UNIVERSITY OF CALIFORNIA,
IRVINE

Mathematical modeling of malignant myelopoiesis: optimal experimental design and targeted therapy

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematical, Computational and Systems Biology

by

Abdon Iniguez

Dissertation Committee:
Professor John S. Lowengrub, Chair
Professor Jun F. Allard
Professor Dominik F. Wodarz

2019

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, I would like to thank my advisors John Lowengrub and Jun Allard for providing guidance throughout my time here at UC Irvine. It was a great pleasure working with both of you. I would like to thank Fred Wan, Dominik Wodarz, and Richard van Etten for being on my committee. I would like to thank my collaborators Luis Martinez Lomeli and Prasanthi Tata for many useful discussions.

I would like to thank all of my friends and family for all of their love and support. Lastly, I would like thank Karen Sujata for all of her support throughout graduate school.

The work is supported by the National Science Foundation Graduate Research Fellowship under Grant No. 16-588, National Institutes of Health through grant U54-CA21737801A1, NSF-Simons Center grants NSF DMS1763272, and a Simons Foundation grant (594598, QN)

CURRICULUM VITAE

Abdon Iniguez

EDUCATION

Doctor of Philosophy in
Mathematical, Computational, and Systems Biology **2019**
University of California, Irvine *Irvine, California*

Bachelor of Science in Applied Mathematics **2014**
University of California, Irvine *Irvine, California*

RESEARCH EXPERIENCE

Graduate Student Researcher **2014–2019**
University of California, Irvine *Irvine, California*

REFEREED JOURNAL PUBLICATIONS

Iniguez A, Allard J. **Spatial pattern formation in microtubule post-translational modifications and the tight localization of motor-driven cargo** **2017**
Journal of Math Biology;74(5):1059-1080

Stipp SR, **Iniguez A**, Wan F, Wodarz D. **Timing of CD8 T cell effector responses in viral infections** **2016**
Royal Society Open Science;3(2):150661

ABSTRACT OF THE DISSERTATION

Mathematical modeling of malignant myelopoiesis: optimal experimental design and targeted therapy

By

Abdon Iniguez

Doctor of Philosophy in Mathematical, Computational and Systems Biology

University of California, Irvine, 2019

Professor John S. Lowengrub, Chair

Chronic myeloid leukemia (CML) is a blood cancer in which there is dysregulation of maturing myeloid cells (granulocytes) driven by a chromosomal mutation which creates the fusion gene, BCR-ABL1. Although there has been much progress in the treatment of CML using tyrosine kinase inhibitors (TKI), there are still unmet clinical needs. For example, there is still a small cohort of patients who, for reasons that are still unknown, do not respond to TKI treatment. Further, a significant proportion of patients who appear to have a complete molecular remission while on TKIs experience a relapse of CML when TKI treatment is discontinued. Mathematical modeling of CML hematopoiesis can provide insight on these processes. Current mathematical models of CML are highly simplified (e.g., linear) and have been used to estimate treatment-related model parameters, but fail to describe the mechanisms that underlie primary resistance and treatment-free remission. Here, we explore how more physiologically accurate, data-driven mathematical models of CML hematopoiesis that incorporate feedback control and lineage branching can provide such insight. Although it is recognized that feedback plays a role in CML hematopoiesis, the interactions are poorly understood. In many cases, we don't know which cell types are providing and receiving the feedback, what signals are used, and what aspects of proliferative cell behavior they influence (i.e. cell cycle speed, renewal probability, or progeny fate choice). Here, we propose to use

mathematical modeling combined with data from single cell transcriptomics to help sort this out. We develop an automated method for model selection that integrates data gleaned from experiments and single cell RNA sequencing to select plausible classes of feedback models. We first apply this approach to normal hematopoiesis and identify models that have desired system properties, e.g., stable equilibria, and make predictions about system behavior upon perturbation. New experiments are presented that validate model predictions. We use a Bayesian utility theory approach to determine the optimal experimental design that allows for identification of model parameters with maximum precision. When extended to incorporate CML hematopoiesis, our initial assessment shows that feedback/branching models are more robust, have a better fit to alternative patterns of patient response than simple linear models, and suggest new treatment strategies.

Chapter 1

Introduction

The hematopoietic system represents the prototypical model of a mammalian tissue organized in a hierarchy with pluripotent stem cells at the apex. This system produces billions of mature myeloid and erythroid cells on a daily basis, and accommodates massive increases in production of individual cell types in response to pathological stresses. An ordered hierarchy of differentiation is shown in figure 1.1.

The control of production of red cells and platelets is reasonably well understood, with lineage-restricted erythroid and megakaryocytic progenitors being selectively responsive to specific cytokines (erythropoietin and thrombopoietin, respectively). By contrast, the regulation of production of neutrophils and other granulocytes in healthy and disease is poorly understood. Two cytokines [granulocyte colony-stimulating factor (G-CSF) and granulocyte-macrophage colony-stimulating factor] can pharmacologically increase neutrophil production, but mice lacking both cytokines maintain baseline neutrophil levels and can still increase neutrophil production in response to infection [6]. Feedback circuits involving phagocytic cells and IL-17/23 production or granule proteins such as neutrophil elastase have been implicated in regulation of neutrophil production but do not completely explain the process [71, 29].

Chronic myeloid leukemia (CML) is a myeloproliferative neoplasm (MPN) characterized by autonomous overproduction of granulocytes (principally neutrophils, along with basophils and eosinophils) and myeloid progenitor cells (Neu, Eos, Bas, GMP, and CMP shown in Figure 1.1) with three distinct disease phases [15]. In the initial "chronic" phase, the differentiation of myeloid progenitors is essentially normal, resulting in excessive levels of mature post-mitotic neutrophils and their immediate precursors. The genetic hallmark of CML is

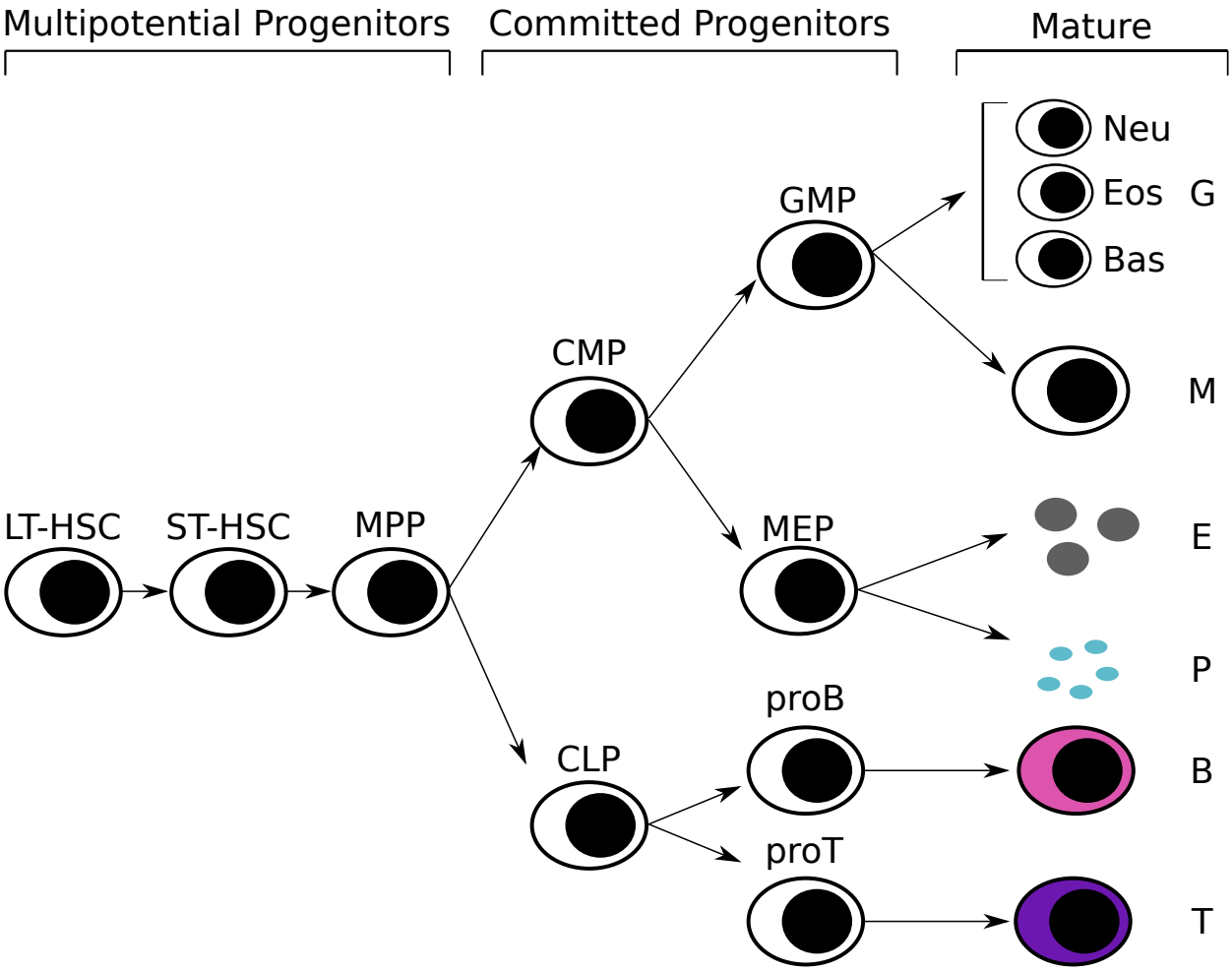

Figure 1.1: Lineage schematic of hematopoiesis. The organization of the hematopoietic lineage beginning with the long term hematopoietic stem cell (LT-HSC) on the left. ST-HSC: Short term hematopoietic stem cell, MPP: Multi Potent Progenitor, CMP: Common Myeloid progenitor, CLP: Common Lymphoid Progenitor, GMP: Granulocyte-Macrophage Progenitor, MEP: Megakaryocte-Erythroid Progenitor, Neu: Neutrophil, Eos: Eosinophils, Bas: Basophils, G: Granulocyte, M: Macrophage, E: Erythroid, P: Platelet, B: B-Cell, T: T-Cell

the Philadelphia (Ph) chromosome, the result of a balanced translocation between chromosomes 9 and 22 [t(9;22)(q34;q11)] that creates a fusion of the genes for BCR and ABL1. This translocation occurs somatically in a hematopoietic stem cell (HSC), with the Ph chromosome found not just in the malignant neutrophils but also in cells of monocytic, erythroid, megakaryocytic, and lymphoid lineage. The product of the BCR-ABL1 fusion gene is a dysregulated cytoplasmic protein-tyrosine kinase, BCR-ABL1. CML thus represents a natural model of dysregulated granulocytopoiesis [61].

The prognosis in CML has been changed dramatically by the development of selective small-molecule inhibitors of the BCR-ABL1 kinase, hereafter referred to as ABL1 tyrosine kinase inhibitors (TKIs). The first ABL1 TKI, imatinib mesylate was approved by the FDA in May 2001, and we now have clinical experience with this drug in thousands of patients spanning nearly 20 years. The response of CML to TKI therapy is assessed initially by peripheral blood counts, subsequently by the frequency of the Ph chromosome or BCR-ABL1 fusion gene in blood or marrow cells via quantification of BCR-ABL1 transcripts by RT-PCR [5]. The majority of CML patients treated with imatinib in chronic phase achieve a cytogenetic remission (disappearance of the Ph chromosome) [24].

Today, there are two principal areas of unmet clinical need in CML. The first is resistance to TKI therapy, which can be either acquired or primary. Acquired resistance is defined as progression of disease after an initial response, while primary resistance is failure of an initial response to TKIs. A major mechanism of acquired resistance to TKIs is development of point mutations in BCR-ABL1 kinase domain that decrease sensitivity to the drug, and this can be managed clinically to some extent by switching from one TKI to another. Primary resistance affects about 10-15% of newly diagnosed CML patients and is defined by failure to achieve an "early molecular response", a 10-fold decrease in BCR-ABL1 transcript levels at 3 months [32, 55]. Clinical data indicate that switching TKIs may not benefit these patients, suggesting that this group is destined to do poorly regardless of the specific inhibitor used

[81]. The mechanism underlying primary resistance is unclear despite early attempts at NGS profiling, but BCR-ABL1 mutations are not present.

The second area of clinical need in CML concerns discontinuation of TKI therapy, which presently is intended to be life-long. For example, figure 1.2 shows a patient response curve for a year and three months. The specific patient is shown to relapses after going off therapy after 1 year. The BCR-ABL1 fraction is shown to exceed the initial fraction at the start of therapy. Cell biological studies suggest that the most primitive leukemic stem cell compartment in CML is not eliminated by TKI treatment [28, 7] despite effective inhibition of BCR-ABL1 kinase activity [18]. Rather, exposure to TKIs drives CML stem cells into quiescence where they are resistant to apoptosis, whereas treatment with G-CSF can partially reverse quiescence and TKI resistance [42]. In the clinic, there is a need to predict which patients will sustain CMR upon TKI discontinuation, and for strategies to increase those who maintain TFR and might be cured of their disease [2].

The use of mathematical models in hematopoiesis and CML is part of a growing effort to understand blood development and its diseased state. In particular, CML has been modeled using unbranched lineages with 3-4 subpopulations representing stem cells, progenitor cells, and terminally differentiated post-mitotic cells with constant model parameters and logistic-like terms [56, 74, 72, 78, 35, 64]. The models collapse the lymphoid and myeloid lineages into a single population. The models are then fit to therapy response data to suggest how the subpopulations are affected by therapy. This modeling fails to suggest mechanisms that underlie primary resistance or therapeutic options to increase the frequency of TFR. Thus, there is a need to make more physiologically accurate, data driven mathematical model.

In the following chapters, we use mathematical modeling to study general properties of normal hematopoiesis. We develop an automated model selection method to select plausible classes of feedback/feedforward models using biological observations. Using our selected model of normal hematopoiesis, we extend our model to include CML cell types. Using our

differential equations of hematopoiesis, we propose a Bayesian optimal design framework to suggest the best experiments based on optimizing information gain. The framework allows us to infer parameters that cannot be easily obtained through experiments such as feedback strengths.

This dissertation is organized as follows. Chapter 2 summarizes design space analysis and goes into detail on how this analysis can be used for an ODE lineage model. Chapter 3 is a paper in preparation with Prasanthi Tata, Richard van Etten, and John Lowengrub on modeling chronic myeloid leukemia. Chapter 4 summarizes how one can fit ODE parameters using Bayesian inference. Chapter 5 is a paper in preparation regarding Bayesian optimal experimental design on data driven models of hematopoiesis with Luis Martinez Lomeli, Prasanthi Tata, Babak Shahbaba, John Lowengrub, Vladimir Minin, and Richard van Etten.

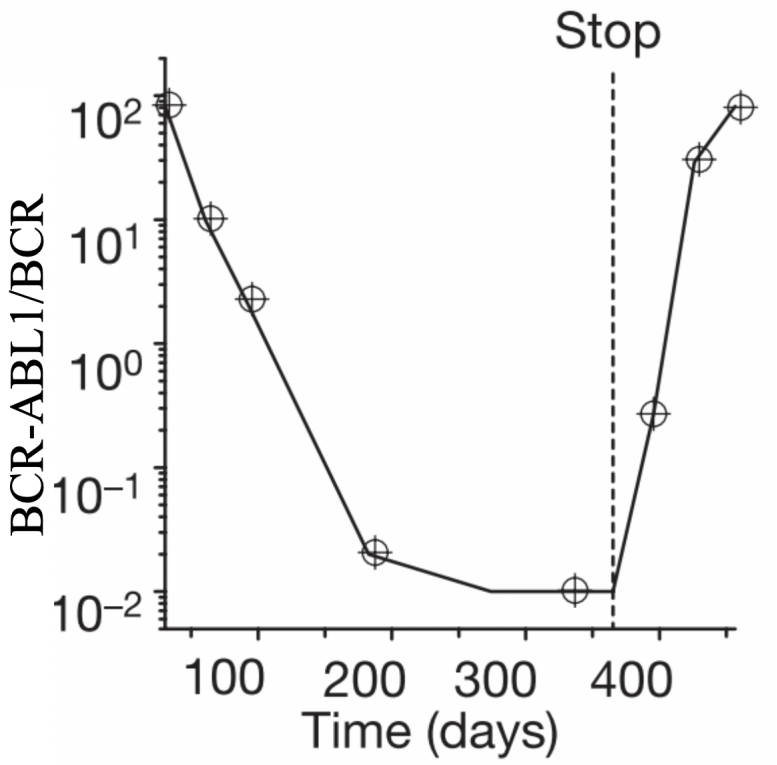

Figure 1.2: Patient relapse. TKI therapy response for a single patient taken from [56].

Chapter 2

Design Space Analysis

2.1 Introduction

In classical dynamical systems, linear stability analysis, bifurcation analysis, and sensitivity analysis are done to understand system behaviors. These types of analyses are much easier on simpler systems, but analyzing highly nonlinear systems is often times difficult to interpret analytically. Moreover, specific parameter values need to be known to search parameter space. As an alternative, we can use an analysis developed by Michael Savageau named design space analysis (DSA) to analyze nonlinear dynamical systems at steady state by characterizing dominant terms in the equations [69, 48, 21]. The analysis also allows us to divide parameter space into regions in where we can find common qualitative behavior, where the qualitative behavior corresponds to stability and sensitivity of the system. Applying this analysis allows us to (1) ignore specific parameter values, (2) obtain analytical steady states, (3) reduce the search area by defining boundaries, and (4) easily automate this process. This analysis has been applied to biochemical systems in which they effectively search for boundaries in parameter space by characterizing model dynamics.

In this chapter, we review how one can construct a design space given an ODE model. We then apply this analysis to a simplistic ODE lineage model and show how this model can be used to define regions of stability in parameter space.

2.2 Boundaries of design space

In order to apply DSA, the ODE must be a generalized mass action system, shown below in equation 2.1- 2.2,

$$\frac{dX_i}{dt} = \sum_{k=1}^{r} \alpha_{ik} \prod_{j=1}^{n+m} X_j^{g_{ijk}} - \sum_{k=1}^{r} \beta_{ik} \prod_{j=1}^{n+m} X_j^{h_{ijk}} \tag{2.1}$$

$$X_i(0) = X_{i0} \tag{2.2}$$

for i=1,..,n. The n corresponds to the number dependent variables and m corresponds to the number of independent variables. The α_{ik} and β_{ik} parameters correspond to rate constants of the differential equation and r corresponds to the number of associated rate constants.

DSA takes advantage of the above form by creating a system of deconstructed ODEs where one source term and one sink term in the differential equation dominate, known as a sub-system (S-system). The following is the generalized form of a S-system:

$$\frac{dX_i}{dt} = \alpha_{ip} \prod_{j=1}^{n+m} X_j^{g_{ijp}} - \beta_{iq} \prod_{j=1}^{n+m} X_j^{h_{ijq}} \tag{2.3}$$

$$X_i(0) = X_{i0} \tag{2.4}$$

where p and q correspond to the number of positive and negative terms of the differential

equation, respectively. We are interested in solving these solutions at steady state, and therefore solve the system by setting the time derivative to zero. We take advantage of the form shown in equation 2.3 and take the log of the system:

$$\log(\alpha_{ip}) + \sum_{j=1}^{n+m} g_{ijp} \log(X_j) = \log(\beta_{iq}) + \sum_{j=1}^{n+m} h_{ijq} \log(X_j) \tag{2.5}$$

thus, making this a linear solve in log space. In defining the S-systems, we must make assumptions about the model and its parameters. To satisfy the S-systems, we impose inequality constraints to satisfy the dominating source and sink terms of the S-systems by the following,

$$\alpha_{ip} \prod_{j=1}^{n+m} X_j^{g_{ijp}} > \alpha_{i\bar{p}} \prod_{j=1}^{n+m} X_j^{g_{ij\bar{p}}} \text{for i = 1,..,n; } \bar{p}\text{=1,...,p-1,p+1,...,r} \tag{2.6}$$

$$\beta_{iq} \prod_{j=1}^{n+m} X_j^{h_{ijq}} > \beta_{i\bar{q}} \prod_{j=1}^{n+m} X_j^{h_{ij\bar{q}}} \text{for i = 1,..,n; } \bar{q}\text{=1,...,q-1,q+1,...,r} \tag{2.7}$$

We can then log transform these inequalities to obtain

$$\log(\alpha_{ip}) + \sum_{j=1}^{n+m} g_{ijp} \log(X_j) > \log(\alpha_{i\bar{p}}) + \sum_{j=1}^{n+m} g_{ij\bar{p}} \log(X_j) \tag{2.8}$$

$$\log(\beta_{iq}) + \sum_{j=1}^{n+m} h_{ijq} \log(X_j) > \log(\beta_{i\bar{q}}) + \sum_{j=1}^{n+m} h_{ij\bar{q}} \log(X_j) \tag{2.9}$$

These inequalities become our boundaries in parameter space in which qualitative behavior is shared once the X_j's are evaluated at steady state, where the steady states are obtained from solving the log-linear S-system. Not all S-systems will have a unique solution or satisfy

the inequality constraints. The S-systems that do not satisfy these constraints will be thrown out. The analysis can be summarized by figure 2.1.

2.3 Analysis on a two cell lineage model

To illustrate, we apply this analysis to a simple lineage model with two cell types (stem cells (s) and terminal cells (t)). The stem cells have the ability self renew at a probability p_0 and differentiate with probability $1 - p_0$. Terminal cells are considered post-mitotic cells and die at a rate d. We rescale time by the stem cell division. We add negative feedback on the self renewal probability of the stem cells from the differentiated cells. Figure 2.2 shows the lineage with their parameters. The corresponding differential equations are

$$x_s' = (2p_0 - 1)x_s \tag{2.10}$$

$$x_t' = 2(1 - p_0)x_s - dx_t \tag{2.11}$$

where $p_0 = \bar{p_0}/(1 + \gamma x_t)$. $\bar{p_0}$ is defined as the max probability of stem cell self renewal and γ is the feedback strength.

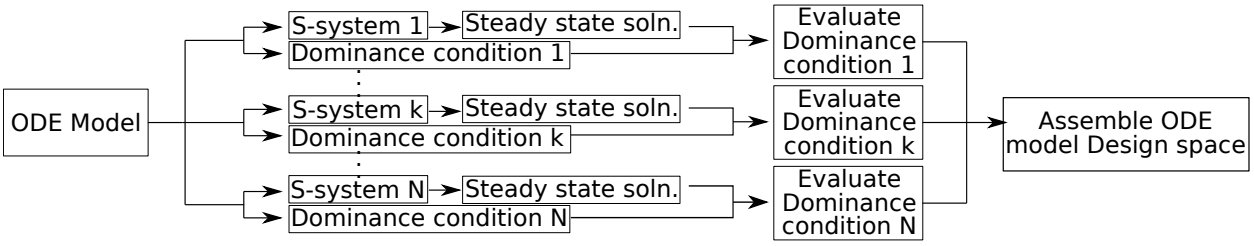

Figure 2.1: Flow chart for DSA. Given an ODE, we can obtain the design space by obtaining all S-systems, steady states, and evaluated dominance conditions.

	S-system 1	S-system 2	S-system 3	S-system 4
x'_s	$2\bar{p_0}x_{new}^{-1}x_s - x_s$	$2\bar{p_0}x_{new}^{-1}x_s - x_s$	$2\bar{p_0}x_{new}^{-1}x_s - x_s$	$2\bar{p_0}x_{new}^{-1}x_s - x_s$
x'_t	$2x_s - dx_t$	$2x_s - dx_t$	$2x_s - 2\bar{p_0}x_{new}^{-1}x_s$	$2x_s - 2\bar{p_0}x_{new}^{-1}x_s$
0	$1 - x_{new}$	$\gamma x_t - x_{new}$	$1 - x_{new}$	$\gamma x_t - x_{new}$
boundary 1	$dx_t > 2\bar{p_0}x_{new}^{-1}x_s$	$dx_t > 2\bar{p_0}x_{new}^{-1}x_s$	$dx_t < 2\bar{p_0}x_{new}^{-1}x_s$	$dx_t < 2\bar{p_0}x_{new}^{-1}x_s$
boundary 2	$1 > \gamma x_t$	$1 < \gamma x_t$	$1 > \gamma x_t$	$1 < \gamma x_t$

Table 2.1: S-systems from equations 2.12-2.14

We begin by rewriting the equations in the form shown in equations 2.1.

$$x'_{sc} = 2\bar{p_0}x_{new}^{-1}x_{sc} - x_{sc} \tag{2.12}$$

$$x'_{tc} = 2(1 - \bar{p_0}x_{new}^{-1})x_{sc} - dx_{tc} \tag{2.13}$$

$$0 = 1 + \gamma x_{tc} - x_{new}. \tag{2.14}$$

Note, we introduce a new variable $x_{new} = 1 + \gamma x_{tc}$, which is necessary to satisfy the form of equation 2.1.We next find all combinations in which one source term and one sink term dominates the system of differential equations.

Using the newly defined system (equations 2.12- 2.14), we obtain 4 S-System combinations (Shown in table 2.1).

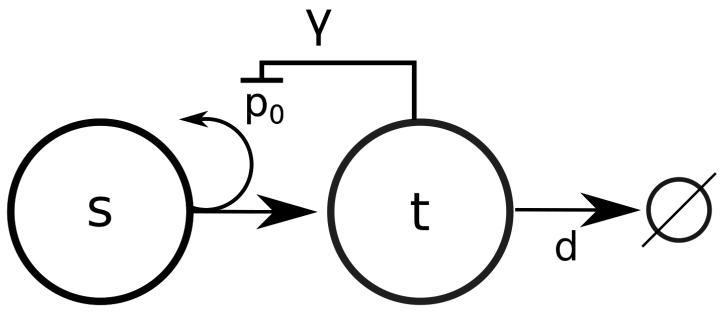

Figure 2.2: Lineage schematic.

We continue the analysis with S-system 2 from table 2.1. Using S-system 2, we set the time derivatives to zero and rearrange the equations such that we obtain $A\bar{x} = \bar{b}$, where

$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} \log(\bar{x_s}) \\ \log(\bar{x_t}) \\ \log(\bar{x_{new}}) \end{bmatrix} = \begin{bmatrix} \log(2\bar{p_0}) \\ \log(\frac{d}{2}) \\ \log(\gamma) \end{bmatrix}$$

such that $\bar{x_s}$ and $\bar{x_t}$ are the steady state solutions for the S-system and $\bar{x_{new}}$ is the solution of the newly defined variable at steady state. Solving the linear equation gives us:

$$\begin{bmatrix} \log(\bar{x_s}) \\ \log(\bar{x_t}) \\ \log(\bar{x_{new}}) \end{bmatrix} = \begin{bmatrix} \log(\frac{d\bar{p_0}}{\gamma}) \\ \log(\frac{2\bar{p_0}}{\gamma}) \\ \log(2\bar{p_0}) \end{bmatrix}$$

. We can now construct the boundaries for this S-System. We substitute our steady state solutions obtained above into the logged inequalities constraints. Thus, the inequalities in log space become

$$0 < \log(2\bar{p_0}) \tag{2.15}$$

Out of the 4 possible S-systems, only S-system 2 has a unique steady state with parameters that satisfy the constraints. We can plot the design space by varying $\bar{p_0}$ and γ. The design space shows one region where $\bar{p_0} > 0.5$, a requirement for a positive steady state in the full system. The domain in parameter space corresponds to S-system 2 in table 2.1.

