picture, solve several research questions at once, and gain robust results and predictions. The ultimate goal is to go a step beyond and not only gain mechanistic insights to understand biological processes better, but also to provide concrete solutions to health and environmental issues.

A promising therapeutic approach is cell reprogramming and repair. Reprogramming describes reverse differentiation, i.e. the transition from a specialized cell type to a stem cell. A major breakthrough was made in 2006 by identifying the cell culture conditions by which differentiated cells could be genetically reprogrammed to embryo-like stem cells [Takahashi and Yamanaka, 2006]. These induced pluripotent stem cells were generated from embryonic and adult fibroblast cultures by defined factors. Several similar projects followed and researchers succeeded for instance in converting astrocytes into neural stem or progenitor cells and into specific lineages of neurons in human cell cultures [Corti et al., 2012]. Recently this could even be implemented in a living mouse brain [Ma et al., 2018]. By modelling healthy and diseased homoeostasis of a specific tissue in detail, one could upon model validation predict possible reprogramming strategies and subsequently test the most promising ones in the lab experimentally [Folguera-Blasco et al., 2018].

In conclusion, steadily improved experimental techniques will on the long run lead to more detailed, accurate, and less noisy measurements. Inference from such experimental data will be increasingly accurate and allow for fine-grained models which incorporate an increasing number of model parameters to uncover mechanisms in homeostatic processes in healthy and diseased. In addition, the cost for performing large-scale experiments is steadily decreasing, which leads to much larger data sets. This taken together with newly developed experimental techniques will ultimately require the constant development of new, or the extension and tailoring of existing mathematical modeling and inference approaches. The inference of such complex multivariate relationships will in most cases require the combination of mechanistic mathematical models with machine learning techniques, which can handle big data sets easily [Zhang et al., 2020].

Abbreviations

Abbreviation	Explanation
A	asymmetric divisions
ABC	approximate Bayesian computation
AIC	Akaike information criterion
aS	active stem cells
BIC	Bayesian information criterion
BM	bone marrow
C	constrained divisions
CHIP	clonal hematopoiesis of indeterminate potential
CME	chemical master equation
CMP	common myelocyte progenitor
CRN	chemical reaction network
DNA	desoxyribose nucleic acid
dpl	days post labelling
dS	dormant stem cells
FACS	fluorescence activated cell sorting
GMP	granulocyte monocyte progenitors
GRN	gene regulatory network
HSC	hematopoietic stem cell
HSPC	hematopoietic stem and progenitor cells
IVP	initial value problem
LRT	likelihood ratio test
M	mature cells and late progenitors
MAP	maximum a posteriori estimate
MCMC	markov chain monte carlo
MDS	myelodysplastic syndromes
MEP	megakaryocyte erythrocyte progenitor
MLE	maximum likelihood estimation
MLP	multipotent lymphocyte progenitor
MPP	multipotent progenitor cell
N	neuron
NB	neuroblast
ODE	ordinary differential equation
PC	principal component
PL	profile likelihood
qS	quiescent stem cells
S	symmetric divisions
SEZ	subempendymal zone
SMC	sequential monte carlo
SSA	stochastic simulation algorithm
TAP	transitampifying progenitor
U	unconstrained divisions
VAF	variant allele frequency
WHO	World Health Organization

List of Figures

List of Tables

Bibliography

J. Adolfsson, R. Månsson, N. Buza-Vidas, A. Hultquist, K. Liuba, C. T. Jensen, D. Bryder, L. Yang, O. Borge, L. A. M. Thoren, K. Anderson, E. Sitnicka, Y. Sasaki, M. Sigvardsson, and S. E. W. Jacobsen. Identification of flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential: A revised road map for adult blood lineage commitment. *Cell*, 121(2):295 – 306, 2005.

H. Akaike. Information theory and an extension of the maximum likelihood principle. *Proceedings of the Second International Symposium on Information Theory*, pages 267–281, 1973.

K. Akashi, D. Traver, T. Miyamoto, and I. L. Weissman. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature*, 404(6774):193–197, Mar 2000.

B. Alberts, D. Bray, J. H. W. Hunt, J. Lewis, M. Raff, K. Roberts, and J. D. Watson. *Molecular Biology of the Cell.* 2. Courier Corporation, 1989.

S. J. Altschuler and L. F. Wu. Cellular heterogeneity: Do differences make a difference? *Cell*, 141 (4):559–563, May 2010.

G. Andrey and S. Mundlos. The three-dimensional genome: regulating gene expression during pluripotency and development. *Development*, (144):3646–3658, 2017.

A. Auger, P. Chatelain, and P. Koumoutsakos. R-leaping: Accelerating the stochastic simulation algorithm by reaction leaps. *The Journal of Chemical Physics*, 125(8):084103, Aug 2006.

M. Bachar, J. J. Batzel, and S. Ditlevsen. *Stochastic biomathematical models: with applications to neuronal modeling*, volume 2058 of *Lecture notes in mathematics. Mathematical biosciences subseries.* Springer, Heidelberg, 2013. ISBN 9783642321566 (alk. paper).

O. Basak, C. Giachino, E. Fiorini, H. R. MacDonald, and V. Taylor. Neurogenic subventricular zone stem/progenitor cells are notch1-dependent in their active but not quiescent state. *Journal of Neuroscience*, 32(16):5654–5666, Apr 2012.

L. Bast, F. Calzolari, M. K. Strasser, J. Hasenauer, F. J. Theis, J. Ninkovic, and C. Marr. Increasing neural stem cell division asymmetry and quiescence are predicted to contribute to the age-related decline in neurogenesis. *Cell Reports*, 25(12):3231–3240.e8, Dec 2018.

