January 2015

MODELING, ANALYSIS, AND CONTROL OF SYK-MEDIATED SIGNALING EVENTS FOR B CELLS AND ASSOCIATED CELLULAR RESPONSE FOR B CELLS

Reginald McGee
Purdue University

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Reginald L. McGee II

Entitled
MODELING, ANALYSIS, AND CONTROL OF SYK-MEDIATED SIGNALING EVENTS FOR B CELLS
AND ASSOCIATED CELLULAR RESPONSE FOR B CELLS

For the degree of   Doctor of Philosophy

Is approved by the final examining committee:

Dr. Gregery Buzzard
Chair
Dr. Ann Rundell

Dr. Robert Geahlen

Dr. Zhilan Feng

To the best of my knowledge and as understood by the student in the Thesis/Dissertation
Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32),
this thesis/dissertation adheres to the provisions of Purdue University's "Policy of
Integrity in Research" and the use of copyright material.

Approved by Major Professor(s):   Dr. Gregery Buzzard

Approved by: Dr. David Goldberg   8/21/15

Head of the Departmental Graduate Program   Date

MODELING, ANALYSIS, AND CONTROL OF SYK-MEDIATED SIGNALING

EVENTS FOR B CELLS

AND ASSOCIATED CELLULAR RESPONSE FOR B CELLS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Reginald L. McGee

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2015

Purdue University

West Lafayette, Indiana

This dissertation is dedicated to my grandparents, parents, and the ancestors on whose shoulders I stand.

ACKNOWLEDGMENTS

I must thank Kevin Rotz, one of my best friends throughout this process; I undoubtedly would not be here without the many office discussions and peer mentorship over the past six years. His friendship is invaluable.

I would like to thank Dr. Vu Dinh, Jeffrey Perley, Dr. Mariya Krisenko, and the Rundell Lab for their research insights, help, and fruitful discussions. I am also very thankful to Dr. Peter Hollenbeck for allowing me to take BIOL231, a class that has been proven to crucial for my biological understanding.

I would like to thank Dr. Edray Goins, Dr. David Goldberg, Dr. Rodrigo Bañuelos, Dr. Candice Price, Dr. Syvillia Averett, and my mathematical godmother Dr. Talitha Washington for all of their mentorship and investment in seeing me succeed.

I am truly indebted to Dr. Zenephia Evans who since my first week on campus has been a mother away from home to me and so many other underrepresented students at Purdue. I must also thank institutions and initiatives like AGEP, NIMBioS, MBI, Math@SACNAS for their support.

I am forever grateful to my undergraduate advisors Drs. Desmond and Sonya Stephens for being two of the best role models, mentors, and surrogate parents anyone could ever ask for. Additionally, I am appreciative of all the professors over the years who saw something in me, encouraged me, and held me to a standard higher than I had set for myself; particularly, Dr. Philip Kutzko, Dr. Roselyn Williams, Dr. Juan Gatica, Dr. Ajith Gunaratne, Dr. Kbenesh Blayneh, Mr. James Rossio, Ms. Trisha Baker, Mrs. Felicia Seals-Buchanan, and Mr. Johnny Chatman.

Finally, I would like to thank my other MATH 609 officemates since 2009, Mrs. Rebecca Lank, the Purdue math community, and my friends throughout Purdue's campus and in the Greater Lafayette community. Last, but not least, I must thank my family for their undying support and for always being in my corner.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

McGee, Reginald L. PhD, Purdue University, December 2015. Modeling, analysis, and control of Syk-mediated signaling events for B cells and associated cellular response for B cells. Major Professor: Dr. Gregery T. Buzzard.

Understanding the immune system and its responses to foreign threats (antigens) is a matter of understanding the immune cells involved, their individual responses, and chemicals responsible for intracellular and intercellular communication. The overall immune response is driven by individual actions of neutrophils, antigen-presenting cells, and lymphocytes (T cells and B cells), among other cells. Intercellular communication is the means by which immune cells develop coordinated response while intracellular signals determine responses within a cell; both depend on systems of chemical reactions at their respective scales. The perspective taken in this dissertation is that of understanding B cells at the intracellular scale and the signaling molecules responsible for its responses.

B cells, a type of white blood cell in the immune system, identify antigens by binding to them via B cell receptors (BCRs). After identifying an antigen, mechanisms in the B cell membrane initiate a system of chemical interactions that propagate an intracellular signal and thereby determining the cell's response. In the first part of this thesis, we present a model for B cell signaling using dynamical systems and motivated by the desire to understand the role of the protein Syk. Syk is intricately involved in the early signaling events and is required for proper response to an antigen. The importance of this protein has led to mutant variants being genetically engineered to manipulate its impact. This mutant variant is one of the primary novelties of our model, and allows us to investigate the role of feedback loops involving Syk in produc-

ing responses. This mutant model is used to develop testable hypothesis regarding the B cell mutant kinase known as Syk-AQL.

It is often difficult to resolve questions regarding complicated biological systems through experimentation alone; this has led to the rise in the use of mathematical modeling in systems biology. Experimentation is still important, however, as data is needed to refine models, and designing experiments to most efficiently refine models is an important topic of research. This is a motivation for an interest in model-based experimental design, where experiments can be systematically chosen to reduce dynamic uncertainty in a given model. In the second part of this thesis, we provide background on methods of experiment design and discuss the Maximal Informative Next Experiment (MINE) method in greater detail. In particular, we provide a theoretical foundation for this method and prove a convergence result for MINE with nonlinear models. Design criteria have been developed to sequentially provide maximal reduction in uncertainty and one criterion has been rigorously justified. We will extend this analysis to other design criteria and in more general contexts. Experimental design results will be useful for work on B cell modeling as well as other applications. This project is a step towards better understanding cellular response and creating tools useful modeling biological systems.

1. INTRODUCTION

1.1   OBJECTIVES

This dissertation presents a model of B cell receptor signaling pathways and introduces results regarding criteria for experimental design that are chosen to reduce dynamic uncertainty. Dynamical systems and numerical analysis were utilized for development of the model and its subsequent analysis. Due to the the limited amount of experimental data available while determining model parameters, the design of experiments became of interest. The work on experimental design was primarily theoretical analysis.

The aim of this work is investigating cell response, and in particular to investigate the role of the kinase Syk in B cell response. In the case of B cells, there has been a growing amount of experimentation with genetically engineering mutant proteins, but there was no model which featured the mutant protein. A reliable computational model can be useful for in silico experimentation, which can be done expediently and inexpensively. Furthermore, pairing models with experiments is powerful as effectively designed experiments can help refine models and predictive models can produce testable hypotheses.

1.2   BACKGROUND

In this section, we provide background on the biological and mathematical material presented throughout this thesis.

1.2.1 LYMPHOCYTE BIOLOGY

A lymphocyte is a type of white blood cell and is a part of the adaptive immune system. Lymphocytes are divided into two types of cells, known as T cells and B cells. T cells appear in several different forms (helper, cytotoxic, memory, regulatory, and natural killer); they coordinate many aspects of the immune response. B cells are distinct from T cells in their ability to bind to specific antigens (substances foreign to the body). Also, while T cells are generated in the thymus, B cells are generated inside bone marrow and can work either together with T cells or independent of T cells to combat antigens (foreign threats).

Antigens are recognized through the B cell receptor (BCR). Once an antigen is bound, signaling mechanisms on and around the B cell membrane are activated, mostly through tyrosine phosphorylation. Tyrosine phosphorylation is the process of a protein gaining or losing a phosphate group at one of the protein's tyrosine amino acids. Due to the the negative charge of a phosphate group, its presence or removal leads to a conformational shift in the protein, the various conformations are the activated and inactivated states for signaling components and they lead to a intricate system of molecular interactions initiates. There are many kinases involved in this process; two kinases which govern early signaling events in B cells are Lyn and Syk. Syk plays a central role in the response of a B cell. In particular, there are at least three pathways from Syk to the enzyme PLC2γ, which acts as a second messenger. Further downstream events lead to the production of molecules such as Erk, NFAT, NF-κB, their translocation into the cell nucleus and then transcription begins. These events regulate whether the final cell response is activation (proliferation), apoptosis (cell death), or anergy (chronic unresponsiveness).

T cells and B cells share similarities in their signaling, particularly in their downstream pathways. Since there are many existing T cells models but few B cell models, the most natural course is to modify a T cell model to reflect B cell dynamics. This project builds on an existing T cell model by Zheng [1] for B cell study. We describe

the structure of this model, give preliminary simulation results, and discuss future aims for tuning of the model and analysis of dominant interactions.

1.2.2   MASS ACTION KINETIC MODELING

The models considered herein are formulated with the Law of Mass Action [2], a kinetic scheme used frequently to model chemical reactions. Mass action kinetics provide mathematical relationships between reactants, catalysts, and byproducts, and applies to many chemical and biological processes. Reactions in the lymphocyte antigen receptor signaling pathway include conversions in proteins, binding events, and localization and movement of proteins. These processes can be represented with several mass action kinetic schemes which we will now discuss.

A standard conversion reaction where A converts to B is written as $A \rightarrow B$ and the Law of Mass Action tells us that the concentration of B changes at a rate $\frac{d[B]}{dt} = k[A]$. When reactants are involved in multiple reactions the rate of change can be decomposed into terms according to the reactions. For instance, the reaction $A \leftrightarrow B \leftrightarrow C$ implies that the concentration of B changes at rate $\frac{d[B]}{dt} = k_1[A] - k_2[B] + k_3[C] - k_4[B] = v_{AB} + v_{AC}$. Here $v_{AB} = k_1[A] - k_2[B], v_{BC} = k_3[C] - k_4[B]$ are terms corresponding to the reactions $A \leftrightarrow B$ and $B \leftrightarrow C$, respectively.

Many processes can be represented by the same or similar kinetic schemes. The degradation of a reactant A with byproduct B is modeled by the standard conversion scheme described above, and the concentration of B changes at a rate $v_{standard} = k[A]$. The degradation process is observed in the B cell model after receptors are internalized; there the reactant A represents the internalized receptors and the byproduct B represents the receptors marked for degradation by lysosomes.

A ubiquitous reaction is the combination of reactants A and B that results in a byproduct C. A combination reaction is written as $A + B \leftrightarrow C$ and the concentration of C changes at a rate $v_{combination} = k_1[A][B] - k_2[C]$. Binding is the first combination to occur in the model, when the signaling receptor is engaged by a ligand. In the

case of receptor engagement, the free receptor is viewed as a reactant A and the ligand is viewed as a reactant B. Similarly, this scheme can be used to represent the combination of reactants. The binding of ligand to the B cell receptor (BCR) and the binding of BTK to the linker protein BLNK are modeled in this fashion as well. For instance in the latter, $A =$BLNK, $B =$Btk, and $C =$Btk-bound.

Tyrosine phosphorylation is one of the more prevalent processes in intracellular signaling in lymphocytes and can be modeled as a reaction acted on by a catalyst. The removal and addition of phosphates at specific tyrosines can promote or inhibit conversions and activate or inactivate proteins. A catalyzed reaction is written as $A \overset{C \rightarrow}{\leftrightarrow} B$ and the concentration of B also changes at a rate $v_{catalyzed} = k_1[A][C] - k_2[B]$. An example of promotion and inhibition of a conversions in the model is the phosphatase SHP1* removing phosphate groups from other signaling components. Activation and inactivation of proteins like Syk and Lyn occur when other enzymes phosphorylate particular tyrosine sites on these proteins. These processes are viewed, and thus modeled, as catalyzed reactions. The catalyzed reaction framework also extends to other reactions like the localization of clathrin to the cell membrane, here $A =$Clathrin, $B =$Clathrin-localized, and $C =$Syk-catalytically-active.

Each model state is a distinct form of a protein in the signaling network and the grouping of reaction terms and the conservation of mass allow the differential equations for each model state to be constructed compactly. For a reaction $A \leftrightarrow B$ the differential equation for A and B will contain associated reaction terms $v_{AB} = k_1[A] - k_2[B]$ and $-v_{AB}$, respectively. The appearance of negated reaction terms is due to a need for conservation of mass in the system. Thus, differences of reaction terms are used to construct the overall differential equation for each model state.

1.2.3   T CELL ANTIGEN RECEPTOR SIGNALING MODEL

The discovery of a feedback from the kinase Erk impacting early events in T cell receptor signaling motivated the development of the original Zheng T cell receptor

signaling model [1]. The Zheng model, formulated with mass action kinetics, captures events from initial receptor engagement through nuclear transcription. Developed in phases, the original Zheng model was a 17 state system focused on signaling events that occur near the cell membrane soon after receptor engagement. Despite an assumption of a constant level of receptor phosphorylation, the initial model dynamics suggested the possibility of previously unrecognized feedbacks.

This was confirmed by the immunoblotting experiments inspired by model predictions. These experiments revealed a new feedback downstream from Erk onto the early signaling events. To include the Erk-mediated feedback in the model, the system was expanded past the early signaling events to include the Erk-MAPK pathway that extends to the nucleus. Extending the model required a determination of parameters associated with the added downstream dynamics, which in turn required experimental data. To discern the degree of Erk's mediation of early signaling, the enzyme was suppressed in certain experiments by targeting and inhibiting the preceding enzyme MEK; normal cells and cells responsive to the MEK inhibitor were used to collect data.

Time-dependent and dose-dependent T cell signaling responses were measured in vitro and this diverse data was used to fit model parameters. To understand the levels of uncertainty underlying the choice of parameters several local and global sensitivity analysis methods were performed [1]. This comparative study of the analyses yielded information on key reactions and lead to further hypotheses on feedbacks. Gradually refined to 19 and finally 22 states, the Zheng model is an example promoting the power of the interplay between computational models and experimentation.

In order to study gain more understanding of regulatory T cells, Perley revised the model [3] to include the NFAT and NF-κB pathways and has since expanded it further [4] to include features such as the MTOR pathway. The assumption of constant receptor phosphorylation was relaxed in Perley's revision of the Zheng model and has allowed for the study of the impact on signal strength on the system. Due to its predictive abilities and success with open-loop control the model shown in [3]

served as the starting point for development of B cell model. One notable revision that was required for the B cell model was the removal of the Erk-mediated feedback, a feedback which is not present in B cells. The ERK-MAPK and NF-κB pathways are presented with the B cell model in this thesis, though a NFAT pathway has not been released due to a lack of data.

1.2.4 MODEL-BASED EXPERIMENTAL DESIGN

The interplay between models and data in system biology makes the design of effective experiments of extreme importance. Mathematically, an effective design allows modelers to circumvent tedium during model refinement and focuses attention on how to make the most impact. An effective design can save experimentalists both monetary and time expenses.

We consider a model f_θ for a physical process, depending on parameters θ. Due to limits on the dynamics that can be included in the model and a lack of parameter information, there is often epistemic uncertainty of whether f_θ is the best representation for the process. This is uncertainty that can be reduced by additional measurements to gain more understanding of the underlying process. In cases where parameters can be determined, there is aleatoric uncertainty underlying the choice of parameters. This is uncertainty that is due to the randomness of measurements. These uncertainties are further discussed in [5] and its references.

Classical model-based experimental design is commonly used to reduce the uncertainty in the choice of model parameters. In the classical case, we assume we can conduct an experiment any time point, or more generally any multidimensional control point, and we assume independent identically distributed Gaussian errors in measurement. These ideas are discussed more fully in these texts [6, 7].

In the classical scenario, it is often assumed that f_θ is linear in the parameters θ. Linearity of parameters implies that f_θ can be expressed as a linear combination of specified, possibly nonlinear, basis functions with coefficients that are the components

of θ. Basis functions evaluated at the multiple points yield a Vandermonde-type matrix. Using this matrix together with the measured data we can estimate parameters θ via the method of Generalized Least Squares.

This Vandermonde matrix leads to different assessments for the estimate $\hat{\theta}$ is for the true parameters θ, corresponding to distinct objectives in a convex optimization problem. The assumptions above imply that $\hat{\theta}$ is a multidimensional Gaussian random variable, so the assessments for $\hat{\theta}$ focus its covariance matrix. One common assessment the D-optimality criterion is the determinant of the covariance matrix for $\hat{\theta}$, which is related to the volume of the 95% confidence ellipsoid for $\hat{\theta}$. The goal of D-optimal experimental design is to choose experiments that minimize the volume of this confidence ellipsoid. These ideas are also explored more in [6, 7].

Recent model-based experimental design techniques are less reliant on the linearity assumption and focus on reducing dynamic uncertainty in the model. Intuitively, if one is to reduce dynamic uncertainty it is reasonable to seek experiments at time points where there is the most variance in the model. More formally, the Maximally Informative Next Experiment (MINE) criteria, introduced originally in [8] and studied further in [9, 10], consider objectives that are functions of the model variance, as opposed to parameter variance in the classical setting. The experiments suggested by the MINE criteria induce a probability distribution on parameter space through a likelihood function. When considering a probability distribution on parameter space, one can use a Bayesian framework where the probability distribution is updated following each set of experiments. Given the sequential reductions of dynamic uncertainty, it is natural to ask whether the series of experiments will give convergence to true biological dynamics.

Again a key contrast between MINE and classical experimental designs is the lack of an assumption of linear dependence on parameters in the former. An additional distinction is the sequential determination of experiments with MINE, while in the classical framework the collection of experiments is produced through a single opti-

mization problem. Chapter 3 is presents the MINE framework in more detail and gives conditions under which the estimated dynamics converge to the true dynamics.

1.3 ORGANIZATION

Chapter 2 presents the B cell antigen receptor signaling model and demonstrates some of its abilities for generating hypotheses. Chapter 3 culminates in a convergence result using a recent experimental design criteria. Additionally, there are numerical results for the validity of this result using a toy model.

2. B CELL ANTIGEN RECEPTOR SIGNALING MODEL

2.1 PREFACE

The work presented in this chapter was originally published in the Processes:

.

(with Mariya Krisenko, Robert Geahlen, Ann Rundell, and Gregery Buzzard) *A Computational Study of the Effects of Syk Activity on B Cell Receptor Signaling Dynamics*, Processes **3** (2015), no. 1, pp. 75-97. doi:10.3390/pr3010075

This article has been modified to conform to the format required.

2.2 INTRODUCTION

Signaling through the B cell receptor (BCR) involves an intricate network of molecular reactions necessary for B cells to generate an immune response. The signaling network involves a variety of proteins including kinases and phosphatases and is particularly dependent on the protein-tyrosine kinase (PTK) Syk. To better understand the network, it is imperative to examine the roles of key signaling components like Syk and their most influential interactions. We will employ a computational approach to quantify the impact of Syk and other key enzymes and factors such as the effect of the amount of antigen on the B cell response.

Catalytically active Syk has been shown experimentally to play a central role in BCR signaling, but questions regarding its behavior still exist and the time frames in which critical interactions must occur have yet to be completely characterized. Experimentally, a mutated version of Syk called analog-sensitive Syk or Syk-AQL has been engineered to accept orthogonal inhibitors, *i.e.*, inhibitors that have been syn-

thesized to render the mutant kinase inactive almost immediately [11]. Furthermore, by replacing wild-type Syk with Syk-AQL creates B cells whose signaling capacity can be reduced or interrupted completely by the addition of the orthogonal inhibitor, experimentalists could then control the time that Syk remains active following receptor engagement, which helped to confirm how BCR signaling is modulated by the actions of Syk. Recently, a Syk-deficient B cell line was generated in which Syk-AQL expression can be induced in response to the drug tetracycline. Thus, in addition to being able to turn Syk off when desired, its expression level before activation can be adjusted if needed.

Computational modeling allows us to gain insight into Syk's impact that was not previously possible with experimentation alone. We have developed a model built on a T cell receptor (TCR) signaling model originally created by Zheng *et al.* [1] and later expanded and used by Perley *et al.* [3] for cellular level control. Perley's success in using Zheng's model for prediction and open-loop control made it an ideal candidate to adapt for our B cell study. The signaling of B cell and of T cell can be divided into early interactions, which occur proximal to the membrane, and downstream interactions, which occur in the cytoplasm and ultimately lead to the nucleus. The dynamics of the downstream signaling are nearly identical between the cells, and thus this part of the Zheng model remained largely unchanged. The signaling dynamics of T cell and of B cell differ the most in their early signaling, which is where most model revisions were required.