It is possible to relate the S-system back to the true ODE system with classical techniques. It has been shown by [69, 48, 21] that if we are to perform a linear stability analysis on the S-system, the stability results hold for the full system under the bounded domain. Similarly,

we can do a parameter sensitivity analysis to gain insight on the full system.

Performing a linear stability analysis on the S-system that corresponds to the region in parameter space shown in figure 2.3, we obtain the following eigenvalues:

$$\lambda = [-0.5d - \frac{0.125\gamma\sqrt{(d(-64+16d)\bar{p_0}^4)/\gamma^2}}{\bar{p_0}^2}, \tag{2.16}$$

$$-0.5d + \frac{0.125\gamma\sqrt{(d(-64+16d)\bar{p_0}^4)/\gamma^2}}{\bar{p_0}^2}] \tag{2.17}$$

which suggests a stable spiral or stable node depending on the value of d. To validate, we

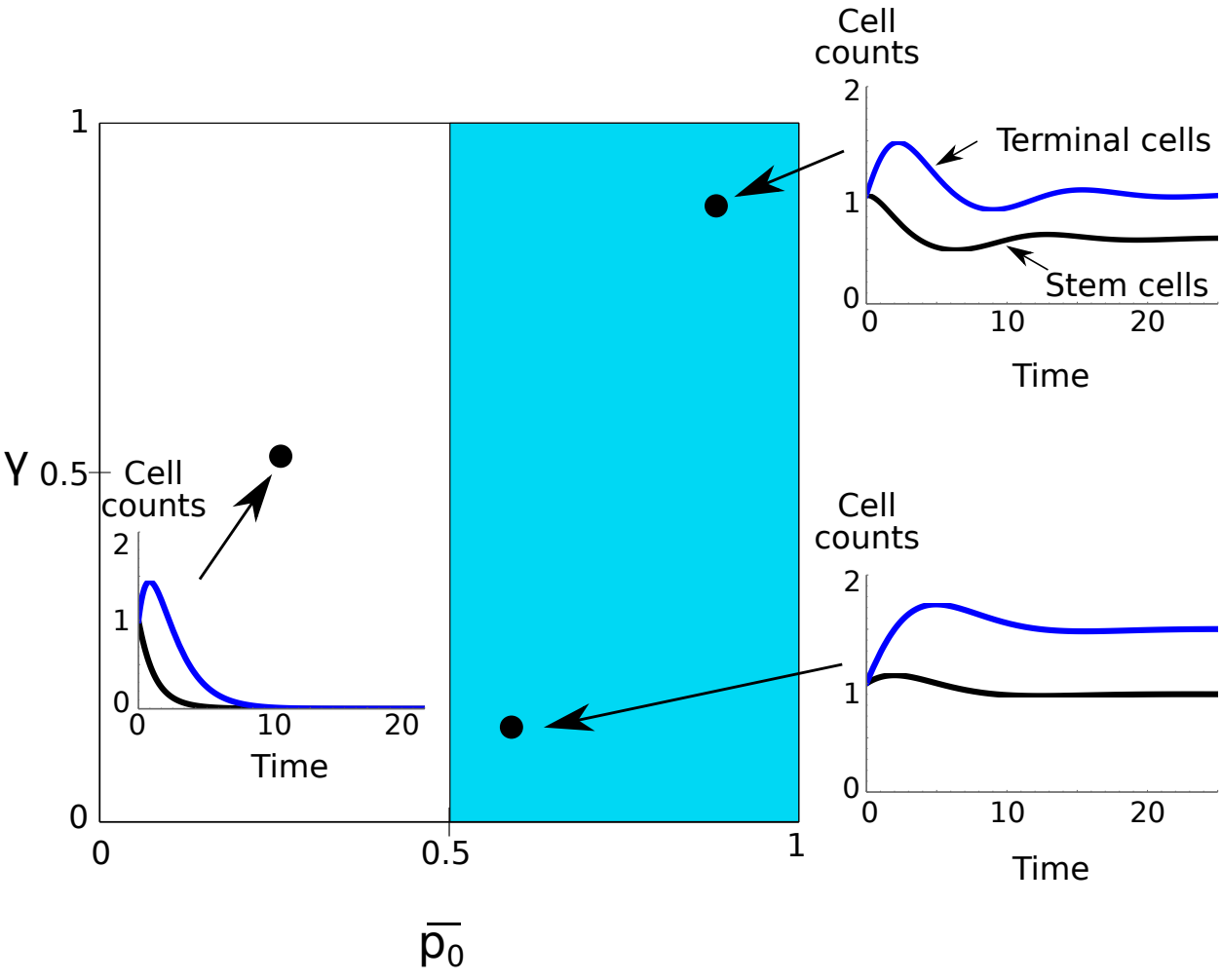

Figure 2.3: Design space for two cell lineage model. The design space is showing a slice of parameter space varying the max self renewal probability ($\bar{p_0}$) and feedback gain for stem cells (γ). If we are to sample parameters sets (shown by points), we observe oscillatory behavior in the full system, as shown by the time evolution plots.

do the analysis on the full system and obtain the following eigenvalues:

$$\lambda = [\frac{-0.25d\bar{p_0} - 0.125\sqrt{d\bar{p_0}^2(4d + (32 - 64p_0)\bar{p_0})}}{\bar{p_0}^2}, \tag{2.18}$$

$$\frac{-0.25dp_0 + 0.125\sqrt{d\bar{p_0}^2(4d + (32 - 64\bar{p_0})\bar{p_0})}}{\bar{p_0}^2}] \tag{2.19}$$

and under these parameter ranges, we also obtain the same behavior. Using the full system, we can also plot the dynamics in the domain. When plotting the full system in figure 2.3, we can observe the stem and progenitor populations oscillating before reaching steady state, characteristics of a stable spiral. The S-system and the true system eigenvalues are not identical, but under the same parameters in the bounded domain, the systems will have the same behavior. If we are outside the domain, we cannot conclude anything about the full systems dynamics. In fact, looking at figure 2.3, being outside of the domain corresponds to solutions going to the zero steady state due to $\bar{p_0} < 0.5$.

Chapter 3

Mathematical modeling of malignant myelopoiesis and targeted therapy

3.1 Introduction

3.1.1 Hematopoiesis

The hematopoietic system is a hierarchical model in which cells are organized to successfully create and maintain new blood cells. The system is a collection of progressively more differentiated cells starting from a hematopoietic stem cell (HSC) located in the bone marrow (BM) and ending with post-mitotic terminal myeloid and lymphoid cells located in the peripheral blood (PB). This robust system creates $10^{10} - 10^{11}$ blood cells a day, which consist of both myeloid and lymphoid type cells and accommodates massive increases in production of individual cell types in response to pathological stresses [50]. The control of production of red cells and platelets is reasonably well understood, with lineage-restricted erythroid and megakaryocytic progenitors selectively responsive to specific cytokines (erythropoietin and

thrombopoietin, respectively)[44, 73]. In contrast, the regulation of production of granulocytes (neutrophils, basophils, and eosinophils) in healthy and diseased systems is poorly understood. For example, two cytokines [granulocyte colony-stimulating factor (G-CSF) and granulocyte-macrophage colony-stimulating factor] can pharmacologically increase neutrophil production, but mice lacking both cytokines maintain baseline neutrophil levels and can still increase neutrophil production in response to infection [6]. Other feedback factors such as IL-17/23 production or granule proteins such as neutrophil elastase have been implicated in regulation of neutrophil production but do not completely explain the process [71, 29]. Factors such as IL-6 have been shown to upregulate stem cell division and to bias cell fates towards the myeloid lineage [51].

3.1.2 Chronic myeloid leukemia

Chronic Myeloid Leukemia (CML) is a blood cancer that causes an overproduction of granulocyte cells. The origin of the disease is derived from a translocation between chromosome 9 and 22, which creates a fusion gene BCR-ABL1 in what is referred to as the Philadelphia (Ph) chromosome. This translocation occurs somatically in a hematopoietic stem cell (HSC), with the Ph chromosome found not just in the malignant neutrophils but also in cells of monocytic, erythroid, megakaryocytic, and lymphoid lineage. The product of the BCR-ABL1 fusion gene is a dysregulated (e.g., constitutively upregulated) cytoplasmic protein-tyrosine kinase, BCR-ABL1, which promotes cell division and differentiation. CML thus represents a natural model of dysregulated granulocytopoiesis [61].

The disease is considered to have three stages: chronic, accelerated, and blast crisis [15]. Chronic CML begins with the Ph chromosome translocation, believed to originate in hematopoietic stem cells (HSC). Transformed HSCs, known as leukemic stem cells do not significantly increase their numbers in the chronic phase of CML; indeed, in newly diagnosed CML pa-

tients most HSC and their surrogate long term culture-initiating cells (LTC-IC) are non-malignant (Ph–) [40, 1, 59]. The action of BCR-ABL1 in promoting dysregulated granulopoiesis is instead manifested mainly at the level of committed progenitors, which drive the over-production of granulocytes. In the accelerated and blast crisis phases of the disease the progenitor compartments (common myeloid progenitors (CMP), and the granulocyte-macrophage progenitors (GMP)) expand significantly [40]. Although recent studies in mice suggest that normal hematopoiesis is regulated primarily through multipotential progenitors (MPPs), studies of MPP frequency in human CML have not been reported [1].

CML has one of the simplest of cancer genomes [38]. In early chronic phase, the sole abnormality detected is the BCR-ABL1 gene fusion, with very few copy number or somatic nucleotide variants. Progression to accelerated phase and blast crisis is characterized by acquisition of additional clonal karyotypic abnormalities and of a limited number of somatic mutations at the genomic level [20].

3.1.3 CML therapy and unmet clinical needs

In the past, a combination of interferon-alpha treatment with a stem cell transplant was the main form of therapy [33]. As of now, the first line of treatment for chronic phase CML has been imatinib mesylate, a tyrosine kinase inhibitor (TKI). Imatinib works by inhibiting the phosphorylation of the tyrosine kinase in the fusion gene BCR-ABL1 to stop the signaling cascade that promotes leukemogenesis [33]. Recently, second and third generation TKIs (dasatinib and nilotinib) have been approved for initial treatment. Along with these first line of treatment TKIs, there are TKIs that were created to treat CML at later stages (accelerated/blast crisis) and to target BCR-ABL point mutations (Ponatinib) [39].

The response to the TKI is measured by the BCR-ABL1 transcripts in the peripheral blood using qPCR. The TKI response curves give rise to a biphasic behavior. Refs. ([56, 74]) have

speculated through mathematical modeling that the first initial steep slope arises from a large Ph+ progenitor death due to TKI therapy, followed by a less steep slope representing the leukemic stem cell death.

Although treatment of CML through TKIs has lowered the death rate of this disease, there are still unmet clinical needs. Currently, TKI therapy is intended to be life-long. A recent study [52] showed that even though patients on TKI therapy exhibit complete molecular remission (CMR; defined as undetectable BCR-ABL1 transcripts), a significant fraction of them (60%) experience molecular relapse of CML within 6 months of discontinuation, indicating that TKI monotherapy is not curative in most patients [52, 66]. Cell biological studies suggest that the most primitive leukemic stem cell compartment in CML is not eliminated by TKI treatment despite effective inhibition of BCR-ABL1 kinase activity [28, 7, 18]. Rather, exposure to TKIs drives leukemic stem cells into quiescence where they are resistant to apoptosis, whereas treatment with G-CSF can partially reverse quiescence and TKI resistance [42]. In the clinic, there is a need to predict which patients will sustain CMR upon TKI discontinuation, and for strategies to increase those who maintain treatment-free remission (TFR) and might be cured of their disease [2].

Resistance is another issue when treating CML. The resistance can come in two ways: primary or acquired. In acquired resistance, the therapy is initially effective but eventually CML progresses. A major mechanism in acquired resistance to TKIs is point mutations in the BCR-ABL1 kinase domain. The point mutation decreases the sensitivity to the drug and can be managed to some extent by using additional second or third generation TKIs. In primary resistance, there is no initial response to TKI therapy. Primary resistance affects 10-15% of the newly diagnosed CML patients [81, 55]. Clinical data indicates that switching TKIs may not benefit these patients, suggesting that this group is destined to do poorly regardless of the specific inhibitor used. It is unclear what is driving this primary resistance.

3.1.4 Current mathematical models of hematopoiesis and chronic myeloid leukemia

While mathematical modeling of leukemia has a long history, most modelers have favored highly simplified lineage architectures and minimal feedback [22, 16]. In particular, CML has been modeled using unbranched lineages with 3-4 subpopulations representing stem cells, progenitor cells, and terminally differentiated post-mitotic cells with constant model parameters and logistic-like terms [56, 74, 72, 78, 35, 64]. The models collapse the lymphoid and myeloid lineages into a single population. The models are then fit to therapy response data to suggest how the subpopulations are affected by therapy. This modeling fails to suggest mechanisms that underlie primary resistance or therapeutic options to increase the frequency of TFR.

By using constant model parameters, current models do not incorporate feedback regulation of the system dynamics. There is convincing evidence now that like normal hematopoiesis, feedback regulation also plays a role in CML hematopoiesis and that the normal and Ph+ cells can influence each others behavior [60, 77, 62, 63]. For example, IL-6 expressed in CML cells in mice has been shown to influence leukemic exposed healthy cell by altering progenitor dynamics and differentiation bias towards myeloid cells [60].

3.2 Methods

3.2.1 Mathematical model of hematopoiesis

The classical depiction of hematopoiesis is a hierarchy of cell types starting with the hematopoietic stem cell at the top, followed by progenitors and ultimately ending with mature cells located in the peripheral blood. Therefore, we model hematopoiesis using a lineage model

[46, 10] to describe cellular growth dynamics. The modeling allows us to follow the similar hierarchical structure by creating an order of differentiation. Our branched lineage model of hematopoiesis is simplified and only models hematopoietic stem cells (HSCs), progenitor cells (MPPs), and terminally differentiated cells (Myeloid and Lymphoid cells). The model can easily include additional cell types, such as intermediate progenitor cell types, which will provide an additional level of detail. The model of interest consists of two dividing cell types consisting of S (HSC) and P (MPP) cells with a division rate associated to the cells (η_1 and η_2 respectively). The S cells have the ability to self renew with probability (p_0) or differentiate ($1 - p_0$). The P cells have the ability to self renew with probability (p_1) or differentiate into either TD_l (lymphoid) or TD_m (myeloid) cells (q_1 or $1 - p_1 - q_1$ respectively). The terminal cells form the majority of the hematopoietic system and consist of TD_l and TD_m cells. TD_m and TD_l cells only have the ability to die at rates d_m and d_l, respectively. The lineage schematic is shown in figure 3.2. The equations below (Eq. 3.1-3.4) describe the dynamics of the system:

$$x'_S = (2p_0 - 1)\eta_1 x_S \tag{3.1}$$

$$x'_P = 2(1 - p_0)\eta_1 x_S + (2p_1 - 1)\eta_2 x_P \tag{3.2}$$

$$x'_{TD_L} = 2q_1\eta_2 x_P - d_l x_{TD_l} \tag{3.3}$$

$$x'_{TD_M} = 2(1 - p_1 - q_1)\eta_2 x_P - d_m x_{TD_m} \tag{3.4}$$

3.2.2 Mathematical model of CML

We extend our model of hematopoiesis to CML by assuming a parallel lineage with the same feedback structure. There is evidence to believe a strong interaction between normal and CML cell types [77]. Thus, we assume the feedback is dependent on both the normal and CML cell types. In doing so, the system of 4 ODEs becomes 8 ODEs, where equations 3.5-3.8

correspond to the normal/healthy hematopoietic lineage. The normal lineage is denoted as x_i such that $i = S, P, TD_l, \text{and } TD_m$.

$$x'_S = (2p_0 - 1)\eta_1 x_S \tag{3.5}$$

$$x'_P = 2(1 - p_0)\eta_1 x_S + (2p_1 - 1)\eta_2 x_P \tag{3.6}$$

$$x_{TD_l}{}' = 2q_1\eta_2 x_P - d_l x_{TD_l} \tag{3.7}$$

$$x_{TD_m}{}' = 2(1 - p_1 - q_1)\eta_2 x_P - d_m x_{TD_m} \tag{3.8}$$

The leukemic lineage is described in equations 3.9-3.12 and is denoted as x_i^L such that $i = S, P, TD_l, \text{and } TD_m$.

$$x_S^{L\prime} = (2\tilde{p_0} - 1)\tilde{\eta_1} x_S^L \tag{3.9}$$

$$x_P^{L\prime} = 2(1 - \tilde{p_0})\tilde{\eta_1} x_S^L + (2\tilde{p_1} - 1)\tilde{\eta_2} x_P^L \tag{3.10}$$

$$x_{TD_l}^L{}' = 2\tilde{q_1}\tilde{\eta_2} x_P^L - d_l x_{TD_l}^L \tag{3.11}$$

$$x_{TD_m}^L{}' = 2(1 - \tilde{p_1} - \tilde{q_1})\tilde{\eta_2} x_P^L - d_m x_{TD_m^L}. \tag{3.12}$$

where feedback/feedforward for the normal lineage is defined as

$$p_0 = \frac{\bar{p_0}}{1 + \gamma_1(x_P + x_P^L)} \tag{3.13}$$

$$p_1 = \frac{\bar{p_1}}{1 + \gamma_3(x_{TD_m} + x_{TD_m}^L)} \tag{3.14}$$

$$q_1 = \frac{\bar{q_1}}{1 + \gamma_4(x_{TD_m} + x_{TD_m}^L)} \tag{3.15}$$

$$\eta_1 = \frac{\bar{\eta_1}}{1 + \gamma_2(x_S + x_S^L)} \tag{3.16}$$

$$\eta_2 = \frac{\bar{\eta_2}}{1 + \gamma_5(x_S + x_S^L)} \tag{3.17}$$

and the feedback/ feedforward for the leukemic lineage is defined as

$$\tilde{p_0} = \frac{\bar{p_0}}{1 + \mu_1\gamma_1(x_P + x_P^L)} \tag{3.18}$$

$$\tilde{p_1} = \frac{\bar{p_1}}{1 + \mu_3\gamma_3(x_{TD_m} + x_{TD_m}^L)} \tag{3.19}$$

$$\tilde{q_1} = \frac{\bar{q_1}}{1 + \mu_4\gamma_4(x_{TD_m} + x_{TD_m}^L)} \tag{3.20}$$

$$\tilde{\eta_1} = \frac{\bar{\eta_1}}{1 + \mu_2\gamma_2(x_S + x_S^L)} \tag{3.21}$$

$$\tilde{\eta_2} = \frac{\bar{\eta_2}}{1 + \mu_5\gamma_5(x_S + x_S^L)}. \tag{3.22}$$

We minimize the amount of differences between the normal and leukemic lineages. Thus, the only difference lies in the feedback/feedforward form, allowing CML cells to be less sensitive to feedback/feedforward which gives rise to a competitive advantage. This is modeled by introducing a μ parameter for each feedback/feedforward as shown in equations 3.18 - 3.22. The μ parameters are constrained between 0 and 1 and represent a fractional change in the feedback/feedforward strength. We are aware there are many ways to incorporate cross talk between the normal and leukemic populations, but for simplicity we assume this form.

Modeling CML development and TKI therapy

We use equations 3.9-3.22 to model CML development using the following approach:(1) we model normal hematopoiesis in equilibrium, (2) introduce 10^3 CML stem cells, (3) allow the full system to evolve over a set amount of time (t_{TKI}), and (4) model TKI therapy by introducing a decay term for dividing cells in the leukemic lineage proportional to the

division rate (see equations 3.5-3.5 and 3.23-3.26), as TKI therapy targets dividing cells.

$$x_S^{L\prime} = (2p_0^* - 1)\eta_1^* x_S^L - \delta_1\eta_1 x_S^L \tag{3.23}$$

$$x_P^{L\prime} = 2(1 - p_0^*)\eta_1^* x_S^L + (2p_1^* - 1)\eta_2^* x_P^L - \delta_2\eta_2 x_P^L \tag{3.24}$$

$$x_{TD_l}^{L\prime} = 2q_1^*\eta_2^* x_P^L - d_l x_{TD_l}^L \tag{3.25}$$

$$x_{TD_m}^{L\prime} = 2(1 - p_1^* - q_1^*)\eta_2^* x_P^L - d_m x_{TD_m^L} \tag{3.26}$$

with the feedback/feed-forward unchanged. See Eqs. 3.13-3.22. A schematic of the modeling is shown in figure 3.1.

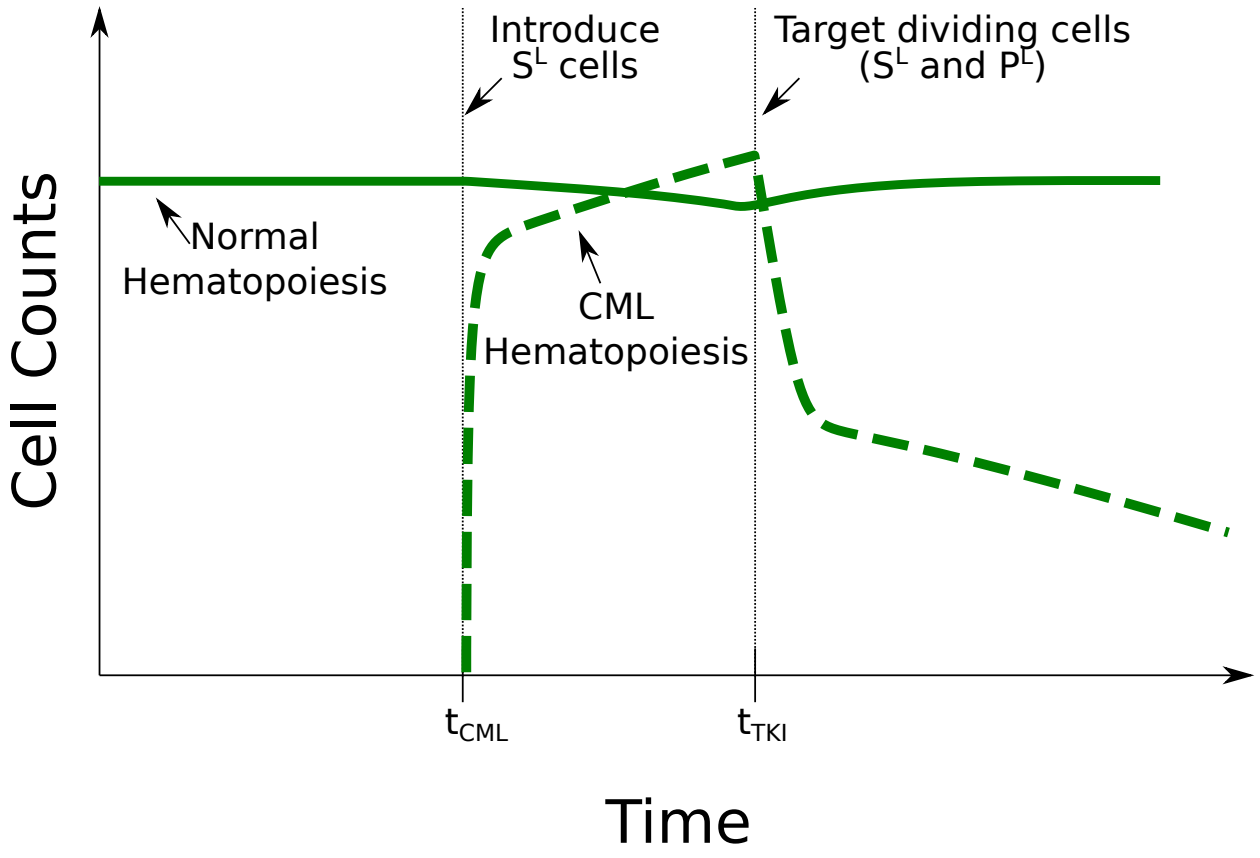

Figure 3.1: Schematic of CML evolution and TKI therapy.

Obtaining BCR-ABL/BCR reads

In practice, the response to the TKI is measured by the BCR-ABL1 transcripts using qPCR. To model this, we can use the terminal cells ($x_{TD_l}, x_{TD_m}, x^L_{TD_l}, x^L_{TD_m}$) in equations 3.5-3.5 and 3.23-3.26 to obtain a similar ratio. The response (BCR-ABL1/BCR) considers normal and leukemic terminal cells as the following:

$$\frac{x^L_{TD_l} + x^L_{TD_m}}{2(x_{TD_l} + x_{TD_m}) + x^L_{TD_l} + x^L_{TD_m}} \tag{3.27}$$

3.2.3 Design space analysis

We use an automated method developed by Savageau [21, 48, 49, 69] that separates models by distinct qualitative behaviors at steady state. The strategy is to deconstruct the model of interest at steady state to focus on cases where one production term and one loss term dominate, which gives a dominant subsystem (S-System). The method then imposes inequalities to satisfy the S-System dynamics. The inequalities are evaluated at the S-System's steady state to assess self-consistency. If the inequalities are satisfied, the system is self-consistent and the regions where equality holds form boundaries that pertain to a particular qualitative behavior associated with the system. The interior region (where strict inequality holds) is termed a domain in design space. The benefits of this method are that it does not require prior knowledge about parameter values and it can enumerate the different types of qualitative dynamics a certain system may have. By eliminating subsets of parameters for which the equilibrium is unstable, this approach will automatically select models that are robust to parameter variation due to stability.

3.3 Results

3.3.1 Model selection

Experimental evidence suggests there is feedback regulation between cell types in normal hematopoiesis through secreted signals, e.g. cytokines [43, 60, 63]. Although a great deal has been learned especially in recent years about the structure of hematopoietic lineages and evidence points to the presence of feedback control at many stages [68, 45], our knowledge of the details of feedback regulation in hematopoiesis is still relatively scant, especially for granulopoiesis, where even late-stage feedback interactions are poorly understood. In many cases we don't know which cell types are providing and receiving the feedback, what signals are used, whether they are humoral or local, and what aspects of proliferative cell behavior they influence (i.e. cell cycle speed, renewal probability, or progeny fate choice). We can incorporate feedback control by having the model parameters potentially subject to positive or negative feedback/feedforward from any cell type. Then, we can distinguish among the models accounting for (1) systems that follow fundamental principles such as stability and robustness, (2) existing animal model data on normal hematopoiesis, and (3) molecular data . Using this approach, we believe we can make more physiologically accurate models of CML with substantial predictive power.

We begin by applying this analysis to our proposed model described in equations 3.1-3.4. We assume that the self-renewal probabilities (p_0, p_1), fate switching probability (q_1), and division rates (η_1, η_2) are potentially subject to feedback regulation, which can be either positive or negative and can arise from any of the cells, shown in figure 3.2A. The different lines correspond to potential positive or negative feedback or feedforward from any cell types. In proposing this, we get to 5^9 or 59,049 different candidate feedback models.