L. Bast, M. Buck, J. S. Hecker, R. A. J. Oostendorp, K. S. Götze, and C. Marr. Computational modeling of stem and progenitor cell kinetics identifies plausible hematopoietic lineage hierarchies. *iScience*, 24(102120), Feb 2021.

I. Beerman and D. J. Rossi. Epigenetic control of stem cell potential during homeostasis, aging, and disease. *Cell Stem Cell*, 16(6):613–625, 2015.

H. S. Bhat and N. Kumar. On the derivation of the Bayesian Information Criterion. *School of Natural Sciences, University of California*, 2010.

C. Blanpain and B. D. Simons. Unravelling stem cell dynamics by lineage tracing. *Nature Reviews Molecular Cell Biology*, 14(8):489–502, Jul 2013.

B. Bonev and G. Cavalli. Organization and function of the 3d genome. *Nature Reviews Genetics*, 17(11):661–678, Nov 2016.

M. Bouab, G.N. Paliouras, A. Aumont, K. Forest-Bérard, and K.J.L. Fernandes. Aging of the subventricular zone neural stem cell niche: evidence for quiescence-associated changes between early and mid-adulthood. *Neuroscience*, 173:135–149, Jan 2011.

S. P. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge University Press, Cambridge, UK, 2004. ISBN 0521833787.

M. Brendel, D. Bonvin, and W. Marquardt. Incremental identification of kinetic models for homogeneous reaction systems. *Chemical Engineering Science*, 61(16):5404 – 5420, 2006.

K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach.* Springer, 2003.

K. P. Burnham and D. R. Anderson. Multimodel inference. *Sociological Methods Research*, 33(2): 261–304, Nov 2004.

K. Busch, K. Klapproth, M. Barile, M. Flossdorf, T. Holland-Letz, S. M. Schlenner, M. Reth, T. Höfer, and H. Rodewald. Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature*, 518(7540):542–6, Feb 2015.

J. C. Butcher. *Numerical Methods for Ordinary Differential Equations.* John Wiley Sons, third edition edition, 2016.

B. Calderhead and M. Girolami. Estimating bayes factors via thermodynamic integration and population mcmc. *Computational Statistics Data Analysis*, 53(12):4028 – 4045, 2009.

F. Calzolari, J. Michel, E. V. Baumgart, F. J. Theis, M. Götz, and J. Ninkovic. Fast clonal expansion and limited neural stem cell self-renewal in the adult subependymal zone. *Nature Neuroscience*, 18(4):490–2, 2015.

Y. Cao, D. T. Gillespie, and L. R. Petzold. Efficient step size selection for the tau-leaping simulation method. *The Journal of Chemical Physics*, 124(4):044109, Jan 2006.

C. Chakraborty and G. Agoramoorthy. Stem cells in the light of evolution. *Indian Journal of Medical Research*, 135(6):813–819, 2012.

F. Chamroukhi and B. T. Huynh. Regularized maximum likelihood estimation and feature selection in mixtures-of-experts models. arXiv:1810.12161 [stat.ML], Oct 2018.

C. Cheyuo, M. Aziz, and P. Wang. Neurogenesis in neurodegenerative diseases: Role of MFG-e8. *Frontiers in Neuroscience*, 13, jun 2019.

O. Chis, J. R. Banga, and E. Balsa-Canto. Methods for checking structural identifiability of nonlinear biosystems: A critical comparison. *IFAC Proceedings Volumes*, 44(1):10585 – 10590, 2011. 18th IFAC World Congress.

S. S. Chung and C. Y. Park. Aging, hematopoiesis, and the myelodysplastic syndromes. *Blood Advances*, 1(26):2572–2578, Dec 2017.

W. G. Cochran. Sampling techniques (john wiley sons, 3rd edition). 1977.

S. Corti, M. Nizzardo, C. Simone, M. Falcone, C. Donadoni, S. Salani, F. Rizzo, M. Nardini, G. Riboldi, F. Magri, and et al. Direct reprogramming of human astrocytes into neural stem cells and neurons. *Experimental Cell Research*, 318(13):1528–1541, Aug 2012.

M. R. Costa, F. Ortega, M. S. Brill, R. Beckervordersandforth, C. Petrone, T. Schroeder, M. Gotz, and B. Berninger. Continuous live imaging of adult neural stem cell division and lineage progression in vitro. *Development*, 138(6):1057–1068, Feb 2011.

G. Craciun and C. Pantea. Identifiability of chemical reaction networks. *J Math Chem*, 44:244–259, 2008.

F. P. Davidescu and S. B. Jørgensen. Structural parameter identifiability analysis for dynamic reaction networks. *Chemical Engineering Science*, 63(19):4754 – 4762, 2008. Model-Based Experimental Analysis.

M. Daynac, L. Morizur, A. Chicheportiche, M. Mouthon, and F. D. Boussin. Age-related neurogenesis decline in the subventricular zone is associated with specific cell cycle regulation changes in activated neural stem cells. *Scientific reports*, 6:21505, 2016.

P. Del Moral, A. Doucet, and A. Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, Jun 2006.

S. Doulatov, F. Notta, K. Eppert, L. T. Nguyen, P. S. Ohashi, and J. E. Dick. Revised map of the human progenitor hierarchy shows the origin of macrophages and dendritic cells in early lymphoid development. *Nature Immunology*, 11(7):585–593, jun 2010.

S. Doulatov, F. Notta, E. Laurenti, and J. E. Dick. Hematopoiesis: A human perspective. *Cell Stem Cell*, 10(2):120 – 136, 2012.

C. Droste. *Uncertainty in parameter estimation for nonlinear dynamical models.* PhD thesis, 1998.

B. Dykstra, S. Olthof, J. Schreuder, M. Ritsema, and G. de Haan. Clonal analysis reveals multiple functional defects of aged murine hematopoietic stem cells. *The Journal of Experimental Medicine*, 208(13):2691–2703, Nov 2011.