In the past decade there have been a number of computational models, both stochastic and deterministic varieties, focusing on various aspects of B cell signaling, but none have considered impairment to Syk and the resulting effect on cell response. Stochastic simulations have been used by Tsourkas *et al.* [12] and Mukherjee *et al.* [13] while considering spatial dynamics of BCR signaling. The impact of affinity discrimination was considered by Tsourkas in their study, while Mukherjee investigated the roles of Syk and Lyn in immunoreceptor tyrosine-based activation motif (ITAM) phosphorylation. A deterministic model by Chaudhri *et al.* [14] consid-

ers a scope similar to Zheng's T cell model, with the model covering both membrane proximal, early signaling events and downstream signaling events. This model pays particular interest to the role of phosphatases in the signal transduction. In 2012, Barua *et al.* [15] developed a deterministic model of B cell early signaling in order to study the feedback loops involving Lyn and how varying stimulation to the BCR leads to a range of dynamics in Syk. Impressively, the model incorporates every phosphorylation event for all six signaling components considered.

Our model is novel in its incorporation of Syk-AQL dynamics and given its scope, the inclusion of both early and downstream signaling, this allows us to investigate the impact of Syk modulation on a large number of signaling components. Instead of considering all possible phosphorylations for our 32 signaling components, our model considers only the most critical events in order to represent relevant physiological behavior and minimize model complexity. Understanding the means by which cell responses are determined is also of particular interest and the model will allow us to investigate the impact of both the amount of antigen and the level of Syk activity on the response. In this initial study we are particularly interested in the regulation of Erk and NF-κB activity since both contribute to determining cell fate.

We will study how Erk and NF-κB phosphorylation change with modulation of receptor affinity. Using parameter values derived largely by fitting the Perley model [3] to T cell data as nominal points, we use B cell data from Healy *et al.* [16] to then determine points in parameter space that allow us to reproduce data from cellular assays. Then, using the difference of Erkp, the sum of singly and doubly phosphorylated Erk, and NF-κB as our metrics, we consider how cell response changes with receptor characteristics in both wild-type and mutant cells.

One interesting prediction of our model is that activation of Erkp and NF-κB depend on ligand binding rates in a way that is nearly independent of the reverse rate for low values of the reverse rate and in a way that depends on a fixed ratio of powers of forward and reverse binding rates for higher values of the reverse binding rate. This is illustrated and discussed in Section 4.3 and is one indication that affinity

(the ratio of forward and reverse binding rates) alone is not sufficient to characterize the response of a B cell to a given antibody.

The cell line used and experimental procedures are described in the experimental section. In the model section, we describe model construction and explicitly show equations and a diagram for the signaling dynamics. In the methods section, we discuss the sensitivity analysis used and criteria used to screen the parameter space. The discussion section includes biological background for the model and findings of the sensitivity analysis, parameter screening, and contour analysis. We also present a comparison of our model output with a dataset from Chaudhri *et al.* [14], and finish with some discussion of future direction and limitations.

2.3 MODEL DEVELOPMENT

2.3.1 BIOLOGICAL BACKGROUND

Since our B cell model is derived from an existing T cell model, we note here some of the primary components of B cell signaling, with a focus on aspects that are unique to B cells. Conventional T cells bind peptide antigens presented by major histocompatibility molecules whereas B cells can bind multiple molecular species through polymorphic cell surface immunoglobulins that serve as antigen receptors. The B cells work collaboratively with T cells to respond to monomeric antigens or independently of T cells to respond to polymeric antigens that cluster the BCR.

Once an antigen is bound and the BCR is aggregated, the signaling mechanisms at the B cell membrane are activated and an intricate system of molecular interactions initiates [17]. There are many kinases involved in this process; two connected with events proximal to the receptor in B cells are the Src-family PTK Lyn and the PTK Syk [18]. Syk plays a central role in the overall response of a B cell [19]. Unlike signaling in T cells, which depends on the Src-family PTK Lck to phosphorylate the first tyrosine of the ITAM of the TCR, the first ITAM tyrosine in B cells can be phosphorylated by Syk when Lyn, which is homologous to Lck in T cells, is not

present [20]. Furthermore, if Syk is not expressed, BCR signaling cannot proceed. Following the initiation of BCR signaling, the regulation of BCR, Syk, and Lyn activity is orchestrated by feedback loops involving the aforementioned PTKs and a collection of regulatory enzymes.

The regulatory enzymes considered are the tyrosine-phosphatase SHP1, the C-terminal Src kinase Csk and its binding protein Cbp, and the phosphatase CD45 [21, 22]. In addition to regulating fully activated Lyn along with CD45, SHP1 also dephosphorylates the ITAMs of the BCR and tyrosines Y342 and Y346 of Syk, thus reducing their activity. Note that SHP1 does not complete these actions until it has been activated itself by a Lyn, which has been dephosphorylated at its inhibitory site by CD45. After binding with a phosphorylated Cbp, activated Csk counteracts the dephosphorylation of Lyn promoted by CD45. The phosphorylation of Cbp is also promoted by CD45 activity. Gaining a better grasp of the timing of the interplay in these feedback loops is an important task as it provides insight into exactly how the BCR and primary PTKs are regulated and thus illuminates the overall sequence of events for signal transduction.

Once active, Syk phosphorylates several substrates and the resulting signals propagate into several downstream pathways that lead to the activation of downstream targets such as Erk, NFAT, and NF-κB [17–19]. Following the translocation of these molecules into the nucleus, transcription begins and cell fate activation (proliferation), apoptosis (cell death), or anergy (chronic unresponsiveness) are determined. Interestingly, these cell responses have been found by Healy *et al.* [16] to correspond to characteristic combinations of the aforementioned targets. For example, anergic B cells exhibit signaling activity in the Erk and NFAT pathways, but not in the NF-κB pathway [16]. Again, Erk and NF-κB are of particular interest in this study due to their role in cell fate determination. The affinity of a receptor for an antibody $K_{\text{affinity}} = \frac{k_{\text{forward}}}{k_{\text{reverse}}}$ is a measure of how tightly a ligand binds to a receptor. However, for a given affinity, larger forward and reverse rates can allow the ligand to bind and unbind repeatedly and rapidly. Allowing this association and disassociation with the

receptor to occur over prolonged periods of times and in the proper concentrations could be sufficient to simulate the low, chronic exposure of the BCR to antigen, which has been seen to induce anergy [23]. The causes for anergy have not been completely characterized, and we hope to use the model to gain a better understanding of the molecular triggers leading to anergy and the associated nonresponsiveness of B cells.

2.3.2 MODEL
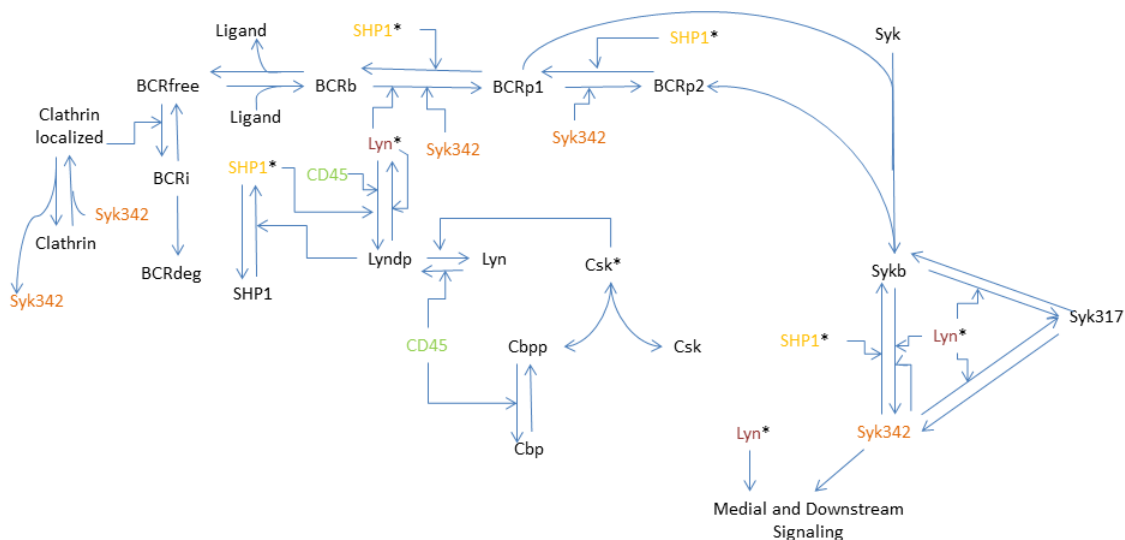
Fig. 2.1.: A depiction of the early signaling events induced by binding between B cell receptor and ligand, as described in the biological background section. Jagged arrows denote stimulations, curved arrows denote binding, straight arrows denote conversions, and color denotes species to appear repeatedly in the diagram. [24]

The model we present was developed based on the deterministic model for the TCR MAPK pathway created by Zheng [1] and extended by Perley [3]. Due to similarities in the signaling network, much of the model structure for the medial and downstream pathways required little modification. In particular, the model structure and equations for the MAPK pathway, which contains Erk, and the NF-κB pathway,
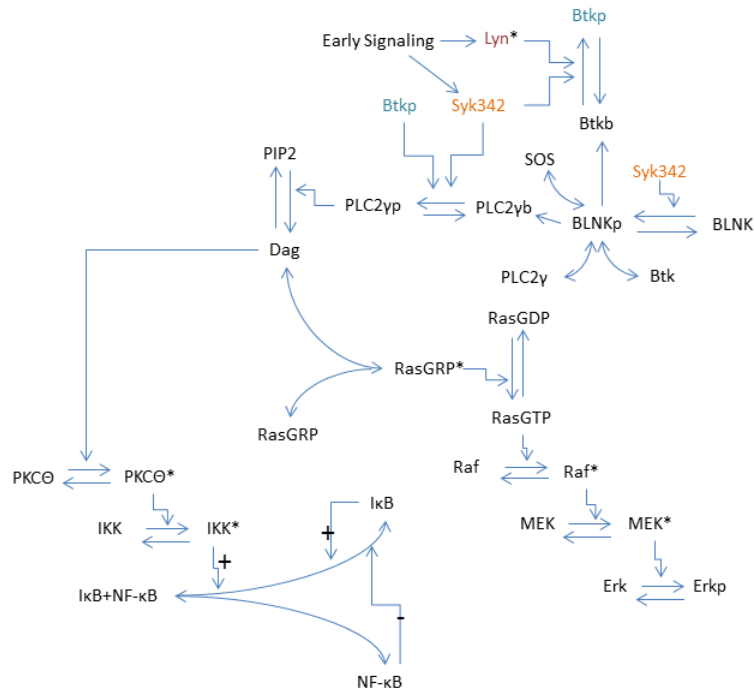
Fig. 2.2.: A depiction of the medial and downstream signaling events induced by binding between B cell receptor and ligand, as described in the biological background section. Jagged arrows denote stimulations, curved arrows denote binding, straight arrows denote conversions, and color denotes species to appear repeatedly in the diagram. The plus and minus marks near the IκB-NF-κB disassociation reaction indicate which are positive feedbacks and which are negative feedbacks. [24]

are analogous to those found in [4]. A diagram focusing on the signaling dynamics for our system is shown in Figure 2.1 and Figure 2.2.

The model tracks the concentrations of 22 distinct species with the different forms of these species represented by individual variables. The model equations were formulated with mass action kinetics; conservation laws were used to reduce the number of variables in the system. The resulting model consists of 32 ordinary differential equations and has 114 parameters. Based on a sensitivity analysis described in the methods section, there are 12 kinetic parameters whose impact we will investigate in this study. Another important parameter is the ligand concentration, which is an external input to the model. We present equations for the revised early signaling dynamics only and refer to [3] for the remaining equations. We assume that there is no downstream activity prior to receptor stimulation; hence, our model is structured to be at zero at steady state for downstream variables like Erkp and NF-κB. Additionally, we do not have any feedback between our downstream and upstream components, so setting these variables to zero is more of a baseline value than an absolute value.

2.3.3   MODEL EQUATIONS

BCR Activation: The receptor dynamics considered here include engagement of the BCR, ITAM tyrosine phosphorylation, Syk binding to the BCR, and BCR internalization, recycling and degradation. The key variables include free BCR $x_{BCRfree}$, BCR bound by ligand x_{BCRb}, singly-phosphorylated BCR x_{BCRp1}, and doubly-phosphorylated BCR x_{BCRp2}. The model formulation reflects how the kinase Syk can bind to either form of the phosphorylated BCR. Due to the positive promotion of ITAM tyrosine phosphorylation by membrane proximal PTKs [19], we assume that if Syk binds to a singly-phosphorylated BCR, that receptor will become doubly-phosphorylated before the kinase can unbind. Thus there is no term for unbinding

from Sykb to BCRp1 in any of the model equations. Receptor internalization x_{BCRi} is promoted by clathrin, which is localized $x_{Clathrin_{local}}$ to the membrane by Syk.

$$\frac{dx_{BCRfree}}{dt} = [\text{BCR recycling}] - [\text{BCR internalization}]...$$

$$+ [\text{Ligand disassociation}] - [\text{Ligand association}]$$

$$= r_{Recycling} \cdot x_{BCRi} - r_{Internalization} \cdot x_{Clathrin_{local}} x_{BCR_{Free}}...$$

$$+ r_{disassociation} \cdot x_{BCRb} - r_{association} \cdot [\text{Ligand}] \cdot x_{BCR_{Free}}$$

$$\frac{dx_{BCRb}}{dt} = [\text{Ligand binding}] - [\text{Ligand disassociation}]...$$

$$+ [\text{ITAM1 dephosphorylation}] - [\text{ITAM1 phosphorylation}]$$

$$= r_{association} \cdot [\text{Ligand}] \cdot x_{BCR_{Free}} - r_{disassociation} \cdot x_{BCRb}...$$

$$+ (r_{BCRp1dephosphorylation} \cdot x_{SHP1*}...$$

$$+ r_{BCRp1dephosphorylation\ by\ phosphatases}) x_{BCRp1}...$$

$$- (r_{BCRp1phosphorylation1} \cdot x_{Lyn*} + r_{BCRp1phosphorylation2} \cdot x_{Syk342}) x_{BCRb}$$

$$\frac{dx_{BCRp1}}{dt} = [\text{ITAM1 phosphorylation}] - [\text{ITAM1 dephosphorylation}]...$$

$$+ [\text{ITAM2 dephosphorylation}] - [\text{ITAM2 phosphorylation}]$$

$$- [\text{Syk-BCR binding1}]$$

$$= (r_{BCRp1phosphorylation1} \cdot x_{Lyn*} + r_{BCRp1phosphorylation2} \cdot x_{Syk342})x_{BCRb}...$$

$$- (r_{BCRp1dephosphorylation} \cdot x_{SHP1*}...$$

$$+ r_{BCRp1dephosphorylation \ by \ phosphatases})x_{BCRp1}...$$

$$+ (r_{BCRp2dephosphorylation} \cdot x_{SHP1*}...$$

$$+ r_{BCRp2dephosphorylation \ by \ phosphatases}) \cdot x_{BCRp2}...$$

$$- r_{BCRp2 \ phosphorylation} \cdot x_{Syk342}x_{BCRp1}...$$

$$- r_{Syk-BCR \ binding1} \cdot x_{Syk}x_{BCRp1}$$

$$\frac{dx_{BCRp2}}{dt} = [\text{ITAM2 phosphorylation}] - [\text{ITAM2 dephosphorylation}]$$

$$+ [\text{Syk-BCR unbinding}] - [\text{Syk-BCR binding2}]$$

$$= r_{BCRp2 \ phosphorylation} \cdot x_{Syk342}x_{BCRp1}...$$

$$- (r_{BCRp2dephosphorylation} \cdot x_{SHP1*}...$$

$$+ r_{BCRp2dephosphorylation \ by \ phosphatases}) \cdot x_{BCRp2}...$$

$$+ r_{Syk-BCR \ unbinding} \cdot x_{Sykb} - r_{Syk-BCR \ binding2} \cdot x_{Syk}x_{BCRp2}$$

$$\frac{dx_{BCRi}}{dt} = [\text{BCR internalization}] - [\text{BCR recycling}] - [\text{BCR degradation}]$$

$$= r_{Internalization} \cdot x_{Clathrin_{local}}x_{BCR_{Free}}...$$

$$- r_{Recycling} \cdot x_{BCRi} - r_{Degradation} \cdot x_{BCRi}$$

$$\frac{dx_{Clathrin_{local}}}{dt} = [\text{Clathrin localization via Syk342}] - [\text{Clathrin delocalization}]$$

$$= r_{Clathrin \ localization} \cdot x_{Syk342}x_{Clathrin} - r_{Clathrin \ delocalization} \cdot x_{Clathrin_{local}}$$

Syk Activation: We consider four forms of Syk, three of which have been modified through binding or phosphorylation. The variable x_{Syk} represents the amount of kinase that has not been activated and is unbound. The variable x_{Sykb} is the basally active form of the kinase that has been bound to the BCR. The catalytically active

form of Syk that has been phosphorylated at tyrosine Y342 and Y346 is denoted by x_{Syk342} and is responsible for enhancing signaling propagation. If either active form of the kinase becomes phosphorylated at tyrosine Y317 it is rendered inactive. This inactive form is denoted by x_{Syk317}. The forms represented by x_{Syk342} and x_{Syk317} are still assumed to be bound to the BCR. As discussed below, each of these four forms of Syk can bind to an orthogonal inhibitor; this binding also renders Syk inactive.

$$
\begin{aligned}
\frac{dx_{Sykb}}{dt} = & \ [\text{Syk-BCR binding}] - [\text{Syk-BCR unbinding}]... \\
& + [\text{Syk dephosphorylation at Y342}] - [\text{Syk phosphorylation at Y342}]... \\
& + [\text{Syk dephosphorylation at Y317}] - [\text{Syk phosphorylation at Y317}] \\
= & \ (r_{Syk-BCR\ binding1} \cdot x_{BCRp1} + r_{Syk-BCR\ binding2} \cdot x_{BCRp2})x_{Syk}... \\
& - r_{Syk-BCR\ unbinding} \cdot x_{Sykb}... \\
& + (r_{Syk342\ dephosphorylation} \cdot x_{SHP1*}... \\
& + r_{Syk342\ dephosphorylation\ by\ phosphatases})x_{Syk342}... \\
& - (r_{Syk342\ via\ Lyn*} \cdot x_{Lyn*} + r_{Syk342\ autophosphorylation} \cdot x_{Syk342})x_{Sykb}... \\
& + r_{Syk317\ dephosphorylation} \cdot x_{Syk317} - r_{Syk317\ phosphorylation1}x_{Lyn*}x_{Sykb}
\end{aligned}
$$

$$
\begin{aligned}
\frac{dx_{Syk342}}{dt} = & \ [\text{Syk phosphorylation at Y342}] - [\text{Syk dephosphorylation at Y342}]... \\
& + [\text{Syk dephosphorylation at Y317}] - [\text{Syk phosphorylation at Y317}] \\
= & \ (r_{Syk342\ via\ Lyn*} \cdot x_{Lyn*} + r_{Syk342\ autophosphorylation} \cdot x_{Syk342})x_{Sykb}... \\
& - (r_{Syk342\ dephosphorylation} \cdot x_{SHP1*}... \\
& + r_{Syk342\ dephosphorylation\ by\ phosphatases})x_{Syk342}... \\
& + r_{Syk317\ dephosphorylation} \cdot x_{Syk317} - r_{Syk317\ phosphorylation2} \cdot x_{Lyn*}x_{Syk342}
\end{aligned}
$$

$$
\begin{aligned}
\frac{dx_{Syk317}}{dt} = & \ [\text{Syk phosphorylation at Y317}] - [\text{Syk dephosphorylation at Y317}] \\
= & \ (r_{Syk317\ phosphorylation1} \cdot x_{Sykb} + r_{Syk317\ phosphorylation2} \cdot x_{Syk342})x_{Lyn*}... \\
& - 2r_{Syk317\ dephosphorylation} \cdot x_{Syk317}
\end{aligned}
$$

Syk-AQL dynamics: As discussed in the Introduction, Syk-AQL allows for Syk activity to be modulated through the addition of the orthogonal inhibitor (OI). The binding of the OI to the mutant Syk-AQL is also modeled using mass action kinetics and can be seen in Figure 2.3.