In constructing our data driven model, we first sort through all feedback models that have

stable equilibrium. Using design space analysis allows us to sort for stable models and group model with similar qualitative behaviors. We applied the method [69] on all of the candidate feedback models and eliminate models that cannot satisfy the constraints that arise from the analysis. Accordingly, we arrive to the 4 model classes that are shown in Figure 3.2B. The model classes each have qualitatively different behaviors. The models within the classes all

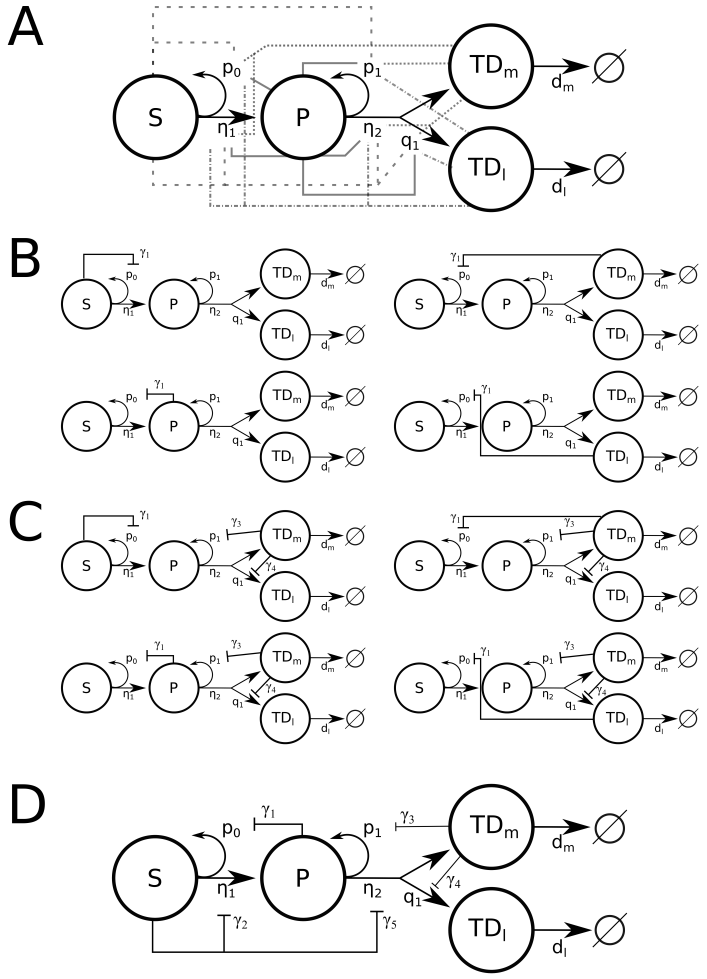

Figure 3.2: Model reduction based on design space analysis and biological prior knowledge. A. Branched lineage consisting of stem (S), progenitor (P), and terminal cells (TD_m: Myeloid, TD_l: Lymphoid). Feedback can affect self renewal probabilities (p_0 and p_1), division rates (η_1 and η_2), and fate switching probability (q_1) from any cell types. The different line styles correspond to a particular cell type. B. Using design space analysis, we arrive to 4 model families. Each model family is grouped by having at least one S-System in common. C. Using biological prior knowledge, we reduce model space by hypothesizing IL-6 directs the fate for progenitors. We incorporate this by adding fate control [10]. D. Proposed model of hematopoiesis based on design space analysis and incorporating biological prior knowledge.

share common qualitative behaviors and their differences lie in whether or not the model has feedback/feedforward on the rest of the parameters, where the regulation can be positive, negative, or no regulation. Interestingly, the model classes only differ by where the stem cell self renewal probability receives its feedback from. In searching for stability, our model space reduces to 26,244 candidate models.

To narrow down the possible candidate models, we incorporate biological prior knowledge. There is evidence showing cytokines, such as IL-6, have the ability to bias MPP differentiation towards a myeloid state [51, 77]. Thus, we argue IL-6 is a candidate feedback factor expressed in our myeloid compartment (TD_m) and has the ability to negatively regulate MPP self renewal (p_1) and fate switching probability(q_1). This negative feedback circuit, known as fate control [10] has been shown to provide an effective strategy for dealing with perturbations in branching models which is a feature we look for in our model selection process. Fate control in this instance is shown to bias differentiation to the myeloid compartment at the expense of progenitor self renewal and lymphoid differentiation. In hypothesizing fate control, we further reduce our candidate model space to 324 models, as shown in figure 3.2C. The remaining models differ in positive, negative, or no feedback from any cell types on either stem or progenitor division rates.

Finally, we can use scRNA-seq data to identify putative feedback loops and sender and receiver cells. We re-analyzed data from a scRNA-seq study of normal hematopoiesis that identified many interesting cell clusters whose transcriptomes suggested pairwise combinations of cells expressing feedback ligands and their receptors [58]. Although the study did not cleanly separate out different kinds of early stem/progenitor cells, the early stem/progenitors did cluster into two groups, perhaps representing HSCs and MPPs. In these two groups, we were able to recognize several ligands and receptors (such as ANGPT1 and CCL3 and their receptors), although we could not be certain about the sender and receiver cell types. Following references [3, 58], we hypothesized that ANGPT1 is secreted by HSCs and negatively

regulates MPP division rates (η_2) and that CCL3 is produced by MPPs and negatively regulates HSC self-renewal (p_0). In proposing the additional feedback/feedforward, we reduce our model space to 9 models, where the differences between the remaining candidate models lie in whether or not the HSC division rate has positive, negative, or no feedback from any cell type.

3.3.2 Experimental validation

Feedforward

To test the hypothesized negative feedforward on MPP division (η_2) suggested by scRNA-seq data, we perturbed the hematopoietic system. The experiment consisted of irradiation of healthy mice with 50 cGy to induce HSC loss. The data is obtained using flow cytometry with the gating strategy shown in figure 3.3A. Data was collected on day 1, 3, and 7 and figure 3.3B shows the comparison of control vs irradiated mice for both the HSC and MPP compartments measured by %LSK (Lin-Ska1+cKit-), where LSK is used to differentiate stem and progenitors from the rest of the hematopoietic cells in the bone marrow. At day one, the 50 cGy mice HSCs reduce in fraction and ultimately return to equilibrium values by day 7. As for the MPPs, there is no significant change between the control and 50 cGy mice. To validate the existence of the negative feedforward loop on MPP division rate, we are able to measure the proliferation percentage of both HSCs and MPP. We observe a 3 fold increase in the percent of HSCs and MPPs which suggest in addition to the negative feedforward on MPP division rate (η_2), there may also be a negative regulation of HSCs onto their own division rate (η_1) (See figure 3.3C). Thus, we arrive to our final feedback model shown in figure 3.2D.

Feedback

To further validate our model, we designed myeloid depletion experiments to observe responses in the HSC and MPP compartments. In perturbing the myeloid compartment, we are disrupting the feedback loop that affects MPP self renewal probability.

The depletion experiment uses an RB68C5 antibody at different concentrations to reduce the number of myeloid cells (See figure 3.4A, C). The analysis is 24hrs post treatment. To measure the response, we use % LSK as shown in figure 3.4C. A reduction of myeloid cells shows an increase in the MPP compartment and a decrease in the HSC compartment, showing consistent dynamics for both the negative feedback on MPP self renewal (p_1) and HSC self renewal (p_0). Although we do observe consistent dynamics, there is additional work that needs to be done to validate the feedbacks.

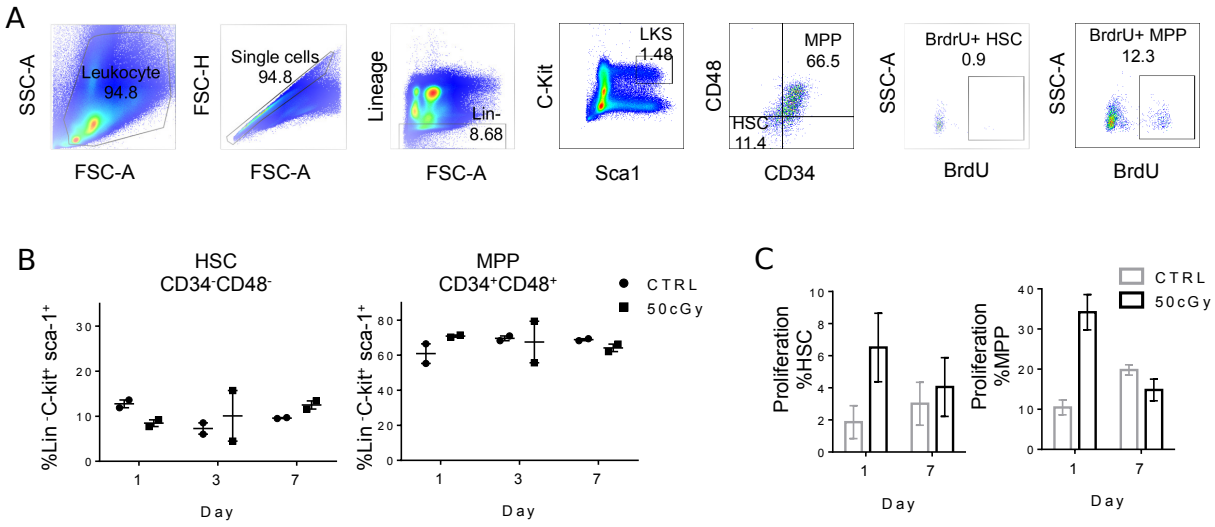

Figure 3.3: Flow cytometric analysis of mouse bone marrow cells for frequency and proliferation of Hematopoietic Stem Cells (HSC) and Multi Potential Progenitors (MPP) A. Gating Schema for phenotyping Hematopoietic Stem Cells (HSC) and Multi Potential Progenitors (MPP), and BrdU incorporation in respective compartments B. HSC and MPP compartment size on days 1, 3, and 7 following the 50cGy radiation in wild type control mice (circles) and mice that were irradiated (squares) C. Frequency of HSC and MPP proliferation in CTRL (gray bars) and irradiated (black bars) mice measured by BrdU incorpation.

3.3.3 Design space analysis on proposed model

Using design space analysis, we can analyze our model of interest (figure 3.2D) and explore parameter space. We identify 6 regions/domains and take a slice of parameter space and vary HSC and MPP self renewal feedback gains (γ_1 and γ_3 respectively). If we are to sample parameters from each domain, shown as a different shade in figure 3.5, we can observe general qualitative dynamics in the different regions. Under these parameter constraints,

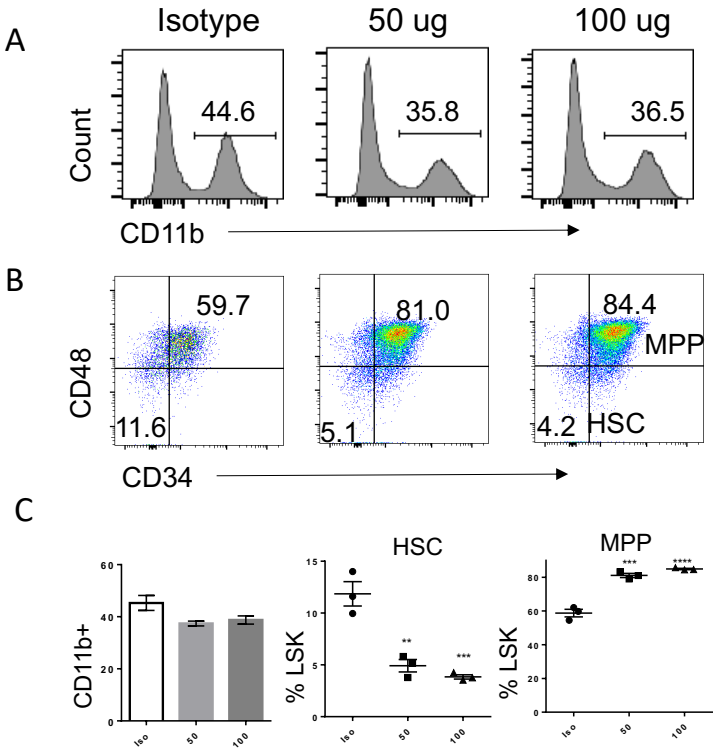

Figure 3.4: Flow cytometric analysis of HSC and MPP compartments 24hrs following myeloid cell depletion using RB68C5 antibody. A. Representative histograms depicting frequency of myeloid cells as measured by CD11b expression in mice treated with Isotype control, 50ug, and 100ug of RB68C5 antibody by i.p administration B. Frequency of Hematopoietic Stem Cells (HSC) and Multi Potential Progenitors (MPP) as percent of Lin-c-kit+sca1+ cells are displayed in representative scatter plots C. Bar graph shows frequency of CD11b+ cells in mice that were dosed with isotype control (white bar), 50ug (light gray bar) , and 100ug (dark gray bar) RB68C5 antibody respectively. HSC and MPP frequency from mice that received isotype (circles), 50ug (square), and 100ug (triangles) of RB6 antibody. Data are mean $\pm$ SEM. HSC, ANOVA $P < 0.005$, MPP, ANOVA $P < 0.0001$

all domains are stable. If we perform a linear stability analysis on the S-systems associated with the domains, the top 3 domains (i,iii,v) are stable nodes. The bottom 3 domains (ii, iv, vi) are stable spirals. We can also observe the boundaries have specific steady state ranges when looking at the time evolution plots in figure 3.5i-vi. As we decrease γ_1, we increase the steady state values. The corresponding S-systems are shown in table A.1 and equations A.1-A.4.

3.3.4 CML extension

To model CML development in the presence of normal hematopoietic cells, we assume the only difference between the two lineages is the feedback strength for leukemic HSC, as shown by the addition of $\tilde{p_0}$ in the schematic in figure 3.7A. This small difference gives the leukemic

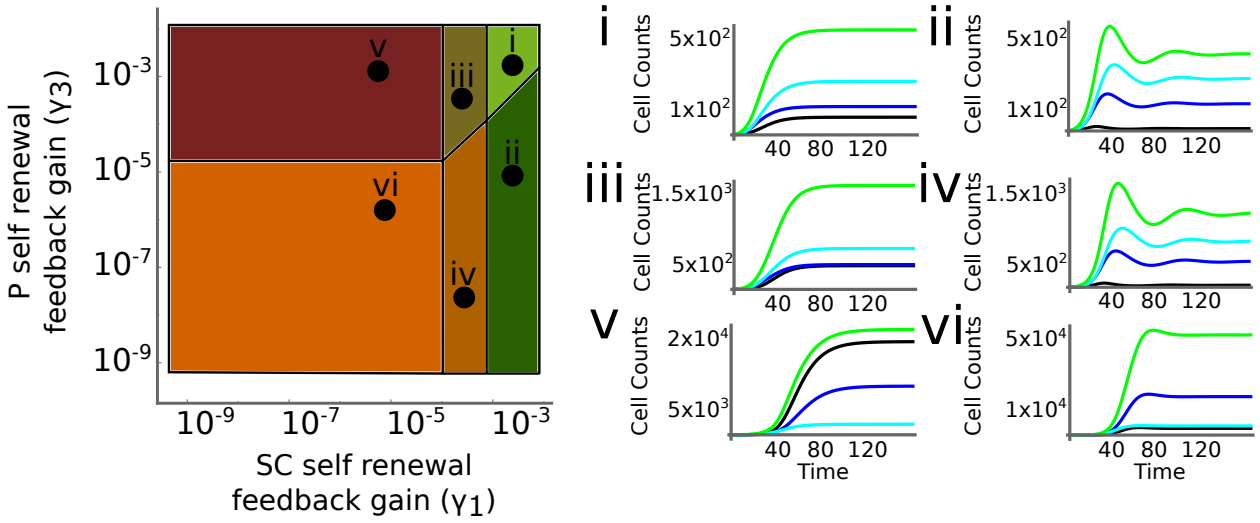

Figure 3.5: Design space analysis applied to proposed model. We apply the analysis to our proposed model from 3.2D. Left: We focus on a slice of parameter space and analyze self renewal gain for stem cell vs self renewal gain for progenitors. Our analysis leads to 6 bounded domains, where we sample their dynamics, denoted by the roman numeral. Right: The 6 plots correspond to the sampled parameter values from the parameter space to the left. Using linear stability analysis, we can characterize the boundaries by their stability. From the analysis, we observe the top 3 designs (i, iii, v) all share the same qualitative behaviors. Similarly, the bottom 3 designs (ii, iv, vi) all share similar qualitative behaviors. Parameters used: p_0=0.6, p_1 = 0.48, q_1 = 0.1, d_m=0.6, d_l=0.2, γ_2=1e-4, γ_4=1e-4, γ_5=1e-4, η_1=1.0, η_2=2

lineage a competitive advantage for growth. Our assumption is based on the ability for leukemic HSCs having the ability to initiate CML [63].

To obtain biologically realistic parameters, our first goal is to obtain equilibrium values. We do an extensive parameter sweep and find equilibriums of cell types consistent with the literature [53]. We also aim to obtain behaviors associated to CML experiments. Specifically, we focus on a perturbation in which HSCs and MPPs were transplanted into CML mice and readouts of myeloid and lymphoid percentages were obtained from the transplanted cells [63]. Reynaud et al. [63] showed when HSCs were transplanted, the transplanted myeloid percentage after 30 days was much higher than the lymphoid. Moreover, if MPPs were transplanted, the opposite occurred and instead the transplanted lymphoid percentage dominated. We were able to recapitulate these dynamics and restrict parameter space using design space analysis. We also gained insight on the reason for this response. Looking at the cell populations in figure 3.6A, transplanting HSCs allows engraftment and thus we get a stable populations of transplanted cells where the fraction of myeloid cells being produced is higher than lymphoid. On the other hand, transplanting MPPs does not allow for engraftment and leads to eventual extinction of the transplanted population. Thus, the transplanted lymphoid cell percentage after 30 days is shown to be higher due to the death rate of myeloid cells being greater than the lymphoid cells (See figure 3.6A-B).

Leukemic stem cell load highly influences TKI Therapy

Using our proposed feedback model, we explore the effects of TKI therapy on CML development (See figure 3.1 and section *Modeling CML development and TKI therapy* on how CML and therapy is modeled). Specifically, we want to look at how applying therapy at different times can effect the response dynamics. To implement, we vary t_{TKI}, a parameter responsible for how much we allow CML to develop. Varying the parameter t_{TKI} is analogous to changing the leukemic load fraction in an individual run.

We apply therapy at three different times associated to three different leukemic loads. The three different t_{TKI} times are defined as early treatment (6 months), mid treatment (1.5 years), and late treatment (3 years). When therapy is applied at 6 months and 1.5 years (Figure 3.7D,E), a subtle drop in the normal cells is observed due to growth in CML cells, but ultimately the normal cells recover back to equilibrium causing the CML cells to decrement. The biphasic response is observe in figure 3.7C for both early (Black, Solid) and mid (Black, Dashed) treatments. Applying therapy at 3 years gives a much different response (Figure 3.7F). The late treatment leads to a depletion of normal cells and TKI therapy loses its effectiveness to leukemic stem cells (See figure 3.7C, Black Dash Dot). If we look at the response curve, zero response is observed, a behavior observed in patients with primary resistance. We can analyze the feedback interactions to better understand the observed behavior. Allowing the leukemic stem cell compartment to grow decreases the division rate of the HSC and MPP compartments and increases the self renewal probability (Figure 3.8). Under these dynamics, the HSC compartment decreases the flux which results in TKI therapy being less effective in targeting HSCs and MPPs as it only targets dividing cells. This mechanism suggests why some patients are destined to do poorly in TKI therapy.

3.3.5 Model comparisons

Using our proposed model defined in figure 3.2D, we compare our results to an existing model of CML [56], as shown in the red inset lineage model in figure 3.7C. We will refer to the inset model as Model A. We first generated a synthetic response curve from our proposed feedback model (solid black) to which we fit Model A (solid red), as shown in Figure 3.7C. While there is general agreement between the two models, we observe a clear difference when we use our feedback model to generate therapy curves at different stages along the progression of CML (i.e. when the stem cell compartment has been overrun by mutant stem cells to different degrees), we observe very different behaviors (compare upper

and lower black curves in figure 3.7C, representing late (Dot Dashed line) versus early (Solid line) therapy). With Model A, therapeutic intervention at a later stage of disease does not change the response. Clearly, Models A and our model make very different predictions about the outcomes of treatment. In particular, under our model, some patients are expected to fail to respond well to TKI therapy, which is indeed what is observed clinically [32, 81]. Moreover, our model predicts that those patients who do respond poorly are more likely to be those with a higher proportion of CML stem cells in the BM.

3.4 Discussion

In this work, we develop an automated method for model selection that integrates biological observations to select plausible classes of feedback models. We first apply this approach to normal hematopoiesis and identify models that have desired system properties. When extended to incorporate CML hematopoiesis, our initial assessment shows that feedback/branching models are more robust and have a better fit to alternative patterns of patient response than simple linear models.

Applying design space analysis allows us to systematically search for regions of stability in thousands of proposed models in an efficient way and remove model that have unstable equilibrium without specifying parameter values. The analysis is used to group specific models that have similar quantitative behavior to obtain minimal models. In addition, using biological prior knowledge to narrow down the amount of candidate models gives us the ability to hypothesize how cells might interact. In doing so, we can test specific hypotheses and observe if behaviors can be predicted. In our two validation experiments involving radiation and myeloid depletion, we were able to predict the behavior with our proposed model. Although the two experiments did validate our model, this method is an iterative process to maximize the predictability of true hematopoietic development and maintenance.

Extending our data driven model to CML allows us to model development of CML hematopoiesis in the presence of normal hematopoiesis. In doing so, we speculate a driving force in primary resistance is the leukemic HSC compartment. The time dynamics show as the leukemic stem load is increased, TKI therapy is shown to be less effective. This can be explained by the effective feedbacks/feedforwards. If the leukemic stem load is higher, the effective stem cell and progenitor division decreases and the stem cell self renewal increases. Knowing this, we can hypothesize exhausting the leukemic stem cell load can make the TKI more effective. There have been many efforts for combination therapies, but understanding the mechanism for resistance allows us to suggest plausible targets to increase the efficacy of TKIs.

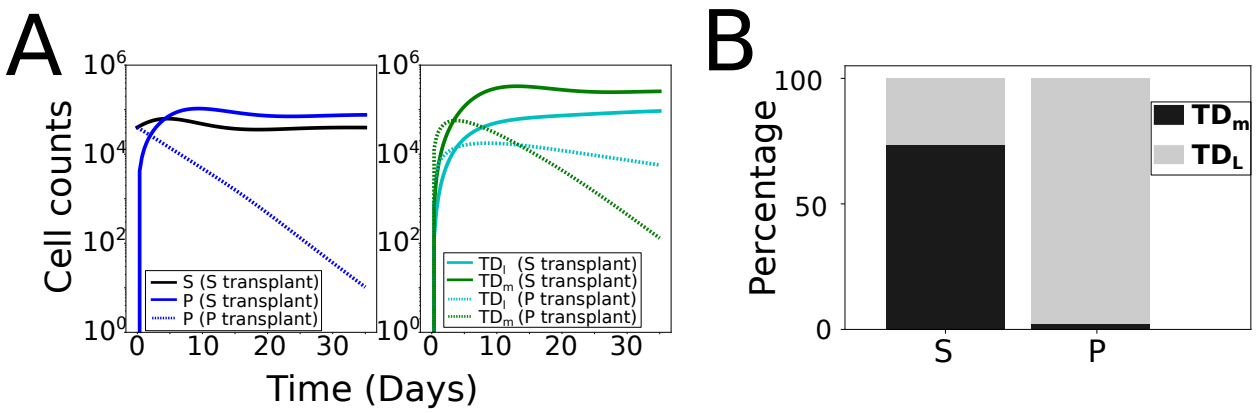

Figure 3.6: Qualitative dynamics recovered from transplant experiment administered by [63]. A. Time evolution of S, P, TD_l, and TD_m cells when 4000 S (Solid) or 4000 P (Dashed) cells are transplanted . B. Bar chart representing the percentage of myeloid and lymphoid cells after 30 days when stem (S) or progenitor (P) cells are transplanted.

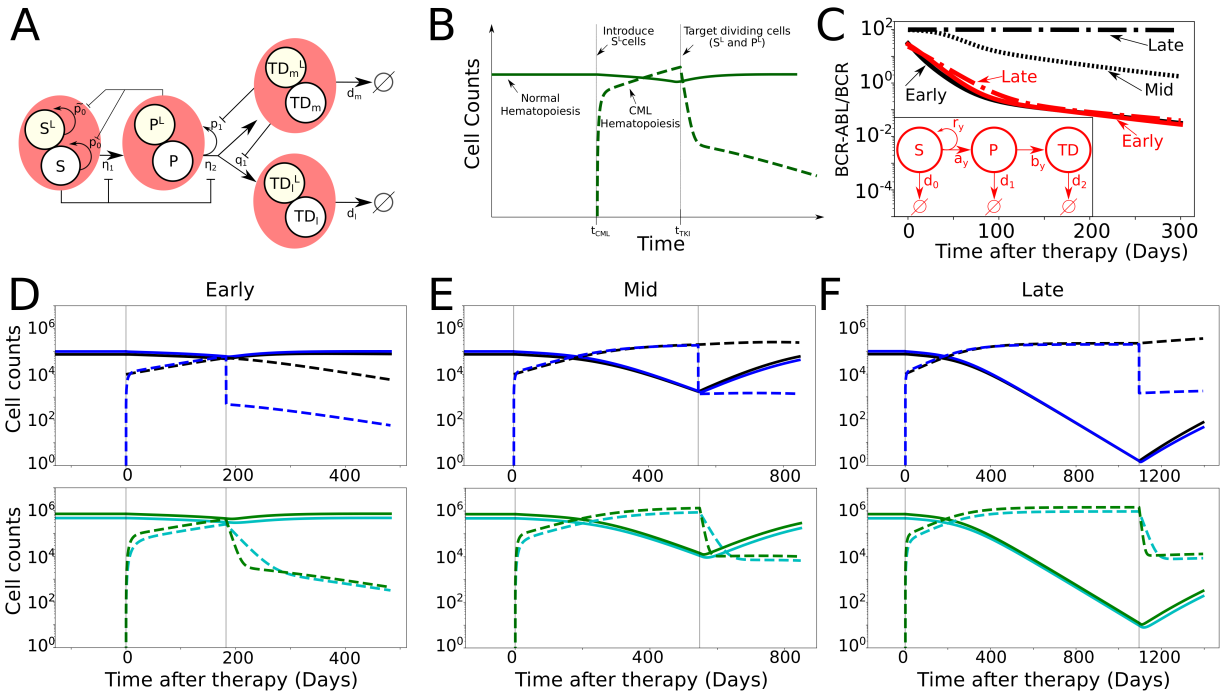

Figure 3.7: Applying data driven model of hematopoiesis to chronic myeloid leukemia. A. Schematic of two branched lineages consisting of normal and CML cell types. The two model lineages share the same feedback architecture. The difference between the two lineages is the leukemic HSC self renewal is less affected by negative feedback, denoted by $\tilde{p_0}$. B. Schematic showing how CML is modeled. We begin with having normal hematopoiesis at equilibrium. At time t_{CML}, S^L cells are added and at time t_{TKI}, therapy is added killing dividing leukemic cells. C. Therapy response curves corresponding to the application of TKI at different time points. We compare the proposed model with a linear model found in the literature [56], shown by the red inset model and red response curves. D-F. Cell dynamics of normal and leukemic cells in response to an early, mid, and late treatment of TKI. Parameters used: p_0=0.43, p_1 = 0.42, q_1 = 0.11, d_m=0.17, d_l=0.05,γ_1=8e-7, γ_2=4e-6, γ_3=7e-9, γ_4=1e-9, γ_5=1e-8, η_1=0.37, η_2=1.31

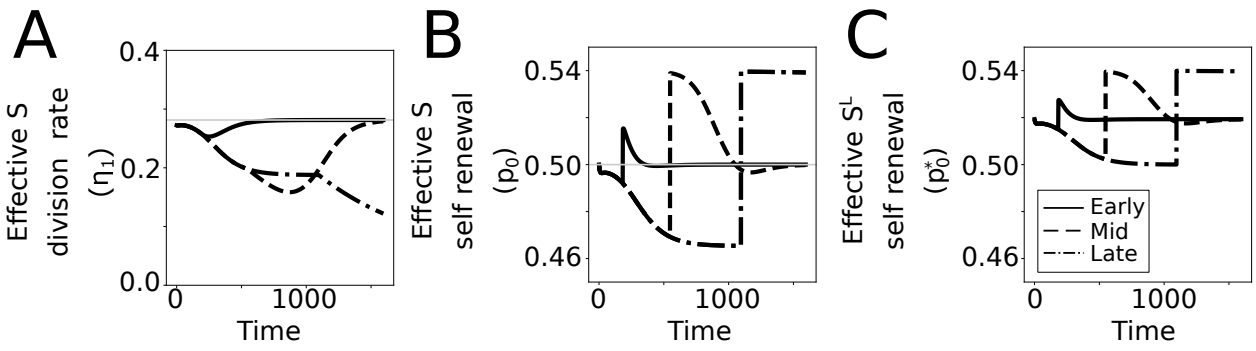

Figure 3.8: Effective rates for stem cell self renewal and division rate. A. The stem cell division rate decreases as we delay TKI treatment. At early and mid treatments, the stem cell division rate recovers B. - C. Effective self renewal probabilities for different therapy times. As we apply therapy later in time, the stem cell self renewal probability increases past 50% and starts creating more stem cell. Parameter used are the same as figure 3.7.