J. Eliáš and M. Vořechovský. Modification of the audze–eglājs criterion to achieve a uniform distribution of sampling points. *Advances in Engineering Software*, 100:82–96, 2016.

S. Engblom. Computing the moments of high dimensional solutions of the master equation. *Applied Mathematics and Computation*, 180:498–515, 2006.

A. Filipczyk, C. Marr, S. Hastreiter, J. Feigelman, M. Schwarzfischer, P. S. Hoppe, D. Loeffler, K. D. Kokkaliaris, M. Endele, B. Schauberger, O. Hilsenbeck, S. Skylaki, J. Hasenauer, K. Anastassiadis, F. J. Theis, and T. Schroeder. Network plasticity of pluripotency transcription factors in embryonic stem cells. *Nature cell biology*, 17(10), October 2015.

S. Filippi, C. P. Barnes, J. Cornebise, and M. P.H. Stumpf. On optimality of kernels for approximate bayesian computation using sequential monte carlo. *Statistical Applications in Genetics and Molecular Biology*, 12(1), jan 2013.

D. S. Fischer, A. K. Fiedler, E. M. Kernfeld, R. M. J. Genga, A. Bastidas-Ponce, M. Bakhti, H. Lickert, J. Hasenauer, R. Maehr, and F. J. Theis. Inferring population dynamics from single-cell rna-sequencing time series data. *Nature Biotechnology*, 37(4):461–468, Apr 2019.

T. M. Fliedner, D. Graessle, C. Paulsen, and K. Reimers. Structure and function of bone marrow hemopoiesis: Mechanisms of response to ionizing radiation exposure. *Cancer Biotherapy and Radiopharmaceuticals*, 17(4):405–426, Aug 2002.

N. Folguera-Blasco, E. Cuyàs, J. A. Menédez, and T. Alarcón. Epigenetic regulation of cell fate reprogramming in aging and disease: A predictive computational model. *PLoS Comput Biol.*, 14 (3), Mar 2018.

E. C. Forsberg, T. Serwold, S. Kogan, I. L. Weissman, and E. Passegué. New evidence supporting megakaryocyte-erythrocyte potential of flk2/flt3+ multipotent hematopoietic progenitors. *Cell*, 126(2):415–426, Jul 2006.

N. Friel and A. N. Pettitt. Marginal likelihood estimation via power posteriors. *Royal Statistical Society*, 70(3), 2008.

F. Fröhlich, P. Thomas, A. Kazeroonian, F. J Theis, R. Grima, and J. Hasenauer. Inference for stochastic chemical kinetics using moment equations and system size expansion. *PLoS computational biology*, 12(7):e1005030, 2016.

F. Fröhlich, C. Loos, and J. Hasenauer. Scalable inference of ordinary differential equation models of biochemical processes. In *Methods in Molecular Biology*, pages 385–422. Springer New York, 2018.

A. Gábor, A. F. Villaverde, and J. R. Banga. Parameter identifiability analysis and visualization in large-scale kinetic models of biosystems. *BMC Systems Biology*, 11(1), May 2017.

C. W. Gardiner. *Stochastic methods: a handbook for the natural and social sciences.* Springer series in synergetics. Springer, Berlin, 4th ed edition, 2009. ISBN 9783540707127.

D. F. Gatz and L. Smith. The standard error of a weighted mean concentration—i. bootstrapping vs other methods. *Atmospheric Environment*, 29(11):1185 – 1193, 1995.

G. Genovese, A. K. Kähler, R. E. Handsaker, J. Lindberg, S. A. Rose, S. F. Bakhoum, K. Chambert, E. Mick, B. M. Neale, M. Fromer, and et al. Clonal hematopoiesis and blood-cancer risk inferred from blood dna sequence. *New England Journal of Medicine*, 371(26):2477–2487, Dec 2014.

R. Gesztelyi, J. Zsuga, A. Kemeny-Beke, B. Varga, B. Juhasz, and A. Tosaki. The hill equation and the origin of quantitative pharmacology. *Archive for History of Exact Sciences*, 66(4):427–438, 2012.

B. Giebel, T. Zhang, J. Beckmann, J. Spanholtz, P. Wernet, A. D. Ho, and M. Punzel. Primitive human hematopoietic cells give rise to differentially specified daughter cells upon their initial cell division. *Blood*, 107(5):2146–2152, Mar 2006.

D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403 – 434, 1976.

D. T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1-3):404–425, 1992.

D. T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4):1716–1733, 2001.

N. Goardon, E. Marchi, A. Atzberger, L. Quek, A. Schuh, S. Soneji, P. Woll, A. Mead, K.A. Alford, R. Rout, S. Chaudhury, A. Gilkes, S. Knapper, K. Beldjord, S. Begum, S. Rose, N. Geddes, M. Griffiths, G. Standen, A. Sternberg, J. Cavenagh, H. Hunter, D. Bowen, S. Killick, L. Robinson, A. Price, E. Macintyre, P. Virgo, A. Burnett, C. Craddock, T. Enver, S. Jacobsen, C. Porcher, and P. Vyas. Coexistence of lmpp-like and gmp-like leukemia stem cells in acute myeloid leukemia. *Cancer Cell*, 19(1):138–152.

P. Greulich and B. D. Simons. Dynamic heterogeneity as a strategy of stem cell self-renewal. *Proceedings of the National Academy of Sciences*, 113(27):7509–7514, Jun 2016.

D. Griffiths and D. J. Higham. *Numerical Methods for Ordinary Differential Equations.* Springer Undergraduate Mathematics Series, 2010.

R. Gunawan, K. G. Gadkar, and F. J. Doyle. Methods to identify cellular architecture and dynamics from experimental data. *System Modeling in Cellular Biology*, pages 221–242, Mar 2006.