Fig. 2.3.: Orthogonal Inhibitor (OI) binding kinetics. Sykj denotes any of the modeled forms of Syk. [24]

This binding results in the term $r_{Inhibitor\ association} \cdot [\text{OI}] \cdot x_{Syk_j}$ where j denotes any one of the forms of Syk modeled. These terms are subtracted from each of the equations for their respective forms of Syk. By conservation of mass, we have the following equation for orthogonally inhibited Syk:

$$
\begin{aligned}
\frac{dx_{Syk-inh}}{dt} =\ & [\text{Inhibitor\ \ association - Syk bound}]... \\
& + [\text{Inhibitor\ \ association - Syk Y342}]... \\
& + [\text{Inhibitor\ \ association - Syk Y317}]... \\
& + [\text{Inhibitor\ \ association - free Syk}] \\
=\ & r_{Inhibitor\ association} \cdot [\text{OI}] \cdot x_{Sykb} + r_{Inhibitor\ association} \cdot [\text{OI}] \cdot x_{Syk342}... \\
& + r_{Inhibitor\ association} \cdot [\text{OI}] \cdot x_{Syk317} + r_{Inhibitor\ association} \cdot [\text{OI}] \cdot x_{freeSyk}.
\end{aligned}
$$

Lyn Activation: For the Src-family PTK Lyn (x_{Lyn}) to become fully activated (x_{Lyn*}), it must be dephosphorylated at Y508 (x_{Lyndp}) and then go through an autophosphorylation reaction. We consider both events with the following equations:

$$\frac{dx_{Lyndp}}{dt} = [\text{Lyn dephosphorylation}] - [\text{Lyn phosphorylation}]...$$

$$+ [\text{Lyn de-autophosphorylation}] - [\text{Lyn autophosphorylation}]$$

$$= r_{Lyn\ dephosphorylation} x_{CD45} x_{Lyn} - r_{Lyn\ phosphorylation} x_{Csk*} x_{Lyndp}...$$

$$+ r_{Lyn\ de-autophosphorylation} x_{Lyn*}...$$

$$- (r_{Lyn*\ phosphorylation} + r_{Lyn*\ autophosphorylation} x_{Lyn*}) x_{Lyndp}$$

$$\frac{dx_{Lyn*}}{dt} = [\text{Lyn autophosphorylation}] - [\text{Lyn de-autophosphorylation}]$$

$$= (r_{Lyn*\ phosphorylation} + r_{Lyn*\ autophosphorylation} x_{Lyn*}) x_{Lyndp}...$$

$$- (r_{Lyn\ de-autophosphorylation1} \cdot x_{CD45}...$$

$$+ r_{Lyn\ de-autophosphorylation2} \cdot x_{SHP1*}) x_{Lyn*}...$$

$$- r_{Lyn\ de-autophosphorylation\ by\ phosphatases} \cdot x_{Lyn*}$$

Regulatory Enzyme Dynamics: Following the initiation of BCR signaling, the regulation of BCR, Syk, and Lyn activity is orchestrated by feedback loops involving the aforementioned PTKs and a collection of regulatory enzymes. The dynamic members of the regulatory subsystem are SHP1, Csk and Cbp, with their dynamics being driven by the amount of CD45. The activated forms of SHP1, Csk and Cbp are denoted by the variables x_{SHP1*}, x_{Csk*} and x_{Cbpp*}, respectively, and are modeled with the following equations:

$$\frac{dx_{SHP1*}}{dt} = [\text{SHP1 activation}] - [\text{SHP1 inactivation}]$$

$$= r_{SHP1\ activation} \cdot x_{Lyndp} x_{SHP1} - r_{SHP1\ inactivation} \cdot x_{SHP1*}$$

$$\frac{dx_{Csk*}}{dt} = [\text{Csk activation}] - [\text{Csk disassociation}]$$

$$= r_{Csk\ activation} \cdot x_{Cbpp} x_{Csk} - r_{Csk\ disassociation} \cdot x_{Csk*}$$

$$\frac{dx_{Cbpp*}}{dt} = [\text{Cbp phosphorylation}] - [\text{Cbp dephosphorylation}]$$

$$= r_{Cbp\ phosphorylation} \cdot x_{Lyndp} x_{Cbp} - r_{Cbp\ dephosphorylation} \cdot x_{CD45} x_{Cbpp}$$

Medial Signaling Dynamics (BLNK, BTK, PLC2γ): The second messenger PLC2γ is critical for transducing a signal downstream following Syk activation. Before becoming fully activated, PLC2γ must bind to the linker protein BLNK and be phosphorylated by Syk and the Bruton's tyrosine kinase (BTK). Here BTK must also bind to BLNK, and it is phosphorylated by Syk and Lyn before it becomes fully activated. These events are modeled by the following equations:

$$\frac{dx_{BLNKp}}{dt} = [\text{BLNK phosphorylation}] - [\text{BLNK dephosphorylation}]$$

$$= r_{BLNK \ phosphorylation} \cdot x_{Syk342} x_{BLNK} \cdots$$

$$- (r_{BLNK \ dephosphorylation} \cdot x_{SHP1*} \cdots$$

$$+ r_{BLNK dephosphorylation \ by \ phosphotases}) x_{BLNKp}$$

$$\frac{dx_{BTKb}}{dt} = [\text{BLNK-BTK binding}] - [\text{BLNK-BTK unbinding}]$$

$$= r_{BLNK-BTK \ binding} \cdot x_{BTK} x_{BLNKp} - r_{BLNK-BTK \ unbinding} \cdot x_{BTKb}$$

$$\frac{dx_{BTKp}}{dt} = [\text{BTK \ phosphorylation \ by \ Syk342}]...$$

$$+ [\text{BTK \ phosphorylation \ by \ Lyn*}]...$$

$$- [\text{BTK dephosphorylation}]$$

$$= (r_{BTK \ phosphorylation \ by \ Syk342} \cdot x_{Syk342} \cdots$$

$$+ r_{BTK \ phosphorylation \ by \ Lyn*} \cdot x_{Lyn*}) x_{BTKb} \cdots$$

$$- r_{BTK \ dephosphorylation} \cdot x_{BTKp}$$

$$\frac{dx_{PLC2\gamma b}}{dt} = [\text{BLNK-PLC2}\gamma \text{ binding}] - [\text{BLNK-PLC2}\gamma \text{ unbinding}]$$

$$= r_{BLNK-PLC2\gamma \ binding} \cdot x_{PLC2\gamma} x_{BLNKp} - r_{BLNK-PLC2\gamma \ unbinding} \cdot x_{PLC2\gamma b}$$

$$\frac{dx_{PLC2\gamma p}}{dt} = [\text{PLC2}\gamma \text{ \ phosphorylation \ by \ Syk342}]...$$

$$+ [\text{PLC2}\gamma \text{ \ phosphorylation \ by \ BTK*}] - [\text{PLC2}\gamma \text{ \ dephosphorylation}]$$

$$= (r_{PLC2\gamma \ phosphorylation \ by \ Syk342} \cdot x_{Syk342} \cdots$$

$$+ r_{PLC2\gamma \ phosphorylation \ by \ BTK*} \cdot x_{BTK*}) x_{BTKb} \cdots$$

$$- r_{PLC2\gamma \ dephosphorylation} \cdot x_{PLC2\gamma p}$$

Described in [3, 4] are the remaining equations not shown here, *i.e.*, equations for Erkp, IκB and NF-κB, which are referenced below.

2.4   MATERIALS AND METHODS

In this section we describe the experimental methods used to obtain Erk phosphorylation data and the algorithmic methods used for sensitivity analysis and parameter screening.

2.4.1   EXPERIMENTAL PROTOCOLS

CELL LINES

Chicken DT40 B-cells lacking Syk were obtained from Dr. Tomohiro Kurosaki. Cells were cultured in RPMI 1640 media supplemented with 10% fetal calf serum, 1% chicken serum, 50 μM 2-mercaptoethanol, 1 mM sodium pyruvate, 100 IU/mL penicillin G, and 100 μg/mL streptomycin. Stable DT40 cell lines expressing analog sensitive Syk-AQL-EGFP (R428Q/M429L/M442A, referred to as Syk-AQL) were constructed using the Lenti-X Tet-On Advanced Inducible 105 Expression System (Clontech, Mountain View, CA, USA). To constitutively express the tetracycline-controlled transactivator, rtTA, in the Tet-On inducible system, the HEK293 cells were first infected with viral particles containing the pLVX Tet-On Advanced Regulator. Lentiviral particles were generated by co-transfecting HEK293T cells with 4 μg of pLVX-Tet-On, 4 μg of pHR'-CMV-ΔR8.20 vpr, and 2 μg of pHR'-CMV-VSVG using Lipofectamine 2000 (Invitrogen, Carlsbad, CA, USA). The supernatants containing viral particles were harvested 48 h post-transfection and used to infect Syk-deficient DT40 cells. Two days after infection, cells were selected with 500 μg/mL G418 and screened for rtTA expression. Cells constitutively expressing rtTA protein were infected with lentiviral particles packaged with pLVX-Tight-Puro-Syk-AQL-EGFP as described above. After 48 h, cells were selected with 1 μg/mL puromycin and these cells were treated with 1 μg/mL doxycycline to induce Syk-AQL expression followed by screening for expression by Western blotting.

CELLULAR ACTIVATION ASSAY

For the analysis of Erk phosphorylation, DT40 Syk-AQL-EGFP cells were treated with or without goat-anti-mouse IgM (10 μg/mL) for the indicated periods of time at 37 °C and then lysed in buffer containing 25% sucrose, 2.5% SDS, 25 mM Tris/2.5 mM EDTA, 2.5 mg pyronin Y, and 2% 2-mercaptoethanol. The DNA in lysates was sheared by passing through a 26 G × 1/2 in needle. Proteins in the lysate were separated by SDS-PAGE, transferred to polyvinylidene difluoride membrane, and analyzed by Western blotting with anti-pERK (Cell Signaling p44/p42 MAPK (T202/Y204) rabbit 4370S), and anti-Syk (Santa Cruz N-19 rabbit) antibodies. The results of these assays were used for parameter screening and will be referenced in the results section.

2.4.2 SENSITIVITY ANALYSIS

Our first objective was to identify parameters that produce behavior that fits B cell data. This is important as our nominal parameters largely came from parameters estimated by fitting to T cell data. In order to obtain a computationally tractable search, we first conducted a sensitivity analysis to identify the parameters with the greatest influence on model output associated with our available data.

We were concerned with fitting the model output to Erkp and IκB data reported by Healy *et al.* in [16], Erkp data obtained as described in the previous section, and NF-κB data reported by Oh *et al.* in [11]. Recall that Erkp denotes the sum of singly and doubly phosphorylated Erk. The experimental conditions we sought to simulate were administrations of ligand at time $t = 0$ in doses ranging from 5.5 to 150 μg/mL. The data from Healy *et al.* was used for model fitting through parameter screening, and so the form of this data was taken into consideration during sensitivity analysis. The measurements taken by Healy *et al.* were reported relative to the basal or unstimulated phosphorylation of a species. Thus, to evaluate the fitness of a set of parameters, we ran the model to steady state and recorded the value of x_{out}^{Basal},

where $out = Erkp$ or $I\kappa B$ throughout all subsequent sections, before continuing the simulation with the addition of ligand.

Fitness was quantified using the objective

$$J_{out} = (y_{obs} - y_{simulated})/\sigma_{out} \tag{2.1}$$

Given that the basal and simulated values both depend on parameters, for our initial sensitivity analysis we chose to compute the sensitivity of

$$y_{sim} = \frac{x_{out}^{Stimulation}(t_{obs}, p_k)}{x_{out}^{Basal}(p_k)} \tag{2.2}$$

with respect to variations in parameters p_k. Note that p_k is the kth point in our parameter screening. To estimate the uncertainty in the data σ_{out} for Equation (2.1), we assumed a linear dependence of σ_{out} on y_{obs} and conducted a linear regression using the information in Figure 2C of [16] . We found that the error in the measurements could be reasonably approximated by

$$\sigma_{out} = 0.0127 + 0.3084 \cdot y_{obs} \tag{2.3}$$

where y_{obs} is an observed measurement. Given the total number of model parameters and the cost associated with varying them, we partitioned parameters into seven distinct groups and conducted a sensitivity analysis with respect to each group when determining which parameters to screen initially. These groups of parameters were detrained using natural divisions such as BCR dynamics, Syk dynamics, regulatory enzyme and Lyn dynamics, Erk pathway dynamics, *etc.* A study by Zheng [1] comparing local derivative-based sensitivity methods and global variance-based methods found that global parameter sensitivities were necessary to capture model behavior when considering a large parameter space, but that there were no significant difference between Sobol analysis and the other variance based methods considered. Given the relative independence of these groups, we calculated only primary Sobol sensitivities [25] to estimate the sensitivity of the outputs, normalized as in Equation (2.2), to each specified parameter. The sensitivity, $S_{p_k}(t) = \frac{Var_{p_k}(E_{p\neq p_k}[x_{out}^{Stimulation}|\ p_k])}{Var(x_{out}^{Stimulation})}$, for

a given parameter was computed at integer values $t = 0, \ldots, 30$ using the method based on sparse-grid interpolation as described in [26]. This expression is designed to capture the relative sensitivity of the output as a function of one particular parameter p_k, averaged over the other parameters. That is, if we fix p_k, we can determine the average behavior as we vary the remaining parameters, and then determine how this average changes as a function of p_k. These calculations were carried out in log space in each parameter, with a range of one order of magnitude above and below the nominal value for each parameter.

To match the conditions in much of that in [16], we used 20 µg/mL for the stimulation amount at time $t = 0$. For each parameter considered, we calculated the median of the sensitivities for that parameter over the times considered. The value 0.15 was found to be a natural threshold for each group, and if the median was less than 0.15, we concluded the parameter was insensitive and excluded it from future parameter screening. This criterion left us with 12 parameters to consider for the parameter screening. Plots of these sensitivity values are included in Supplementary Materials.

2.4.3 PARAMETER SCREENING

Using Latin Hypercube Sampling (LHS), we screened parameter space for points that we consider acceptable if they produce simulations satisfying $|J_{out}| \leq \eta$ or equivalently

$$y_{obs} - \eta\sigma \leq y_{sim} \leq y_{obs} + \eta\sigma \tag{2.4}$$

The first screening used the following data from Healy *et al.*: an Erk measurement at time $t = 5$ minutes and dose responses for IκB all measured at time $t = 15$ minutes. The doses provided in the dose response experiment were 5.5, 16.5, 50, and 150 µg/mL.

The second screening was with respect to our data and also used the condition Equation (2.4) to determine acceptability. However, in the acceptability condition for

this screening, we modified the calculation of y_{sim} in that we calculate phosphorylation relative to the ending value rather than the basal value. That is, the signal intensities for the Western blots from our data were normalized by their ending phosphorylation levels to avoid the magnification of errors that would result from a small initial value. Applying the same normalization procedure to simulated data gives the form

$$y_{sim} = \frac{x_{out}^{Stimulation}(t_{obs}, p_k)}{x_{out}^{Stimulation}(t_{final}, p_k)} \qquad (2.5)$$

In this case the uncertainty in our experimental measurements was determined by calculating the standard deviation of the three replicates.

The data from Oh *et al.* [11] consisted of data for wild-type and mutant B cells, which featured Syk-AQL. The mutant data was reported relative to wild-type activity. The wild-type data was used to ensure that we achieved reasonable behavior in NF-κB after using data from Healy *et al.* to fit IκB, the model variable that directly preceded NF-κB. Using sensitive parameters relating to Syk dynamics as a guide to select a small subset of parameters to tune manually, the mutant data was used to determine a separate parameter set to reproduces this mutant (Syk-AQL) behavior.

2.4.4 CONTOUR ANALYSIS

In order to investigate the dependence of Erk and NF-κB activation on ligand–receptor binding rates, we simulated the model at a dose of 20 µg/mL anti-BCR over a product grid of forward and reverse binding rates. This was done for each of wild-type, mutant with no orthogonal inhibitor, and mutant with a dose of 1 µM orthogonal inhibitor. The parameter grid was constructed using evenly spaced points in log scale over ranges for forward and reverse binding rates found in literature [12, 14].

In order to avoid numerical inaccuracies associated with overly stiff parameters, we halted any simulation that took longer than fifteen minutes during the contour analysis. For these grid points, we used the built-in MATLAB function *griddata* in order to interpolate the corresponding values.

Table 2.1.: Sensitive reaction parameters

	Reactions	Parameters
Group 1	BCR dynamics	$rw0_{kf}$
Group 2	Syk activation	$rw7_{kr}, rw9_{kf}$
Group 3	Regulatory enzyme dynamics	N/A
Group 4	Medial signaling	$rw15_{kf}, rw16_{kf}, rw16_{kr}$
Group 5	Medial signaling	$r12s_{kf}, r13_{kf}, r13_{kr}$
Group 6	Erk pathway dynamics	$r18_{kf}, r19_{kf}$
Group 7	NF-κB pathway dynamics	$r38_{kf}$

The parameter screening, sensitivity analysis, and contour analysis scripts were implemented in MATLAB 2012a. The script was parallelized to run on a Sun Server X3-2 server with two Intel Xeon E5-2690 processors and 160 GB RAM. For the parameter screening and sensitivity analyses, parameter ranges were set to two orders of magnitude on either side of the nominal parameters for each group except the fourth group (see Section 4.2 and Table 2.1).

2.5 RESULTS

2.5.1 SENSITIVE PARAMETERS

Primary Sobol sensitivities [25] were calculated for each output and we analyzed the distribution of the sensitivities over time. The following parameters were considered during the parameter screening. Inclusion in the parameter screening meant that the parameter did not violate the criterion that median $S(t) < 0.15$. The box and whisker plots for parameter sensitivities leading to this criterion are included in the supplementary information figure S1. We ultimately sought for the model to produce a graded response to increasing dosages of anti-BCR stimulation in Syk342 and then

let that gradation propagate downstream; thus, the results of the sensitivity analysis match what one would expect as they correspond to key signaling reactions.

The parameters in group one correspond to BCR dynamics, and $rw0_{kf}$ specifically represents the forward rate of the ligand binding reaction to the BCR. Group two contains parameters related to Syk activation, and $rw7_{kr}$ is the reverse rate of the phosphorylation reaction for the Y342 tyrosine on Syk. Parameter $rw9_{kf}$ is the forward rate in the phosphorylation reaction for the Y317 tyrosine on Syk that has already been phosphorylated at Y342. Group three is comprised of parameters from the regulatory enzyme subsystem. A sensitivity analysis was not conducted with respect to this group due to issues with stiffness. We describe next steps to examine this stiffness and future plans to expand the regulatory enzyme subsystem to become fully dynamic in Section 2.6.

Groups four and five consisted of parameters relating to rates for reactions involving medial signaling components BLNK, PLCγ, Bruton's Tyrosine Kinase (BTK). For group four, parameter $rw15_{kf}$ is the rate at which BLNK is phosphorylated by Syk342, while $rw16_{kf}$ and $rw16_{kr}$ are the forward and backward rates for the binding of PLCγ to the linker protein BLNK. In group five, $r12s_{kf}$ is the forward rate at which Syk342 phosphorylates bound PLCγ. Parameters $r13_{kf}$ and $r13_{kr}$ represent the rate at which $PLC\gamma$ phosphorylates the phospholipid PIP$_2$ and the corresponding rate of dephosphorylation. Finally, group six is made up of medial signaling parameters for reactions involving the kinase PKC and also the downstream MAPK pathway leading to Erk. Parameter $r18_{kf}$ is the rate at which Erk is phosphorylated by MEK. Parameter $r19_{kf}$ is the rate at which the enzyme SOS binds to phosphorylated BLNK. Finally, Group seven consists of parameters for reactions related to the NFκB pathways. Here $r38_{kf}$ is the rate of phosphorylation of IκB by the kinase IKK.

2.5.2   PARAMETER SCREENING AND FITTING

To find a set of parameters that qualitatively match a variety of data, we first screened with respect to the data from Healy *et al.* [16] and required $|J_{Erkp}| \leq 1$ and $|J_{I\kappa B(dose1)}| \leq 2$. We found seven parameter vectors that met the criteria among the 1800 candidates considered. Due to large per-simulation time requirements, large objective values for doses #3 and #4 of the dose response experiments, and tradeoffs between Erk costs and IκB costs, we determined we would need to manually tune parameters related to IκB to achieve reasonable fits at all four doses.