Chapter 4

Parameter estimation using Bayesian inference

4.1 Introduction

Two sentences to time previous work

Dynamic population models using ODEs have been traditionally fit using simple techniques, such as sum of squares [8, 11, 4]. Specifically, models of hematopoiesis have been fit under system equilibrium and point estimates for parameters have been estimated [9, 13]. Using these techniques allow only point estimates and assume normality assumptions. Bayesian methods allow us to quantify uncertainty in parameter estimations, relax the normality assumption, allow a hierarchical structure in parameters, and gives the ability to provide prior knowledge. Parameter estimation in biology using Bayesian inference is not a new concept and has been used to fit parameters of ODE models in pharmacokinetics/pharmacodynamics, disease modeling, hematopoiesis modeling, ecological models, and biochemical pathway models[26, 79, 57, 65, 23, 27]. Using ODE models allows us to elucidate potential feed-

back mechanisms to better understand the biological system. Specifically, in our context, we can gain insight on potential feedback interactions in immature hematopoieitic cells through cytokine secretion. Fitting our model also allows us to administer computational experiments to make predictions and further propose experiments.

In this study, we use a simplified model, consisting of dividing cells (s) and terminal cell (t) to fit synthetic data. The data is created using the ODE solution and thus the ground truth is known. We will review the data generation process and some general techniques to analyze the generated posterior.

4.2 Model fitting and inference

Bayes theorem gives us the ability to combine information about our data and incorporate prior knowledge. Given Bayes thorem, the posterior distribution is obtained by the following

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} \tag{4.1}$$

where $\boldsymbol{y}$ is our observed data, $\boldsymbol{\theta}$ is our parameter or parameters of interest, $p(\boldsymbol{\theta})$ is our prior, $p(y|\boldsymbol{\theta})$ is our likelihood, and $p(\boldsymbol{y})$ is our marginal likelihood. All of these components are used to find the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$. Since the denominator $p(\boldsymbol{y})$ does not depend on the parameters $\boldsymbol{\theta}$, it is normal to express the posterior as

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(y|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{4.2}$$

where we obtain the posterior up to a constant of proportionality. Thus, we need to specify the likelihood function and prior distributions for the parameters.

The first requirement is to choose a model that explains the data generation process. In

choosing one, we are making assumptions on how the data is generated and is only an approximation. The second requirement is choosing prior distributions for your parameters. It is important to choose priors before looking at data. In general, we assume independence in prior distributions. If there is previous data supporting parameter values (e.g. cell division rates, death rates), it is often incorporated by placing mass of the distribution at the known value. If little is known, it is often times reflected by choosing a prior that covers a large parameter range. In most cases in biology, parameter values might not be known, but orders of magnitudes can often be suggested. Gelman provides a extensive analysis on what types of priors one should use [25].

To obtain the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$ in equation 4.2, the samples of the posterior distribution are obtained by Markov Chain Monte Carlo (MCMC) using Stan [12, 76]. At each iteration of the MCMC algorithm the ODE is solved multiple times (one per leapfrog step for HMC and NUTS [12, 76])

4.2.1 Mechanistic model

We consider a two cell lineage to show how one can fit parameters under this framework. The cell types of interest are stem cells and terminal cells. The schematic of this simple lineage was shown earlier in figure 2.2.

The model can be described by the following set of differential equations:

$$x_s' = (2p_0 - 1)x_s \tag{4.3}$$

$$x_t' = 2(1 - p_0)x_s - dx_t \tag{4.4}$$

where $p_0 = \frac{\bar{p_0}}{1+\gamma x_t}$, $\bar{p_0}$ is the max self renewal probability, d is the death rate of the terminal cells, and γ is the feedback strength. In defining this model, we must also include parameter

constraints to impose stability of the system. The constraints for the model: $0.5 < \bar{p_0} < 1$ and $d, \gamma > 0$. The populations x_s and x_t are time dependent and are greater than or equal to 0. For the following equations, we will define $\boldsymbol{\theta} = (\bar{p_0}, d, \gamma)$ to be the vector of parameters.

4.2.2 Data generation process

The simulation study shows how one would fit time series data to a mechanistic ode model. We can assume this data is coming from an experiment where a system is perturbed and the results of the perturbation are observed by measuring the number of cells over a set number of time points.

For our purposes, the data generation process assumes the cellular levels are coming from some known initial conditions and follow a trajectory according to the ode model defined in equations 4.3-4.4. The distribution of the data is given as

$$\boldsymbol{y}_j(t_j) \sim N(\boldsymbol{x}(\boldsymbol{\theta}, t_j), \sigma^2 \cdot \boldsymbol{I}), j = 1, \ldots, M \tag{4.5}$$

where we define $\mathbf{x} = (x_s, x_t)$ and M is the amount of data points we have observed.

4.2.3 Defining the likelihood and prior knowledge

Using the Bayesian framework, we can link our model by defining the likelihood as a product of normal distributions.

$$p(\boldsymbol{y} \mid \boldsymbol{\Theta}) = \prod_{j=1}^{M} N\left(\boldsymbol{y}_j \mid \ \boldsymbol{x}(\boldsymbol{\theta}, t_j), \sigma^2\right) \tag{4.6}$$

$$= \prod_{j=1}^{M} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(y_j - \boldsymbol{x}(\boldsymbol{\theta}, t_j))^2}{2\sigma^2}} \tag{4.7}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}, \sigma)$, a vector of all parameters we want to infer. We can then provide measures of uncertainty in all parameters contained in $\boldsymbol{\Theta}$ by calculating the following posterior:

$$p(\boldsymbol{\Theta} \mid \boldsymbol{y}) \propto p\left(\boldsymbol{y} \mid \boldsymbol{x}(\boldsymbol{\Theta})\right) p\left(\boldsymbol{\Theta}\right) \tag{4.8}$$

where we assume independence in the priors allowing us to define $p(\boldsymbol{\Theta})$ by the following:

$$p\left(\boldsymbol{\Theta}\right) = p\left(\bar{p_0}\right) p\left(d\right) p\left(\gamma\right) p\left(\sigma\right). \tag{4.9}$$

For our example, we defined $\bar{p_0} \sim U(0.5, 1)$ and $d, \gamma, \sigma \sim LN(0, 1)$. We set hard constraints on $\bar{p_0}$ because this parameter is a probability and we know in order for stability, the parameter must be between these ranges. The lognormal distributions are chosen because our parameters are strictly greater than 0, and thus guarantees sampling positive values. An alternative would be to set a truncated normal where we only look at the positive values.

4.3 Results

Diagnostics It is important to check the summary diagnostics in order to check if there was convergence in the distribution. STAN reports convergence by an $\hat{R}$ statistic, which measures the ratio of the average variance of samples within each chain to the variance of all the samples across chains [75]. It is recommended the $\hat{R}$ value is below 1.1. In our example, all of our parameters have an $\hat{R}$ value of 1 (See table 4.1), showing convergence in the chains. In addition to the $\hat{R}$, we can plot the trace plots, which represent the values over sampling iterations for each parameter after the warm up iterations. If converged, the plots should show the samples close to the true value, as shown in figure 4.1. The true value is shown in black. In addition to convergence, we need to measure the quality of the samples. Due to

MCMC sampling being a Markovian process, we must deal with issues of autocorrelation, which introduce additional uncertainty to the fit. STAN reports an n_{eff} statistic that is defined by the autocorrelations within the sequence at different lags [75]. In our example, our $n_{eff} > 5000$ for all parameters showing a good effective sample size (See table 4.1).

Posteriors Using MCMC methods, we obtain the posterior. A summary comparing the prior vs posterior distributions are shown in table 4.1. In all cases, the true parameter is contained within the 95% credible intervals for the posterior. One can plot the prior vs. posterior to visualize how much information is gained from the data. The goal of these plots are to see whether or not the posterior was updated with the given data by a shift and/or reduction in variance. In our example, we observe a shift with the 3 ODE parameters and the error term $(\bar{p_0}, \gamma, d, \sigma)$, shown in figure 4.2. The posterior's mass is concentrated at the true value, shown in green.

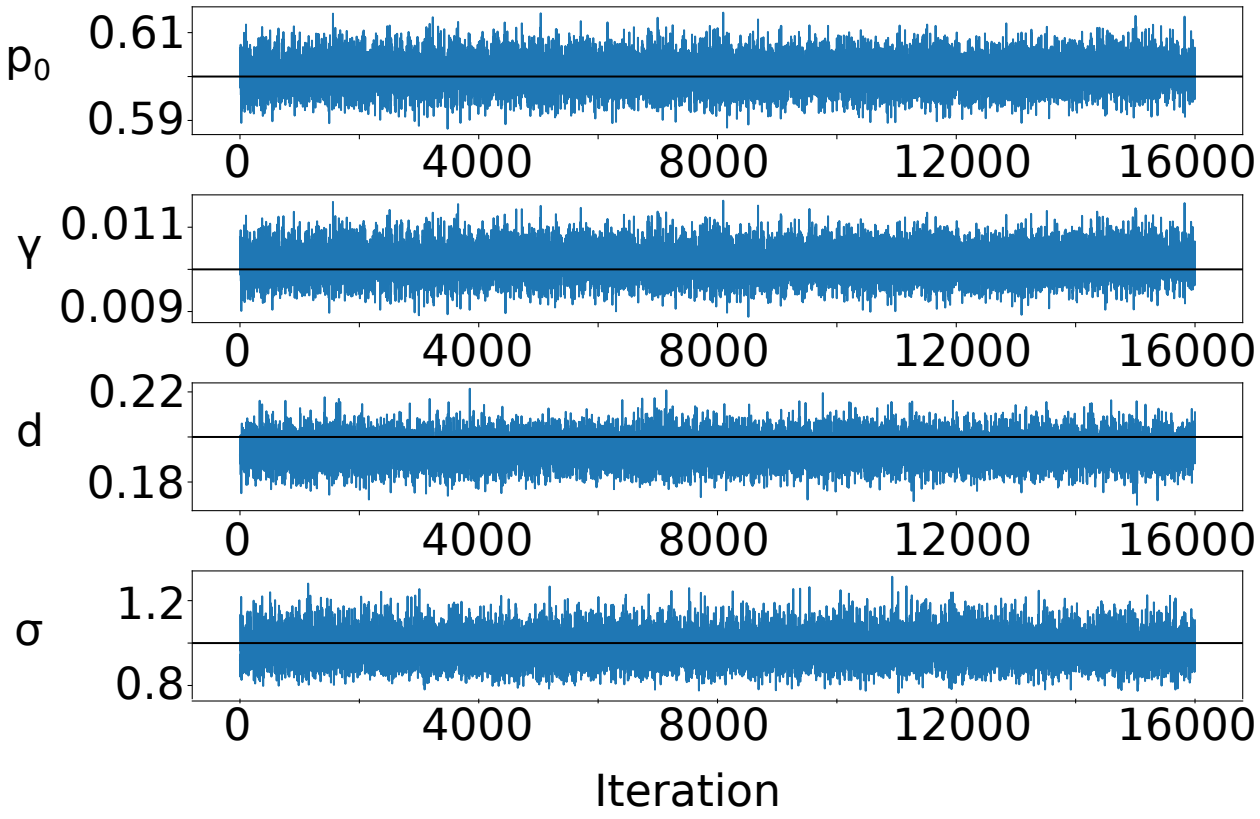

Figure 4.1: Trace plots for simple ODE lineage.

	Prior			Posterior			True	n_{eff}	$\hat{R}$
Parameter	2.5%	50%	97.5%	2.5%	50%	97.5%			
$\bar{p_0}$	0.51	0.75	0.99	0.59	0.6	0.61	0.6	5691	1
γ	0.14	1	7.10	9.5e-3	0.01	0.01	0.01	5713	1
d	0.14	1	7.10	0.18	0.19	0.21	0.2	8138	1
σ	0.14	1.0	7.1	0.85	1.03	1.13	1	8604	1

Table 4.1: Prior vs posterior distributions for the two cell lineage model. We observe clear information gain as shown in the reduction of the 95% credible intervals and the shift of the posterior medians. The MCMC inputs used: 8000 iterations, 4000 warmup, 4 chains.

Another great tool to use in analyzing the posteriors is constructing a pairs plot, which can be used to visualize if parameters of the ODE model have a relationship. The pair plot is a scatter plot where one posterior is plotted against another. In our example, we can see a linear relationship between $\bar{p_0}$ and γ in figure 4.3. This relationship can be explained by looking at how the feedback is defined, $\bar{p_0}/(1+\gamma x_s)$ and the steady state of the x_t population $(x_t^* = (2\bar{p_0} - 1)/\gamma)$.

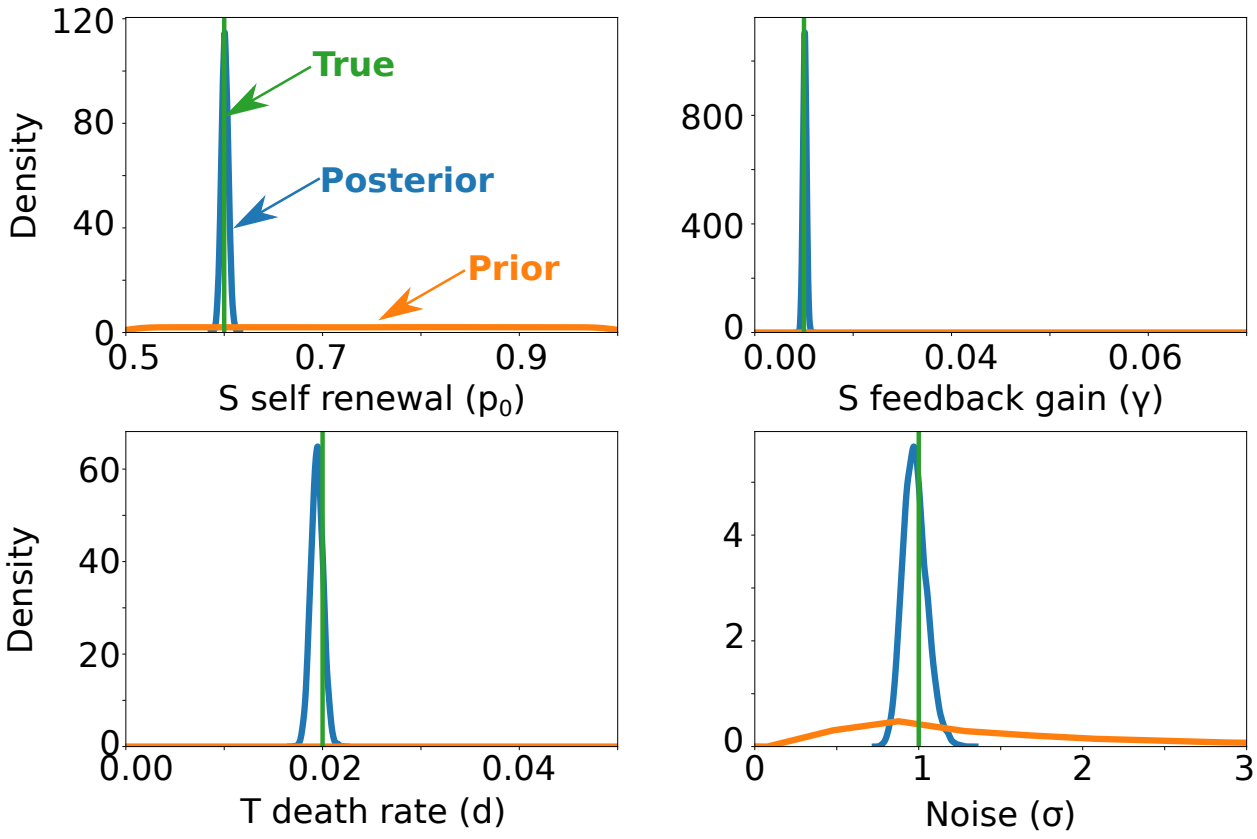

Figure 4.2: Prior vs posterior for simple ODE lineage.

Using the posterior distribution also gives the opportunity to present the fit using point estimations. The idea is to sample many times from the posterior and generate ODE solutions. From these solutions, we can obtain a median ODE solution and credible regions for our solutions. In doing this, we can observe how well of a fit it is and observe the variability in the solutions. Figure 4.4 shows the data, the ODE median solution and the 95% credible intervals, where the credible intervals summarize the uncertainty related to the ode fit when

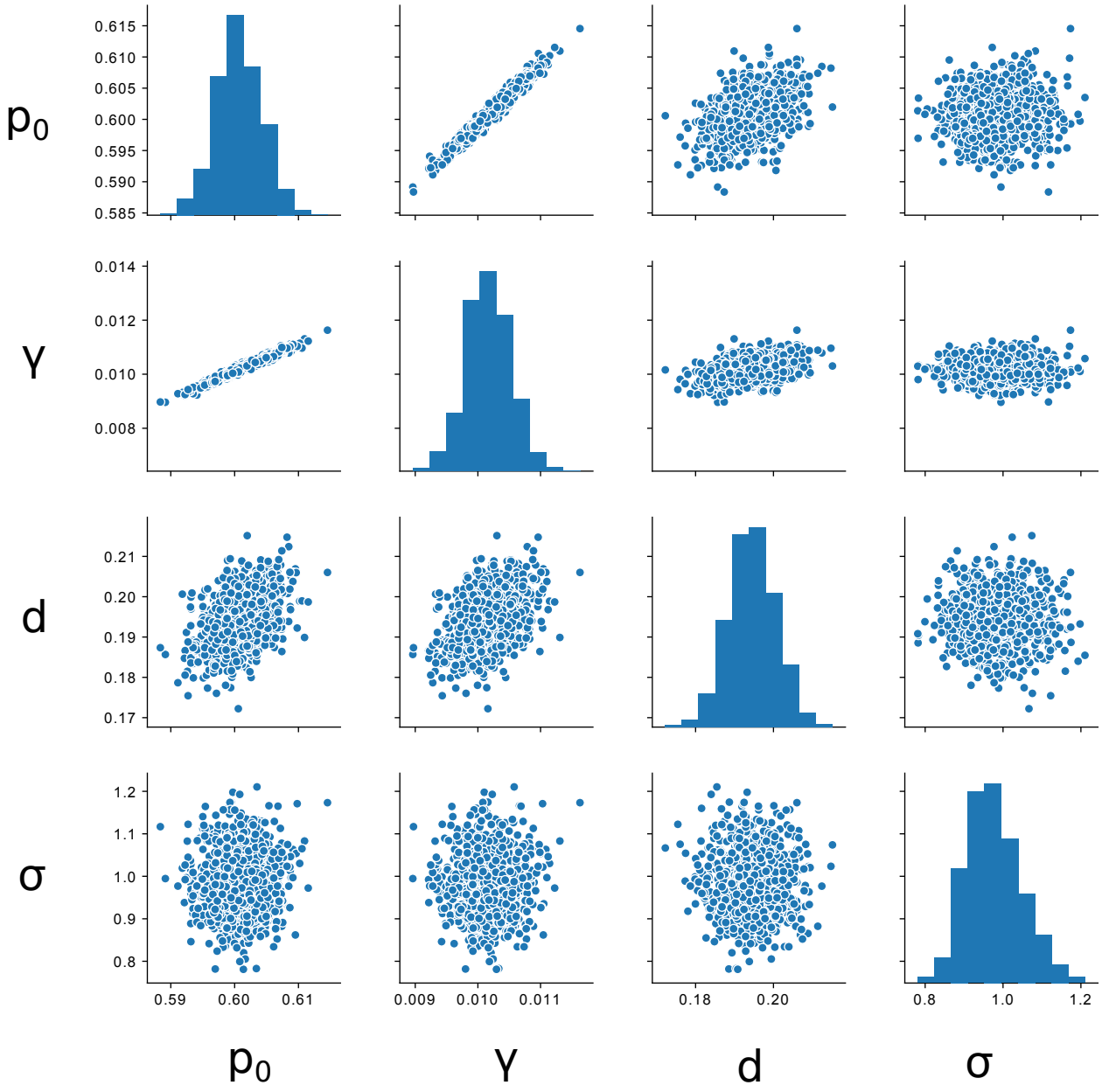

Figure 4.3: Pair plot for simple ODE lineage.

sampling parameter values from the posterior.

Posterior predictive checks We can use posterior predictive checks to assess the goodness of fit. We do this by sampling many times from the posterior and generating synthetic data from our differential equation with noise to then compare it back to the original data. In doing this many times, we obtain a distribution of where the synthetic data lies and if the model is correct, the original data should likely be under the posterior predictive distri-

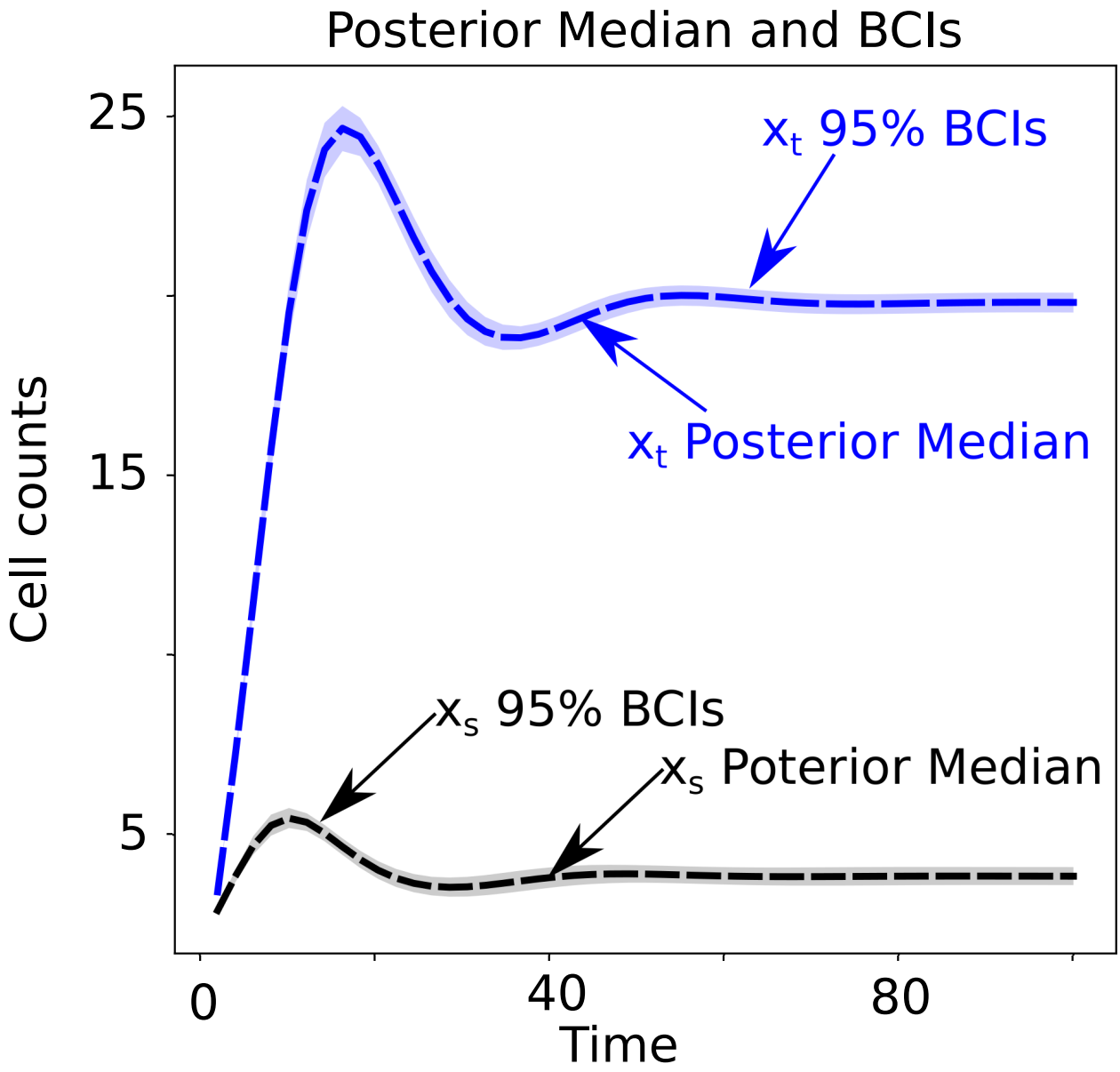

Figure 4.4: Point estimate and Bayesian credible intervals for simple ODE lineage.

bution. In our example, we add 95% predictive bands and show a majority of the generated data is contained in the band (see figure 4.5).