F. S. Guthery. *A Primer on Natural Resource Science.* Texas AM University Press, 2008.

S. Haas, A. Trumpp, and M. D. Milsom. Causes and consequences of hematopoietic stem cell heterogeneity. *Cell Stem Cell*, 22(5):627 – 638, 2018.

E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II. Springer Series in Computational Mathematics*, volume 14. Springer, Berlin, Heidelberg, 1996.

Q. L. Hao, J. Zhu, M. A. Price, K. J. Payne, L. W. Barsky, and G. M. Crooks. Identification of a novel, human multilymphoid progenitor in cord blood. *Blood*, 97(12):3683–3690, 2001.

J. Hasenauer, 2020. URL `https://www.limes-institut-bonn.de/en/research/research-departments/unit-2/hasenauer-lab/hasenauer-lab-home/`.

H. Hass, C. Loos, E. Raimúndez Álvarez, J. Timmer, J. Hasenauer, and C. Kreutz. Benchmark problems for dynamic modeling of intracellular processes. *Bioinformatics*, Jan 2019.

M. Herberg and I. Roeder. Computational modelling of embryonic stem-cell fate control. *Development*, 2015.

A. Hima Bindu and B. Srilatha. Potency of various types of stem cells and their transplantation. *Journal of Stem Cell Research Therapy*, 01(03), 2011.

J. O. Hirschfelder, C. F. Curtis, and R. B. Bird. *The Molecular Theory of Gases and Liquids.* Wiley, New York, 1954.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

S. Hross and J. Hasenauer. Analysis of cfse time-series data using division-, age- and label-structured population models. *Bioinformatics*, 32(15):2321–2329, Mar 2016.

S. Hug, M. Schwarzfischer, J. Hasenauer, C. Marr, and F. J. Theis. An adaptive scheduling scheme for calculating bayes factors with thermodynamic integration using simpson's rule. *Statistics and Computing*, 26(3):663–677, Mar 2015.

S. Hug, D. Schmidl, W. B. Li, M. B. Greiter, and F. J. Theis. *Uncertainty in Biology. Studies in Mechanobiology, Tissue Engineering and Biomaterials*, volume 17, chapter Bayesian Model Selection Methods and Their Application to Biological ODE Systems. Springer, 2016.

E. G. Hughes, S. H. Kang, M. Fukaya, and D. E. Bergles. Oligodendrocyte progenitors balance growth with self-repulsion to achieve homeostasis in the adult brain. *Nature Neuroscience*, Apr

2013.

H. Hui, Y. Tang, M. Hu, and X. Zhao. Stem cells: General features and characteristics. *Stem Cells in Clinic and Research*, Aug 2011.

B. Hwang, J. H. Lee, and D. Bang. Single-cell rna sequencing technologies and bioinformatics pipelines. *Experimental Molecular Medicine*, 50(8), Aug 2018.

M. Jagannathan-Bogdan and L. I. Zon. Hematopoiesis. *Development*, 140(12):2463–2467, Jun 2013.

T. Jahnke and W. Huisinga. Solving the chemical master equation for monomolecular reaction systems analytically. *J Math Biol*, 54(1):1–26, Jan 2007.

S. Jaiswal, P. Fontanillas, J. Flannick, A. Manning, P. V. Grauman, B. G. Mar, R. C. Lindsley, C. H. Mermel, N. Burtt, A. Chavez, and et al. Age-related clonal hematopoiesis associated with adverse outcomes. *New England Journal of Medicine*, 371(26):2488–2498, Dec 2014.

T. Kalmar, C. Lim, P. Hayward, S. Muñoz-Descalzo, J. Nichols, J. Garcia-Ojalvo, and A. M. Arias. Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biology*, 2009.

G. Karp. *Cell and molecular biology: concepts and experiments.* John Wiley, 2004. ISBN 0471656658.

A. M. Klein and B. D. Simons. Universal patterns of stem cell fate in cycling adult tissues. *Development*, 138(15):3103–11, 2011.

E. Klipp. *Systems biology.* Wiley-VCH, 2010. ISBN 9783527318742.

M. Klose, M. C. Florian, A. Gerbaulet, H. Geiger, and I. Glauche. Hematopoietic stem cell dynamics are regulated by progenitor demand: Lessons from a quantitative modeling approach. *STEM CELLS*, 37(7):948–957, apr 2019.

N. Komin and A. Skupin. How to address cellular heterogeneity by distribution biology. *Current Opinion in Systems Biology*, 3:154 – 160, 2017.

A. Kremling. *Kompendium Systembiologie. Mathematische Modellierung und Modellanalyse.* Springer, 2012.

C. Kreutz. New concepts for evaluating the performance of computational methods. *IFAC-PapersOnLine*, 49(26):63 – 70, 2016. Foundations of Systems Biology in Engineering - FOSBE 2016.

C. Kreutz, A. Raue, D. Katschek, and J. Timmer. Profile likelihood in systems biology. *the FEBS Journal*, 280:2564–2571, 2013.

A. Kriegstein and A. Alvarez-Buylla. The glial nature of embryonic and adult neural stem cells. *Annual Review of Neuroscience*, 32(1):149–184, Jun 2009.

C. Kuehn. *Control of Self-Organizing Nonlinear Systems. Understanding Complex Systems*, chapter Moment Closure - A Brief Review. Springer, Cham, 2016. ISBN 978-3-319-28028-8.

X. Lan, D. J. Jörg, F. M. G. Cavalli, L. M. Richards, L. V. Nguyen, R. J. Vanner, P. Guilhamon, L. Lee, M. M. Kushida, D. Pellacani, and et al. Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy. *Nature*, 549(7671):227–232, Aug 2017.