We next screened the seven accepted parameter vectors for fitness to the Erkp data obtained as in Section 3.1.2 and normalized as in Equation (2.5) with a threshold of $\eta = 1$. From this screening, we selected one parameter vector p based upon the smoothness of its Erkp time course, time to full Erkp activation, agreement with intermediate Erkp data points, and smoothness of non-degraded IκB time courses. Simulations using p are shown in Figure 2.4. For this Erkp data, we did not do any local optimization, but rather focused on the qualitative response. The right panel in Figure 2.4 indicates both the variability in experimental data and good qualitative agreement between these data and simulation.

From the vector p, we improved our fits for IκB by manually tuning parameters in the IκB pathway. The parameters that were adjusted were the rate of IκB production and the rate of NF-κB production. This final manual tuning led us to the parameter vector we call p_{WT}^*. The final fits for IκB can be seen in Figure 2.5. We were able to achieve qualitative agreement to the wild-type NF-κB data of Oh *et al.* [11] without any further changes to parameters, as seen in the left panel of Figure 2.6.
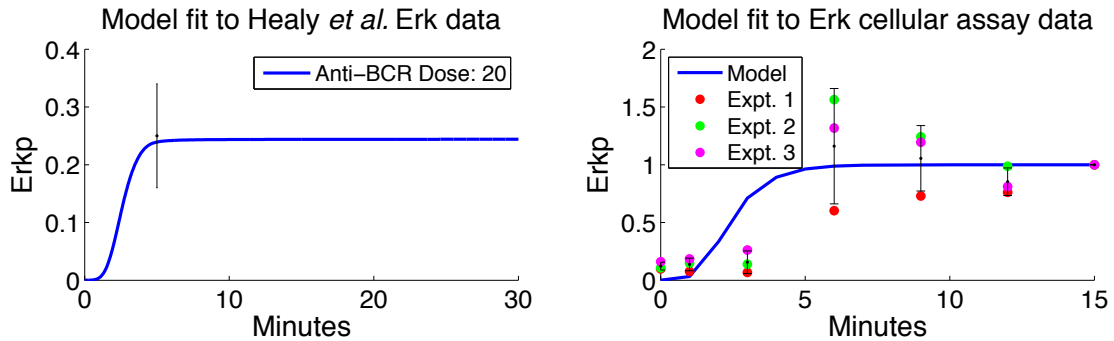
Fig. 2.4.: Simulations using p_{WT}^* compared with experimental data. On the left, a simulation for the Erkp time course (normalized by total Erk) with 20 µg/mL anti-BCR is shown with the mean from Healy *et al.* [16] at time $t = 5$ and one standard deviation interval of uncertainty. On the right, simulations using p_{WT}^* and 10 µg/mL anti-BCR (normalized by Erk at time $t = 15$) are compared with Erkp triplicate data from Section 3.1.2. [24]
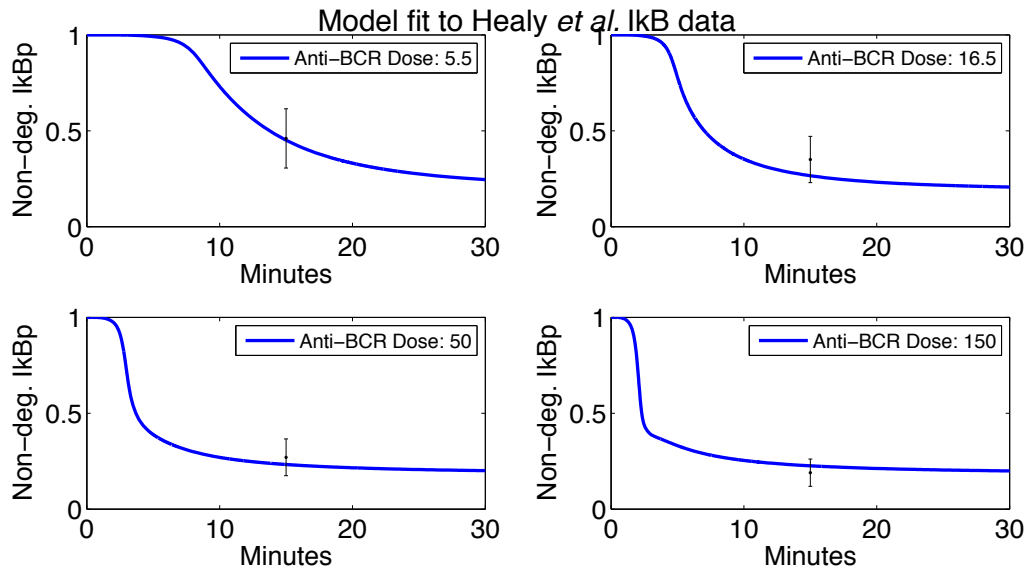
Fig. 2.5.: Simulations using p_{WT}^* compared with IκB data from Healy *et al.* [16]. Simulations for non-degraded IκB (normalized by total IκB) are shown (left to right, top to bottom) for 5.5, 16.5, 50 and 150 µg/mL anti-BCR, with all measurements taken at time $t = 15$ and one standard deviation interval of uncertainty. [24]

Since mutant Syk-AQL has experimentally different NF-κB response compared with WT, we manually tuned the sensitive parameters associated with Syk dynamics. We found that increasing the rate of Y317 phosphorylation $rw9_{kf}$ allowed us to fit two of the three nonzero data points. The agreement to the mutant data with this new parameter vector p^*_{Mutant} can be seen in the right panel of Figure 2.6. Intuitively, this corresponds to inhibiting a larger fraction of Syk, and thus there is less Syk available to propagate a signal. Interestingly, we could also achieve the same fits to mutant data by lowering the total amount of Syk in the cell. This was reminiscent of the effects of the drug tetracycline, which can regulate the amount of kinase prior to stimulation. Note that the measurements used from Oh *et al.* were reported relative to phosphorylation levels observed following an experiment where cells were stimulated using PMA and ionomyocin. We do not simulate the effects of ionomyocin in this work since calcium is not modeled, so our simulated activity in the right panel of Figure 2.6 is relative to the final phosphorylation level observed in simulated wild-type activity.
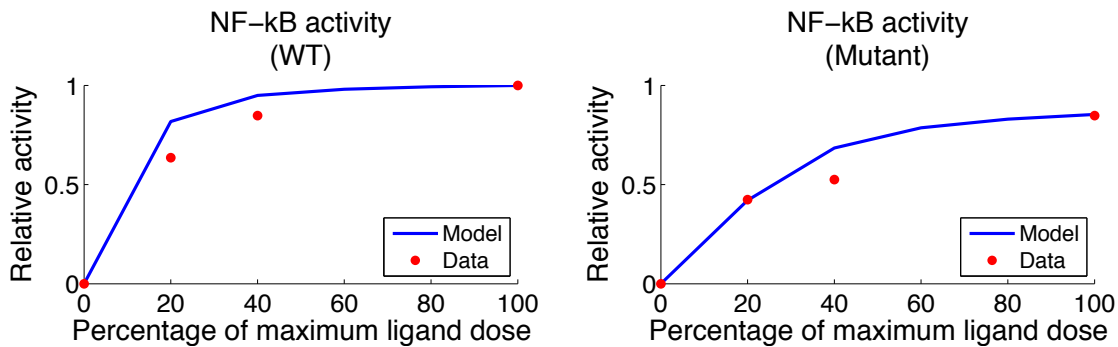
Fig. 2.6.: Anti-BCR dose response curves compared with experimental data from Oh *et al.* [11]. On the left, a simulation using p^*_{WT} (normalized by WT activity at the maximum dose) is shown to qualitatively agree with the wild-type NF-κB data (•). On the right, a simulation with the parameter vector p^*_{Mutant} (also normalized by WT activity at the maximum dose) is shown with NF-κB data (•) from B cells with Syk-AQL activity. [24]
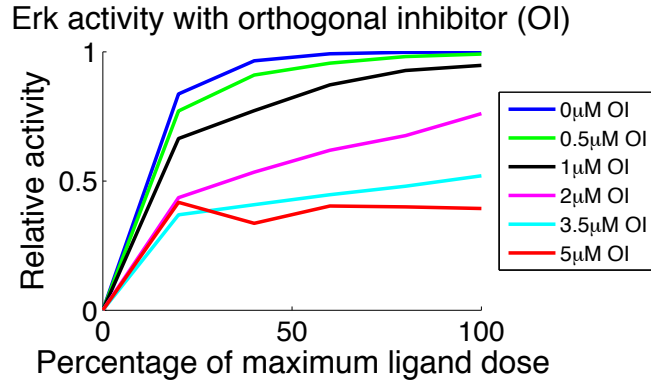
Fig. 2.7.: Anti-BCR dose response curves for baseline Syk-AQL activity and inhibited activities. The curves show simulated relative activity for Erkp measured at $t = 5$ after applying ligand and orthogonal inhibitor (μM) simultaneously. All curves have been normalized by Erk activity at the maximum dose with no orthogonal inhibitor added. The color of the curve corresponds to the amount of orthogonal inhibitor specified in the legend. [24]

In Figure 2.7, we plot predicted dose response curves associated with the parameter vector p^*_{Mutant} as a function of ligand dose, one curve for each of several doses of orthogonal inhibitor (the OI doses are specified in μM in the legend). The simulation values are given at $t = 5$ minutes. To investigate the qualitative response, we express the ligand dose in each case as a percentage of saturating dose. As seen in Figure 2.7, our model exhibits a clear dose response to antigen. Additionally, it is clear in the figure that the orthogonal inhibitor limits the Erkp response; activity level is reduced as the amount of inhibitor increases, suggesting that active Syk is critical to propagate the signal and may be a limiting quantity.

2.5.3 CONTOUR ANALYSIS

As shown by Healy *et al.* in [16], there is full signaling activity through the Erk pathway and limited activity in the NF-κB pathway during an anergic response. To investigate a variety of affinities that could induce anergy, we vary the forward and

reverse kinetic rates for BCR binding and consider the cell activity as a function of the binding rates. We seek to find areas of the grid of binding rates that lead to high Erkp activity and low NF-κB activity. We have constructed contour plots for normalized Erkp activity minus normalized NF-κB activity for several scenarios: WT B cells, mutant B cells without OI added, mutant B cells with 1 μM of OI added. The contour plots allow us to ascertain relationships between the binding rates associated with the responses we found.

As seen in Figure 2.8, for each scenario the response at low values of the reverse binding rate is qualitatively different from the response at higher values of reverse binding rate. At low values, the response depends only on the forward binding rate, while at higher values the response depends more or less linearly in log space on both binding rates. The slope for this linear relationship is not 1, however, which would be the case if the response depended on the standard affinity, $K_a = \frac{rw0_{kf}}{rw0_{kr}}$. As seen in the contour plots, the response above the value $rw0_{kr} > -0.5$ is reasonably described as a function of $\log rw0_{kf} - \alpha \log rw0_{kr}$. This leads to a kind of power law affinity, $K_{a,\alpha} = \frac{rw0_{kf}}{(rw0_{kr})^\alpha}$, where the multiplier $\alpha = 3/4$ is the reciprocal of the slope of the linear relationship in the contour plot. The origin of the power law affinity will be investigated in future analysis of the dynamical system.
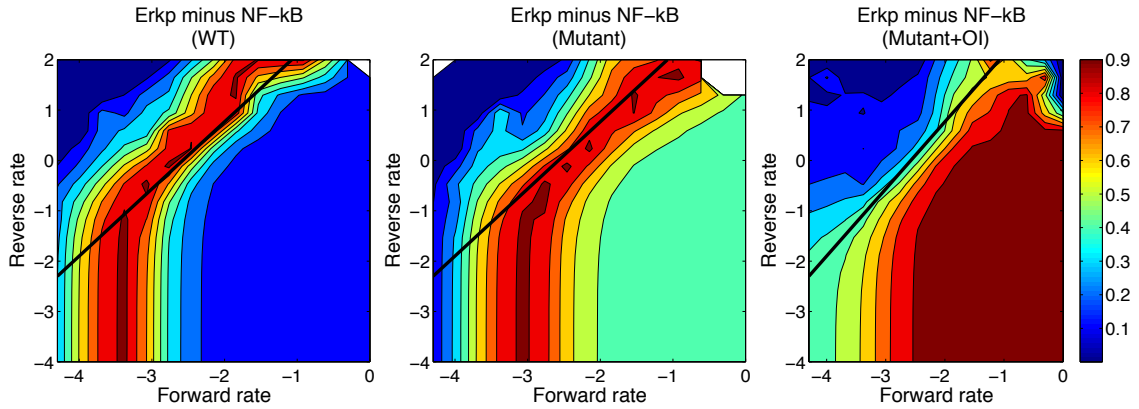
Fig. 2.8.: Contour plots for wild-type (WT), mutant without orthogonal inhibitor and mutant with 1 μM orthogonal inhibitor. The diagonal black line has a slope equal to 4/3. Regions with high values correspond to large Erkp response and small NF-κB response (both responses normalized by their maximum WT activity), and hence possible regions of anergy. Both rates are shown in log scale. [24]

To illustrate these dependencies, we plot in Figure 2.9 the responses in the low reverse rate region against the forward rate $rw0_{kf}$ and the responses in the high reverse rate region against the power law affinity. As expected from the contour plots, the plots in Figure 2.9 show a clear dependence on forward rate alone in the region of low reverse rate and a reasonably clear dependence on the power law affinity in the region of high reverse rate.

There are higher plateaus of Erkp-NF-κB present in the mutant plots (middle and right) of Figure 2.9. Plots of each quantity separately (not shown) demonstrate that plateau levels of Erkp are relatively unchanged while NF-κB is suppressed in these mutants. These higher plateaus lead to the question of whether it is easier to induce and observe anergy in B cells with Syk-AQL than in WT. If so, this could have important implications for attempts to produce mice with these mutant B cells.

In order to further understand the effect of Syk-AQL and OI on Syk, we consider the allocation of Syk in each scenario. Using the power law affinity, we find that the variables x_{Sykb}, x_{Syk342}, and x_{317} all follow a sigmoidal course. Note that a percentage
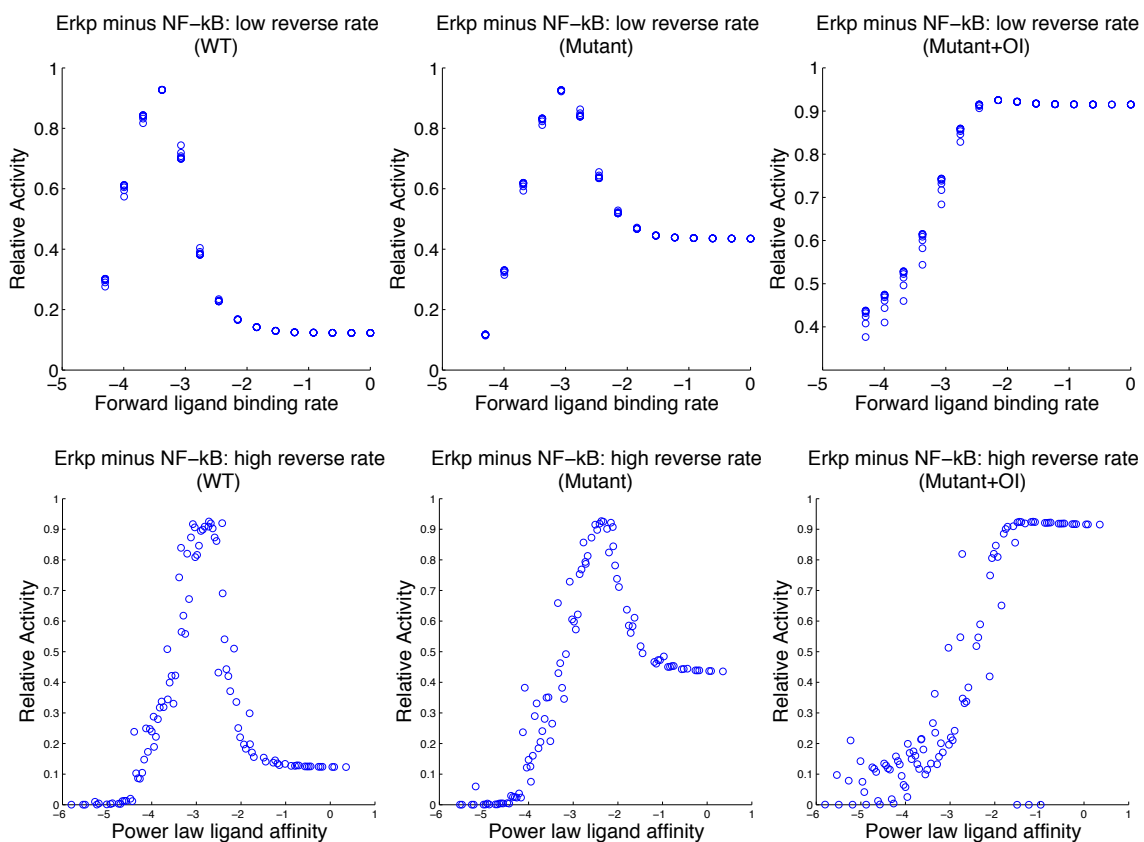
Fig. 2.9.: Plots of normalized Erkp minus normalized NF-κB (each normalized by their maximum WT activity) over a product grid of forward and reverse binding rates as in the contour plots above, but separated into regions of high and low reverse rates. The first column is wild-type simulation, the second column is mutant simulation without orthogonal inhibitor, and the third column is mutant simulation with 1 μM orthogonal inhibitor. Rates are shown in log scale. [24]

of total Syk is also allocated to other variables, such as free, unbound Syk, and to Syk bound to clathrin; since our focus is on the active forms of Syk, we omit these other forms. We find that Syk-AQL with no orthogonal inhibitor mimics fairly closely the response of wild-type, except that Syk342 is somewhat reduced. As expected, Syk-AQL with orthogonal inhibitor shows a marked decrease in these three forms of Syk, with the balance migrating to inhibited Syk.
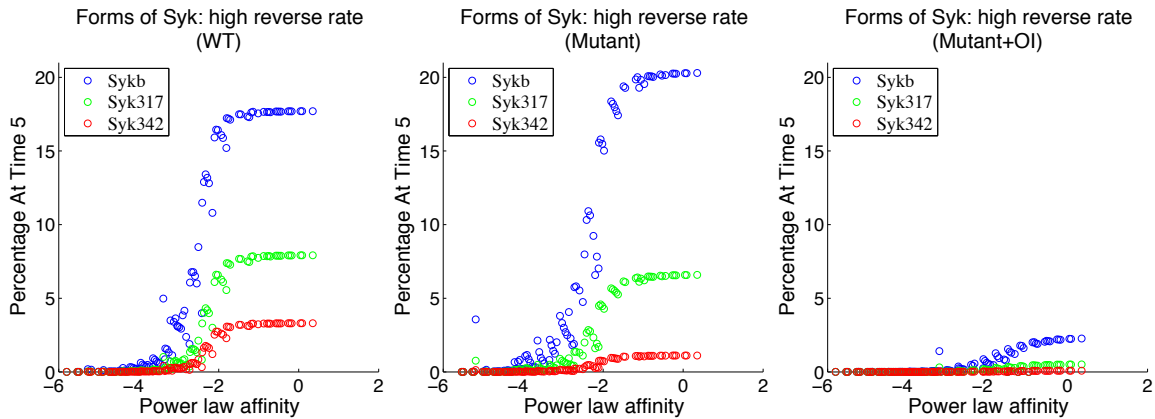
Fig. 2.10.: Plots for three forms of Syk in the model as a function of the power law affinity constant for wild-type and mutant behavior. We notice lower phosphorylation levels at both tyrosine Y317 and Y342 in the mutant cells. After the addition of 1 μM orthogonal inhibitor to the mutant cell there is the expected decrease in overall activity; the balance is accounted for by inactive forms of Syk. [24]

The analysis in this section has several possible biological implications. The moderate reduction in Syk-AQL activity compared with wild-type suggests that the level of NF-κB activation, which is reduced in Syk-AQL cells relative to wild-type, is more sensitive to small reductions in kinase activity than Erkp activation, which is essentially unchanged. Since this profile of high Erkp and reduced NF-κB is consistent with anergy, this implies that anergy is particularly sensitive to small changes in catalytic activity. A second possible implication derives from the observed dependence of Erkp and NF-κB on the power law affinity. This implies that B cells may respond to certain details of binding and unbinding rates for ligands rather than simple affinity

alone. These observations provide a platform upon which to plan future experimental approaches and to predict experimental outcomes to further evaluate the role of Syk and changes in its catalytic activity in determining cell fate decisions following BCR engagement.