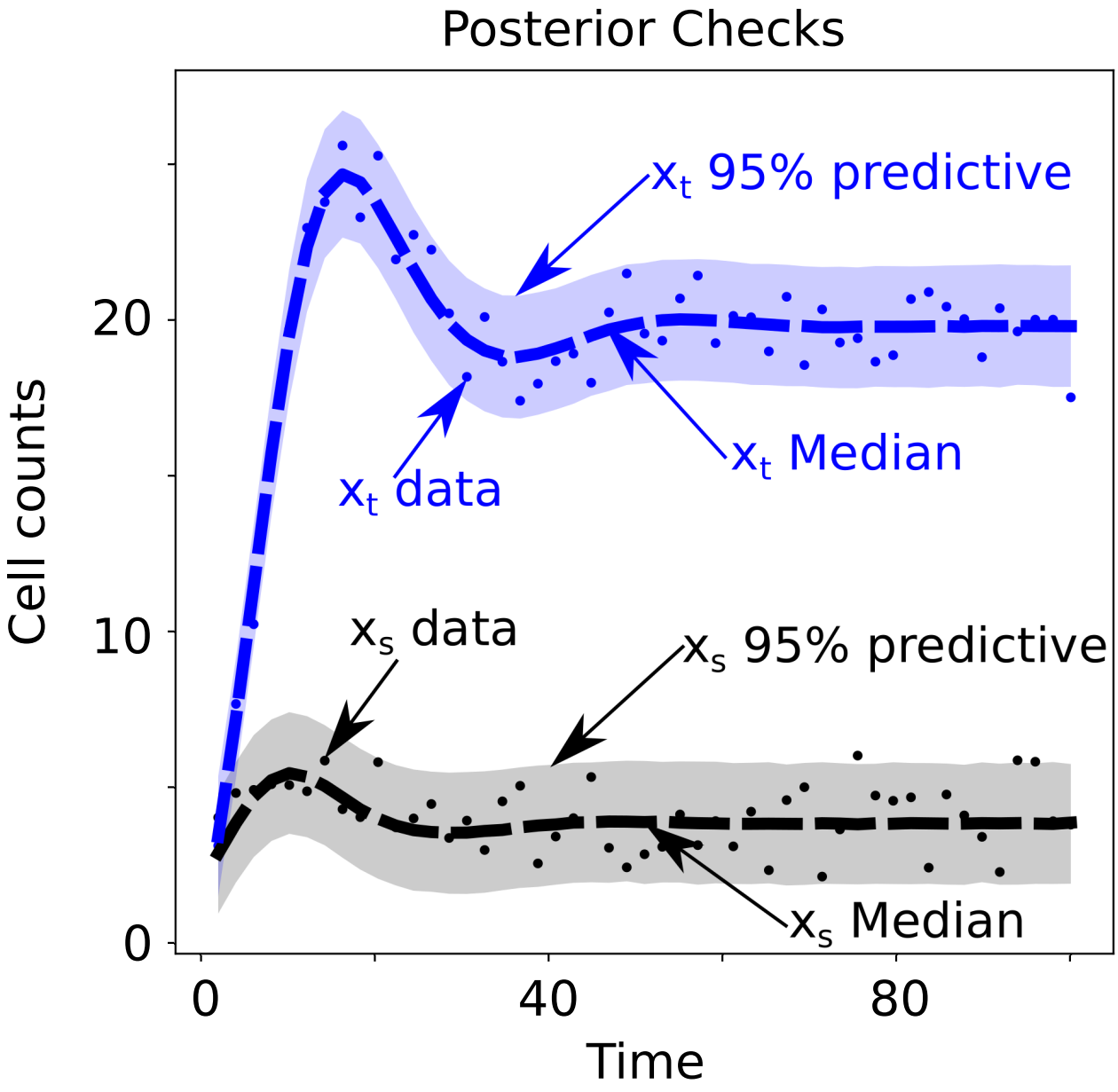

Figure 4.5: Posterior checks for simple ODE lineage.

4.4 Discussion

In this chapter, we reviewed Bayesian statistics as it pertains to parameter inference for ODEs. We showed we are able to recover the ODE parameters and the variability of the data within the 95% credible intervals of the posteriors. Fitting a mechanistic model gives us the ability to understand underlying mechanisms, which allows for new hypotheses.

In practice, there are many improvements and extensions one can do. For example, we can perform a sensitivity analysis on the priors by shifting the mean, variance, or even change the distribution to observe how it effects the posterior. In doing this, we can check how robust the fit is based on our priors. We can also extend our model to fit additional time series data, where each set of data corresponds to an individual. If we are to assume this, we can then extend our model to have a hierarchical component where we consider each time series data set to have its own set of parameters coming from an underlying distribution [37].

In this basic problem, there were no issues in the fitting or the sampling due to the simplicity of the model and only fitting 3 ODE parameters (p_0, γ, d) and an error term (σ). As we increase the amount of parameters we want to infer, the more computationally expensive it will be to sample effectively. The MCMC sampler can also begin exploring regions of parameter space where the numerical solutions are not valid, which can cause issues in the sampling. All in all, it is important to not only consider the likelihood and priors, but to also understand the general dynamics of the mechanistic model so that the user is aware of potential issues when fitting.

Chapter 5

Bayesian optimal experimental design for mathematical models of hematopoiesis

5.1 Introduction

The hematopoietic system represents a prototype model of a mammalian tissue organized in a hierarchy with pluripotent stem cells at the apex. This system produces billions of mature myeloid and lymphoid cells on a daily basis in the basal state, and accommodates to massive cell increase in response to pathological stresses [50]. The system must have in place a tightly regulated feedback control mechanism at multiple levels to ensure an appropriate proportion of stem, progenitor, and mature cells. However, the nature of the feedback regulation and how it plays a role in cell maintenance is still not well understood.

The use of mathematical models in hematopoiesis is part of a growing effort to understand blood development. Previous works have developed ODE lineage models using linear models

with constant rates and incorporate little to no feedback regulation of the system dynamics [56, 54, 53, 9]. Although these models have provided some insight, they should be improved to reflect the feedback mechanisms of blood development. Incorporating feedback control to our proposed mathematical model allows us to capture richer dynamics which are more representative of blood development and maintenance. In some cases, these simplistic ODE models have been fit under system equilibrium and point estimates for the parameters have been estimated [9, 13] and uncertainty in parameter estimations containing feedback has only been considered in some cases [23, 27].

Perturbation experiments targeting hematopoietic cells are needed to observe recovery dynamics and identify feedback mechanisms. For example, several animals/subjects can be exposed to a non-lethal agent to induce a system response where cellular counts can be quantified. In some cases, the subjects can be tracked longitudinally by taking repeated measures of the cellular counts, but in others, the animals need to be harvested at each observation point. For the latter case, bone marrow extraction experiments in mice are common, and they represent a statistical challenge for quantifying the system response since each data point belongs to a single animal and all the pre-observation cellular counts are latent. In particular, it is not clear what conditions are needed for correct statistical inference for feedback mechanisms with this kind of data and how sensitive statistical estimates are with respect to the choice of experimental design.

The hematopoietic experimental designs include multiple variables that are fixed before the data is collected, such as the number of animals per observation time, timing of the measurements, nature of the perturbation, etc. In choosing an experimental design, it is important in optimizing resources to minimize cost and maximize information in order to identify regulation mechanisms and infer parameters. To be able to discriminate among different designs, one common option is to use a utility function that quantifies the information gain about a model of an experiment for each design. The design that provides higher expected utility

(maximum information gain) is considered the optimal design. However, choosing the best design is generally a non-trivial task since the space of all possible designs is generally large and a significant amount of computation is needed to calculate the expected utility for each design. Previous work has shown how this could be applied in different branches of biology to optimize certain parameters in experimental setups [19, 82, 31]. Other works in the literature have shown the usefulness of Bayesian utility theory for maximizing information gain about model parameters [14, 47, 17, 36]. Although this theory has been used in many fields, this application has not been applied in the hematopoietic system using a cross sectional study.

In this work, we propose a framework to provide an optimal experimental design in hematopoiesis with latent components, eg. bone marrow extraction experiments. We propose a mechanistic differential equation model to provide a quantitative description of feedback regulation in hematopoiesis through feedback parameters, division rates, and self-renewal probabilities. We bridge mathematical modeling with statistical inference using a Bayesian hierarchical model based on a latent variables approach. This method allow us to fit the mechanistic model to experimental data consisting of cell counts when the individuals are measured once. Our framework allows for quantitative parameter identification and prediction of system response when perturbation data experiments is available. For this data, we bring a stochastic component by modeling biological variability among the animals in the experiment and the variability introduced by measurement error. We aim to optimize information gain in ODE parameters, specifically the feedback rates, among multiple candidate experimental designs using an utility metric. We use the expected KL divergence between the prior and posterior distribution for the parameters and we apply this to quantify information across all the parameters simultaneously, as well as information gain for individual parameters marginally. Our method is flexible enough to be applied to more complex dynamic models with extra cellular types, stochastic formulations and to incorporate covariates information into the optimal decision process.

5.2 Methods

5.2.1 Experimental setup

We are interested in an experimental set up where the hematopoietic system is perturbed and results of the perturbation are observed by measuring the number of cell types of interest, such as hematopoietic stem cells (HSCs), progenitors (MPPs, CMPs, CLPs), and mature myeloid or lymphoid cells. Throughout this chapter, we will use simulated and real data based on the following experiment. We start with M genetically identical mice that are kept under the same laboratory conditions. Each mouse is exposed to an external perturbation, e.g. sub-lethal radiation, with the purpose of targeting the number of HSCs. These mice are sacrificed at different times post perturbation and the counts of HSCs and MPPs are obtained from the bone marrow of each individual mice using flow cytometry.

5.2.2 Data generation process

We postulate that at time t_0 HSC and MPP counts come from a distribution that encapsulates normal biological variation among mice. For mathematical convenience, we assume that this distribution is lognormal, or equivalently, that log-transformed unobserved true HSC and MMP counts come from two independent normal distributions with means $\log(\mu_{\text{HSC}}), \log(\mu_{\text{MPP}})$ and the same variance σ_b. Denoting these latent log-transformed counts with a bivariate vector $\mathbf{u}_i$ and their means with a vector $\boldsymbol{\mu}$ for each mouse $i = 1, \ldots, M$, we can write our initial condition assumption as

$$\mathbf{u}_i \sim N\left(\log(\boldsymbol{\mu}), \sigma_b^2 \cdot \mathbf{I}\right), i = 1, \ldots, M. \tag{5.1}$$

We are now ready to specify distribution of observed HSC and MPP counts. First, we order mice in such a way that mice indexed by $1, \ldots, k$ correspond to animals, whose bone marrow is sampled immediately post perturbation. We assume that conditionally on the true log-counts of HSCs and MPPs, the observed log-transformed counts $\mathbf{y}_i = \log(\mathbf{y}_i^*)$, where $\mathbf{y}_i^*$ represents the raw counts, are normally distributed, with the mean being equation to the true counts and variance σ_t^2 that represents technical variation that arises due to the measurement error/noise: come from a subset of mice that are sacrificed right after the perturbation experiment. In this case, the exact distribution model for these records is known since it only depends on the log mean IC values but are subject to technical error in the

$$\mathbf{y}_i(t_0) \mid \mathbf{u}_i \sim \quad N\left(\mathbf{u}_i, \sigma_t^2 \cdot \mathbf{I}\right), i = 1, \ldots, k. \tag{5.2}$$

The unconditional distribution for the HSC and MPP counts at the initial time t_0 is derived from equations (5.1) and (5.2) as

$$\mathbf{y}_i(t_0) \sim \quad N(\log(\boldsymbol{\mu}), (\sigma_t^2 + \sigma_b^2) \cdot \mathbf{I}), i = 1, \ldots, k. \tag{5.3}$$

To model cell counts measured at time points after the initial time t_0, we assume that the cellular population levels in each mouse start from latent initial conditions and follow deterministic latent trajectories according to mechanistic ODEs. The latent population levels trajectories evolve for all times $0 < t \leq t_j$ until the moment t_j when the mouse is harvested and the cell counts are measured with noise (Figure 5.1A). The conditional distribution of the observed HSC and MPP counts after the initial time given latent initial conditions is

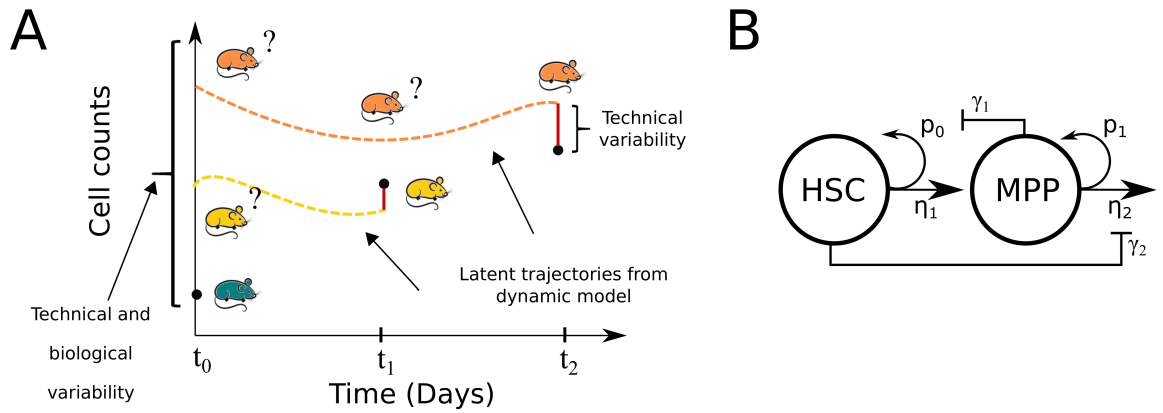

Figure 5.1: Illustration of the proposed latent variables approach and mechanistic model. A. Proposal of latent variables approach. In mouse experiments where the bone marrow is obtained, it is protocol to sacrifice the mouse. We model this by assuming each data point at at time t_0 to be subject to technical and biological variability. At times greater than t_0, we assume each data point has a latent trajectory subject to technical variability (shown in dashed). These latent trajectories are modeled using our mechanistic ODE model. B. ODE lineage model consisting of HSC and MPPs. HSC and MPPs have the ability to self renew (p_0 and p_1) with some division rate (η_1 and η_2). The HSC self renewal is negatively regulated by the MPPs population and MPPs division is negatively regulated by the HSC population.

given by

$$\mathbf{y}_j(t_j) \mid \mathbf{u}_j \sim N(\log(\mathbf{x}(\mathbf{u}_j, \boldsymbol{\theta}, t_j)), \sigma_t^2 \cdot \mathbf{I}), j = k+1, \ldots, M \tag{5.4}$$

where $\mathbf{x}(\mathbf{u}_j, \boldsymbol{\theta}, t_j)$ is a bivariate vector of HSC and MPP counts for mouse j that started with latent counts $\mathbf{u}_j$, and evolved according to a mechanistic process with parameters $\boldsymbol{\theta}$ up to time t_j. Note that equation (5.4) does not include the biological noise term explicitly since the model for initial cell counts already includes it. This approach implies that the HSC and MPP trajectories are latent and can be observed cross-sectionally only once, in contrast to other experiments where repeated measures are taken from a cohort of animals/subjects followed over time.

5.2.3 Mechanistic model of the mean process

The dynamical model for latent trajectories is based on a lineage model for describing cell dynamics [46, 10]. Our approach follows the classical depiction of hematopoiesis as a hierarchy of cells starting with the hematopoietic stem cell at the top and more differentiated cells following. We assume a similar hierarchical structure by creating an order of differentiation, as shown in Figure 5.1B. For our purpose, we only model HSC and MPPs. The model can be easily extended to include cells downstream, such as myeloid or lymphoid type cells.

The lineage model of interest describes fundamental properties of cellular growth. The cell populations (HSC and MPP) have the ability to divide at rates η_1 and η_2^* respectively, to undergo symmetric self-renewal with some probabilities p_0^* and p_1, and to differentiate with probabilities $1 - p_0$ and $1 - p_1$. Due to the increasing evidence of feedback/feed-forward playing a role in hematopoiesis development and maintenance [71, 29], we include this mechanism in the model using qualitative mathematical and biological prior knowledge (Figure 5.1B) [51, 58, 83, 3]. We incorporate feedback/feed-forward control by making the model parameters functions of the cell populations where the parameters γ_1 and γ_2 measure the strength of the feedback/feed-forward regulation. The first control is a negative feedback on stem cell self renewal from progenitor cells. The second is a negative feed-forward on the progenitor division rate from stem cells. The corresponding equations governing the time evolution of latent HSC (x_{HSC}) and MPP (x_{MPP}) counts are

$$\begin{aligned} x'_{HSC} &= (2p_0^* - 1)\eta_1 x_{HSC} \\ x'_{MPP} &= 2(1 - p_0^*)\eta_1 x_{HSC} - \eta_2^* x_{MPP}, \end{aligned} \tag{5.5}$$

where $p_0^* = \frac{p_0}{1+\gamma_1 x_{MPP}}$ and $\eta_2^* = \frac{\eta_2}{1+\gamma_2 x_{HSC}}$ represent the maximum values that the HSC self

renewal probability and MPPs division can take correspondingly. We group all the ODE model parameters in the vector $\boldsymbol{\theta} = (p_0, \eta_1, \eta_2, \gamma_1, \gamma_2)^t$. The non-linear terms of this model allow for a rich range of dynamics for the HSC and MPP populations. For example, the populations will remain at steady state dynamics under no perturbation or the populations can respond to an external stimulus in a recovery phase, usually characterized by oscillatory behavior, before returning to steady state.

5.2.4 Hierarchical Bayesian framework

The goal of our statistical framework is to link the dynamic mathematical model with empirical data. We use a latent variables approach where the ODE model allows us to interpolate all the latent trajectories for the records missing before a mouse is harvested. In principle, we can also extrapolate and forecast under the same model but we focus here on the dynamics until the steady state is reached. Following the data generation process, we define the likelihood function as the product of normal densities at time zero and at later times. The first records at time zero are log transformed, $y_1, \ldots, y_k$, and modeled with a Normal distribution centered at the mean IC values with technical and biological noise. The remaining post-initial time records $y_{k+1}, \ldots, y_M$ follow a Normal density where its mean depends on the latent trajectory provided by the ODE model under technical error,

$$\begin{aligned} p(\mathbf{y} \mid \boldsymbol{\Theta}) = & \prod_{i=1}^{k} N\left(\mathbf{y}_i \mid \log \boldsymbol{\mu}, \sigma_t^2 + \sigma_b^2\right) \\ & \times \prod_{j=k+1}^{M} N\left(\mathbf{y}_j \mid \log \mathbf{x}(\boldsymbol{\theta}, t_j), \sigma_t^2\right) \end{aligned} \tag{5.6}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}, \mathbf{u}, \sigma_b^2, \sigma_t^2, \boldsymbol{\mu})^t$ includes all the parameters of the hierarchical model. Using a Bayesian approach, we provide measures of uncertainty to all model parameters by calculat-

ing the posterior distribution of all the parameters using Bayes theorem,

$$p(\mathbf{\Theta} \mid \mathbf{y}) \propto p\left(\mathbf{y} \mid \mathbf{x}(\mathbf{\Theta})\right) p\left(\mathbf{\Theta}\right) \tag{5.7}$$

where the likelihood function takes the form of (5.6). We assume independence among the model parameters, ie, the prior distribution can be decomposed as

$$p(\mathbf{\Theta}) = p(\boldsymbol{\theta}) \cdot p(\mathbf{u}) \cdot p(\sigma_b^2) \cdot p(\sigma_t^2) \cdot p(\boldsymbol{\mu}). \tag{5.8}$$

and for each of the parameters $\theta_i \in \mathbf{\Theta}$ we reparametrize its prior as $\log \theta_i \sim N(m_{\theta_i}, \sigma^2_{\theta_i})$. Note that p_0 requires a logit reparametrization, $\text{logit}(\frac{p_0 - \frac{1}{2}}{2}) \sim N(m_{p_0}, \sigma^2_{p_0})$, since it is constrained between 0.5 and 1. We obtain samples from the posterior distribution (5.7) using Markov Chain Monte Carlo simulation.

5.2.5 Experimental design

In the laboratory, there are multiple variables that can be modified when an experiment is performed. Each combination of these variables defines an experimental design which is part of the set of all possible designs D. The optimal experimental design goal is to discriminate between all the possible designs by using a suitable utility metric $U(\mathbf{y}, d)$ that quantifies the amount of information gain about the model parameters for a dataset $\mathbf{y}$ under the design $d \in D$. The optimal design is determined by comparing the expected utility over all datasets,

$\mathbf{y}$, and model parameters, $\mathbf{\Theta}$, for the design d,

$$u(d) = \mathbb{E}_{\mathbf{\Theta},\mathbf{y}}\left[U(\mathbf{y}, d)\right] = \int_{\mathbf{y}} \int_{\mathbf{\Theta}} U(\mathbf{y}, d) p(\mathbf{y}|\mathbf{\Theta}, d) p(\mathbf{\Theta}) d\mathbf{\Theta} d\mathbf{y}. \tag{5.9}$$

For the utility $U(\mathbf{y}, d)$, we use the Kullback-Leibler (KL) divergence function that quantifies the information gain from prior vs posterior distribution of the model parameters and it is defined as,

$$U(\mathbf{y}, d) = \int_{\mathbf{\Theta}} \log\left(\frac{p(\mathbf{\Theta}|\mathbf{y}, d)}{p(\mathbf{\Theta})}\right) p(\mathbf{\Theta}|\mathbf{y}, d) d\mathbf{\Theta}. \tag{5.10}$$

and this utility is equivalent to the Shannon information gain [67]. A similar approach allows to calculate the marginal expected utility for each parameter $\theta_j \in \mathbf{\Theta}$ individually. This approach requires to approximate the posterior density $p(\theta_i|\mathbf{y}, d)$ and obtain the marginal utility value as

$$U_j(\mathbf{y}, d) = \int_{\theta_j} \log\left(\frac{p_j(\theta_j|\mathbf{y}, d)}{p(\theta_j)}\right) p(\theta_j|\mathbf{y}, d) d\theta_j. \tag{5.11}$$

5.2.6 Computational implementation

Estimating $U(\mathbf{y}, d)$

Each design $d \in D$ specifies the timing of the measurements and the number of mice replicates in a hematopoiesis experiment. When the design is fixed, one set of true parameters is sampled from their prior distributions and a synthetic dataset, $\mathbf{y}$, that contains the cellular population records for HSCs and MPPs is generated using the latent variables approach (equation 5.6). At the initial time, the initial conditions (IC) for the latent trajectories are determined by sampling values for μ_{HSC} and μ_{MPP}, log transformed and random biological and technical noise $\sigma_b^2 + \sigma_t^2$ is added simultaneously. From the IC, the ODE model (equation 5.5) is forward solved until the time when a trajectory is observed (mouse is sacrificed), log transformed and technical noise is added only, σ_t^2. This process is repeated for all the latent trajectories to obtain the dataset $\mathbf{y}$.

For the inverse problem, the samples of the posterior distribution of the parameters are obtained by Markov Chain Monte Carlo using Stan [12, 76]. At each iteration of the MCMC algorithm the ODE is solved multiple times (one per leapfrog step for HMC and NUTS [12, 76]) which can be computationally expensive for complex mechanistic models. Also, parameter dependencies can impose a complex geometry that make the posterior sampling difficult.

The posterior samples are used to compute the value of the KL utility function (5.9) by

Monte Carlo integration using all the MCMC iterations $i = 1, \ldots, n$,

$$\begin{aligned} U(\mathbf{y}, d) \approx \quad \widehat{U}(\mathbf{y}, d) = & \quad \tfrac{1}{n} \sum_{i=1}^{n} \log \left(\tfrac{p(\boldsymbol{\Theta}^{(i)}|\mathbf{y},d)}{p(\boldsymbol{\Theta}^{(i)})} \right) \\ = & \quad \tfrac{1}{n} \sum_{i=1}^{n} \log \left(\tfrac{p(\mathbf{y}|\boldsymbol{\Theta}^{(i)},d)p(\boldsymbol{\Theta}^{(i)})}{p(\mathbf{y}|d)} \right) - \log p(\boldsymbol{\Theta}^{(i)}) \\ = & \quad \tfrac{1}{n} \sum_{i=1}^{n} \log p(\mathbf{y}|\boldsymbol{\Theta}^{(i)}, d) - \log p(\mathbf{y}|d), \end{aligned} \tag{5.12}$$

where $\boldsymbol{\Theta}^{(i)}$ is the i-th sample from the posterior distribution which is evaluated at the prior and posterior densities. Note that the KL ratio (equation 5.12) requires an estimate of the marginal likelihood $p(\mathbf{y}|d)$ by Bridge Sampling, which is a thermodynamics integration method for calculating marginal likelihood of single models based on the posterior distribution samples [30]. The marginal KL utility value for each parameter θ_j is also approximated via Monte Carlo,

$$U_j(\mathbf{y}, d) \approx \widehat{U}_j(\mathbf{y}, d) = \frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{\widehat{p}(\theta_j^{(i)}|\mathbf{y}, d)}{p(\theta_j^{(i)})} \right) \tag{5.13}$$

where $\widehat{p}(\theta_j|\mathbf{y}, d)$ is an approximation for the true marginal density and it is estimated by a Gaussian kernel density using parameter posterior samples [41, 70]. This step is required since the analytical marginal density is generally not available.

Estimating $u(d)$.

The mean utility value (equation 5.9) is estimated using the KL utility using N different datasets and their corresponding true parameter sets. For each dataset the utility value is

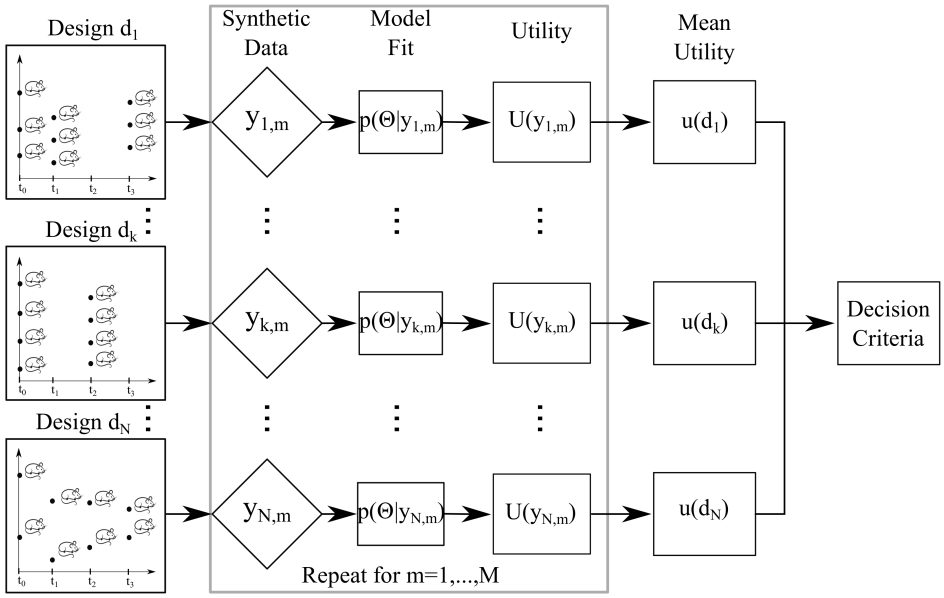

Figure 5.2: Flow chart for computational implementation. Given a design d_k, we sample a set of true parameter values from the priors to obtain a synthetic data set generated from our mechanistic model using the latent variables approach. Using the synthetic data, we obtain the posterior distribution and we calculate the utility value associated to the nth synthetic data set. Once we have repeated this process N times, we calculate the mean utility for design d_k. If we are to do this for M designs, we are able to use this process to choose the design with the optimal mean utility based on a decision criteria which is usually based on several constraints. For example, these constraints can very from minimizing the number of days or having a mouse limit analogous to a budget.

calculated and all the values are averaged,

$$u(d) \approx \widehat{u}(d) = \frac{1}{N} \sum_{k=1}^{N} \widehat{U}(\mathbf{y}_k, d) \tag{5.14}$$

where the utility $\widehat{U}(\mathbf{y}_k, d)$ is approximated using equation 5.12. For individual parameters,

the mean utility value is computed similarly,

$$u_j(d) \approx \widehat{u}_j(d) = \frac{1}{N} \sum_{k=1}^{N} \widehat{U}_j(\mathbf{y}_k, d) \tag{5.15}$$

where $\widehat{U}_j(\mathbf{y}_k, d)$ is estimated using equation 5.13.