N. Lartillot and H. Philippe. Computing bayes factors using thermodynamic integration. *Systematic Biology*, 55(2):195–207, 2006.

E. Laurenti, C. Frelin, S. Xie, R. Ferrari, C. F. Dunant, S. Zandi, A. Neumann, I. Plumb, S. Doulatov, J. Chen, and et al. From haematopoietic stem cells to complex differentiation landscapes. *Nature*, 553:418 – 426, Jan 2018.

J. Lee, S. R. Yoon, I. Choi, and H. Jung. Causes and mechanisms of hematopoietic stem cell aging. *International Journal of Molecular Sciences*, 20(6):1272, Mar 2019.

S. C. Lee and O. Abdel-Wahab. The mutational landscape of paroxysmal nocturnal hemoglobinuria revealed: new insights into clonal dominance. *Journal of Clinical Investigation*, 124(10):4227–

4230, Sep 2014.

J. Lei, S. A. Levin, and Q. Nie. Mathematical model of adult stem cell regeneration with crosstalk between genetic and epigenetic regulation. *PNAS; Proceedings of the National Academy of Sciences*, 111(10):E880–E887, 2014.

F. Lewis, A. Butler, and L. Gilbert. A unified approach to model selection using the likelihood ratio test. *Methods in Ecology and Evolution*, 2(2):155–162, August 2010.

S. M. Lewis and A. E. Raftery. Estimating bayes factors via posterior simulation with the laplace-metropolis estimator. *Journal of the American Statistical Association*, 92(438):648–655, Jun 1997.

W. A. Link and R. J. Barker. Model weights and the foundations of multimodel inference. *Ecology*, 87(10):2626–35, 2006.

J. Lipková, G. Arampatzis, P. Chatelain, B. Menze, and P. Koumoutsakos. S-leaping: An adaptive, accelerated stochastic simulation algorithm, bridging τ-leaping and r-leaping. *Bulletin of Mathematical Biology*, 81(8):3074–3096, Jul 2018.

L. Ljung and T. Glad. On global identifiability for arbitrary model parametrizations. *Automatica*, 30(2):265 – 276, 1994.

C. Loos, S. Krause, and J. Hasenauer. Hierarchical optimization for the efficient parametrization of ode models. *Bioinformatics*, Dec 2018.

M. D. Luecken and F. J. Theis. Current best practices in singlecell rnaseq analysis: a tutorial. *Molecular Systems Biology*, 15(6), Jun 2019.

J. Lueckmann, G. Bassetto, T. Karaletsos, and J. H. Macke. Likelihood-free inference with emulator networks. *PMLR*, 96:32–53, 2019.

V. Lupperger, F. Buggenthin, P. Chapouton, and C. Marr. Image analysis of neural stem cell division patterns in the zebrafish brain. *Cytometry Part A*, 93(3):314–322, Nov 2017.

K. Ma, X. Deng, X. Xia, Z. Fan, X. Qi, Y. Wang, Y. Li, Y. Ma, Q. Chen, H. Peng, J. Ding, C. Li, Y. Huang, C. Tian, and J. C. Zheng. Direct conversion of mouse astrocytes into neural progenitor cells and specific lineages of neurons. *Transl Neurodegener.*, Jul 2018.

C. Maier, C. Loos, and J. Hasenauer. Robust parameter estimation for dynamical systems from outlier-corrupted data. *Bioinformatics*, page btw703, Jan 2017.

R. Månsson, A. Hultquist, S. Luc, L. Yang, K. Anderson, S. Kharazi, S. Al-Hashmi, K. Liuba, L. Thorén, J. Adolfsson, and et al. Molecular evidence for hierarchical transcriptional lineage priming in fetal and adult stem cells and multipotent progenitors. *Immunity*, 26(4):407–419, Apr 2007.

J. H. Matis and T. E. Wehrly. Generalized stochastic compartmental models with erlang transit times. *Journal of Pharmacokinetics and Biopharmaceutics*, 18(6):589–607, Dec 1990.

V. Menon. Clustering single cells: a review of approaches on high-and low-depth single-cell rna-seq data. *Briefings in Functional Genomics*, 17(4):240–245, Dec 2017.

R. G. Miller. *Simultaneous Statistical Inference.* Springer, 2nd edition, 1966.

D. F. Moore. Asymptotic properties of moment estimators for overdispersed counts and proportions. *Biometrika*, 73(3):583–588, Apr 1986.

J. Moré and D. Sorensen. Computing a trust region step. *SIAM Journal on Scientific and Statistical Computing*, 4(3):553–572, 1983.

Y. Murakami. Bayesian parameter inference and model selection by population annealing in systems biology. *PLoS ONE*, 9(8):e104057, Aug 2014.

T. Niederberger, H. Failmezger, D. Uskat, D. Poron, I. Glauche, N. Scherf, I. Roeder, T. Schroeder, and A. Tresch. Factor graph analysis of live cell–imaging data reveals mechanisms of cell fate decisions. *Bioinformatics*, 31(11):1816–1823, Jan 2015.

J. Ninkovic and M. Götz. Fate specification in the adult brain–lessons for eliciting neurogenesis from glial cells. *Bioessays*, 35(3):242–52, Mar 2013.

F. Notta, S. Zandi, N. Takayama, S. Dobson, O. I. Gan, G. Wilson, K. B. Kaufmann, J. McLeod, E. Laurenti, C. F. Dunant, and et al. Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science*, 351(6269):aab2116–aab2116, Jan 2016.

M. A. Nowak, F. Michor, and Y. Iwasa. The linear process of somatic evolution. *PNAS; Proceedings of the National Academy of Sciences*, 100(25):14966–14969, 2003.