2.5.4 INDEPENDENT DATASET COMPARISON

We compared the model to an independent dataset from Chaudhri *et al.* [14]. These data were not used in screening the parameters; the comparison is presented in Figure 2.11. The Chaudhri data include ligand concentrations that are much smaller than those available in our training data and indicate a relatively large activation even at very small ligand concentrations. Our model displays significantly smaller activity levels than those seen in the Chaudhri data at these low ligand concentrations. We believe that further parameter screening could produce better agreement to these data, but the underlying question is somewhat deeper in view of the phenomenon of anergy, in which B cells display reduced response to higher levels of ligand concentration. Experiments have shown that a low constant signal [23] can drive a B cell to become anergic and thus relatively unresponsive to the presence of antigen. Hence the question is not only what is the effective level of phosphorylation of Erk at low doses of ligand but also what is the effect of such low doses over extended periods of time. This is consistent with our model predictions of relatively high levels of Erkp activity and low levels of NF-κB activity in response to small amounts of active Syk. However, our model also suggests that the details of forward and reverse binding rates may also play a role in anergy.
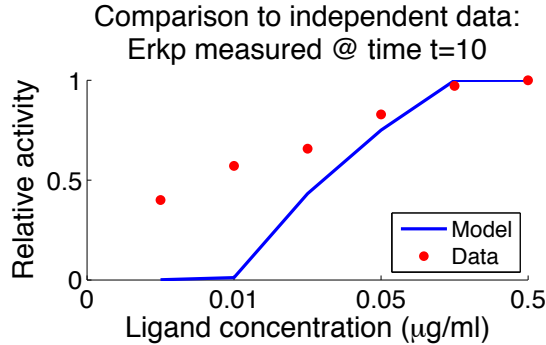
Fig. 2.11.: Anti-BCR dose response curves resulting from p_{WT}^*; the figure shows ligand dose response for Erkp resulting from p_{WT}^* as compared with data from Chaudhri *et al.* [14]. As with the data, the simulation curve is normalized by the simulated value at the maximum dose 0.5 µg/mL. [24]

2.6   CONCLUSIONS AND FUTURE DIRECTIONS

We have constructed a deterministic model of B cell signaling, with a focus on the role of Syk in modulating the activity of Erk and NF-κB. In particular, we include dynamics for the mutant kinase Syk-AQL, which experimentally displays dynamics that are qualitatively similar to wild-type dynamics in the absence of orthogonal inhibitor but can be modulated through the addition of orthogonal inhibitor. With the correct choice of parameters, our model reproduces data from recent cellular assays and qualitatively matches trends from datasets in the literature.

We sought to explore the kinetic rate constants associated with ligand binding that produced high relative activation of Erkp and low relative activation of NF-κB. These signaling conditions have been previously associated with anergy. We found that at different levels of $rw0_{kr}$ our responses actually depended on quantities other than the standard affinity constant. For low levels of $rw0_{kr}$, the model predicts that the response depends only on the forward rate of BCR binding $rw0_{kf}$. At higher levels of $rw0_{kr}$, the model predicts that the response depends on a power law form of the affinity constant, $K_{a,\alpha} = \frac{rw0_{kf}}{(rw0_{kr})^\alpha}$. These predictions were robust for WT

and mutant simulations. Given the complexity of the dynamical system, a model reduction will likely be necessary in order to analytically investigate the origin of the power law affinity underlying the model response.

Insight into the model prediction that NF-κB is more sensitive than Erkp to changes in signaling activity is found when considering the relative amplification in each pathway. For both Erkp and NF-κB, we considered the relative change in response between wild-type and mutant with orthogonal inhibitor simulations. The relative changes were both with respect to the signaling component DAG, the last signaling component to influence both pathways. We calculated the difference between Erkp in wild-type and mutant+OI simulations and then divided by wild-type Erkp simulation to get the normalized change in Erkp. We made a similar calculation using DAG, normalized by wild-type DAG simulation, and then took the ratio of the normalized change in Erkp to the normalized change in DAG. This gives us a measure of the amplification of the DAG signal in the response of Erkp. We likewise calculated the simulated amplification of DAG in the response of NF-κB. We found the amplification for Erkp to be ≈ 0.28 and the amplification for NF-κB to be ≈ 0.99. That is, the response of NF-κB to DAG is nearly 1:1, while the response of Erkp to DAG is reduced to roughly one-fourth of the incoming signal. These estimates agree with the findings in contour analysis that if there is a reduction in signaling activity to Syk, and thus DAG, then NF-κB will be more affected than Erkp. The mechanisms and parameters in these two pathways are structurally distinct: the Erkp pathway is based on mass-action kinetics, while the NF-κB pathway includes promotion of PKCΘ^* by DAG and a feedback loop involving NF-κB. Further experiments are needed to validate these predictions. One approach to this might be to use the DAG analog PMA as a means of effectively altering the level of DAG and investigate the resulting changes in Erkp and NF-κB experimentally.

Planned expansions to the model include stimulation by ionomycin, the addition of Ca^{2+} dynamics, and the addition of the NFAT pathway. We plan also to restructure the dynamics of CD45, which is constant in the current version of the model; this

modification will impact the regulatory enzyme dynamics as they are driven by CD45 activity.

One of the difficulties with this model is the stiffness of the differential equations; for a large subset of parameter space the model takes one to tens of minutes for a simulation of 30 minutes. This stiffness limits our ability to explore the parameter space fully. Model stiffness prevented sensitivity analysis with respect to the parameters for regulatory enzyme dynamics, which made up group three of Table 2.1. Stiffness also presented issues during other sensitivity analysis trials and during the parameter screening and so we will seek to address this issue in future studies. We believe the improvements to CD45 dynamics will alleviate at least some of the issues with stiffness.

As seen in the right panel of Figure 2.7, there is a discrepancy between our model and the activity observed by Chaudhri *et al.* [14]. It is not clear whether this limitation can be resolved via the tuning of ligand binding parameters or if there are additional mechanisms needed to capture the response to lower levels of ligand.

In general, the ways in which the modulation of Syk changes the response of Erk, NFAT, and NF-κB is an important question of interest for our group. Our model is an early attempt to disentangle the behavior of Syk from these downstream responses. While there is much left to be improved in our model, we believe that it will be an important tool in our search to understand the mechanisms underlying the onset of anergy in B cells. Beyond that, we believe that our model may be used as in [3] as the basis for model-informed control strategies to achieve desired cellular responses.

3. EXPERIMENTAL DESIGN USING MAXIMALLY INFORMATIVE NEXT EXPERIMENT (MINE)

3.1 INTRODUCTION

A common problem in systems biology is determining the appropriate parameters and structure for a model of a biological process. One method for reducing the uncertainty associated with model development is model-based experimental design. Model-based experimental design is an approach by which a collection of time points for measurements is produced to and the points are the best in some given sense.

We consider a more standard approach to model-based design of experiments (MBDOE) in subsection 3.1.1. In subsection 3.1.2, we consider a MBDOE approach developed more recently and that will be used for the work in this chapter.

3.1.1 EXPERIMENT DESIGN AND INFORMATION MATRICES

The idea behind classical experimental design is to find a set of experiments via an optimization problem that reduces uncertainty in parameters θ for a model f_θ. In this setting, the optimal solution is the collection of experiments that provides the best insight, in some sense, into the true value of θ. When considering these problems it is often assumed that f_θ is linear in θ; so f_θ can be expressed as a linear combination of chosen basis functions $\{\Psi_0(\hat{u}), \Psi_1(\hat{u}), ..., \Psi_{d_p}(\hat{u})\}$ and with the parameters $\{\theta^{(0)}, ..., \theta^{(d_p)}\}$ as coefficients. Note that there is no assumption that restricts Ψ_i from being nonlinear functions.

Assume that we can measure m times and that at any point in our interval of consideration $[a, b]$ our measurements at $\hat{u}$ are taken with independent error $\epsilon_{\hat{u}} \sim N(0, \sigma^2)$. In order to reduce measurement error, it may be advantageous to measure certain

points more than once. To allow for multiple measurements we introduce weights $\omega_{\hat{u}} \geq 0$ at any point of measurement $\hat{u}$. Hence, for any $\hat{u} \in [a, b]$ we have that $\epsilon_{\hat{u}} \sim N(0, \omega_{\hat{u}}\sigma^2)$, where $\sum \omega_{\hat{u}} = 1$ and $\omega_{\hat{u}}$ can be interpreted as the fraction of the m total measurements that are conducted at $\hat{u}$.

Intuitively, the goal is to reduce some measure of the variance-covariance matrix D for the parameters θ, but the analysis operates more naturally when considering the information matrix. Instead of minimizing the variance in parameters, one considers the dual problem of maximizing the information. The information matrix is defined as $M(\xi) = \sum_{\hat{u} \in supp\xi} \omega_{\hat{u}} \Psi_{\hat{u}}^T \Psi_{\hat{u}}$, where $\Psi_{\hat{u}} = [\Psi_0(\hat{u}), \Psi_1(\hat{u}), ..., \Psi_{d_p}(\hat{u})]^T$ and ξ is the particular collection of experiments. The information matrix is related to the variance-covariance matrix through the Cramer-Rao lower bound, a result of statistical theory which states $D \geq M^{-1}$, in the sense that their difference is positive semidefinite. For linear models, we have $D = M^{-1}$. There is a version of this idea that extends to nonlinear models and more general probability of data as a function of parameters known as the Fisher Information Matrix.

An experimental design ξ is specified by the number of points m, the points themselves $\hat{u}_i$, and the weights $\omega_{\hat{u}_i}$. In classical theory, ξ is allowed to be a probability measure which implies that an optimal design is supported on a finite number of points. An objective for the optimization problem is generally a convex, monotone function R of the parameter variance-covariance matrix $D = Cov(\theta)$. The notion of the best insight into the true value of θ varies, as each R specifies a different optimization problem $\arg\min_{\mathcal{C}} R(D)$. In general the solution to this optimization gives a parallel design, i.e., specifying all measurements and respective weights to be used, and is deemed to be best or optimal in some sense. Several optimality conditions are A-Optimality, D-optimality, and E-Optimality and their respective objective functions are presented in Table 3.1.

Table 3.1.: Selected Optimality Conditions

Condition	Objective R
A-optimal	Tr $Cov(\hat{\theta})$
D-optimal	det $Cov(\hat{\theta})$
E-optimal	$\lambda_{\max}(Cov(\hat{\theta}))$

There are many classic results in the field of experimental design. For each of these criteria and others one has an upper bound on the number of measurements m required in an optimal design [6]. We state the theorem precisely below.

Theorem 3.1.1 [6] *For any admissible information matrix M there exists a design ξ that contains $m \leq \frac{n(n+1)}{2}$ measurement points with non-zero weight and $M(\xi) = M$.*

We define the regression range χ to be the image of $\Psi_{\hat{u}}$ over all possible $\hat{u}$. The regression range can be used in an alternate check for optimality and is described in an equivalence theorem by Kiefer-Wolfowitz. In the case of D-optimality, the Kiefer-Wolfowitz theorem is stated as follows.

Theorem 3.1.2 (Kiefer-Wolfowitz, [7]) *Assume that the regression range $\chi \subset \mathbb{R}^k$ contains k linearly independent vectors. Then for every admissible information matrix M that is positive definite we have that the following are equivalent:*

i. M is D-optimal for θ.

ii. $xM^{-1}x^T \leq k$ for all $x \in \chi$.

Condition (ii) is often used to give an iterative algorithm for experiment design. With a given design, the algorithm looks for a regression vector for which (ii) is violated. The corresponding control point is added to the design. One problem with this algorithm is the difficulty in specifying the appropriate weight for the new point and adjusting the weights of the previous points accordingly. In essence, this

algorithm is a gradient-descent algorithm, and determining the appropriate step size can be difficult. For additional equivalent criteria see Section 9.4 in [7]. An analogous version of this theorem can be shown for each of the other optimality conditions.

One of the more common examples of linear regression is polynomial regression. When doing regression with polynomial one has that their basis functions are $\{1, \hat{u}, ..., \hat{u}^{d_p}\}$ and the parameters to be estimated are the coefficients $\{\theta^{(0)}, ..., \theta^{(d_p)}\}$. In the following example we illustrate how D-optimality can be used to determine the best points at which to measure when doing linear regression.

Example 3.1.1 (D-optimality in linear regression) *In the case of linear regression we have $p(\hat{u}) = \theta^{(0)} + \theta^{(1)}\hat{u}$ on $\hat{u} \in [-1, 1] = I$ with $\theta = [\theta^{(0)} \ \theta^{(1)}]^T \in \mathbb{R} \times \mathbb{R}$ and let $\beta(\hat{u}) = [1 \ \hat{u}]^T$. The goal of linear regression is to determine points to measure in our interval of interest I that help to improve our estimates of model parameters.*

We seek measurement points $u = \{a, b\}$, where $a, b \in I$, then we define the measurement matrix $\Psi_u = \begin{bmatrix} 1 & a \\ 1 & b \end{bmatrix}$.

Now let $\omega_a, \omega_b \geq 0$ s.t. $\omega_a + \omega_b = 1$ and $W_u = \begin{bmatrix} \omega_a & 0 \\ 0 & \omega_b \end{bmatrix}$ then the information matrix is defined as

$$
\begin{aligned}
M_\omega &= \Psi_u^T W_u \Psi_u \\
&= \begin{bmatrix} 1 & 1 \\ a & b \end{bmatrix} \begin{bmatrix} \omega_a & 0 \\ 0 & \omega_b \end{bmatrix} \begin{bmatrix} 1 & a \\ 1 & b \end{bmatrix} \\
&= \begin{bmatrix} 1 & \omega_a a + \omega_b b \\ \omega_a a + \omega_b b & \omega_a a^2 + \omega_b b^2 \end{bmatrix} \\
&= \sum_{x \in u} \omega_x \beta(x) \beta^T(x).
\end{aligned}
$$

Recall that the D-optimality criterion requires we maximize the Moment Matrix and in the classical case of linear regression this corresponds to

$$\max_u M_\omega = \max_u \begin{vmatrix} 1 & \omega_a a + \omega_b b \\ \omega_a a + \omega_b b & \omega_a a^2 + \omega_b b^2 \end{vmatrix} = \max_u \frac{\omega_a \omega_b}{2}(b-a)^2 \implies a = -1, b = 1.$$

Thus, D-optimality suggests that in order to best determine parameters for p we should measure at the endpoints of our interval I. Additionally, this shows that even weight should be given to both of the points since $\omega_a = \omega_b$.

The result in Example 3.1.1 is extended in Section 9.5 of [7] to give the D-optimal design for degree d polynomial regression over the interval $[-1, 1]$. In this design, there are $d + 1$ points: the roots of the derivative of the dth degree Legendre polynomial together with the endpoints 1 and -1. In this design, the weights of the optimal design are all $1/(d+1)$. The example above is the case when $d = 1$ so the end points are the only points suggested for measurement.

The mathematical and statistical theory underlying these optimality conditions can be read in [6]. More on FIM-based design of experiments is described in [27].

3.1.2 MINE CRITERIA

An alternative to classical experimental design techniques like those in described in [6,27] are the Maximally Informative Next Experiment (MINE) criteria [8,28]. The MINE criteria focus on identifying points of large uncertainty in the model over the interval considered. The criteria are again model-based, but are distinct from the convex objectives in the last subsection; MINE criteria consider metrics related to differences in the model output. The objectives for this consideration lie in what we call the difference space, which we define more formally in Section 3.2.

The criteria use the following related, but distinct metrics:

1. Maximal distance in difference space

2. Maximal volume in difference space

3. Maximal observational independence

The first MINE criterion is considered in [10] and identifies the time point of largest dynamic uncertainty over all possible measurement times. That is, with a given probability distribution on parameters, the model output varies more at some time points than others, and we choose the measurement time corresponding to the maximum variance.

The second criterion is a higher dimensional extension of the first and is the object of interest for this study. In this case we look for multiple measurement points to be identified simultaneously. The problem here is that we cannot rely solely on the variance at a single time point to determine multiple measurement points because we would end up measuring the same point multiple times. Instead we use the covariance matrix of multiple measurements, which allows us to identify points with large variance that are relatively uncorrelated. The volume in this criterion arises when one considers the covariance in difference space, or, as we demonstrate, the covariance of differences in model values at particular time points. The time points corresponding to the maximal volume encompass uncorrelated points with large dynamic uncertainty in the model. The second criterion has also been studied in [28] for the linear case. In that paper, they show that the second MINE criterion is equivalent to D-optimality, but with some assumptions regarding orthonormality of the rows or columns of the design matrix Ψ_u. However, for a given model, this assumption often does not apply, so we cannot generally directly apply results of D-optimality directly to MINE designs. The third criterion is essentially a normalization of the second criterion.

In contrast to the classical optimal design of experiments, the MINE criteria do not generally produce a fully parallel design. Instead, the MINE framework can be used to produce a sequential design, in which the distribution on parameters is updated at every step according to the experiments chosen. For instance, the first MINE criterion chooses a single experiment at each step. Dinh *et al.* [10] used this sequential approach and proved a convergence result that, with appropriate hypothe-

ses, ensures convergence of model dynamics to true dynamics. The convergence here is with respect to the Expected Dynamics Estimator (EDE). After the nth round of experiments, we have a probability distribution p_n depending on those choice of experiments. The expectation of the model output $\mathbb{E}_{p_n}(f_\theta)$ is defined as the EDE. If we know the true dynamics to be g then we seek a reduction in dynamic uncertainty such that $\mathbb{E}_{p_n}(f_\theta) \to g$ as $n \to \infty$.

We begin by presenting the mathematical framework for MINE. In the third section we provide rigorous justification for results on the parameters for the probability distribution present on the difference space considered in MINE. In the fourth section, we prove the necessary lemmas for the main results on convergence when choosing design points according to the second MINE criterion. We then present some the numerical illustrations of the convergence result and finally present some limitations of the MINE criterion in the linear case.

3.2 MATHEMATICAL FRAMEWORK

Let Ω be a bounded, connected subset of $\mathbb{R}^{d_p}$ and $\mathcal{U} \subset \mathbb{R}^c$. We define a model, depending on parameters from Ω, to be a map $f : \Omega \times \mathcal{U} \to \mathbb{R}$. For a given set of parameters $\Theta \in \Omega$, we can also write the model as $f_\Theta : \mathcal{U} \to \mathbb{R}$. We think of $\mathcal{U}$ as the part of the model that we can control, such as the time points we can measure or a point in space in a PDE model. Moreover, we extend this map to allow for a vector of control points by $f : \Omega \times \mathcal{U}^d \to \mathbb{R}^d$.

The MINE framework operates from a Bayesian point of view, in which the parameters Θ are fixed, but our knowledge of them is incomplete and expressed as a probability distribution on Ω. We typically have some physical process in mind when considering these systems. In particular, we are interested in systems that describe biological processes. We simply require $f_\Theta \in C_b(\mathcal{U})$, the space of continuous bounded functions. For dynamical systems applications one might assume more, namely that

$\frac{df_\Theta}{dt}$ is $C^1(\mathcal{U})$ and $\mathcal{U}$ is compact in $\mathbb{R}^c$. In this case we regard f_Θ as the vector of state variables as a function of time.

When f_Θ represents a physical process it is advantageous to conduct experiments to constrain model dynamics and reduce uncertainty. This motivates the need for the experimental design framework where the aim is to determine measurements points $v \in \mathcal{U}$ to reduce epistemic uncertainty in f_Θ.

Let Θ be a random variable that represents the model parameters. We assume that there is an existing probability distribution for Θ which may be an imposed prior or one arising after some prior measurements $v \in \mathcal{U}$. We denote the probability density as $Q_{\mu,\Sigma}(\theta)$ on Ω, with distribution parameters $\mathbb{E}_Q(\Theta) = \mu$ and $Cov_Q(\Theta) = \Sigma$. The subscripts for $Q_{\mu,\Sigma}(\theta)$ are suppressed in some places for brevity. Often one assumes the stronger condition that $\Theta \sim N(\mu, \Sigma)$, and in this case $Q_{\mu,\Sigma}$ is the multivariate Gaussian. Experimental design methods are used to reduce the uncertainty underlying a choice of Θ. Formally, we define the control vector $u = (u^{(1)}, \cdots, u^{(d)}) \in \mathcal{U}^d$. If $d > 1$, the model evaluated at the control vector is $f(\Theta, u) = [f(\Theta, u^{(1)})\ f(\Theta, u^{(2)}) \cdots f(\Theta, u^{(d)})]^T$.