Finding the optimal design

We consider a finite grid of designs $d_1, d_2, \ldots, d_M$ that are relevant for the hematopoiesis experiment. The optimal experimental design is determined by computing the expected utilities (5.14) and (5.15) for all the designs. For this purpose, N different sets of parameter values are sampled from their prior distributions (Table B.3) and the corresponding N full synthetic datasets are generated. These datasets contain cellular subpopulation records for the HSC and MPPs subpopulations and they are generated according to the latent variables approach (equation 5.6). For each design, the expected utility is calculated based on the N datasets where a subset of the records for each dataset is taken according to the specified number of mice replicates and observation days in the design. We show the optimal design process on Figure 5.2, where N was fixed to 60 datasets.

Calculation of Coverage

We determine the coverage of the estimated parameter (accuracy), by calculating how often the true parameter values are included in the 95% credible intervals (equation 5.19) and we use the magnitude of these intervals as a metric for precision (equation 5.21). We determine the relative bias for the MCMC simulations by using the posterior medians as point estimates

and we calculate the normalized residuals using the true parameter values (equation 5.18).

Relative bias, relative width and coverage

In our simulation study, for each design d we used several synthetic datasets and their corresponding MCMC simulation to infer back the true parameter values. For each of these simulations, we calculated the mean relative bias, mean relative width of the credible intervals and the coverage.

First, we use the posterior median as a point estimate for each parameter i in the simulation run j,

$$\hat{\theta}_{i,j}^{(d)} = median(p(\theta_{i,j}|y_j^{(d)})) \tag{5.16}$$

where $y_j^{(d)}$ represents the j-th synthetic dataset for the design d. The relative bias is calculated by normalizing the difference between the posterior median and the true value, ie, for the i-th parameter

$$\text{RelBias}_{i,j}^{(d)} = \frac{\hat{\theta}_{i,j}^{(d)} - \theta_{i,j}}{\theta_{i,j}} \tag{5.17}$$

where $\theta_{\cdot,j}$ represents the true parameter values for simulation j. The mean relative bias is calculated for each design averaging over all the relative bias estimates for all the simulations

of design d as

$$\text{Mean RelBias}_i^{(d)} = \frac{1}{N}\sum_{j=1}^{N}\text{RelBias}_{i,j}^{(d)} \tag{5.18}$$

We define the coverage of a credible interval with a similar interpretation as a frequentist confidence interval is calculated as the proportion of the time that the interval contains the true value of interest. In this case we calculate the number of times the true parameter value $\theta_{i,j}$ is included in the corresponding 95% credible interval $C_{i,j}$ for the design d.

$$\text{Coverage}_i^{(d)} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{1}_{\theta_{i,j}\in C_{i,j}} \tag{5.19}$$

The relative width of the posterior distribution of a parameter is given by the magnitude of the 95% credible interval, ie, the distance between the 97.5th and 2.5th percentiles as:

$$\text{RelWidth}_{i,j}^{(d)} = q_{0.975}(p(\theta_{i,j}|y_j^{(d)})) - q_{0.025}(p(\theta_{i,j}|y_j^{(d)})) \tag{5.20}$$

and the mean relative width is given by

$$\text{Mean RelWidth}_i^{(d)} = \frac{1}{N}\sum_{j=1}^{N}\text{RelWidth}_{i,j}^{(d)} \tag{5.21}$$

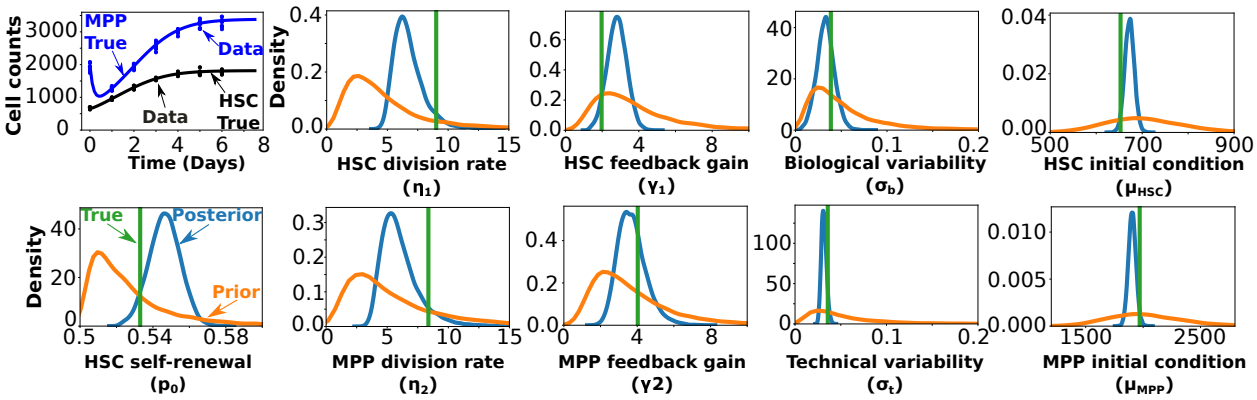

Figure 5.3: Parameter identification for synthetic data set. Top Left. Sampling from our prior distributions, we generate a synthetic data set from our ODE solution (shown in solid black and blue) where technical and biological noise is added according to our latent variables framework (equations 5.3 and 5.4), where the HSCs and MPPs are observed on 7 consecutive days with 7 mice replicates per day. Top-Bottom Right. Using this synthetic data, we show the prior vs posterior distributions comparisons for the ODE parameters and mean of the latent initial conditions (in original scale). The true values of the parameters used in the simulation are shown in green. The true parameters, prior distributions and 95% credible intervals are shown in Appendix Table B.3.

5.3 Results

5.3.1 Successful parameter identification

The complexity of the mechanistic model may give rise to issues in parameter identification alone. Adding the latent variables approach on top of our mechanistic model gives rise to latent trajectories with unknown initial conditions which might cause additional identifiability problems. Thus, we pose the question of parameter identifiability with this new model.

We show an illustration of our method in Figure 5.3. We generated a synthetic data set where the HSC and MPPs are observed 7 consecutive days with 7 replicates per day (49 mice). We inferred 5 ODE parameters (p_0, η_1,η_2,γ_1, γ_2), 2 initial conditions (μ_{HSC} and μ_{MPP}), 2 error terms (σ_t and σ_b), and 84 latent trajectories (93 parameters in total). We used Stan [12, 76] to obtain the posterior distribution of the model parameters and we observed a significant change from prior to posterior distributions (Figure 5.3). Also, we were are able to recover

7x7 Design	p_0	η_1	η_2	γ_1	γ_2	σ_t	σ_b	μ_1	μ_2
Mean Relative Bias	0.00	0.10	0.05	-0.08	-0.47	0.07	-0.31	-0.03	0.05
Mean Relative Width	0.04	0.60	0.72	1.34	2.52	0.36	1.98	0.10	0.11
Coverage	0.97	0.94	0.94	1.00	0.89	1.00	1.00	0.94	0.86

Table 5.1: Coverage metrics for 60 simulations using 7 mice replicates during 7 days. The relative bias, relative width and coverage are calculated according to equations 5.18, 5.21 and 5.19 shown on the Appendix.

the true parameter values within the 95% posterior probability intervals.

To show that parameter identification is achieved under this design (7 days x 7 replicates), we generated 60 synthetic data sets and we recovered the posterior distribution for the model parameters. We then calculated the mean relative bias, mean relative width and coverage of the 95% credible intervals for all the parameters as shown on Table 5.1. We observed high mean relative bias for γ_2 and σ_b parameters induced by the prior distribution selected.

We extended this approach to multiple designs in our simulation study as explained in our flow chart shown in figure 5.2. We show the identifiability metrics comparing designs with 3 mice replicates vs 7 mice replicates at different observation times as shown in Appendix Tables B.5 and B.6. Our results show that the model parameters are included in the 95% credible intervals. We also observe that designs with more data points allow for narrower posteriors, as quantified by the mean relative width, but higher bias is observed in some parameters. Since we show that we are able to identify the model parameters, we are now interested in showing and quantifying information gain using Bayesian utility theory. In particular, we are interested in finding the optimal experimental setup that provides the highest information gain for inferring the model parameters.

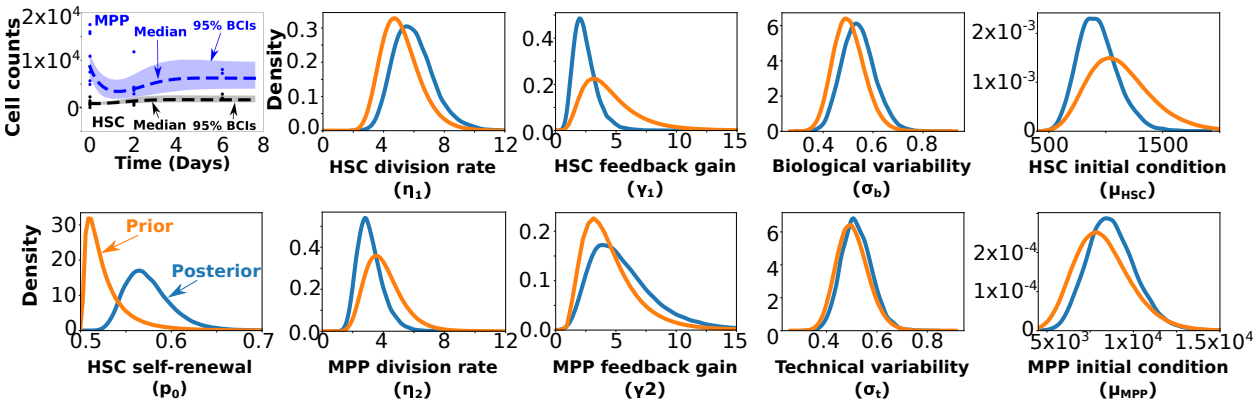

Figure 5.4: Parameter estimation for real data set. Top Left Preliminary data obtained from mice that were irradiated with 50 cGy and the records were obtained over a span of 7 days. We fit our hierarchical model to this dataset and we show the posterior distribution for the ODE solutions where the median (dashed) and 95% Bayesian credible intervals (bands) are shown. Top-Bottom Right. Prior vs posterior distributions for the fitted ODE parameters and the mean of the latent initial conditions. The summaries from priors and posterior distributions can be found on Appendix Table B.4

5.3.2 Model fit to perturbation experiment

As a proof of concept, we use our hierarchical model to fit and obtain posterior distributions from preliminary data in which stem cells are perturbed. In this experiment, low dose radiation (50 cGy) is applied to mice to observe the response dynamics. Mice were obtained at three time points, where each mouse was sacrificed and the bone marrow was extracted. The data points correspond to cell counts for HSCs and MPPs which were sorted by flow cytometry. Our data suggests the HSCs and MPPs return to equilibrium in under one week. Figure 5.4 shows the fit to our data for HSCs and MPPs over a 7 day interval (top left) and the prediction for the systems median dynamics (dashed). The 95% credible region (BCI) are also plotted. Although our fit can be shown to pass through the data points, most parameters did not exhibit a significant shift in the posterior distributions compared to the priors (Figure 5.4, Table B.4), suggesting little information gain for the parameter inference. We have shown the ability for this model to be identifiable through the coverage analysis shown in table 5.1. Knowing this, we can begin to ask which experimental setups have the ability optimize our information gain.

5.3.3 Utility grid search

To explore the amount of data necessary to maximize information gain in an experiment, we explore different experimental setups by varying the amount of mice collected per day and the timing of the measurements. Therefore, we define a design to represent some amount of mice observed at some sampling frequency. We sampled a finite number of experimental designs (70) to include a varying amount of mice over different observation days. Our design space is defined as following. The first observation day starts at day 0 right after radiation and more times are added until day 6. We also varied the number of mice observed per day: 3, 4, 5, 6, and 7 mice replicates.

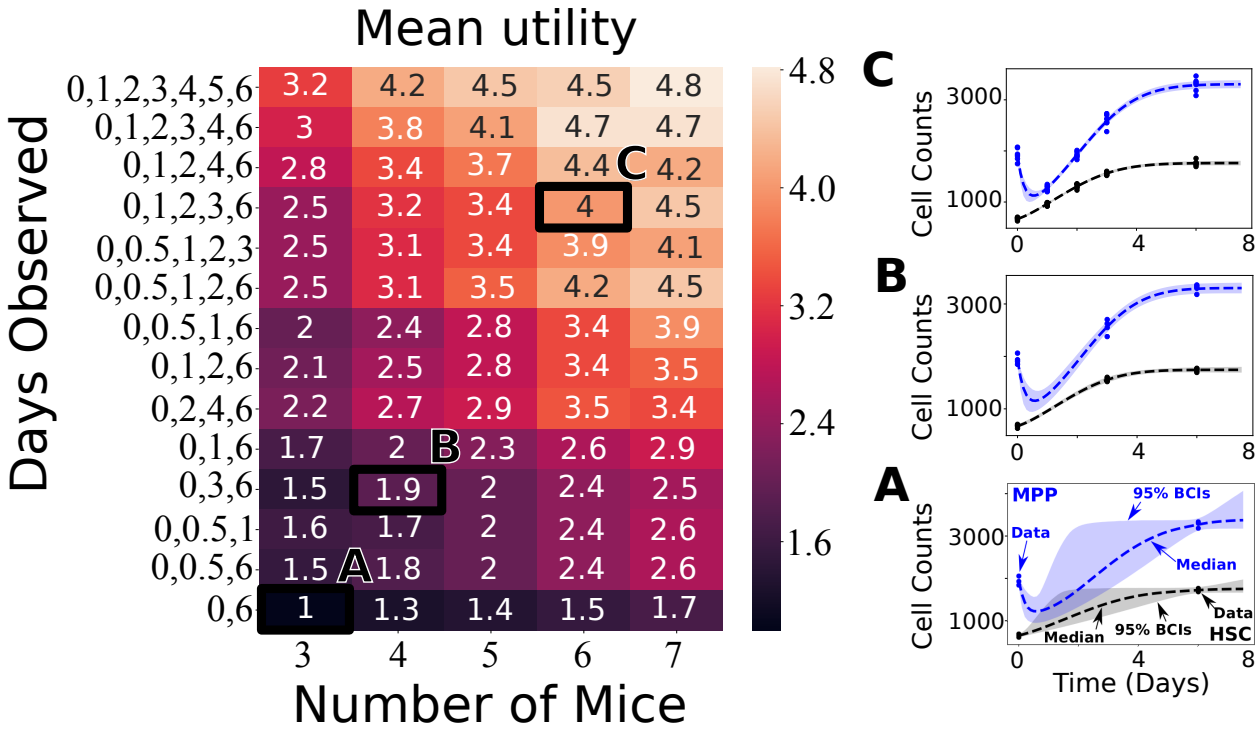

Figure 5.5: Mean utility values for sampled designs. Left Column. We show the mean utility heat map across all model parameters following the simulation process presented on Figure 5.2. Utility values correspond to a fold change to the minimum utility value (3 x (0,6)). The higher utility fold changes correspond to a greater information gain. The boxes show 3 designs with increasing mean utility value. Right Column. We show an illustration of the posterior distribution for the ODE solutions using the same synthetic dataset for the selected designs. We show the medians (dashed) and 95% Bayesian Credible Intervals (bands). The parameter values, priors and posterior summaries are available in the Appendix Table B.3

Exploring this finite experimental design space leads us to the calculation of the expected utilities by using 60 different synthetic datasets for each design (equation 5.14 and 5.15). We see in Figure 5.5 that the mean utilities show an increasing trend when the number of replicates and the frequency of sampling increases. The values correspond to a fold change with respect to the minimum utility (3 mice observed at day 0 and 6). To better understand the scale, we choose several designs (boxed), compare the mean utilities and plot the ODE credible regions.The plots provide a visual explanation on how higher information gain, as quantified by the mean utility values, correspond to designs with higher number of data points which provide narrower credible regions (Figure 5.5).

5.3.4 Parameter utility

The expected utility averages over parameters to give an average quantitative value for information gained. To get a better understanding of particular parameter contributions, we can also obtain the individual parameter utilities. In doing this, we can better understand how data affects identification of certain model parameters.

Obtaining parameter utility values also allows us to quantify whether or not a more data rich design is better. For example, if we focus on SC division (η_1), the parameter utilities suggest data should reside near the initial condition. If we look at the top four design rows for (η_1) in Figure 5.6, the marginal mean utility value does not change even though we are adding more mice and more observation days. This shows that for (η_1), the main contribution for the information gained is coming from times up to day 3. Similarly, if we focus on the HSC self renewal probability feedback parameter (γ_1), we observe a higher utility when data is obtained at later times. For the same parameter, we observe a lower marginal mean utility when looking at designs where day 6 is not included. We also observe that only a small amount of data is necessary to identify the initial conditions (μ_1 and μ_2). For the full set of

parameter utilities, refer to Figure 5.8.

As an example to utilize the parameter utilities, we can assume we've been given a budget to only buy 20 mice for our perturbation experiment. If we are to look at the boxed B, C, and D marginal mean utilities in figure 5.6, we can begin to compare how the distribution of mice affects the information gain for η_1 and γ_1. All of these designs sum up to 20 mice and we can compare prior vs posterior plots to visualize information gain. As mentioned previously, designs with data near the initial condition increases the utility for η_1 but has the opposite effect on γ_1. We can find a compromise design, shown in box C. For the full set of parameter prior vs posterior, please refer to figure 5.9.

5.4 Discussion

In this paper we present a new method for optimal experimental design for mechanistic mathematical models of hematopoiesis. Our method combines a latent variables approach with Bayesian utility theory for quantifying information gain about the model parameters using the KL-utility function between the prior vs posterior distributions. Our approach proved to be a suitable way to model empirical data from bone marrow experiments in mice, where each animal is measured one time, and successful parameter identification of the ODE parameters is possible on synthetic data. Using a small dataset from a perturbation experiment, we showed that our feedback model provides an adequate fit and it can predict the system response to the perturbation. Most elements of the ODE model can be partially identified.

Since information gain depends on the design chosen, we considered a finite grid of possible designs relevant for mice perturbation experiments and we determined that having more observation times with fewer mice replicates is more favorable for obtaining higher infor-

mation gain as measured by the overall expected utility values. Similarly, by using kernel density estimation, we found that some designs provide higher information gain for certain parameters only. For example, division rates are better informed by designs that include records at the beginning, right after the radiation was applied, but the feedback parameters are better identified by designs that have records closer to equilibrium.

As more data becomes readily available in the future, there are many avenues we can take in the optimal design framework. For example, using additional data on downstream cell types can provide insight on the response dynamics to a more complete hematopoietic system with additional feedback mechanisms. This provides more flexibility for incorporating new variables into our design of experiments approach like the depletion of downstream cell types. Furthermore, incorporating richer sources of data, such as *in vivo* experiments can produce serially sampled barcoded single cells which allows for lineage tracking and fate determination where multiple compartments and different scales can be considered.

From a modeling perspective, our framework can be extended to stochastic formulations of the hematopoietic system with explicit feedback terms that generalize well-known branching processes models [80]. One advantage is the possibility to model directly the cellular stochasticities and the heterogeneity of the components of a system. However, stochastic model fitting is generally a hard problem since the likelihood function is generally untractable and more research in this direction is needed.

In an experimental design, the budget, the number of days in the lab, and/or the types of experiments strongly influence how experiments are administered. The present study elucidates advantages in how one can gain information in mechanistic model parameters by placement of data. The constraints in the lab gives us ability to differentiate between designs through utility values and marginal utility values to optimize how much we learn from an experiment. In doing so, we have the ability to have a better understanding of the recovery mechanisms. In our work, we focused on the hematopoietic lineage under a perturbation

involving low dose radiation. More generally, our framework has the ability to extend to any mechanistic model to better understand experimental design under a cross sectional study assumption.

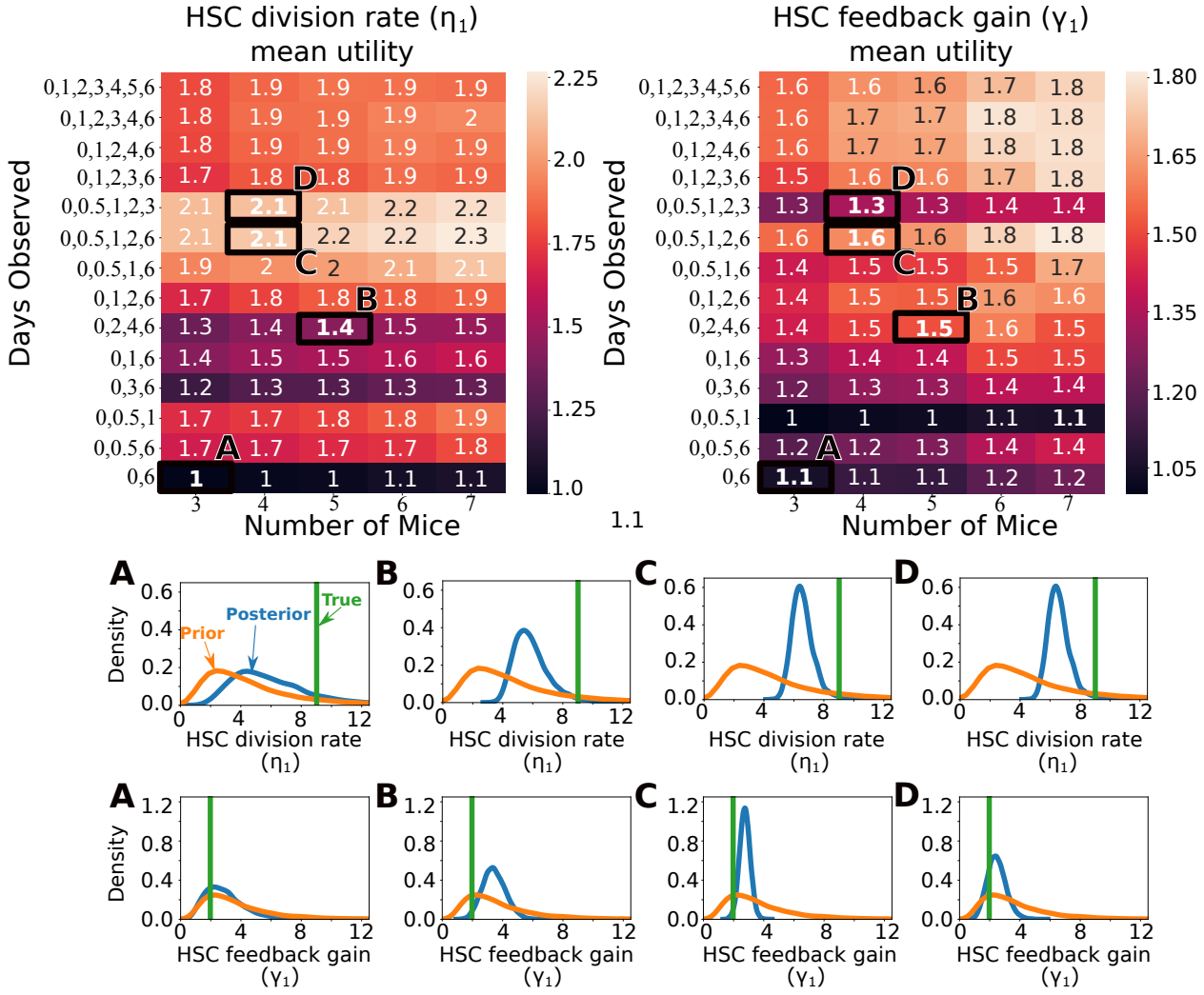

Figure 5.6: Individual parameter utility for HSC division rate and HSC feedback gain. Using the utility theory can also approximate the utility values for the individual ODE parameters. Top Row. We show two mean utility heat maps for the corresponding parameters. The mean utility values are given in fold change with respect to the design with the lowest mean utility value for in each heat map. Four designs are selected with increasing mean utility value where the sampling frequency varies (A-D). Bottom Row Prior vs posterior plots corresponding to changes in the boxed mean utilities (A-D). The true parameters values are shown in green

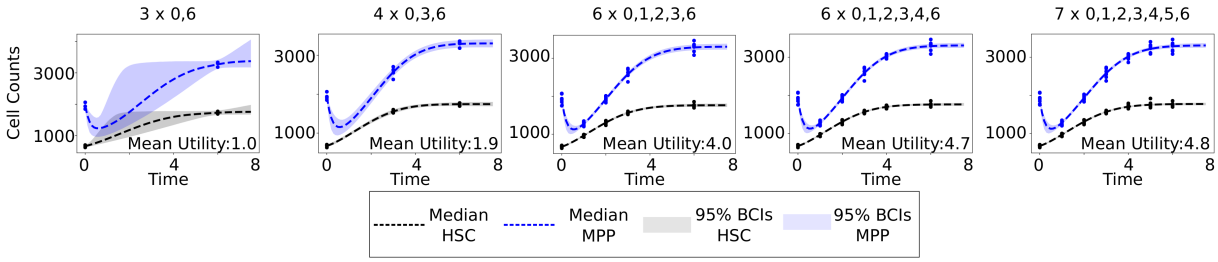

Figure 5.7: Sampling designs and comparing utility values. Designs are sampled from figure 5.5 and sample solutions are plotted to check the goodness of fit. The plots are showing the median ODE solution (dashed) with the 95% bayesian credible intervals (bands). The mean utilities correspond to values shown in figure 5.5.