K. Obernier, A. Cebrian-Silla, M. Thomson, J. I. Parraguez, R. Anderson, C. Guinto, R. J. Rodas, J. Garcia-Verdugo, and A. Alvarez-Buylla. Adult neurogenesis is sustained by symmetric self-renewal and differentiation. *Cell Stem Cell*, 22(2):221–234.e8, Feb 2018.

W. W. Pang, E. A. Price, D. Sahoo, I. Beerman, W. J. Maloney, D. J. Rossi, S. L. Schrier, and I. L. Weissman. Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age. *PNAS; Proceedings of the National Academy of Sciences*, 108(50): 20012–20017, 2011.

W. W. Pang, J. V. Pluvinage, E. A. Price, K. Sridhar, D. A. Arber, P. L. Greenberg, S. L. Schrier, C. Y. Park, , and I. L. Weissman. Hematopoietic stem cell and progenitor cell mechanisms in myelodysplastic syndromes. *PNAS*, 110(8):3011–3016, 2013.

DS Park, AA Akuffo, DE Muench, HL Grimes, PK Epling-Burnette, PK Maini, ARA Anderson, and MB Bonsall. Clonal hematopoiesis of indeterminate potential and its impact on patient trajectories after stem cell transplantation. *PLoS Comput Biol*, 15(4), Apr 2019.

T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103 (482):681–686, 2008.

R. Passier. Origin and use of embryonic and adult stem cells in differentiation and tissue repair. *Cardiovascular Research*, 58(2):324–335, May 2003.

L. Perié, K. R. Duffy, Kok L., R. J. de Boer, and T. N. Schumacher. The branching point in erythro-myeloid differentiation. *Cell*, 163(7):1655–1662, Dec 2015.

L. Petreanu and A. Alvarez-Buylla. Maturation and death of adult-born olfactory bulb granule neurons: role of olfaction. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 22(14):6106–13, 2002.

S. Pinho and P. S. Frenette. Haematopoietic stem cell activity and interactions with the niche. *Nature Reviews Molecular Cell Biology*, 20(5):303–320, Feb 2019.

J. Platel, K. A. Dave, V. Gordon, B. Lacar, M. E. Rubio, and A. Bordey. Nmda receptors activated by subventricular zone astrocytic glutamate are critical for neuroblast survival prior to entering a synaptic network. *Neuron*, 65(6):859–72, Mar 2010.

H. Pohjanpalo. System identifiability based on the power series expansion of the solution. *Mathematical Biosciences*, 41(1):21 – 33, 1978.

G. Poiana, R. Gioia, S. Sineri, S. Cardarelli, G. Lupo, and E. Cacci. Transcriptional regulation of adult neural stem/progenitor cells: tales from the subventricular zone. *Neural Regen Res*, 15 (10):1773–83, 2020.

G. Ponti, K. Obernier, C. Guinto, L. Jose, L. Bonfanti, and A. Alvarez-Buylla. Cell cycle and lineage progression of neural progenitors in the ventricular-subventricular zones of adult mice. *Proceedings of the National Academy of Sciences of the United States of America*, 110(11):E1045–54, 2013.

D. Posada and T. R. Buckley. Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53(5):793–808, Oct 2004.

T. P. Prescott and R. E. Baker. Multifidelity approximate bayesian computation. *arXiv.org*, 11 2018. URL `https://arxiv.org/pdf/1811.09550`.

C. J. H. Pronk, D. J. Rossi, R. Mansson, J. L. Attema, G. L. Norddahl, C. K. F. Chan, M. Sigvardsson, I. L. Weissman, and D. Bryder. Elucidation of the phenotypic, functional, and molecular topography of a myeloerythroid progenitor cell hierarchy. *Cell Stem Cell*, 1(4):428–442, Oct 2007.

A. Raue, M. Schilling, J. Bachmann, A. Matteson, M. Schelker, M. Schelke, D. Kaschek, S. Hug, C. Kreutz, B. D. Harms, F. J. Theis, U. Klingmüller, and J. Timmer. Lessons learned from quantitative dynamical modeling in systems biology. *PLoS One*, 8(9):e74335, 2013.

A. Regev, S. Teichmann, O. Rozenblatt-Rosen, M. Stubbington, K. Ardlie, I. Amit, P. Arlotta, G. Bader, C. Benoist, M. Biton, B. Bodenmiller, B. Bruneau, P. Campbell, M. Carmichael, P. Carninci, L. Castelo-Soccio, M. Clatworthy, H. Clevers, C. Conrad, R. Eils, J. Freeman, L. Fugger, B. Goettgens, D. Graham, A. Greka, N. Hacohen, M. Haniffa, I. Helbig, R. Heuckeroth, S. Kathiresan, S. Kim, A. Klein, B. Knoppers, A. Kriegstein, E. Lander, J. Lee, E. Lein, S. Linnarsson, E. Macosko, S. MacParland, R. Majovski, P. Majumder, J. Marioni, I. McGilvray, M. Merad, M. Mhlanga, S. Naik, M. Nawijn, G. Nolan, B. Paten, D. Pe'er, A. Philippakis, C. Ponting, S. Quake, J. Rajagopal, N. Rajewsky, W. Reik, J. Rood, K. Saeb-Parsy, H. Schiller, S. Scott, A. Shalek, E. Shapiro, J. Shin, K. Skeldon, M. Stratton, J. Streicher, H. Stunnenberg, K. Tan, D. Taylor, A. Thorogood, L. Vallier, A. van Oudenaarden, F. Watt, W. Weicher, J. Weissman, A. Wells, B. Wold, R. Xavier, X. Zhuang, and Human Cell Atlas Organizing Committee. The human cell atlas white paper. *arXiv.org*, 10 2017. URL `https://arxiv.org/pdf/1810.05192`.

H. Resat, L. Petzold, and M. F. Pettigrew. Kinetic modeling of biological systems. *Methods in Molecular Biology*, 541:311–35, 2009.