Due to uncertainty in model dynamics, our objective is to identify points $u \in \mathcal{U}$ that allow us to most accurately estimate the dynamics of $f(\Theta, u)$. In the MINE framework, after Θ and Θ' are independently selected from Ω the models $f(\Theta, \cdot)$ and $f(\Theta', \cdot)$ are evaluated at points in $\mathcal{U}$ and differences are considered to create a difference space consisting of points

$$\Phi_u = \Delta F(\Theta, \Theta', u) = [f(\Theta, u^{(1)}) - f(\Theta', u^{(1)}), \cdots, f(\Theta, u^{(d)}) - f(\Theta', u^{(d)})]^T.$$

As Θ and Θ' are random variables, we have that Φ_u is also a random variable. Dong *et al.* [8] use the function $Q_\Delta(\phi, u) = \int_\Omega \int_\Omega \delta(\phi - \Delta F(\theta, \theta', u)) Q_{\mu,\Sigma}(\theta) Q_{\mu,\Sigma}(\theta') d\theta' d\theta$ as the probability density on the difference space.

In Section 3.3.1 we rigorously justify the fact that function $Q_\Delta(\phi, u)$ is the probability distribution for Φ_u and also give results pertaining to parameters for the distribution on difference space. An explicit form for the covariance on difference

space is given in Section 3.3.2. Finally, in Section 3.3.3 we demonstrate convergence when sequentially choosing $u_n \in \mathcal{U}^d$ according to the second MINE criterion under the appropriate hypotheses.

3.3 RESULTS

3.3.1 ENSEMBLE DISTRIBUTION FOR THE DIFFERENCE SPACE AND ASSOCIATED PARAMETERS

In this section we show that function Q_Δ defined above is the probability density for the random variable Φ_u, compute $\mathbb{E}_{Q_\Delta}[\Phi_u]$, and connect $Cov_{Q_\Delta}[\Phi_u]$ to $Cov_Q[f(\Theta, u)]$. We begin with lemmas regarding the density before the main proposition on the distribution parameters.

Lemma 3.3.1 *Let* $u = (u^{(1)}, \cdots, u^{(d)}) \in \mathcal{U}^d$ *be a control vector for the model* $f :$ $\Omega \times \mathcal{U} \to \mathbb{R}$. *If* $Q_{\mu,\Sigma}(\theta)$ *is the probability density on* Ω *then*

$$Q_\Delta(\phi, u) = \int_\Omega \int_\Omega \delta(\phi - \Delta F(\theta, \theta', u)) Q_{\mu,\Sigma}(\theta) Q_{\mu,\Sigma}(\theta') d\theta' d\theta,$$

is the probability distribution for

$$\Phi_u = \Delta F(\Theta, \Theta', u) = [f(\Theta, u^{(1)}) - f(\Theta', u^{(1)}), \cdots, f(\Theta, u^{(d)}) - f(\Theta', u^{(d)})]^T.$$

We require a general probabilistic result in order to prove Lemma 3.3.1, which asserts that Q_Δ is the density for the difference space. The Dirac generalized function δ can be utilized for an alternate method of computing densities for transformed random variables $Y = h(X)$. To derive the new density the density defined on the domain is integrated against $\delta(y - h(x))$, ensuring that only probabilities of points x where $y = h(x)$ contribute to the probability of y. The result is known [29], but the proof is often omitted. We give the complete proof now.

Lemma 3.3.2 *Let $X \in \mathbb{R}^n$ be a random variable on Ω_X with probability density $f(x)$. Let $h : \mathbb{R}^n \to \mathbb{R}^m$ be measurable, and let $Y = h(X)$. Then Y has probability density*

$$g(y) = \int_{\mathbb{R}^n} \delta(y - h(x))f(x) \ dx.$$

Proof Let $A \subset \mathbb{R}^m$ be Borel, and note that

$$P(Y \in A) = P(h(X) \in A)$$
$$= P(X \in h^{-1}(A))$$
$$= \int_{h^{-1}(A)} f(x) \ dx$$
$$= \int_{\mathbb{R}^n} \mathbb{1}_{h^{-1}(A)} f(x) \ dx.$$

Note also that for any $x \in \mathbb{R}^n$,

$$\int_A \delta(y - h(x))dy = \begin{cases} 1 & \text{if } x \in h^{-1}(A) \\ 0 & \text{if } x \notin h^{-1}(A) \end{cases}$$
$$= \mathbb{1}_{h^{-1}(A)}.$$

Now combining this result with Tonelli's Theorem we have

$$P(Y \in A) = \int_{\mathbb{R}^n} \left(\int_A \delta(y - h(x)) \ dy \right) f(x) \ dx$$
$$= \int_A \int_{\mathbb{R}^n} \delta(y - h(x))f(x) \ dxdy.$$

Since this is true for any Borel set A, this gives the desired conclusion. ∎

Proof [For Lemma 3.3.1] Let $Y = \Phi_u, X = (\Theta, \Theta'), \Omega_X = \Omega \times \Omega, f(x) = Q(\theta)Q(\theta')$ and $h(X) = \Delta F(\Theta, \Theta', u) = f(\Theta, u) - f(\Theta', u)$. We apply Lemma 3.3.2 and find that Φ_u has probability density $g(\phi) = \int_\Omega \int_\Omega \delta(\phi - \Delta F(\Theta, \Theta', u))Q_{\mu,\Sigma}(\theta)Q_{\mu,\Sigma}(\theta')d\theta' d\theta$, which is exactly Q_Δ. ∎

The MINE criteria were first presented in [8] and were studied recently in [28]. Several results regarding the mean and covariance of Φ_u are stated in these works. We rigorously derive the results in the following proposition as it is needed to gain a convergence result in Section 3.3.3.

Proposition 3.3.3 *If $\Phi_u = \Delta F(\Theta, \Theta', u)$ then*

$$\mathbb{E}_{Q_\Delta}[\Phi_u] = 0$$

and

$$Cov_{Q_\Delta}[\Phi_u] = 2Cov_Q[f(\Theta, u)].$$

Proof We first note that

$$\Delta F(\Theta, \Theta', u) = -(f(\Theta', u) - f(\Theta, u))$$
$$= -\Delta F(\Theta', \Theta, u).$$

Now since Θ and Θ' are arbitrary, and $\Phi_u = \Delta F(\Theta, \Theta', u)$ we can deduce that

$$
\begin{aligned}
Q_\Delta(\phi, u) &= \int_\Omega \int_\Omega \delta(\phi - \Delta F(\theta, \theta', u)) Q_{\mu, \Sigma}(\theta) Q_{\mu, \Sigma}(\theta') d\theta' d\theta \\
&= \int_\Omega \int_\Omega \delta(-\phi + \Delta F(\theta, \theta', u)) Q_{\mu, \Sigma}(\theta) Q_{\mu, \Sigma}(\theta') d\theta' d\theta \\
&= \int_\Omega \int_\Omega \delta(-\phi - \Delta F(\theta', \theta, u)) Q_{\mu, \Sigma}(\theta) Q_{\mu, \Sigma}(\theta') d\theta' d\theta \\
&= Q_\Delta(-\phi, u).
\end{aligned}
$$

Since the probability of Φ_u and $-\Phi_u$ are weighted with equal probability, we can deduce that $\mathbb{E}_{Q_\Delta}[\Phi_u] = 0$.

We now show the second result. Again $\Phi_u = \Delta F(\Theta, \Theta', u)$ and the variance-covariance matrix $Cov_{Q_\Delta}[\Phi_u] = Cov_{Q_\Delta}(\phi^{(i)}, \phi^{(j)})$ on $h(\Omega \times \Omega)$ is defined as

$$Cov_{Q_\Delta}(\phi^{(i)}, \phi^{(j)}) = E_{Q_\Delta}[\phi^{(i)}\phi^{(j)}]$$

$$= \int_{h(\Omega \times \Omega)} \phi^{(i)}\phi^{(j)} Q_\Delta(\phi, u) \, d\phi$$

$$= \int_{h(\Omega \times \Omega)} \phi^{(i)}\phi^{(j)} \left(\int_\Omega \int_\Omega \delta(\phi - \Delta F(\theta, \theta', u))Q_{\mu,\Sigma}(\theta)Q_{\mu,\Sigma}(\theta')d\theta'd\theta \right) \, d\phi$$

$$= \int_{h(\Omega \times \Omega)} \int_\Omega \int_\Omega \phi^{(i)}\phi^{(j)} \delta(\phi - \Delta F(\theta, \theta', u))Q_{\mu,\Sigma}(\theta)Q_{\mu,\Sigma}(\theta')d\theta'd\theta \, d\phi.$$

Now by Fubini's Theorem we have

$$Cov_{Q_\Delta}(\phi^{(i)}, \phi^{(j)}) = \int_\Omega \int_\Omega \int_{h(\Omega \times \Omega)} \phi^{(i)}\phi^{(j)} \delta(\phi - \Delta F(\theta, \theta', u))Q_{\mu,\Sigma}(\theta)Q_{\mu,\Sigma}(\theta') \, d\phi d\theta' d\theta.$$

Since $\phi^{(i)} = f(\theta, u^{(i)}) - f(\theta', u^{(i)}), \phi^{(j)} = f(\theta, u^{(j)}) - f(\theta', u^{(j)})$ it follows that

$$Cov_{Q_\Delta}(\phi^{(i)}, \phi^{(j)}) = \int_\Omega \int_\Omega \int_{h(\Omega \times \Omega)} [f(\theta, u^{(i)})f(\theta, u^{(j)}) + f(\theta', u^{(i)})f(\theta', u^{(j)})$$

$$- f(\theta', u^{(i)})f(\theta, u^{(j)}) - f(\theta, u^{(i)})f(\theta', u^{(j)})]$$

$$\delta(\phi - \Delta F(\theta, \theta', u))Q_{\mu,\Sigma}(\theta)Q_{\mu,\Sigma}(\theta') \, d\phi d\theta' d\theta.$$

Integrating with respect to ϕ while θ, θ' are fixed we find that

$$Cov_{Q_\Delta}(\phi^{(i)}, \phi^{(j)}) = \int_\Omega \int_\Omega [f(\theta, u^{(i)})f(\theta, u^{(j)}) + f(\theta', u^{(i)})f(\theta', u^{(j)})$$

$$- f(\theta', u^{(i)})f(\theta, u^{(j)}) - f(\theta, u^{(i)})f(\theta', u^{(j)})]Q_{\mu,\Sigma}(\theta)Q_{\mu,\Sigma}(\theta') \, d\theta' d\theta$$

$$= \int_\Omega \int_\Omega f(\theta, u^{(i)})f(\theta, u^{(j)})Q_{\mu,\Sigma}(\theta)Q_{\mu,\Sigma}(\theta') \, d\theta' d\theta$$

$$+ \int_\Omega \int_\Omega f(\theta', u^{(i)})f(\theta', u^{(j)})Q_{\mu,\Sigma}(\theta)Q_{\mu,\Sigma}(\theta') \, d\theta' d\theta$$

$$- \int_\Omega \int_\Omega f(\theta', u^{(i)})f(\theta, u^{(j)})Q_{\mu,\Sigma}(\theta)Q_{\mu,\Sigma}(\theta') \, d\theta' d\theta$$

$$- \int_\Omega \int_\Omega f(\theta, u^{(i)})f(\theta', u^{(j)})Q_{\mu,\Sigma}(\theta)Q_{\mu,\Sigma}(\theta') \, d\theta' d\theta$$

$$= \mathbb{E}_Q[f(\Theta, u^{(i)})f(\Theta, u^{(j)})] + \mathbb{E}_Q[f(\Theta, u^{(i)})f(\Theta, u^{(j)})]$$

$$- \mathbb{E}_Q[f(\Theta, u^{(i)})]\mathbb{E}_Q[f(\Theta, u^{(j)})] - \mathbb{E}_Q[f(\Theta, u^{(i)})]\mathbb{E}_Q[f(\Theta, u^{(j)})]$$

$$= 2Cov_Q[f(\Theta, u^{(i)}), f(\Theta, u^{(j)})].$$

Thus, $Cov_{Q_\Delta}[\Phi_u]$ is exactly twice the variance-covariance matrix for the model output.

∎

In [8] Dong *et al.* integrate with respect to the measure $\frac{1}{2}Q_\Delta$. This adjustment to the definition of $Cov_{Q_\Delta}[\Phi_u]$ is to compensate for the factor of 2 appearing in the covariance result above.

The second MINE criterion specifies that $u \in \mathcal{U}^d$ should be chosen such that the determinant of $Cov_{Q_\Delta}[\Phi_u]$, the volume $Cov_{Q_\Delta}[\Phi_u]$ induces, is maximized. Proposition 3.3.3 relates $Cov_{Q_\Delta}[\Phi_u]$ and $Cov_Q[f(\Theta, u^{(i)}), f(\Theta, u^{(j)})]$, and will be the basis of a correspondence between the determinant of $Cov_{Q_\Delta}[\Phi_u]$ and variances in the model output.

3.3.2 MINE IN LINEAR EXPERIMENTAL DESIGN

Now we consider the case of linear experimental design. Let $\hat{u} \in \mathcal{U}$ be an arbitrary time point. We define the measurement matrix $\Psi_u = \{\psi_j(u^{(i)})\}$, where each row of Ψ_u consists of the basis elements $\{\psi_j(\hat{u})\}_{j=1}^{d_p}$ evaluated at a particular component of $u = (u^{(1)}, \cdots, u^{(d)}) \in \mathcal{U}^d$. The basis elements are also called the regression functions in classical experimental design. Additionally, we define the reflected density $Q_{\mu,\Sigma}^-(\theta) = Q_{\mu,\Sigma}(2\mu - \theta)$ on Ω.

In the case of a linear mapping we have $f(\Theta, u) = \sum_{j=1}^{d_p} \Theta^{(j)} \psi_j(u)$ and the elements of our difference space are of the form $\Delta F(\Theta, \Theta', u) = \Psi_u(\Theta - \Theta')$. When $f(\Theta, u)$ is linear we are able to compute an explicit form for the covariance in terms of Ψ_u. We begin with a lemma that will assist in proving the covariance proposition.

Lemma 3.3.4 *The random variable $\Theta - \Theta'$ has probability density function $Q_{\mu,\Sigma} * Q_{\mu,\Sigma}^-$.*

Proof Probability densities for sums and differences of random variables can be constructed using convolution [30]. First, we note that $-\Theta'$ has density $Q_{-\mu,\Sigma}(-\theta)$. After translation by 2μ we find that $-\Theta$ has density

$$Q_{-\mu,\Sigma}(-\theta) = Q_{\mu,\Sigma}(2\mu - \theta)$$
$$= Q_{\mu,\Sigma}^{-}(\theta).$$

Hence, the random variable $\Theta - \Theta' = \Theta + (-\Theta')$ has the desired probability density $Q_{\mu,\Sigma} * Q_{\mu,\Sigma}^{-}$. ∎

Proposition 3.3.5 *If $\Phi_u = \Delta F(\Theta, \Theta', u) = \Psi_u(\Theta - \Theta')$ then*

$$Cov_{Q_\Delta}(\Phi_u) = 2\Psi_u \Sigma \Psi_u^T.$$

Proof By the independence and identical distribution of Θ and Θ' we have $\mathbb{E}_Q[\Theta - \Theta'] = 0$ and $Cov_Q(\Theta - \Theta') = 2\Sigma$.

The first result of Proposition 3.3.3 gives $\mathbb{E}_{Q_\Delta}[\Phi_u] = 0$ and when considering $Cov_{Q_\Delta}(\Phi_u)$ we find

$$Cov_{Q_\Delta}(\Phi_u) = \mathbb{E}_{Q_\Delta}[\Phi_u \Phi_u^T]$$
$$= \mathbb{E}_{Q_\Delta}[\Psi_u(\Theta - \Theta')(\Theta - \Theta')^T \Psi_u^T]$$
$$= \Psi_u \mathbb{E}_{Q_\Delta}[(\Theta - \Theta')(\Theta - \Theta')^T]\Psi_u^T.$$

When considering $\mathbb{E}_{Q_\Delta}[(\Theta - \Theta')(\Theta - \Theta')^T]$ we find that

$$\mathbb{E}_{Q_\Delta}[(\Theta - \Theta')(\Theta - \Theta')^T] = \int_{h(\Omega \times \Omega)} (\theta - \theta')(\theta - \theta')^T Q_\Delta(\phi)d\phi$$
$$= \int_{h(\Omega \times \Omega)} (\theta - \theta')(\theta - \theta')^T \left(\int_\Omega \int_\Omega \delta(\phi - \Delta F(\theta, \theta', u))Q_{\mu,\Sigma}(\theta)Q_{\mu,\Sigma}(\theta') \, d\theta' d\theta \right)d\phi$$

and employing Fubini's Theorem we have

$$\mathbb{E}_{Q_\Delta}[(\Theta - \Theta')(\Theta - \Theta')^T] = \int_\Omega \int_\Omega \int_{h(\Omega \times \Omega)} (\theta - \theta')(\theta - \theta')^T$$
$$\delta(\phi - \Delta F(\theta, \theta', u))Q_{\mu,\Sigma}(\theta)Q_{\mu,\Sigma}(\theta') \, d\phi d\theta' d\theta$$
$$= \int_\Omega \int_\Omega (\theta - \theta')(\theta - \theta')^T Q_{\mu,\Sigma}(\theta)Q_{\mu,\Sigma}(\theta') \, d\theta' d\theta$$
$$= \mathbb{E}_Q[(\Theta - \Theta')(\Theta - \Theta')^T].$$

and hence

$$Cov_{Q_\Delta}(\Phi_u) = \Psi_u \mathbb{E}_Q[(\Theta - \Theta')(\Theta - \Theta')^T]\Psi_u^T$$
$$= 2\Psi_u \Sigma \Psi_u^T.$$

∎

Thus, for a linear model we have that the covariance in difference space can be found by conjugating the covariance on Ω with the measurement matrix Ψ_u. In applications linear dependence on parameters is often assumed and this form for $Cov_{Q_\Delta}(\Phi_u)$ is more amenable to implementation. Moreover, this result illustrates a connection to classical experimental design through the prior and measurement matrix. The result also illustrates a limitation of the method, however. Recall that the Ψ_u is a $d \times d_p$ matrix with $d \le d_p$; the method leads to different analyses for the case $d < d_p$ and $d = d_p$. We discuss the case $d = d_p$ in Section 3.3.5 and show that the measurement points are independent of the prior.

3.3.3 CONVERGENCE OF THE EXPECTED DYNAMICS ESTIMATOR

We consider the variance-covariance matrix found in Section 3.3.1 and demonstrate convergence of expected model dynamics when choosing control points that maximize its determinant. In Dinh *et al.* [10] it is shown that if a single measurement point $u_n = (u_n^{(1)})$ is chosen sequentially according to where the maximum variance in the model occurs, and that if the probability density on parameter space is updated at every selection of u_n then the expected model dynamics will converge to the true system dynamics.

Stated formally, the theorem says the following.

Theorem 3.3.6 (Dinh *et al.* 2014) *Let Ω have finite cardinality. Suppose there exists $\Theta_0 \in \Omega$ such that $f(\Theta_0, u) = g(u)$ for all $u \in \mathcal{U}$ and that $1 \le r < \infty$. Let $Q_n(\theta) = c_n \exp(-\sum_{k=1}^n |f(\theta, u_k) - g(u_k)|^r)$ where c_n is the normalizing constant to*

ensure that Q_n is a probability density on Ω. Suppose also that there exists a $C > 0$ and that $\{u_n\}_{n=1}^{\infty}$ are chosen such that for all $u \in \mathcal{U}$

$$C \cdot Var_{Q_n}(f(\Theta, u)) \leq Var_{Q_n}(f(\Theta, u_{n+1})).$$

Then for all $u \in \mathcal{U}$

$$\lim_{n \to \infty} \mathbb{E}_{Q_n}[f(\Theta, u)] = g(u).$$

In the above theorem and the main theorem later this section we notice that there is an added assumption that our parameter space Ω is finite. This assumption is necessary to rule out convergence to incorrect dynamics that can occur when Ω is an open set, our probability distribution is continuous, and the model is not robust in parameters. The necessity of this assumption is discussed in more detail in [10].