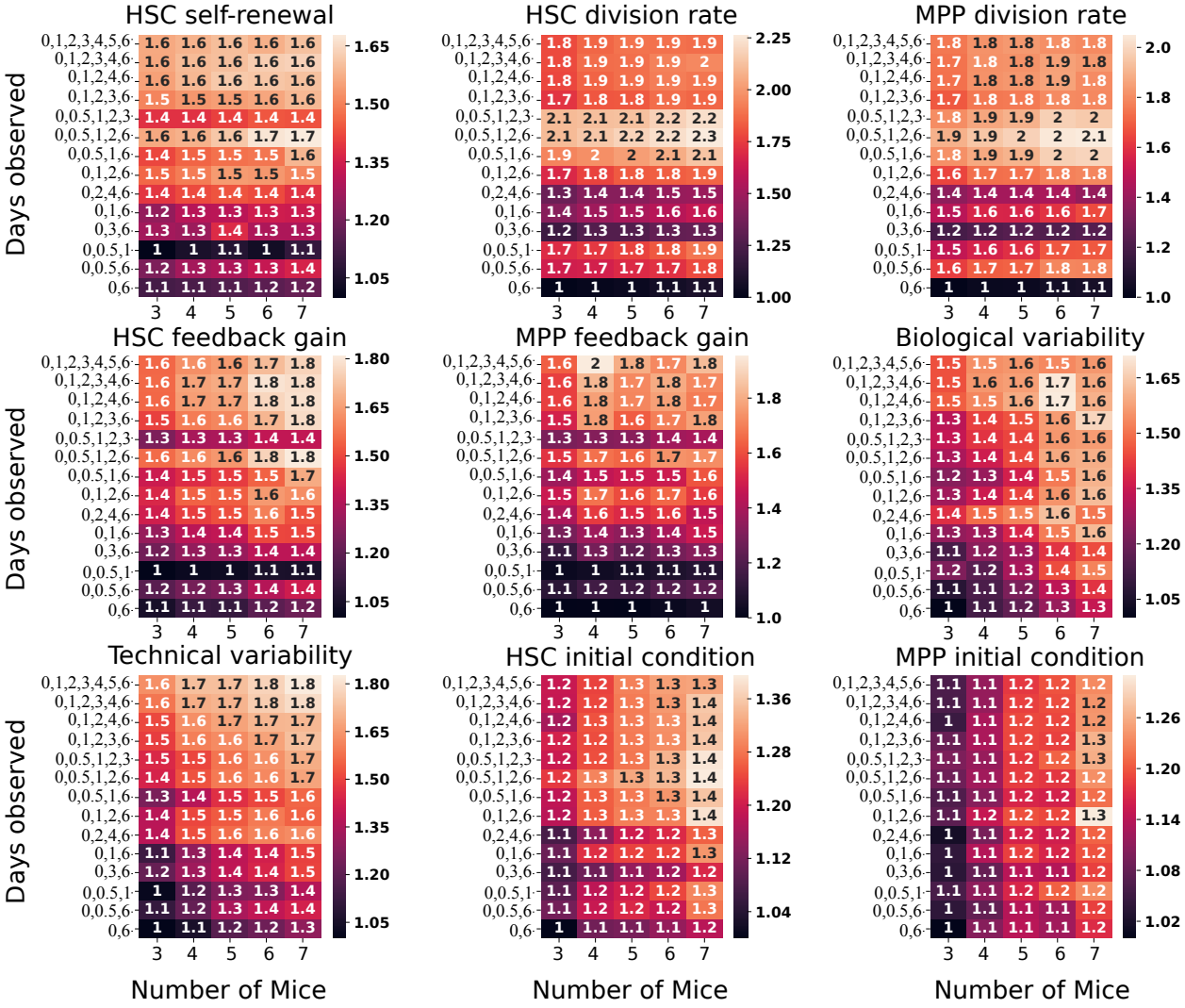

Figure 5.8: Parameter utility values for all ODE model parameters. All heat maps correspond to a parameter utility normalized to the minimum mean utility for each parameter. Each parameter utility grid is scaled differently.

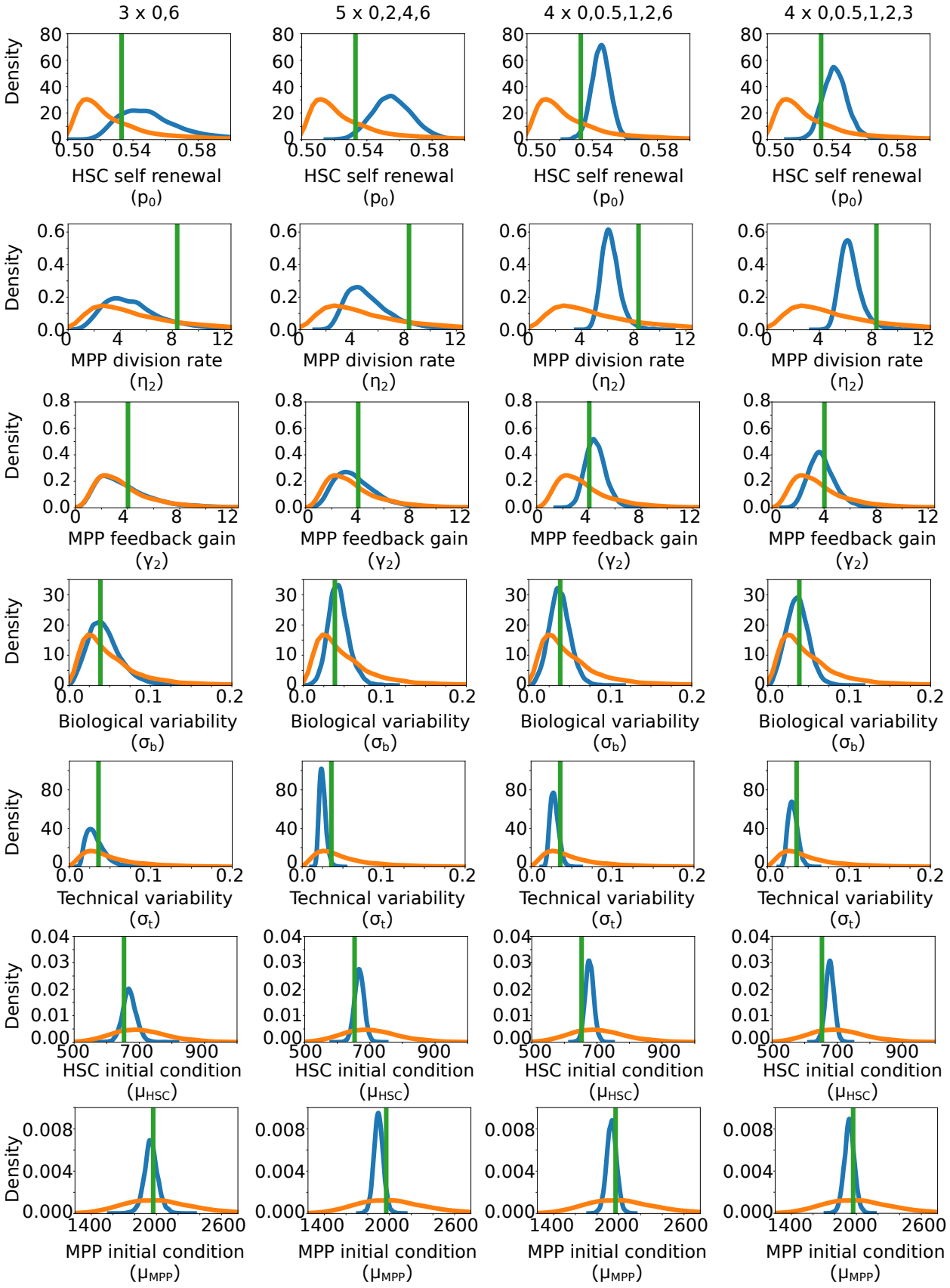

Figure 5.9: Prior vs posterior distributions for sampled designs. The remaining prior vs. posterior plots not shown in figure 5.6.

Chapter 6

Discussion

In this dissertation, we presented two projects that pertain to modeling hematopoiesis through lineage models. We developed a data driven model and showed how it could be extended to CML and how it can be used to optimize experimental design.

The current state of mathematical modeling in CML focuses on fitting therapy data from oversimplified lineage models. In many instances, there is a disjoint in modeling normal and leukemic blood development. Here, we developed more physiologically accurate, data-driven models of normal and CML hematopoiesis that incorporate lineage branching and feedback/feedforward regulation of cell proliferation and fate. Because feedback interactions in hematopoiesis (both normal and CML) are still poorly understood, we developed a general, automated approach to model selection that integrates known biological data to obtain plausible classes of models without knowing parameter values (DSA). We did this for normal hematopoiesis and then extended the models to account for changes introduced by CML and interactions between normal and leukemic cells.

Using our model, we suggest a plausible mechanism for primary resistance; an increase in leukemic HSCs causes an increase in stem cell self-renewal and a decrease in the stem and

progenitor cell division rate which results minimal response to TKI therapy. Using this allows us to better understand the emergence of non genetic resistance and use a strategy to maximize therapy response. For example, IFN-1 treatments have been around for a long time in CML, as it increases the division rate of stem cells. We can propose a combination therapy of TKI+IFN-1 to maximize the amount of stem cells cycling to minimize the stem cell load to give a better response. Interestingly, there is an ongoing trial in Germany in which they do just this [34]. The trial is set to finish in 2020.

Our method for deriving our data driven model is an iterative process and should be continually improved. There are two obvious paths one can take to improve the model predictability. The first would be to extend the DSA to consider systems with mixed positive and negative feedback on model parameters as it has been suggested by [63] to occur in CML. That is, we will consider systems in which feedback can be regulated by more than one cell type and more than one type of regulatory signaling factor. The second would be to include more physiological cell compartments. The minimal model of normal hematopoiesis that we would consider realistic consists of at least 5 compartments and one branch point (the MPP cell). The addition of these cell types will also provide a sort of pseudo spatial aspect where we can assume the terminal post mitotic cells will be in the peripheral blood and the rest will be located in the bone marrow.

Using the family of lineage models, we presented a framework in which we optimize information gain for perturbation experiments of hematopoiesis. Our method combines a latent variables approach in combination with utility theory to guide experimental designs. We have the ability to differentiate designs by quantifying posterior shifts of our model parameters. Using a utility theory approach also gives us the ability to quantify which designs optimize information gain for ODE parameters.

Knowing the mean utility value, we can begin to optimize experiments by minimizing days in the lab, minimizing the mice necessary to successfully identify model parameters. If we

are to look at the marginal utility values, we have the ability to obtain a secondary level of understanding. Knowing these values will give the option to biologists to design experiments based on parameters of interest. For example, we might be interested in understanding a feedback strength. Using the marginal utility values we can distribute resources to better identify the feedback gains.

We understand the designs used are a subset of the design space, but without a constraint, it was difficult to choose which designs to focus on. For example, if we were interested in finding all combinations of 20 mice, within reason, we can search through design space to optimize the experimental constraint.It is also important to point out these types of design studies are highly dependent on the type of priors, likelihoods and ODE models we use.

The process of experimental design allows for many variables when proposing experiments. The variables can vary from number of data points, where the data points are obtained, and what types of experiments. In our study, we focused on how we could optimize given a number of data points placed in a balanced way (distributing mice evenly across time). As an extension to this design, one could add a third or fourth dimension to design space. For example, it would be possible to vary the radiation dose at the initial time. Moreover, changing the target cells will also reflect the response and in turn reflect on utility values.

The projects mentioned provide a general framework for obtaining a data driven model of hematopoiesis/CML using known biological observations to gain insight on unmet clinical needs and experimental optimization. The two arms of the projects have the ability to work together to provide a parameterized data driven model to gain insight on hematopoiesis and CML biology. This will be explored in future work.

Bibliography

[1] A. Abe, Y. Minami, F. Hayakawa, K. Kitamura, Y. Nomura, M. Murata, A. Katsumi, H. Kiyoi, C. H. Jamieson, J. Y. Wang, et al. Retention but significant reduction of bcr-abl transcript in hematopoietic stem cells in chronic myelogenous leukemia after imatinib therapy. *International journal of hematology*, 88(5):471–475, 2008.

[2] W. Ahmed and R. A. Van Etten. Alternative approaches to eradicating the malignant clone in chronic myeloid leukemia: tyrosine-kinase inhibitor combinations and beyond. *ASH Education Program Book*, 2013(1):189–200, 2013.

[3] F. Arai, A. Hirao, M. Ohmura, H. Sato, S. Matsuoka, K. Takubo, K. Ito, G. Y. Koh, and T. Suda. Tie2/angiopoietin-1 signaling regulates hematopoietic stem cell quiescence in the bone marrow niche. *Cell*, 118(2):149–161, 2004.

[4] R. C. Aster, B. Borchers, and C. H. Thurber. *Parameter estimation and inverse problems*. Elsevier, 2018.

[5] M. Baccarani, M. W. Deininger, G. Rosti, A. Hochhaus, S. Soverini, J. F. Apperley, F. Cervantes, R. E. Clark, J. E. Cortes, F. Guilhot, et al. European leukemianet recommendations for the management of chronic myeloid leukemia: 2013. *Blood, The Journal of the American Society of Hematology*, 122(6):872–884, 2013.

[6] S. Basu, G. Hodgson, H.-H. Zhang, M. Katz, C. Quilici, and A. R. Dunn. "emergency" granulopoiesis in g-csf–deficient mice in response to candida albicans infection. *Blood*, 95(12):3725–3733, 2000.

[7] R. Bhatia, M. Holtz, N. Niu, R. Gray, D. S. Snyder, C. L. Sawyers, D. A. Arber, M. L. Slovak, and S. J. Forman. Persistence of malignant hematopoietic progenitors in chronic myelogenous leukemia patients in complete cytogenetic remission following imatinib mesylate treatment. *Blood*, 101(12):4701–4707, 2003.

[8] B. M. Bolker. *Ecological models and data in R*. Princeton University Press, 2008.

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

[10] G. Buzi, A. D. Lander, and M. Khammash. Cell lineage branching as a strategy for proliferative control. *BMC biology*, 13(1):13, 2015.

[11] J. Cao, G. F. Fussmann, and J. O. Ramsay. Estimating a predator-prey dynamical model with the parameter cascades method. *Biometrics*, 64(3):959–967, 2008.

[12] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[13] S. N. Catlin, J. L. Abkowitz, and P. Guttorp. Statistical inference in a two-compartment model for hematopoiesis. *Biometrics*, 57(2):546–553, 2001.

[14] K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

[15] B. Chereda and J. V. Melo. Natural course and biology of cml. *Annals of hematology*, 94(2):107–121, 2015.

[16] G. Clapp and D. Levy. A review of mathematical models for leukemia and lymphoma. *Drug Discov Today Dis Models*, 16:1–6, 2015.

[17] A. R. Cook, G. J. Gibson, and C. A. Gilligan. Optimal observation times in experimental epidemic processes. *Biometrics*, 64(3):860–868, 2008.

[18] A. S. Corbin, A. Agarwal, M. Loriaux, J. Cortes, M. W. Deininger, and B. J. Druker. Human chronic myeloid leukemia stem cells are insensitive to imatinib despite inhibition of bcr-abl activity. *The Journal of clinical investigation*, 121(1):396–409, 2011.

[19] M. B. Dehideniya, C. C. Drovandi, and J. M. McGree. Optimal bayesian design for discriminating between models with intractable likelihoods in epidemiology. *Computational Statistics & Data Analysis*, 124:277–297, 2018.

[20] S. Faderl, M. Talpaz, Z. Estrov, S. O'Brien, R. Kurzrock, and H. M. Kantarjian. The biology of chronic myeloid leukemia. *New England Journal of Medicine*, 341(3):164–172, 1999.

[21] R. A. Fasani and M. A. Savageau. Automated construction and analysis of the design space for biochemical systems. *Bioinformatics*, 26(20):2601–2609, 2010.

[22] C. Foley and M. C. Mackey. Dynamic hematological disease: a review. *Journal of mathematical biology*, 58(1-2):285–322, 2009.

[23] Y. Fong, P. Guttorp, and J. Abkowitz. Bayesian inference and model choice in a hidden stochastic two-compartment model of hematopoietic stem cell fate decisions. *The annals of applied statistics*, 3(4):1696, 2009.

[24] C. Gambacorti-Passerini, L. Antolini, F.-X. Mahon, F. Guilhot, M. Deininger, C. Fava, A. Nagler, C. M. Della Casa, E. Morra, E. Abruzzese, et al. Multicenter independent assessment of outcomes in chronic myeloid leukemia patients treated with imatinib. *Journal of the National Cancer Institute*, 103(7):553–561, 2011.

[25] A. Gelman. Prior Choice Recommendations, 2019.

[26] M. Girolami. Bayesian inference for differential equations. *Theoretical Computer Science*, 408(1):4–16, 2008.

[27] D. Golinelli, P. Guttorp, and J. Abkowitz. Bayesian inference in a hidden stochastic two-compartment model for feline hematopoiesis. *Mathematical Medicine and Biology*, 23(3):153–172, 2006.

[28] S. M. Graham, H. G. Jørgensen, E. Allan, C. Pearson, M. J. Alcorn, L. Richmond, and T. L. Holyoake. Primitive, quiescent, philadelphia-positive stem cells from patients with chronic myeloid leukemia are insensitive to sti571 in vitro. *Blood*, 99(1):319–325, 2002.

[29] D. S. Grenda, M. Murakami, J. Ghatak, J. Xia, L. A. Boxer, D. Dale, M. C. Dinauer, and D. C. Link. Mutations of the ela2 gene found in patients with severe congenital neutropenia induce the unfolded protein response and cellular apoptosis. *Blood*, 110(13):4179–4187, 2007.

[30] Q. F. Gronau, A. Sarafoglou, D. Matzke, A. Ly, U. Boehm, M. Marsman, D. S. Leslie, J. J. Forster, E.-J. Wagenmakers, and H. Steingroever. A tutorial on bridge sampling. *Journal of mathematical psychology*, 81:80–97, 2017.

[31] C. Han and K. Chaloner. Bayesian experimental design for nonlinear mixed-effects models with application to hiv dynamics. *Biometrics*, 60(1):25–33, 2004.

[32] B. Hanfstein, M. Müller, R. Hehlmann, P. Erben, M. Lauseker, A. Fabarius, S. Schnittger, C. Haferlach, G. Göhring, U. Proetel, et al. Early molecular and cytogenetic response is predictive for long-term progression-free and overall survival in chronic myeloid leukemia (cml). *Leukemia*, 26(9):2096, 2012.

[33] M. Henkes, van der Kuip Heiko, and A. W. E. Therapeutic options for chronic myeloid leukemia: Focus on imatinib (glivec®, gleevectm). *Therapeutics and Clinical Risk Management*, pages 163–187, 2008.

[34] A. Hochhaus. Tki and interferon alpha evaluation initiated by the german chronic myeloid leukemia study group - the tiger study (tiger).

[35] M. Horn, I. Glauche, M. C. Müller, R. Hehlmann, A. Hochhaus, M. Loeffler, and I. Roeder. Model-based decision rules reduce the risk of molecular relapse after cessation of tyrosine kinase inhibitor therapy in chronic myeloid leukemia. *Blood*, 121(2):378–384, 2013.

[36] X. Huan and Y. M. Marzouk. Simulation-based optimal bayesian experimental design for nonlinear systems. *Journal of Computational Physics*, 232(1):288–317, 2013.

[37] Y. Huang, H. Wu, and E. P. Acosta. Hierarchical bayesian inference for hiv dynamic differential equation models incorporating multiple treatment factors. *Biometrical Journal*, 52(4):470–486, 2010.

[38] E. Jabbour and H. Kantarjian. Chronic myeloid leukemia: 2016 update on diagnosis, therapy, and monitoring. *American journal of hematology*, 91(2):252–265, 2016.

[39] E. Jabbour, H. Kantarjian, and J. Cortes. Use of second-and third-generation tyrosine kinase inhibitors in the treatment of chronic myeloid leukemia: an evolving treatment paradigm. *Clinical Lymphoma Myeloma and Leukemia*, 15(6):323–334, 2015.

[40] C. H. Jamieson, L. E. Ailles, S. J. Dylla, M. Muijtjens, C. Jones, J. L. Zehnder, J. Gotlib, K. Li, M. G. Manz, A. Keating, et al. Granulocyte–macrophage progenitors as candidate leukemic stem cells in blast-crisis cml. *New England Journal of Medicine*, 351(7):657–667, 2004.

[41] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed ¡today¿].

[42] H. G. Jřrgensen, M. Copland, E. K. Allan, X. Jiang, A. Eaves, C. Eaves, and T. L. Holyoake. Intermittent exposure of primitive quiescent chronic myeloid leukemia cells to granulocyte-colony stimulating factor in vitro promotes their elimination by imatinib mesylate. *Clinical Cancer Research*, 12(2):626–633, 2006.

[43] R. K. *Blood: Physiology and Circulation*. Britannica Educational, 2010.

[44] K. Kaushansky. Thrombopoietin and its receptor in normal and neoplastic hematopoiesis. *Thrombosis journal*, 14(1):40, 2016.

[45] A. Kim, D. Stachura, and D. Traver. Cell signaling pathways involved in hematopoietic stem cell specification. *Experimental cell research*, 329(2):227, 2014.

[46] A. D. Lander, K. K. Gokoffski, F. Y. Wan, Q. Nie, and A. L. Calof. Cell lineages and the logic of proliferative control. *PLoS biology*, 7(1):e1000015, 2009.

[47] D. V. Lindley et al. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

[48] J. G. Lomnitz and M. A. Savageau. Phenotypic deconstruction of gene circuitry. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 23(2):025108, 2013.

[49] J. G. Lomnitz and M. A. Savageau. Rapid discrimination among putative mechanistic models of biochemical systems. *Scientific reports*, 6:32375, 2016.

[50] R. MA and S. T. Hematopoiesis. *Cold Spring Harbor Perspectives in Biology*, 4:12, 2012.

[51] K. Maeda, Y. Baba, Y. Nagai, K. Miyazaki, A. Malykhin, K. Nakamura, P. W. Kincade, N. Sakaguchi, and K. M. Coggeshall. Il-6 blocks a discrete early step in lymphopoiesis. *Blood*, 106(3):879–885, 2005.

[52] F.-X. Mahon, D. Réa, J. Guilhot, F. Guilhot, F. Huguet, F. Nicolini, L. Legros, A. Charbonnier, A. Guerci, B. Varet, et al. Discontinuation of imatinib in patients with chronic myeloid leukaemia who have maintained complete molecular remission for at least 2 years: the prospective, multicentre stop imatinib (stim) trial. *The lancet oncology*, 11(11):1029–1035, 2010.

[53] E. Manesso, J. Teles, D. Bryder, and C. Peterson. Dynamical modelling of haematopoiesis: an integrated view over the system in homeostasis and under perturbation. *Journal of the Royal Society Interface*, 10(80):20120817, 2013.

[54] A. Marciniak-Czochra, T. Stiehl, A. Ho, W. Jäger, and W. Wagner. Modeling of asymmetric cell division in hematopoietic stem cells-regulation of self-renewal is essential for efficient repopulation. *Stem Cells and Development*, 18(3):377–386, 2009.

[55] D. Marin, A. R. Ibrahim, C. Lucas, G. Gerrard, L. Wang, R. M. Szydlo, R. E. Clark, J. F. Apperley, D. Milojkovic, M. Bua, et al. Assessment of bcr-abl1 transcript levels at 3 months is the only requirement for predicting outcome for patients with chronic myeloid leukemia treated with tyrosine kinase inhibitors. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*, 30(3):232–238, 2012.

[56] F. Michor, T. P. Hughes, Y. Iwasa, S. Branford, N. P. Shah, C. L. Sawyers, and M. A. Nowak. Dynamics of chronic myeloid leukaemia. *Nature*, 435(7046):1267, 2005.

[57] A. Nemoto, M. Masaaki, and K. Yamaoka. A bayesian approach for population pharmacokinetic modeling of alcohol in japanese individuals. *Current Therapeutic Research*, 84:42–49, 2017.

[58] A. Olsson, M. Venkatasubramanian, V. K. Chaudhri, B. J. Aronow, N. Salomonis, H. Singh, and H. L. Grimes. Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature*, 537(7622):698, 2016.

[59] A. L. Petzer, C. J. Eaves, P. M. Lansdorp, L. Ponchio, M. J. Barnett, and A. C. Eaves. Characterization of primitive subpopulations of normal and leukemic cells present in the blood of patients with newly diagnosed as well as established chronic myeloid leukemia. *Blood*, 88(6):2162–2171, 1996.

[60] E. M. Pietras, C. Mirantes-Barbeito, S. Fong, D. Loeffler, L. V. Kovtonyuk, S. Zhang, R. Lakshminarasimhan, C. P. Chin, J.-M. Techner, B. Will, et al. Chronic interleukin-1 exposure drives haematopoietic stem cells towards precocious myeloid differentiation at the expense of self-renewal. *Nature cell biology*, 18(6):607, 2016.

[61] A. Quintas-Cardama and J. Cortes. Molecular biology of bcr-abl1–positive chronic myeloid leukemia. *Blood*, 113(8):1619–1630, 2009.

[62] T. Reya, A. W. Duncan, L. Ailles, J. Domen, D. C. Scherer, K. Willert, L. Hintz, R. Nusse, and I. L. Weissman. A role for wnt signalling in self-renewal of haematopoietic stem cells. *Nature*, 423(6938):409, 2003.

[63] D. Reynaud, E. Pietras, K. Barry-Holson, A. Mir, M. Binnewies, M. Jeanne, O. Sala-Torra, J. P. Radich, and E. Passegué. Il-6 controls leukemic multipotent progenitor cell fate and contributes to chronic myelogenous leukemia development. *Cancer cell*, 20(5):661–673, 2011.

[64] 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, 2006.

[65] B. Rosenbaum, M. Raatz, G. Weithoff, G. F. Fussmann, and U. Gaedke. Estimating parameters from multiple time series of population dynamics using bayesian inference. *bioRxiv*, page 392449, 2018.

[66] D. M. Ross, S. Branford, J. F. Seymour, A. P. Schwarer, C. Arthur, D. T. Yeung, P. Dang, J. M. Goyne, C. Slader, R. J. Filshie, et al. Safety and efficacy of imatinib cessation for cml patients with stable undetectable minimal residual disease: results from the twister study. *Blood*, 122(4):515–522, 2013.

[67] E. G. Ryan, C. C. Drovandi, J. M. McGree, and A. N. Pettitt. A review of modern computational algorithms for bayesian optimal design. *International Statistical Review*, 84(1):128–154, 2016.

[68] S. Sarrazin and M. Sieweke. Integration of cytokine and transcription factor signals in hematopoietic stem cell commitment. In *Seminars in immunology*, volume 23, pages 326–334. Elsevier, 2011.

[69] M. A. Savageau, P. M. Coelho, R. A. Fasani, D. A. Tolla, and A. Salvador. Phenotypes and tolerances in the design space of biochemical systems. *Proceedings of the National Academy of Sciences*, 106(16):6435–6440, 2009.

[70] D. W. Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[71] M. A. Stark, Y. Huo, T. L. Burcin, M. A. Morris, T. S. Olson, and K. Ley. Phagocytosis of apoptotic neutrophils regulates granulopoiesis via il-23 and il-17. *Immunity*, 22(3):285–294, 2005.