D. Reynaud, N. Lefort, E. Manie, L. Coulombel, and Y. Levy. In vitro identification of human pro-b cells that give rise to macrophages, natural killer cells, and t cells. *Blood*, 101(11):4313–4321, Jun 2003.

L. Ritsma, S. I. J. Ellenbroek, A. Zomer, H. J. Snippert, F. J. de Sauvage, B. D. Simons, H. Clevers, and J. van Rheenen. Intestinal crypt homeostasis revealed at single-stem-cell level by in vivo live imaging. *Nature*, 507(7492):362–365, Feb 2014.

C. Robers. *Ordinary Differential Equations: Applications, Models, and Computing.* CRC Press, 1 edition, June 2018.

I. Roeder, M. Horn, I. Glauche, A. Hochhaus, M. C. Mueller, and M. Loeffler. Dynamic modeling of imatinib-treated chronic myeloid leukemia: functional insights and clinical implications. *Nature Medicine*, 12(10):1181–1184, Oct 2006.

B. Rösch. Analytic solution to stochastic cell cycle models. Bachelor thesis, April 2018.

D. J. Rossi, C. H. M. Jamieson, and I. L. Weissman. Stems cells and the pathways to aging and cancer. *Cell*, 132(4):681–696, 2008.

Alejandra Sanjuan-Pla, Iain C. Macaulay, Christina T. Jensen, Petter S. Woll, Tiago C. Luis, Adam Mead, Susan Moore, Cintia Carella, Sahoko Matsuoka, Tiphaine Bouriez Jones, and et al. Platelet-biased stem cells reside at the apex of the haematopoietic stem-cell hierarchy. *Nature*, 502(7470):232–236, Aug 2013.

L. Schwartz. On bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 4(1):10–26, 1965.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461 – 464, 1978.

A. Shastri, B. Will, U. Steidl, and A. Verma. Stem and progenitor cell alterations in myelodysplastic syndromes. *Blood*, 129(12):1586–1594, Mar 2017.

B. A. Shook, D. H. Manz, J. J. Peters, S. Kang, and J. C. Conover. Spatiotemporal changes to the subventricular zone stem cell pool through aging. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 32(20):6947–56, 2012.

L. Silberstein, K. A. Goncalves, P. V. Kharchenko, R. Turcotte, Y. Kfoury, F. Mercier,

N. Baryawno, N. Severe, J. Bachand, J. A. Spencer, A. Papazian, D. Lee, B. R. Chitteti, E. F. Srour, J. Hoggatt, T. Tate, C. Lo Celso, N. Ono, S. Nutt, J. Heino, K. Sipila, T. Shioda, M. Osawa, C. P. Lin, G. Hu, and D. T. Scadden. Proximity-based differential single-cell analysis of the niche to identify stem/progenitor cell regulators. *Cell Stem Cell*, 19:530–543, Oct 2016.

V. K. Singh, A. Saini, M. Kalsan, N. Kumar, and R. Chandra. Describing the stem cell potency: The various methods of functional assessment and in silico diagnostics. *Front Cell Dev Biol.*, 4, Nov 2016.

S. A. Sisson. *Handbook of Approximate Bayesian Computation.* Chapman and Hall/CRC, Sep 2018. ISBN 9781315117195.

V. Sotiropoulos and Y. N. Kaznessis. Analytical derivation of moment equations in stochastic chemical kinetics. *Chemical engineering science*, 66(3):268–277, 2011.

A. S. Sperling, C. J. Gibson, and B. L. Ebert. The genetics of myelodysplastic syndrome: from clonal haematopoiesis to secondary leukaemia. *Nature Reviews Cancer*, 17(1):5–19, Jul 2017.

S. Srinath and R. Gunawan. Parameter identifiability of power-law biochemical system models. *Journal of Biotechnology*, 149(3):132 – 140, 2010. Advanced Methods in Molecular Systems Biology.

Paul Stapor, Daniel Weindl, Benjamin Ballnus, Sabine Hug, Carolin Loos, Anna Fiedler, Sabrina Krause, Sabrina Hroß, Fabian Fröhlich, and Jan Hasenauer. Pesto: Parameter estimation toolbox. *Bioinformatics*, page btx676, 2017.

D. P. Steensma, R. Bejar, S. Jaiswal, R. C. Lindsley, M. A. Sekeres, R. P. Hasserjian, and B. L. Ebert. Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood*, 126(1):9–16, Jul 2015.

G. S. Stent. The role of cell lineage in development. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 312(1153):3–19, 1985. URL `http://www.jstor.org/stable/2396299`.

T. Stiehl, N. Baran, A. D. Ho, and A. Marciniak-Czochra. Cell division patterns in acute myeloid leukemia stem-like cells determine clinical course: A model to predict patient survival. *Cancer Research*, 75(6):940–949, Jan 2015.

K. Sudo, H. Ema, Y. Morita, and H. Nakauchi. Age-associated characteristics of murine hematopoietic stem cells. *Journal of Experimental Medicine*, 192(9):1273–1280, Nov 2000.

R. Sutherland. Cell and environment interactions in tumor microregions: the multicell spheroid model. *Science*, 240(4849):177–184, Apr 1988.

K. Takahashi and S. Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663 – 676, 2006.

H. Takano, H. Ema, K. Sudo, and H. Nakauchi. Asymmetric division and lineage commitment at the level of hematopoietic stem cells. *Journal of Experimental Medicine*, 199(3):295–302, 2004.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

T. Toni and M. P. H. Stumpf. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, 26(1):104–110, oct 2009.

T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 6(31):187–202, feb 2009.

C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, Mar 2014.

M. Uhlen, P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester, S. Hober, and et al. Towards a knowledge-based human protein atlas.

Nature Biotechnology, 28(12):1248–1250, Dec 2010.