In preparation for extending this result to the second MINE criterion, recall that $\Psi_u = \{\psi_j(u^{(i)})\}$ is a $d \times d_p$ matrix; Σ, the variance-covariance matrix on Ω with respect to Q, is a $d_p \times d_p$ matrix; and $V(u) = \det Cov_{Q_\Delta}(\Phi_u)$. Finally, for $\hat{u} \in \mathcal{U}$ we denote $Var_Q(f(\Theta, \hat{u})) = \sigma_{\hat{u}}^2$.

In our setting we have a vector of measurements at every step of the sequential design. We use the same probability distribution as in Theorem 3.3.6, but to account for the vector of measurements we sum the d components of the control vector, which leads to a double sum. Our probability distribution is then $Q_n(\theta) = c_n \exp(-\sum_{k=1}^{n} \sum_{i=1}^{d} |f(\theta, u_k^{(i)}) - g(u_k^{(i)})|^r)$. Our first lemma shows that variances with respect to the probability density $Q_n(\theta)$ are uniformly bounded.

Lemma 3.3.7 *Let $f \in C_b(\Omega, \mathcal{U})$. If $Q_n(\theta) = c_n \exp(-\sum_{k=1}^{n} \sum_{i=1}^{d} |f(\theta, u_k^{(i)}) - g(u_k^{(i)})|^r)$ where $1 \leq r < \infty$ then there exists $M > 0$ such that*

$$Var_{Q_n}(f(\Theta, u)) \leq M,$$

for all n and $u \in \mathcal{U}$.

Proof To bound $Var_{Q_n}(f(\Theta, u))$ we first bound $\mathbb{E}_{Q_n}(f(\Theta, u))$ and then $\mathbb{E}_{Q_n}(f(\Theta, u)^2)$. Since f is continuous and bounded we know that $\|f(\theta, u)\|_{L^\infty(\Omega, \mathcal{U})} < \infty$. Bounding the integrands of $\mathbb{E}_{Q_n}(f(\Theta, u))$ and $\mathbb{E}_{Q_n}(f(\Theta, u)^2)$ by the L^∞ norm we find

$$|\mathbb{E}_{Q_n}(f(\Theta, u))| \leq \|f\|_{L^\infty(\Omega, \mathcal{U})}$$

and

$$\mathbb{E}_{Q_n}(f(\Theta, u)^2) \leq \|f\|^2_{L^\infty(\Omega, \mathcal{U})},$$

respectively.

Combining these bounds with the triangle inequality we find that

$$Var_{Q_n}(f(\Theta, u)) = |\mathbb{E}_{Q_n}(f(\Theta, u)^2) - \mathbb{E}_{Q_n}(f(\Theta, u))^2|$$
$$\leq 2\|f\|^2_{L^\infty(\Omega, \mathcal{U})}.$$

We define $M = 2\|f\|^2_{L^\infty(\Omega, \mathcal{U})}$ and we have our desired uniform bound. ∎

In the case $d = 1$, $V(u)$ corresponds to the maximal variance case considered by Dinh *et al.* Notice first that for $d = 1$ we have

$$\Phi_u = f(\Theta, u^{(1)}) - f(\Theta', u^{(1)}),$$

and by Proposition 3.3.3 we have

$$Cov_{Q_\Delta}(\Phi_u) = 2Cov_Q(f(\Theta, u^{(1)}), f(\Theta, u^{(1)}))$$
$$= 2Var(f(\Theta, u^{(1)})),$$

which is the objective function used by Dinh *et al.*

We now consider the case $2 \leq d < d_p$. This case corresponds to choosing multiple measurement points at the nth step in the limiting process; we write the measurement points as a control vector $u_n = (u_n^{(1)}, u_n^{(2)}, ..., u_n^{(d)})$. We prove a proposition by contradiction which will guarantee that at every sequential step at least one diagonal element of $Cov_{Q_\Delta}(\Phi_u)$, a model variance, satisfies Theorem 3.3.6 from Dinh *et*

al. That is, one of our measurements with the second MINE criterion will be within some fixed constant factor of the maximum model variance and thus the expected model dynamics will still converge to the true system dynamics.

Formally, the lemma is stated as follows.

Lemma 3.3.8 *Let* $2 \leq d < d_p$. *Let* Q *be the probability density function on* Ω *and define* $Var_Q(f(\Theta, \hat{u})) = \sigma_{\hat{u}}^2$, $\hat{u} \in \mathcal{U}$. *There exists* $C > 0$ *such that if for all* $u^* = (u^{(1)}, u^{(2)}, ..., u^{(d)}) \in \mathcal{U}^d$ *such that* $\det Cov_{Q_\Delta}(\Phi_{u*}) = \max_u \det Cov_{Q_\Delta}(\Phi_u) > 0$ *and for all* $y \in \mathcal{U}$ *satisfying* $\sigma_y^2 \geq \sigma_{\hat{u}}^2$ *for all* $\hat{u} \in \mathcal{U}$ *then* $\max_i \sigma_{u^{(i)}}^2 \geq C \cdot \sigma_y^2$.

Proof We assume for the sake of contradiction that the hypothesis does not hold, i.e., for all $C > 0$, there exists $u_C^* = (u_C^{(1)}, u_C^{(2)}, ..., u_C^{(d)})$ such that $\det Cov_{Q_\Delta}(\Phi_{u_C^*}) = \max_u Cov_{Q_\Delta}(\Phi_u) = \delta > 0$ and there exists $y_C \in \mathcal{U}$ satisfying $\sigma_{y_C}^2 \geq \sigma_{\hat{u}}^2$, for all $\hat{u} \in \mathcal{U}$ we have $\max_i \sigma_{u_C^{(i)}}^2 \leq C \cdot \sigma_{y_C}^2$. Furthermore, we choose $C_n = \frac{1}{n}$.

By Lemma 3.3.7 we have the uniform bound $\sigma_{y_n}^2 \leq M$ and by choice of C_n we know $\max\{\sigma_{u_{C_n}^{(1)}}^2, \sigma_{u_{C_n}^{(2)}}^2, \because, \sigma_{u_{C_n}^{(d)}}^2\} \leq \frac{1}{n}\sigma_{y_n}^2$. Now we employ Hadamard's Inequality for positive semi-definite matrices. The inequality states that for a positive semi-definite matrix A its determinant satisfies the bound $\det(A) \leq \prod a_{ii}$. Since the variance-covariance matrix is always positive semi-definite and by Proposition 3.3.3 we have $\{Cov_{Q_\Delta}(\Phi_u)\}_{ii} = 2\sigma_{u^{(i)}}^2$ and so

$$\det Cov_{Q_\Delta}(\Phi_{u_{C_n^*}}) \leq 2^d \prod_{i=1}^d \sigma_{u_{C_n}^{(i)}}^2$$
$$\leq \left(\frac{2}{n}\sigma_{y_n}\right)^d$$
$$\leq \left(\frac{2}{n}M\right)^d \to 0 \text{ as } n \to \infty,$$

which is a contradiction. ■

Theorem 3.3.9 *Let* $1 \leq d < d_p$ *and* Ω *have finite cardinality. Suppose there exists* $\Theta_0 \in \Omega$ *such that* $f(\Theta_0, \hat{u}) = g(\hat{u})$ *for all* $\hat{u} \in \mathcal{U}$ *and that* $1 \leq r < \infty$. *Let* $Q_n(\theta) = c_n \exp(-\sum_{k=1}^n \sum_{i=1}^d |f(\theta, u_k^{(i)}) - g(u_k^{(i)})|^r)$ *where* c_n *is the normalizing constant to ensure that* Q_n *is a probability density on* Ω *and* Σ_n *is the variance-covariance*

matrix on Ω with respect to Q_n. Suppose also that $u_{n+1} = (u_{n+1}^{(1)}, u_{n+1}^{(2)}, ..., u_{n+1}^{(d)}) \in \mathcal{U}^d$
are chosen such that

$$\det Cov_{Q_\Delta}(\Phi_u) \leq \det Cov_{Q_\Delta}(\Phi_{u_{n+1}}), \;\; for \; all \; u \in \mathcal{U}.$$

Then for all $\hat{u} \in \mathcal{U}$

$$\lim_{n \to \infty} \mathbb{E}_{Q_n}[f(\Theta, \hat{u})] = g(\hat{u}).$$

Proof   Under the hypotheses, the $d = 1$ case corresponds to exactly to Theorem 3.3.6.

For the case $2 \leq d < d_p$, we choose $u_{n+1} = (u_{n+1}^{(1)}, u_{n+1}^{(2)}, ..., u_{n+1}^{(d)})$ such that $\det Cov_{Q_\Delta}(\Phi_{u_{n+1}})$ is maximal at the $(n+1)^{st}$ step. Again by Proposition 3.3.3, we have $\{Cov_{Q_\Delta}(\Phi_u)\}_{ii} = 2\sigma^2_{u^{(i)}}$. Substituting $Q_{n+1}(\theta)$ for $Q(\theta)$, by Lemma 3.3.8 there exists $u_{n+1}^{(j_{n+1})} \in u_{n+1}$ such that $\sigma^2_{u_{n+1}^{(j_{n+1})}} \geq C \cdot \sigma^2_u$ for all u. Now the points $\{u_n^{(j_n)}\}_{n=1}^\infty$ form a sequence that satisfy Theorem 3.3.6 and thus, $\mathbb{E}_{Q_n}[f(\Theta, \hat{u})] \to g(\hat{u})$ as $n \to \infty$.
∎

3.3.4   NUMERICAL ILLUSTRATION OF THE METHOD.

We illustrate an implementation of the second MINE criterion in the case of regression with Chebyshev polynomials. The regression problem is linear in parameters, so we proceed to maximize the objective function described in Proposition 3.3.5 in order to find points to measure and therefore reduce uncertainty. Recall that the first five Chebyshev Polynomials on $I = [-1, 1]$ [31] are $\{\cos((j-1)\cos^{-1}\hat{u})\}_{j=1}^5$ and these will serve as the basis $\{\psi_j(\hat{u})\}_{j=1}^5$ for regression, note here that $d_p = 5$. We seek to take $d = 2$ measurements at every step in our sequential design.

Let $\sigma > 0$ and Σ_0 be symmetric positive definite and conditional probability of Θ given σ^2 is $p(\theta|\sigma^2) = \mathcal{N}(\mu_0, \sigma^2\Sigma_0^{-1})$. Our linear model is of the form $y = \Psi_u\Theta + \epsilon$, where each component of of ϵ has distribution $\mathcal{N}(0, \sigma^2)$. We assume that the prior distribution on Ω is the multivariate normal distribution. Combining this prior with the probability distribution induced on parameters with $r = 2$, as defined in Section 3.3.3,

we obtain the posterior distribution $Q_n(\theta) = c_n p(\theta|\sigma^2) \exp(-\sum_{k=1}^{n} \sum_{i=1}^{d} |\Psi_{u_k^{(i)}} \theta - g(u_k^{(i)})|^2)$ defined in Section 3.3.3 with $r = 2$ is then our posterior distribution. Using the results found in [31] we find that our posterior distribution Q_n has parameters

$$\mu_n = (\Psi_u^T \Psi_u + \Sigma_0)^{-1}(\Sigma_0 \mu_0 + \Psi_u^T y)$$
$$\Sigma_n = (\Psi_u^T \Psi_u + \Sigma_0).$$

These parameters are necessary to specify the distribution from which we sample at every step in the sequential design. Note that σ is fixed throughout this updating scheme; there is a version of the scheme where σ can be updated as well.

Before beginning the scheme, we choose true parameters θ^{true} so that we have true dynamics $y_u^{true} = \Psi_u \theta^{true}$. In the script, we sample from the current distribution on parameter space, take differences of the parameters, and map the differences forward to the difference space under the transformation $\Psi_u = \{\psi_j(u^{(i)})\}_{i=1,j=1}^{i=2,j=5}$. Each difference space depends on a choice of point u.

We used the built-in MATLAB function *fmincon* to find a minimum to the objective function $-\Psi_u^T \Sigma \Psi_u$. To find initial measurement points to start the optimization, the interval of consideration I was subdivided and the points from the subdivision with the best evaluation in the objective were chosen. The optimization procedure was used iteratively and found measurement points that improved upon the initial estimates at every step n.

For $n = 1$, the difference spaces at the initial and optimal point is shown in the right panel of Figure 3.1. When moving to the optimal measurement point for $n = 1$ there is a visible increase in the volume (area, for $d = 2$) of the difference space. This increase corresponds to greater uncertainty at the optimal points u_1^*, which will be resolved by measuring the dynamics at these points.
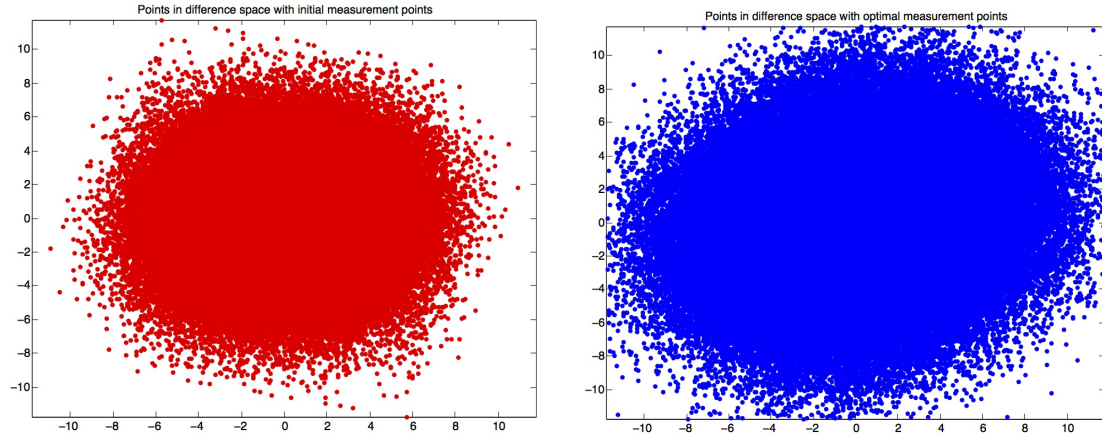
Fig. 3.1.: Plots of the difference space given a random sample of Ω and a choice of measurement points u. On the left, the difference space with initial measurement points is plotted in red. On the right, the difference space with optimal measurement u_1^* points is plotted in blue.

In Figures 3.2-3.7, we indicate how the regression functions evolve as we apply the sequential design. The dynamic uncertainty, or model variance, in the regression functions has been significantly reduced at the 5th iteration. There is not much noticeable improvement between the 6th and $10t^{th}$ iteration. In each of the plots the true dynamics are represented by a solid green curve and the optimal points are solid blue vertical lines. We sample by adding Gaussian noise with standard deviation $\sigma = 0.1$ to y_u^{true}.

At step $n = 1$, there is wide variation throughout the entire interval. The endpoints are found to be the place where experimentation will yield the most reduction in the dynamic uncertainty.

Fig. 3.2.: Regression functions at step $n = 1$. The optimal points for measurement u_1^* were found the be the endpoints.

By step $n = 2$, the variation at the endpoints has been resolved due to the measurements. The next measurement points u_2^* are found to be about the origin and slightly above 0.8, and thereby focusing on uncertainty in the top half of the interval.

Fig. 3.3.: Regression functions at step $n = 2$. The optimal points for measurement $u_2^* \approx (0.02, 0.82)$.

In the plot for $n = 3$ we see that the trajectories have all significantly contracted at the origin and above 0.8. The next measurement points are chosen to be ≈ 0.5 ≈ -0.8

Fig. 3.4.: Regression functions at step $n = 3$. The optimal points for measurement are $u_3^* \approx (-0.82, .5)$.

At step $n = 4$, the variation outside of the interval (-0.8, 0) has been largely constrained. The choice of measurements at this step is interesting. The measurement point at ≈ -0.48 is an obvious candidate since the maximum model variance occurs in this region. The choice of point at ≈ -0.02. is surprising, since there are other points of higher uncertainty and it is not apparent how this point is less correlated to the first chosen point.

Fig. 3.5.: Regression functions at step $n = 4$. The optimal points for measurement $u_4^* \approx (-0.48, -0.02)$.

By step $n = 5$, most areas of large model variance have been reduced by the experiments. The optimal points are chosen to be the endpoints.
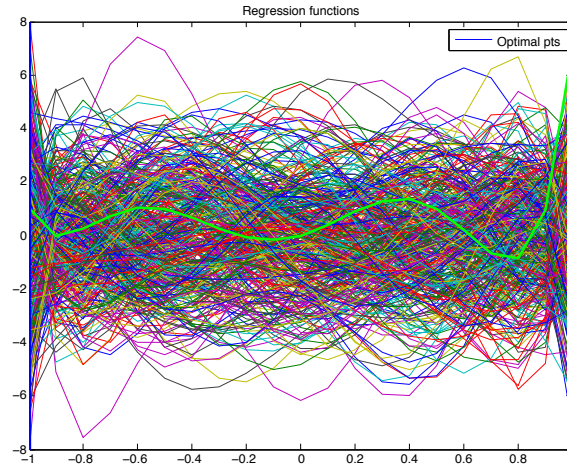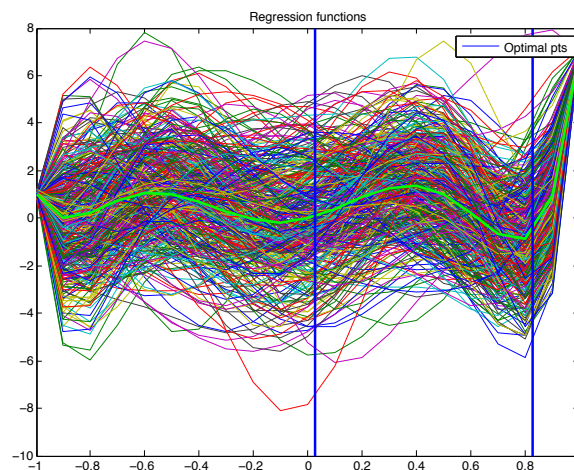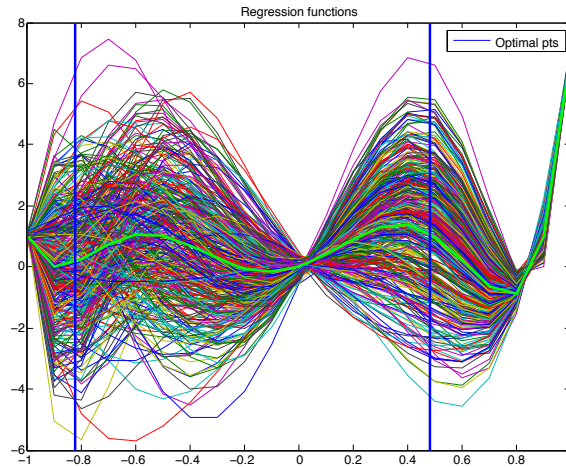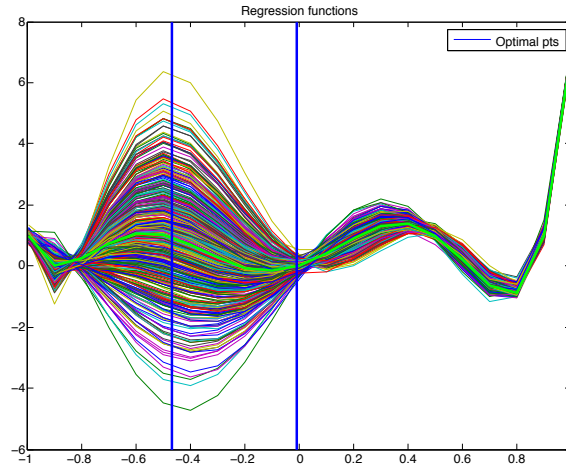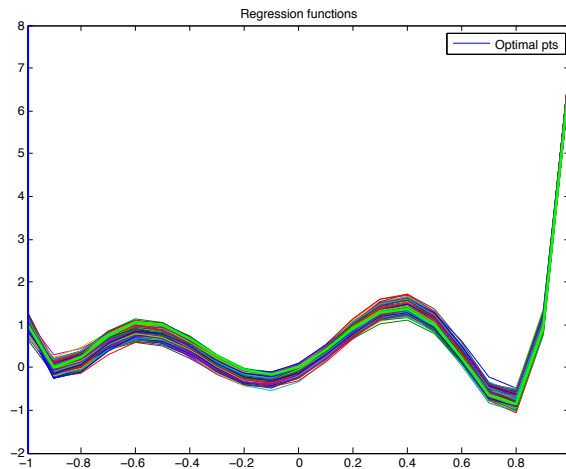

Fig. 3.6.: Regression functions at step $n = 5$. The optimal points for measurement u_5^* were found the be the endpoints.