[72] T. Stiehl, N. Baran, A. D. Ho, and A. Marciniak-Czochra. Clonal selection and therapy resistance in acute leukaemias: mathematical modelling explains different proliferation patterns at diagnosis and relapse. *Journal of The Royal Society Interface*, 11(94):20140079, 2014.

[73] M.-R. Tamadon and S. S. Beladi-Mousavi. Erythropoietin; a review on current knowledge and new concepts. *Journal of renal injury prevention*, 2(4):119, 2013.

[74] M. Tang, M. Gonen, A. Quintas-Cardama, J. Cortes, H. Kantarjian, C. Field, T. P. Hughes, S. Branford, and F. Michor. Dynamics of chronic myeloid leukemia response to long-term targeted therapy reveal treatment effects on leukemic stem cells. *Blood*, 118(6):1622–1631, 2011.

[75] S. D. Team. Stan modeling language: User's guide and reference manual. *Version 2.12*, 2015.

[76] S. D. Team. Pystan: the python interface to stan, version 2.17.1.0., 2018.

[77] R. S. Welner, G. Amabile, D. Bararia, A. Czibere, H. Yang, H. Zhang, L. L. D. F. Pontes, M. Ye, E. Levantini, A. Di Ruscio, et al. Treatment of chronic myelogenous leukemia by blocking cytokine alterations found in normal stem and progenitor cells. *Cancer cell*, 27(5):671–681, 2015.

[78] B. Werner, J. G. Scott, A. Sottoriva, A. R. Anderson, A. Traulsen, and P. M. Altrock. The cancer stem cell fraction in hierarchically organized tumors can be estimated using mathematical modeling and patient-specific treatment trajectories. *Cancer research*, 76(7):1705–1713, 2016.

[79] J. Xu, S. Koelle, P. Guttorp, C. Wu, C. E. Dunbar, J. L. Abkowitz, and V. N. Minin. Statistical inference in partially observed stochastic compartmental models with application to cell lineage tracking of in vivo hematopoiesis. *arXiv preprint arXiv:1610.07550*, 2016.

[80] J. Xu, S. Koelle, P. Guttorp, C. Wu, C. E. Dunbar, J. L. Abkowitz, and V. N. Minin. Statistical inference in partially observed stochastic compartmental models with application to cell lineage tracking of in vivo hematopoiesis. *Annals of Applied Statistics*, in press, 2019.

[81] D. T. Yeung, M. P. Osborn, D. L. White, S. Branford, J. Braley, A. Herschtal, M. Kornhauser, S. Issa, D. K. Hiwase, M. Hertzberg, et al. Tidel-ii: first-line use of imatinib in cml with early switch to nilotinib for failure to achieve time-dependent molecular targets. *Blood*, 125(6):915–923, 2015.

[82] J. F. Zhang, N. E. Papanikolaou, T. Kypraios, and C. C. Drovandi. Optimal experimental design for predator–prey functional response experiments. *Journal of The Royal Society Interface*, 15(144):20180186, 2018.

[83] B. O. Zhou, L. Ding, and S. J. Morrison. Hematopoietic stem and progenitor cells regulate the regeneration of their niche by secreting angiopoietin-1. *Elife*, 4:e05521, 2015.

Appendices

A Mathematical modeling of malignant myelopoiesis and targeted therapy

Design space analysis S-Systems We can summarize the design space shown in figure 3.5 by displaying the S-systems. The corresponding S-systems all share the form shown in equations A.1-A.4. The differences lie in the inequality constraints shown in table 1. A noticeable difference in the domains lie in x_{γ_3}. If $1 > \gamma_3 x_{TD_m}$, we observe the domains are odd and are the bottom 3 in figure 3.5. These three have oscillatory behavior. On the other hand, if we look at $1 < \gamma_3 x_{TD_m}$, the domains correspond to the top 3 boundaries and are stable nodes. This suggests the addition of the feedback adds stability and negates the overshooting we see in the cases without it.

	1	2	3	4	5	6
x_{γ_1}	$\gamma_1 x_P > 1$	$\gamma_1 x_P > 1$	$\gamma_1 x_P > 1$	$\gamma_1 x_P > 1$	$\gamma_1 x_P > 1$	$\gamma_1 x_P > 1$
x_{γ_2}	$1 > \gamma_2 x_S$	$1 > \gamma_2 x_S$	$1 > \gamma_2 x_S$	$1 > \gamma_2 x_S$	$1 < \gamma_2 x_S$	$1 < \gamma_2 x_S$
x_{γ_3}	$1 < \gamma_3 x_{TD_m}$	$1 > \gamma_3 x_{TD_m}$	$1 < \gamma_3 x_{TD_m}$	$1 > \gamma_3 x_{TD_m}$	$1 < \gamma_3 x_{TD_m}$	$1 > \gamma_3 x_{TD_m}$
x_{γ_4}	$1 > \gamma_2 x_{TD_m}$	$1 > \gamma_4 x_{TD_m}$	$1 < \gamma_4 x_{TD_m}$	$1 < \gamma_4 x_{TD_m}$	$1 < \gamma_4 x_{TD_m}$	$1 < \gamma_4 x_{TD_m}$
x_{γ_5}	$1 > \gamma_5 x_S$	$1 > \gamma_5 x_S$	$1 > \gamma_5 x_S$	$1 > \gamma_5 x_S$	$1 < \gamma_5 x_S$	$1 < \gamma_5 x_S$

Table A.1: S-Systems for design shown in figure 3.6

parameter	Before therapy	After therapy
d_0	0.003	No change
d_1	0.0075	No change
d_2	0.05	No change
p_x	0.1	No change
r_x	0.1	No change
a_x	1	No change
b_x	5	No change
p_y	0.01	No change
r_y	0.01	No change
a_y	2	0.02
b_y	10	0.1

Table A.2: Michor Model Parameters used in figure 3.7

$$x'_S = \frac{2\eta_1 p_0 x_S}{x_{\gamma_1} x_{\gamma_2}} - \frac{\eta_1 x_S}{x_{\gamma_2}} \tag{A.1}$$

$$x'_P = \frac{2\eta_1 x_S}{x_{\gamma_2}} - \frac{\eta_2 x_P}{x_{\gamma_5}} \tag{A.2}$$

$$x'_{TD_l} = \frac{2\eta_2 q_1 x_P}{x_{\gamma_4} x_{\gamma_5}} - d_l x_{TD_l} \tag{A.3}$$

$$x_{TD_m} = \frac{2\eta_2 x_P}{x_{\gamma_5}} - d_m x_{TD_m} \tag{A.4}$$

B Bayesian optimal experimental design for mathematical models of hematopoiesis

	Prior			Posterior			True
Parameter	2.5%	50%	97.5%	2.5%	50%	97.5%	
p_0	0.50	0.52	0.60	0.53	0.55	0.54	0.53
η_1^*	1.00	3.84	15.11	5.00	6.51	9.85	9.02
η_2^*	1.00	4.50	20.2	3.98	5.72	10.1	8.37
γ_1	1.00	3.15	10	1.77	2.81	3.84	1.97
γ_2	1.00	3.15	10	2.44	3.64	5.51	4.03
σ_t	0.01	0.04	0.18	0.01	0.03	0.05	0.04
σ_b	0.01	0.04	0.18	0.03	0.03	0.04	0.04
μ_1	549	700	890	652	672	693	653
μ_2	1462	1996	2743	1839	1904	1970	1970

Table B.3: We observe clear information gain as shown in the reduction of the 95% credible intervals and the shift of the posterior medians

Utility Validation

We can measure information gain in other ways to validate a parameter utility value. Two ways we can observe this is by (1) looking at the reduction of variance in the posterior or (2) measuring how much the prior and posterior are overlapping. We can use the coefficient of variation to measure the reduction of variance. To do this, we plot the coefficient of variation and plot vs the utlity value. If our intuition is correct, we will see a negative correlation between these two values. As we increase utility value, we see a decrease in the coefficient of variation (See figure B.1). We can also observe information gain by the reduction of overlap between the prior and posterior. If we are to plot the percent overlap between the prior and posterior, we see similar trends when comparing to the parameter utility (Figure B.2).

	Prior			Posterior		
Parameter	2.5%	50%	97.5%	2.5%	50%	97.5%
p_0	0.50	0.52	0.60	0.54	0.57	0.64
η_1^*	3.06	5	8.16	3.79	5.85	9.07
η_2^*	2.14	3.86	6.94	1.84	3.01	5.07
γ_1	1.50	4	10.7	1.06	2.29	4.88
γ_2	1.50	4	10.7	1.89	5.09	12.9
σ_t	0.39	0.5	0.64	0.41	0.52	0.65
σ_b	0.39	0.5	0.64	0.42	0.54	0.68
μ_1	672	1097	1791	634	915	1317
μ_2	5473	8103	11988	6280	8653	11810

Table B.4: 95% credible intervals for real data

Parameter	3 replicates	0,6	0,0.5,6	0,0.5,1	0,3,6	0,1,6	0,2,4,6	0,1,2,6	0,0.5,1,6	0,0.5,1,2,6	0,0.5,1,2,3	0,1,2,3,6	0,1,2,4,6	0,1,2,3,4,6	0,1,2,3,4,5,6
p_0	Mean RelBias	0.0	-0.0	-0.0	0.0	0.01	0.0	0.0	-0.0	-0.0	-0.01	0.0	0.0	-0.0	-0.0
	Coverage	0.91	0.97	1.0	0.96	0.97	0.93	0.93	1.0	0.93	0.89	0.92	0.91	0.92	0.93
	Mean RelWidth	0.05	0.04	0.05	0.04	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.03
η_1^*	Mean RelBias	-0.12	-0.02	0.09	-0.08	-0.07	-0.01	0.05	0.03	0.08	0.02	0.04	0.06	0.06	0.04
	Coverage	0.97	1.0	1.0	1.0	1.0	0.9	0.88	1.0	0.79	0.84	0.96	0.79	0.87	0.84
	Mean RelWidth	11.07	5.63	6.69	7.11	5.49	5.27	4.45	4.04	3.31	3.69	3.82	3.61	3.37	3.26
η_2^*	Mean RelBias	-0.06	0.03	0.1	-0.06	-0.05	-0.05	0.02	0.04	0.07	-0.01	0.02	0.04	0.04	0.02
	Coverage	0.94	1.0	1.0	0.96	1.0	0.9	0.93	0.97	0.83	0.84	0.96	0.88	0.9	0.89
	Mean RelWidth	9.97	4.56	7.65	9.63	7.02	8.9	6.33	3.56	3.04	2.92	5.95	5.9	4.99	5.36
γ_1	Mean RelBias	-0.29	-0.3	-0.59	-0.28	-0.11	-0.15	-0.1	-0.24	-0.23	-0.4	-0.19	-0.07	-0.16	-0.16
	Coverage	0.91	0.97	0.89	0.96	0.97	0.9	0.9	0.97	0.81	0.91	0.92	0.86	0.88	0.85
	Mean RelWidth	6.98	6.33	6.07	5.81	5.47	4.95	4.94	4.67	4.58	5.69	4.77	4.41	4.18	4.18
γ_2	Mean RelBias	-0.2	-0.41	-0.02	-0.24	-0.2	-0.22	-0.37	-0.31	-0.3	-0.41	-0.25	-0.29	-0.27	-0.3
	Coverage	0.94	0.94	1.0	0.96	0.97	0.9	0.95	0.95	0.95	0.96	0.98	0.95	0.96	0.95
	Mean RelWidth	9.01	8.29	9.48	7.94	7.7	6.97	7.38	7.41	7.19	8.41	6.88	6.63	6.5	6.39
σ_b	Mean RelBias	-0.78	-0.68	-0.22	-0.61	-1.06	-0.73	-0.57	-0.71	-0.51	-0.35	-0.37	-0.64	-0.65	-0.52
	Coverage	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Mean RelWidth	0.12	0.11	0.08	0.1	0.1	0.1	0.1	0.09	0.08	0.08	0.08	0.09	0.08	0.08
σ_t	Mean RelBias	-0.03	0.01	0.19	0.07	0.25	0.24	0.26	0.21	0.2	0.22	0.24	0.29	0.28	0.28
	Coverage	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.93	0.92	0.88	0.85	0.85
	Mean RelWidth	0.08	0.07	0.05	0.06	0.06	0.05	0.05	0.05	0.04	0.04	0.04	0.04	0.03	0.03

Table B.5: Coverage, precision and accuracy metrics for designs with 3 mice replicates per observation time

Parameter	7 replicates	0,6	0,0.5,6	0,0.5,1	0,3,6	0,1,6	0,2,4,6	0,1,2,6	0,0.5,1,6	0,0.5,1,2,6	0,0.5,1,2,3	0,1,2,3,6	0,1,2,4,6	0,1,2,3,4,6	0,1,2,3,4,5,6
p_0	Mean RelBias	0.01	0.01	-0.01	0.0	0.01	-0.0	0.0	0.0	0.0	-0.0	0.0	0.0	0.0	-0.0
	Coverage	0.97	0.97	1.0	0.98	1.0	0.96	0.97	1.0	1.0	1.0	0.94	0.96	0.97	0.97
	Mean RelWidth	0.04	0.04	0.04	0.04	0.03	0.03	0.02	0.02	0.02	0.03	0.02	0.02	0.02	0.02
η_1	Mean RelBias	0.07	0.02	0.05	0.06	0.03	0.1	0.07	0.04	0.06	0.04	0.07	0.07	0.06	0.1
	Coverage	0.97	0.97	1.0	0.94	0.97	0.93	0.97	0.97	0.94	1.0	0.97	0.96	1.0	0.94
	Mean RelWidth	10.31	5.05	4.63	5.41	3.92	4.36	2.94	2.99	2.19	2.6	2.62	3.07	3.26	3.36
η_2^*	Mean RelBias	0.08	0.02	-0.02	0.02	-0.02	0.06	0.03	0.02	0.03	0.01	0.02	0.04	0.02	0.05
	Coverage	0.95	0.97	1.0	0.92	0.97	0.95	0.97	1.0	1.0	1.0	0.97	0.96	1.0	0.94
	Mean RelWidth	7.34	4.35	5.09	8.0	5.19	6.79	4.12	3.08	2.48	3.08	3.67	4.39	4.7	4.38
γ_1	Mean RelBias	-0.07	0.03	-0.46	-0.05	0.05	-0.05	-0.0	-0.0	0.02	-0.36	0.04	0.01	-0.01	-0.08
	Coverage	0.92	0.95	0.89	0.96	0.91	0.96	0.92	0.97	0.98	0.9	1.0	0.98	0.97	1.0
	Mean RelWidth	5.88	4.65	6.14	4.87	4.35	4.29	3.48	3.47	3.18	4.83	3.0	3.22	3.27	3.38
γ_2	Mean RelBias	-0.3	-0.27	-0.36	-0.41	-0.45	-0.37	-0.26	-0.24	-0.27	-0.23	-0.48	-0.31	-0.33	-0.47
	Coverage	0.92	0.97	0.93	0.96	0.88	0.93	0.97	0.94	0.94	0.96	0.88	0.92	0.95	0.89
	Mean RelWidth	9.64	8.33	7.95	7.86	7.62	6.8	5.69	6.22	5.82	6.84	6.2	5.99	6.37	6.1
σ_b	Mean RelBias	-0.74	-0.51	-0.39	-0.63	-0.99	-0.47	-0.7	-0.64	-0.59	-0.41	-0.41	-0.53	-0.54	-0.31
	Coverage	1.0	1.0	1.0	1.0	0.97	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Mean RelWidth	0.1	0.09	0.08	0.08	0.08	0.07	0.06	0.08	0.07	0.06	0.07	0.06	0.07	0.06
σ_t	Mean RelBias	-0.02	-0.06	0.12	0.02	0.2	0.07	0.19	0.13	0.13	0.13	0.15	0.18	0.17	0.07
	Coverage	1.0	1.0	1.0	1.0	0.97	1.0	1.0	1.0	1.0	1.0	1.0	0.98	0.95	1.0
	Mean RelWidth	0.06	0.05	0.04	0.04	0.04	0.03	0.03	0.04	0.03	0.03	0.03	0.02	0.02	0.02

Table B.6: Coverage, precision and accuracy metrics for designs with 7 mice replicates per observation time

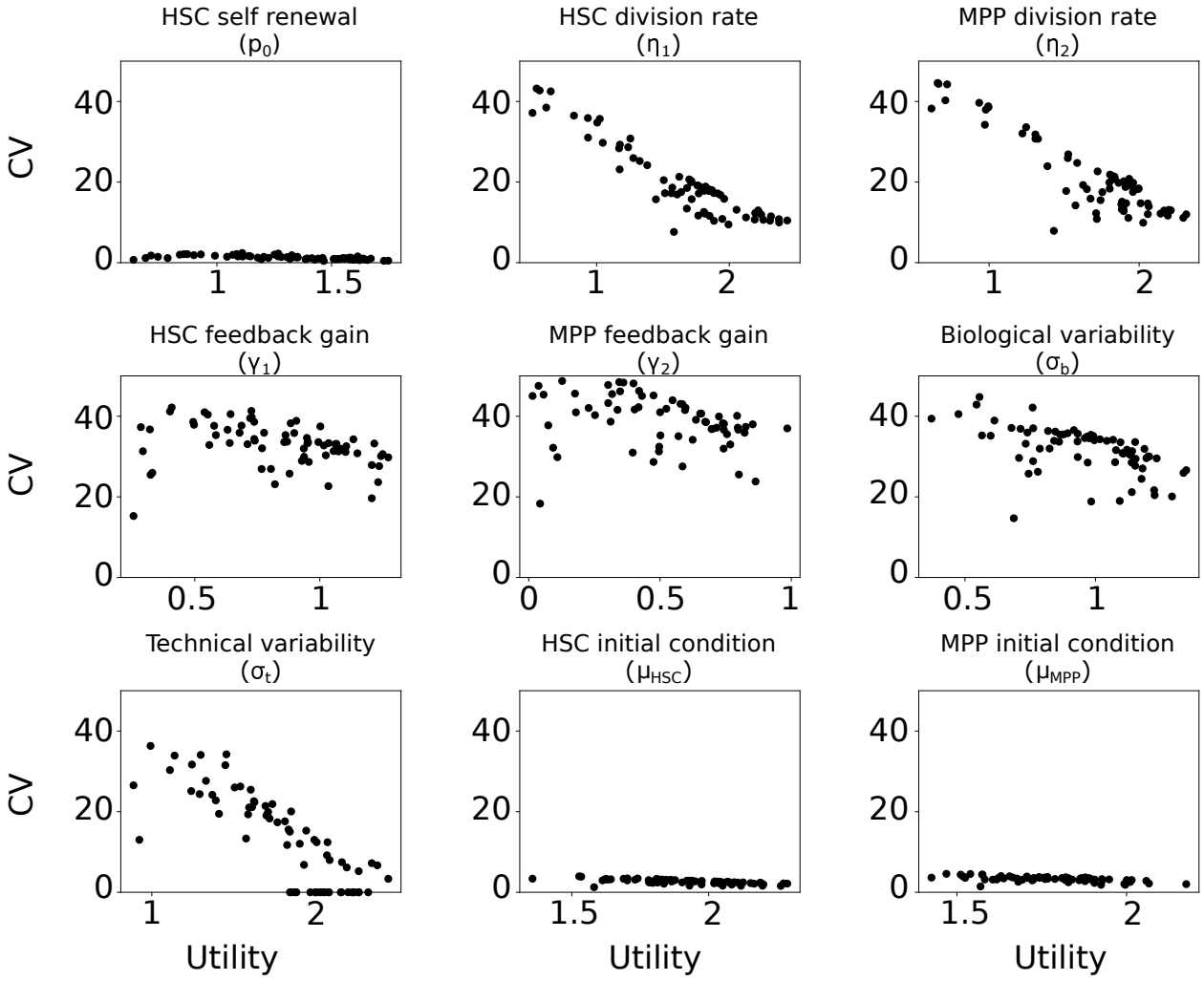

Figure B.1: Coefficient of variation for the posterior decreases as parameter utility increases. The mean coefficient of variation (CV) for all designs are obtained and plotted against the parameter utility value. The utility values are raw values and are not normalized like the utilities in the heat maps.

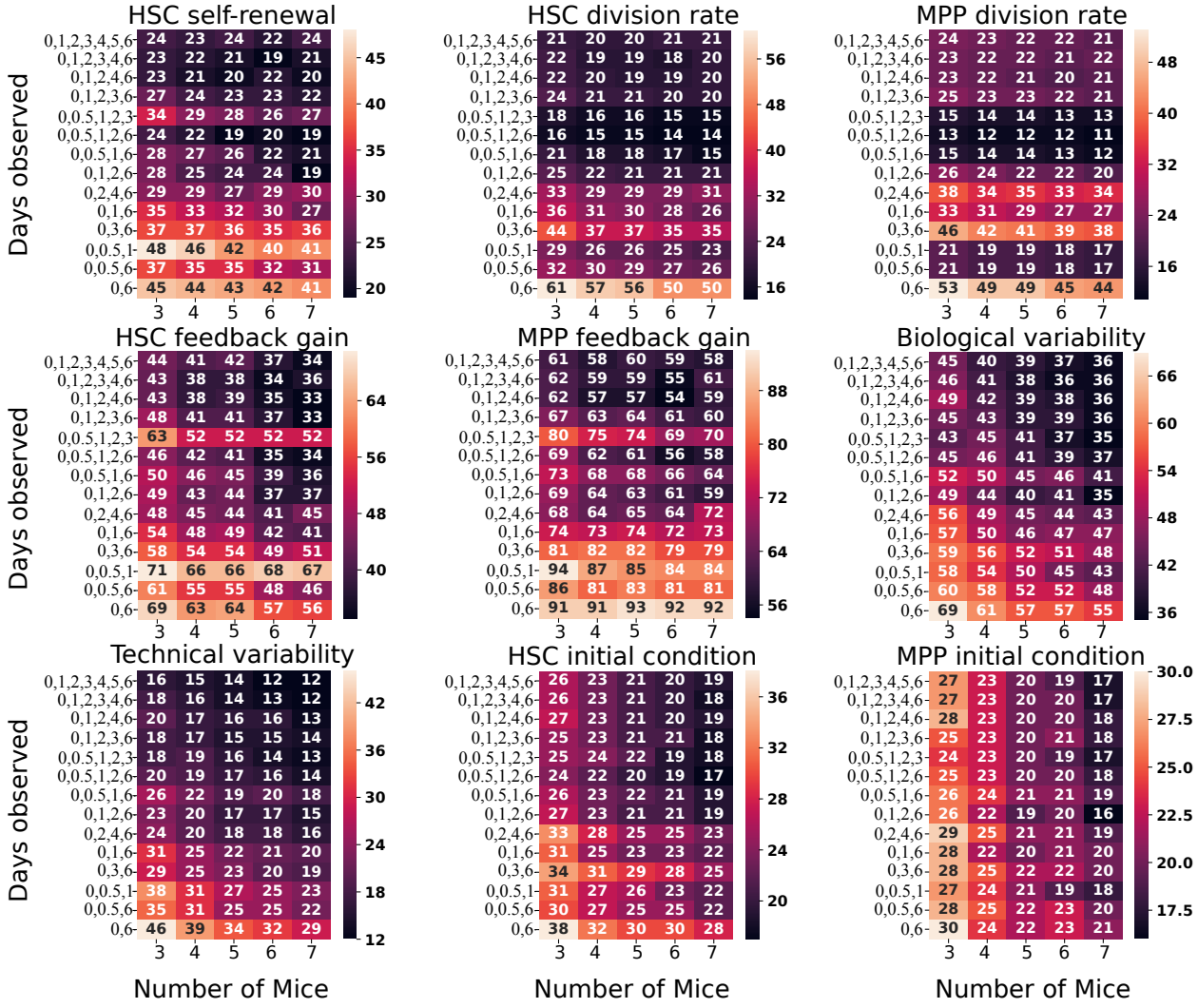

Figure B.2: Prior vs Posterior percent overlap. Using percent overlap as a measure for information gain, we can use use these heat maps to compare to parameter utility values.

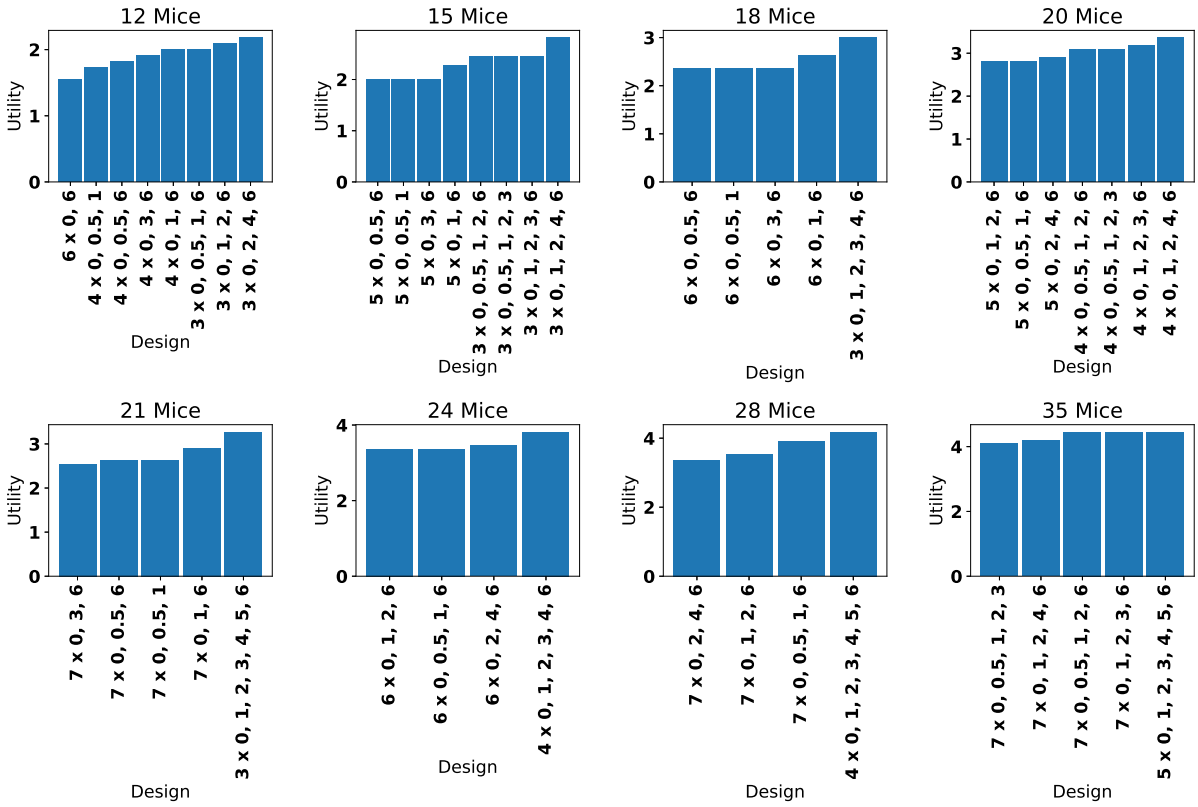

Figure B.3: Comparing designs with fixed mice. The design utility values are compared when the same number of mice are used. The utility value is on the same scale as overall utility in figure 5.5