N. Urbán, D. L. C. van den Berg, A. Forget, J. Andersen, J. A. A. Demmers, C. Hunt, O. Ayrault, and F. Guillemot. Return to quiescence of mouse neural stem cells by degradation of a proactivation protein. *Science*, 353(6296):292–295, Jul 2016.

N. G. van Kampen. *Stochastic processes in physics and chemistry.* North-Holland personal library. Elsevier, Amsterdam, 3rd ed edition, 2007. ISBN 9780444529657 (pbk.).

L. Velten, S. F. Haas, S. Raffel, S. Blaszkiewicz, S. Islam, B. P. Hennig, C. Hirche, C. Lutz, E. C. Buss, D. Nowak, and et al. Human haematopoietic stem cell lineage commitment is a continuous process. *Nature Cell Biology*, 19(4):271–281, Mar 2017.

A. F. Villaverde and J. R. Banga. Dynamical compensation and structural identifiability of biological models: Analysis, implications, and reconciliation. *PLOS Computational Biology*, 13(11): e1005878, Nov 2017.

A. F. Villaverde, A. Barreiro, and A. Papachristodoulou. Structural identifiability of dynamic systems biology models. *PLOS Computational Biology*, 12(10):e1005153, Oct 2016.

A. F. Villaverde, N. D. Evans, M. J. Chappell, and J. R. Banga. Sufficiently exciting inputs for structurally identifiable systems biology models. *IFAC-PapersOnLine*, 51(19):16 – 19, 2018. 7th Conference on Foundation of Systems Biology in Engineering FOSBE 2018.

A. F. Villaverde, N. D. Evans, M. J. Chappell, and J. R. Banga. Input-dependent structural identifiability of nonlinear systems. *IEEE Control Systems Letters*, 3(2):272–277, 2019a.

A. F. Villaverde, F. Fröhlich, D. Weindl, J. Hasenauer, and J. R. Banga. Benchmarking optimization methods for parameter estimation in large kinetic models. *Bioinformatics*, 35(5):830–838, March 2019b.

C. H. Waddington. The epigenotype. 1942. *Endeavour*, pages 18–20, 1942.

T. Walenda, T. Stiehl, H. Braun, J. Fröbel, A. D. Ho, T. Schroeder, T. W. Goecke, B. Rath, U. Germing, A. Marciniak-Czochra, and W. Wagner. Feedback signals in myelodysplastic syndromes: increased self-renewal of the malignant clone suppresses normal hematopoiesis. *PLoS Comput Biol*, 10(4):e1003599, Apr 2014.

E. Walter. *Identifiability of parametric models.* Pergamon Books, 1987.

E. Walter and L. Pronzato. On the identifiability and distinguishability of nonlinear parametric models. *Mathematics and Computers in Simulation*, 42(2):125 – 134, 1996. Mathematical Modelling and Simulation in Agriculture and Bio-Industries Proceedings of the 1st IMACS-IFAC Symposium Msu2SABI.

E. Walter and L. Pronzato. *Identification of Parametric Models.* Springer, 1997.

J. K. Watson, S. Rulands, A. C. Wilkinson, A. Wuidart, M. Ousset, A. Van Keymeulen, B. Göttgens, C. Blanpain, B. D. Simons, and E. L. Rawlins. Clonal dynamics reveal two distinct populations of basal cells in slow-turnover airway epithelium. *Cell Reports*, 12(1):90 – 101, 2015.

C. Weissleder, H. F. North, and C. S. Weickert. Important unanswered questions about adult neurogenesis in schizophrenia. *Current Opinion in Psychiatry*, 32(3):170–178, May 2019.

D. A. Williams. Discrimination between regression models to determine the pattern of enzyme synthesis in synchronous cell cultures. *Biometrics*, 26(1):23–32, 1970. URL `http://www.jstor.org/stable/2529041`.

Q. F. Wills, E. Mellado-Gomez, R. Nolan, D. Warner, E. Sharma, J. Broxholme, B. Wright, H. Lockstone, W. James, M. Lynch, and et al. The nature and nurture of cell heterogeneity: accounting for macrophage gene-environment interactions with single-cell rna-seq. *BMC Genomics*, 18(1), Jan 2017.

C. B. Wilson, K. W. Makar, and M. Pérez-Melgosa. Epigenetic regulation of t cell fate and function. *The Journal of Infectious Diseases*, 185(s1):S37–S45, Feb 2002.

J. Yang, M. V. Plikus, and N. L. Komarova. The role of symmetric stem cell divisions in tissue homeostasis. *PLoS Comput Biol.*, 11, 2015.

W. Y. Yang, W. Cao, T. Chung, and J. Morris. *Applied Numerical Methods Using MATLAB.* Hoboken: John Wiley Sons, Incorporated, 2005.

K. Z. Yao, B. M. Shaw, B. Kou, K. B. McAuley, and D. W. Bacon. Modeling ethylene/butene copolymerization with multisite catalysts: Parameter estimability and experimental design. *Polymer Reaction Engineering*, 11(3):563–588, Jan 2003.

C. A. Yates and G. Klingbeil. Recycling random numbers in the stochastic simulation algorithm. *The Journal of Chemical Physics*, 138(9):094103, Mar 2013.

D. E. Zak. Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: Insights from an identifiability analysis of an in silico network. *Genome Research*, 13(11):2396–2405, Nov 2003.

J. Zhang, S. D. Petersen, T. Radivojevic, A. Ramirez, A. Pérez-Manríquez, E. Abeliuk, B. J. Sánchez, Z. Costello, Y. Chen, M. J. Fero, H. G. Martin, J. Nielsen, J. D. Keasling, and M. K. Jensen. Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism. *Nature Communications*, 11(1), sep 2020.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Statistical Methodology*, 67(2):301–320, April 2005.