Between $n = 6$ and $n = 10$ there is not a large gain in convergence to y^{True}, but the method continues to find points that improve the regression. The choice of measurements at step $n = 5$ are at $\hat{u} = 0$ and $\hat{u} = -0.8$.
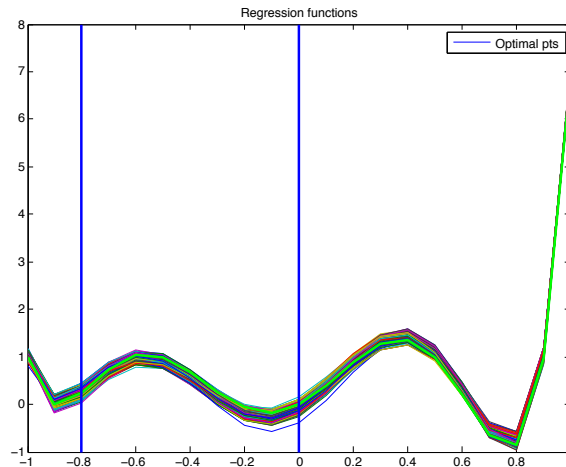
Fig. 3.7.: Regression functions at step $n = 10$. The optimal points for measurement $u_{10}^* = (-0.8, 0)$.

The numerical implementation shows that points chosen according the second MINE criterion are effective in reducing uncertainty in this case of linear regression. The points chosen are not necessarily the points of largest uncertainty at each step and reasonable convergence is still acquired after a small number of iterations.

3.3.5 LIMITATIONS IN THE MINE CRITERION

The MINE method generally relies on the prior distribution on Ω in order to make decisions in difference space; we present how the number of measurements can adversely affect the influence of the prior. We consider two potential cases when choosing two measurement points for a model that has parameters in a two dimensional space. We assume linearity of parameters for the model. We summarize the section with a corollary regarding the arbitrary case $d = d_p$.

Let $I \subset \mathbb{R}$ be an interval that contains the origin, though not necessarily symmetric about the origin. We consider the model $f(\Theta, \hat{u}) = \Theta^{(0)} + \Theta^{(1)} \hat{u}$ on I with $\Theta = [\Theta^{(0)} \ \Theta^{(1)}]^T \in \Omega = [0, 1]^2$. Note that $\Theta^{(0)}$ determines the vertical intercept and $\Theta^{(1)}$ determines the slope. Given measurement points $u = \{a, b\}$, where $a, b \in I$, we define the measurement matrix $\Psi_u = \begin{bmatrix} 1 & a \\ 1 & b \end{bmatrix}$.

Considering narrow ellipsoids in Ω reveals one limitation of the MINE criteria. A narrow ellipsoid parallel to the $\Theta^{(0)}$ axis produces high variability in the intercept $f(\Theta, \hat{u})|_{\hat{u}=0}$ while leaving the slope relatively unchanged. This is illustrated in Figure 3.8. This also corresponds to the variance-covariance matrix $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & \epsilon \end{bmatrix}$ on Ω. In ΔF-space the trajectories become flat due to the narrow range for $\Theta^{(1)}$ and the high variability in the vertical intercept persists.

Fig. 3.8.: On the left, the confidence ellipsoid in Ω parallel to the $\Theta^{(0)}$ axis. On the right, the trajectories produced in difference space when sampling from the ellipsoid, demonstrating high intercept variability and low variability in the slopes.

On the other hand, we see in Figure 3.9 that a narrow ellipsoid parallel to the $\Theta^{(1)}$ axis produces high variability in the slope of $f(\Theta, \hat{u})$ while leaving the vertical intercept $f(\Theta, \hat{u})|_{\hat{u}=0}$ relatively unchanged. Here the variance-covariance matrix is $\Sigma = \begin{bmatrix} \epsilon & 0 \\ 0 & 1 \end{bmatrix}$ on Ω. The trajectories translate into trajectories in ΔF-space still

with high variability in their slope but all are centered around the origin since the range for $\Theta^{(0)}$ was so narrow.
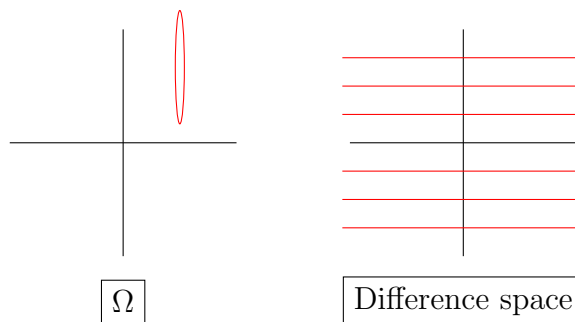

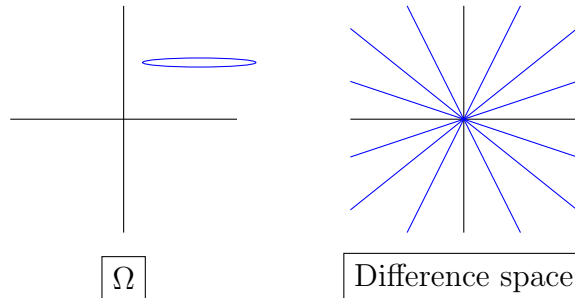Fig. 3.9.: On the left, a thin confidence ellipsoid in Ω parallel to the $\Theta^{(1)}$ axis. On the right, the trajectories produced in difference space when sampling from the ellipsoid, demonstrating high slope variability and low variability in the intercepts.

For an arbitrary diagonal variance-covariance matrix $\Sigma = \begin{bmatrix} c & 0 \\ 0 & d \end{bmatrix}$ we find that the variance in the ΔF-space is $\Psi_u \Sigma \Psi_u^T = \begin{bmatrix} 1 & a \\ 1 & b \end{bmatrix} \begin{bmatrix} c & 0 \\ 0 & d \end{bmatrix} \begin{bmatrix} 1 & 1 \\ a & b \end{bmatrix} = \begin{bmatrix} c + a^2 d & c + abd \\ c + abd & c + b^2 d \end{bmatrix}$

In the case when $c = 1, d = \epsilon$ we have $\Psi_u \Sigma \Psi_u^T = \begin{bmatrix} 1 + a^2\epsilon & 1 + ab\epsilon \\ 1 + ab\epsilon & 1 + b^2\epsilon \end{bmatrix}$.

Now $V(u) = \det \Psi_u \Sigma \Psi_u^T = (1 + a^2\epsilon)(1 + b^2\epsilon) - (1 + ab\epsilon)^2 = (b - a)^2\epsilon$.

In the case when $c = \epsilon, d = 1$ we have $\Psi_u \Sigma \Psi_u^T = \begin{bmatrix} \epsilon + a^2 & \epsilon + ab \\ \epsilon + ab & \epsilon + b^2 \end{bmatrix}$, and here

$V(u) = \det \Psi_u \Sigma \Psi_u^T = (\epsilon + a^2)(\epsilon + b^2) - (\epsilon + ab)^2 = (b - a)^2\epsilon$.

In both cases, the implication is that the volume $V(u)$ is maximized when the distance between a and b is maximized. This is problematic since the resolution in the model dynamics may not be improved depending on the interval I and also since there is no dependence on the prior distribution Σ.

For instance, we know that if we have a narrow ellipsoid parallel to the $\Theta^{(1)}$ axis then our intercepts $f(\Theta, \hat{u})|_{\hat{u}=0}$ are all about the origin. If we chose $I = [0, T]$ then

the maximal volume suggests we should take measures at $a = 0$ and $b = T$. Since the trajectories in this case are about the origin the choice $a = 0$ will not produce information useful in constraining our model trajectories.

In general, if $d = d_p$ we have that Ψ_u is square and our objective becomes $\det \Psi_u \Sigma \Psi_u^T = (\det \Psi_u)^2 \det \Sigma$. The peculiarity here is that our next choice of u which results from maximizing $\det \Psi_u \Sigma \Psi_u^T$ is independent of our prior information Σ. We now state the corollary formally.

Corollary 3.3.10 *If* $\Phi_u = \Delta F(\Theta, \Theta', u) = \Psi_u(\Theta - \Theta')$ *and* $d = d_p$, *then the maximization problem*

$$\max_u V(u) = \max_u \det \Psi_u \Sigma \Psi_u^T,$$

is independent of the prior distribution Σ on Ω.

Moreover, the case $d > d_p$ is not considered as this would lead to a measurement matrix Ψ_u with dependent rows. Thus, there would be no new information gained at that step of sequential rows from at least $d - d_p$ of the measurements.

3.3.6 DISCUSSION

When dealing with epistemic uncertainty modelers often call on experimentalists to gain a better understanding of biological process under consideration. Experiments can often provide more insight into what are reasonable dynamics for a given model. The Maximally Informative Next Experiment criteria specifies when an experiment should be conducted to gain the largest resolution of dynamic uncertainty. An advantage to the MINE criteria is their applicability to nonlinear models; this is appealing for systems biology where the processes and mechanisms often necessitate complicated descriptions.

The second MINE criterion can be used to specify a number of experiments fewer than the dimension of the parameter space. Choosing multiple points for experimentation is a problem of finding points of large uncertainty, but also finding points

with minimal correlation. This problem is solved by considering determinants of covariance matrices; the determinant is an ideal objective as it will promote a lack of correlation during maximization. We have given a theoretical validation for using the second MINE criterion. Our main result in this chapter shows that the method can be used to generate a sequential design of experiments that will, in the limit, lead to convergence of the expected dynamics to the true dynamics. This convergence result applies generally to nonlinear models. Additionally, we have explored limitations of the method and have found that it is useful for a parallel design only when the number of experiments is less than the dimension of the parameter space. This is not true in general for the classic D-optimality condition.

Again, the second MINE criterion was considered by Bouffier *et al.* in [28] for the linear case. Bouffier *et al.* give a sufficient condition for maximization of $\det \Psi_u \Sigma \Psi_u^T$, but still seek a necessary condition for maximization. Results like Lemma 3.3.8 could possibly be used to construct a partial converse to Theorem 1 in [28], as it states that when the determinant of the covariance matrix is maximized at least one diagonal element of $\Psi_u \Sigma \Psi_u^T$ must be within a fixed constant of the maximum model variance.

We have illustrated with numerical examples that it is practical to expect approximations to true dynamics after a reasonable number of iterations in the case of linear regression. Additionally, this method may provide a means for discovering underlying correlations in the model response at different points and insight into model behavior.

4. FUTURE WORK

Future work on this project includes further model development, using the theoretical results on Maximally Informative Next Experiments (MINE) to aid in that development, and extending the theoretical results to other MINE criterion.

4.1 MODEL DEVELOPMENT

The B cell receptor (BCR) signaling model presented in Chapter 2 currently contains tuned Erk-MAPK and NF-κB pathways. There is a structure in place for the NFAT pathway, but the current model parameters have not been fit to B cell data. The NFAT pathway contains five states and the Sobol sensitivity analysis conducted in Section 2.4.2 found six sensitive parameters in the pathway. Identifying values for these parameter that produce reasonable dynamics will be necessary before the pathway can be used to generate testable hypotheses.

There is limited data available in [16] and more time points will be needed to constrain the NFAT dynamics. Further experimentation is thus required and the MINE Criteria can be used to identify time points that will have the most impact on reducing dynamic uncertainty in this part of the model.

4.2 MODEL ANALYSIS

The contour analysis in Section 2.5.3 considered model response as a function of ligand binding rates and found distinct dependencies in regions defined by the reverse binding rate (see Figure 2.8). At higher levels of the reverse rate the responses follow linear contours of slope α; biologically this corresponds to depending on a power law affinity $K_{a,\alpha} = \frac{rw0_{kf}}{(rw0_{kr})^\alpha}$. To trace the mechanism behind the dependencies, we

are constructing motifs that represent characteristic features of the BCR signaling model. The motifs translate into lower dimensional models that are more amenable to analysis.

Motifs have been constructed to reflect features between the receptor and Syk. Each motif focuses on a specific feature, such as Syk's autophosphorylation, binding events, and inhibition. Thus far, the motifs produce similar types contours to those seen in Figure 2.8 and there is interest in proving a limiting result that approximates α, the slope of the linear function at high reverse rate.

4.3   THEORETICAL RESULTS

As mentioned in Section 3.1.2, the third MINE criterion is related to the second criterion; the volume in this case is maximized by increasing the independence between the measurement points. Using a Hilbert Space formulation, the authors view each model output as a vector and the respective vector norms then normalize the determinant from the second criterion. This normalization ensures that increases in volume are due to increases in independence between measurement points and not the magnitude of the outputs. Extending the results in Chapter 3 will be considered in a future works.

REFERENCES

REFERENCES

[1] Y. Zheng and A. Rundell, "Comparative study of parameter sensitivity analyses of the tcr-activated erk-mapk signalling pathway," *IEE Proceedings - Systems Biology*, vol. 153, pp. 201–211, 2006.

[2] J. Keener and J. Sneyd, *Mathematical Physiology*, ser. Interdisciplinary applied mathematics. Springer, 1998. [Online]. Available: https://books.google.com/books?id=-Ist8n0l5LsC

[3] J. P. Perley, J. Mikolajczak, M. L. Harrison, G. T. Buzzard, and A. E. Rundell, "Multiple model-informed open-loop control of uncertain intracellular signaling dynamics," *PLOS Computational Biology*, vol. 10, pp. 1296–1310, 2014.

[4] J. P. Perley, J. Mikolajczak, G. T. Buzzard, M. L. Harrison, and A. E. Rundell, "Resolving early signaling events in t-cell activation leading to il-2 and foxp3 transcription," *Processes*, vol. 2, no. 4, pp. 867–900, 2014. [Online]. Available: http://www.mdpi.com/2227-9717/2/4/867

[5] E. Hofer, M. Kloos, B. Krzykacz-Hausmann, J. Peschke, and M. Woltereck, "An approximate epistemic uncertainty analysis approach in the presence of epistemic and aleatory uncertainties," *Reliability Engineering & System Safety*, vol. 77, no. 3, pp. 229 – 238, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095183200200056X

[6] V. Fedorov and P. Hackl, *Model-Oriented Design of Experiments*. Springer-Verlag, 1997.

[7] F. Pukelsheim, *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics, 2006.

[8] W. Dong, X. Tang, Y. Yu, R. Nilsen, R. Kim, J. Griffith, J. Arnold, and H.-B. Schttler, "Systems biology of the clock in ¡italic¿neurospora crassa¡/italic¿," *PLoS ONE*, vol. 3, no. 8, p. e3105, 08 2008.

[9] J. Bazil, G. Buzzard, and A. Rundell, "A global parallel model based design of experiments method to minimize model output uncertainty," *Bulletin of Mathematical Biology*, vol. 74, no. 3, pp. 688–716, 2012. [Online]. Available: http://dx.doi.org/10.1007/s11538-011-9686-9

[10] V. Dinh, A. E. Rundell, and G. T. Buzzard, "Experimental design for dynamics identification of cellular processes," *Bulletin of Mathematical Biology*, vol. 76, no. 3, pp. 597–626, 2014.

[11] H. Oh, E. Ozkirimli, K. Shah, M. L. Harrison, and R. L. Geahlen, "Generation of an analog-sensitive syk tyrosine kinase for the study of signaling dynamics from the b cell antigen receptor," *Journal of Biological Chemistry*, vol. 282, pp. 33 760–33 768, 2007.

[12] P. K. Tsourkas, C. D. Somkanya, P. Yu-Yang, W. Liu, S. K. Pierce, and S. Raychaudhuri, "Formation of bcr oligomers provides a mechanism for b cell affinity discrimination," *Journal of Theoretical Biology*, vol. 307, pp. 174–182, 2012.

[13] S. Mukherjee, J. Zhu, J. Zikherman, R. Parameswaran, T. A. Kadlecek, Q. Wang, B. Au-Yeung, H. Ploegh, J. Kuriyan, J. Das, and A. Weiss, "Monovalent and multivalent ligation of the b cell receptor exhibit differential dependence upon syk and src family kinases," *Science Signaling*, vol. 6, p. ra1, 2013.

[14] V. K. Chaudhri, D. Kumar, M. Misra, R. Dua, and K. V. S. Rao, "Integration of a phosphatase cascade with the mitogen-activated protein kinase pathway provides for a novel signal processing function," *Journal of Biological Chemistry*, vol. 285, pp. 1296–1310, 2010.

[15] D. Barua, W. S. Hlavacek, and T. Lipniacki, "A computational model for early events in b cell antigen receptor signaling: Analysis of the roles of lyn and fyn," *Journal Immunology*, vol. 189, pp. 646–658, 2012.

[16] J. I. Healy, R. E. Dolmetsch, L. A. Timmerman, J. G. Cyster, M. L. Thomas, G. R. Crabtree, R. S. Lewis, and C. C. Goodnow, "Different nuclear signals are activated by the b cell receptor during positive versus negative signaling," *Immunity*, vol. 6, pp. 419–428, 1997.

[17] B. J. Skaggs and M. R. Clark., "Proximal b cell receptor signaling pathways," *Signal Trransduction*, vol. 4, pp. 173–194, 2004.

[18] T. Kurosaki and M. Hikida, "Tyrosine kinases and their substrates in b lymphocytes," *Immunological Reviews*, vol. 228, pp. 132–148, 2009.

[19] R. L. Geahlen, "Syk and ptyr'd: Signaling through the b cell antigen receptor," *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, vol. 1793, pp. 1115–1127, 2009.

[20] H. Ma, T. M. Yankee, J. Hu, D. J. Asai, M. L. Harrison, and R. L. Geahlen, "Visualization of syk-antigen receptor interactions using green fluorescent protein: differential roles for syk and lyn in the regulation of receptor capping and internalization," *Journal of Immunology*, vol. 166, pp. 1507–1516, 2001.

[21] A. Veillette, S. Latour, and D. Davidson, "Negative regulation of immunoreceptor signaling," *Annual Review of Immunology*, vol. 20, pp. 669–707, 2002.

[22] M. Reth and T. Brummer, "Feedback regulation of lymphocyte signalling," *Nature Reviews Immunology*, vol. 4, pp. 269–278, 2004.

[23] S. F. Andrews and P. C. Wilson, "The anergic b cell," *Blood*, vol. 115, pp. 4976–4978, 2010.

[24] R. L. McGee, M. O. Krisenko, R. L. Geahlen, A. E. Rundell, and G. T. Buzzard, "A computational study of the effects of syk activity on b cell receptor signaling dynamics," *Processes*, vol. 3, no. 1, p. 75, 2015.

[25] A. Saltelli, *Sensitivity Analysis.* J. Wiley & Sons, 2000.

[26] G. Buzzard, "Global sensitivity analysis using sparse grid interpolation and polynomial chaos," *Reliability Engineering & System Safety*, vol. 107, pp. 82–89, 2012.

[27] A. Chakrabarty, G. T. Buzzard, and A. E. Rundell, "Model-based design of experiments for cellular processes," *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, vol. 5, no. 2, pp. 181–203, 2013. [Online]. Available: http://dx.doi.org/10.1002/wsbm.1204

[28] A. M. Bouffier, J. Arnold, and H. B. Schttler, "A mine alternative to d-optimal designs for the linear model," *PLoS ONE*, vol. 9, no. 10, p. e110234, 10 2014.

[29] C. Au and J. Tam, "Transforming variables using the dirac generalized function," *The American Statistician*, vol. 53, no. 3, pp. pp. 270–272, 1999.

[30] S. Ross, *A First Course in Probability.* Pearson Prentice Hall, 2010. [Online]. Available: https://books.google.com/books?id=Bc1FAQAAIAAJ

[31] J. Mason and D. Handscomb, *Chebyshev Polynomials.* CRC Press, 2002. [Online]. Available: https://books.google.com/books?id=8FHf0P3to0UC

VITA

VITA

Reginald L. McGee II was born in Saginaw, MI and graduated from Saginaw High School. He next took his talents to Tallahassee, FL to attend Florida A&M University (FAMU). At FAMU, he majored in mathematical sciences and minored in economics, but rekindled a love for the physical sciences along the way. He reluctantly returned to the midwest for graduate school after accepting a fellowship from Purdue University in West Lafayette, IN. At Purdue, he worked in mathematical biology with an emphasis on modeling, dynamical systems, and immunology. The attainment of his doctoral degree will allow him to make an impact both in the mathematical sciences and in working to improve the number of students from underrepresented groups in science.