A SYSTEMS BIOLOGY APPROACH TO MODEL INTRACELLULAR SIGNALING PATHWAY DYNAMICS IN HETEROGENEOUS CELL POPULATION: APPLICATION TO NFκB SIGNALING PATHWAY AND CHOLERA TOXIN-LIGAND BINDING DYNAMICS

A Dissertation

by

DONGHEON LEE

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Joseph Sang-Il Kwon
Committee Members,	Arul Jayaraman
	Costas Kravaris
	Stratos Pistikopoulos
	Byung-Jun Yoon
Head of Department,	Arul Jayaraman

August 2020

Major Subject: Chemical Engineering

ABSTRACT

Systems biology research employs a quantitative model to interpret experimental measurements and make new predictions of the system to be tested in the new sets of experiments. Therefore, the construction of an accurate model is a key step in the overall systems biology study. Usually, a model developed in the past concerns mostly about the dynamics of cells at the population-level instead of those of individual cells in the population. However, with the recent advances in single-cell experimental techniques, it has been revealed that individual cells in a genetically homogeneous cell population behave heterogeneously in response to an external stimulus. This new discovery requires a different modeling approach to represent and analyze the heterogeneous cell population dynamics.

In this study, a systematic modeling framework is proposed to develop a mathematical model for an intracellular signaling pathway in a heterogeneous cell population. To this end, two sources of the cellular heterogeneity are required to be incorporated: cell-to-cell differences and reaction stochasticity. Specifically, the former and latter sources are taken into account by an individual-based population model (IBPM) and kinetic Monte Carlo (kMC) model, respectively.

First, this study proposes a systematic approach to construct an IBPM to incorporate the cell-to-cell differences. In an IPBM, the dynamics of an intracellular signaling pathway in the population are represented by a set of ordinary differential equations (ODEs), but the model parameters will follow multivariate probability density functions (PDFs) to take into account the cell-to-cell differences among the individual cells. Therefore, the construction of an IBPM requires its ODE model to be developed first. To this end, an integrative approach, which consists of first-principles modeling, identifiability analysis, parameter estimation, and model refinement, is employed to develop and calibrate the ODE model systematically. At the same time, two new methods for developing a semi-mechanistic model have been proposed in order to speed up the overall ODE model development process even when underlying mechanisms are partially known. Both methods improve the prediction accuracy by coupling the first-principle model with the data-driven model inferred

from experimental measurements by rendering model parameters time-varying in the first method and by adding additional correction terms to model states' trajectories in the second method. Once a deterministic ODE model is developed, model parameters' PDFs need to be estimated. In this regard, a numerical scheme is proposed to efficiently estimate the parameters' PDFs. Specifically, the proposed scheme consists of dimension reduction and surrogate modeling to efficiently identify the parameters' PDFs from the available single-cell measurements.

Second, an on-lattice kMC model is developed for incorporating reaction stochasticity, which is the second source of the cellular heterogeneity, as well as the temporal evolution of the membrane configuration. By modeling the multivalent binding kinetics between bacterial toxin and ganglioside expressed on cell membranes as a case system, the accuracy of the kMC model is validated, and how it is different from its corresponding deterministic model is examined.

In summary, this study has proposed a systematic modeling approach to construct a mathematical model for an intracellular signaling pathway by addressing their parameter and structural uncertainty to simulate the cell-to-cell heterogeneity. To validate the proposed methodologies, the NFκB signaling pathway and binding kinetics between ganglioside and bacterial toxins such as cholera toxin are modeled and calibrated as case studies.

DEDICATION

To my parents, and my wife

ACKNOWLEDGMENTS

First, I would like to thank my advisor Dr. Joseph S. Kwon for his support and guidance throughout my time at Texas A&M University. Particularly, I appreciate his patience and providing me ample opportunities to explore new areas of research. Second, I am deeply grateful to Professor Arul Jayaraman for his support and suggestions throughout my Ph.D. study. Also, I would like to thank my committee members, Dr. Costas Kravaris, Dr. Stratos Pistikopoulos, and Dr. Byung-Jun Yoon, for their insightful comments and suggestions for improvement.

I would like to appreciate Professor Hung-Jen Wu and his lab members including Dr. Akshi Singla who have been collaborating with us for valuable discussion and guidance in the field of glycobiology. And, I appreciate Dr. Yufang Ding in Professor Jayaraman's research group providing us experimental results and insightful thoughts on the NFκB signaling pathway.

I appreciate all of my colleagues and friends, Dr. Yuchan Ahn, Mohammed Saad Faizan Bangi, Parth Bhandakkar, Kaiyu Cao, Hyun-Kyu Choi, John Larry Dial, Aaron Green, Pallavi Kumari, Alec Mohr, Dr. Abhinav Narasingam, Silabrata Pahari, Bhavana Bhadriraju Venkata Naga Sai, Parth Shah, Dr. Prashanth Siddhamshetty, Harwinder Singh Sidhu, Niranjan Arvind Sitapure, and Dr. Seeyub Yang, for their help and support during my Ph.D. study.

Most importantly, I would like to express my gratitude to my parents, Hyeon-cheol Lee and Yoonhee Kim, my brother, Dongsub Lee, and my wife, Lan Li, for their sacrifice, support, and love during my Ph.D. study.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Dr. Joseph S. Kwon (chair), Dr. Arul Jayaraman, Dr. Costas Kravaris, and Dr. Stratos Pistikopoulos of Artie McFerrin Department of Chemical Engineering and Dr. Byung-Jun Yoon of Electrical & Computer Engineering.

All work for the dissertation was completed by the student with certain computational contributions made by various members of Dr. Kwon's lab; the students are either explicitly mentioned in the thesis and/or in papers that were published or that will be published in the future. Experiment contributions have been made by two experimental groups, involving Dr. Arul Jayaraman's lab and Dr. Hung-Jen Wu's lab at Texas A&M University. Their names are mentioned in the thesis or in papers that were published or will be published in the future.

All other work conducted for the thesis was completed by the student independently.

Funding Sources

This work was supported by Artie McFerrin Department of Chemical Engineering and Texas A&M Energy Institute at Texas A&M University. Also, this work was also partially supported by National Science Foundation (CHE-1904784).

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

1.1 Background

The central aim for biological research is to gain the quantitative understanding of cellular biochemical processes such as intracellular signaling and metabolic pathways [12]. Due to the intrinsic complexity of an intracellular process, a systems biology approach, which integrates experimental measurements and mathematical modeling, has become indispensable to gain a system-level understanding of the dynamics [13]. Consequently, one of the main tasks in a systems biology study is to develop a mechanistic model that can explain the experimental observations made in the past, provide new hypothesis about the system, and design future experiments to test the hypothesis and to improve the developed mechanistic model.

Since many commonly used experimental techniques such as Western blot measure cellular response at the population-level, most of the systems biology models developed in the past are deterministic and thus describe the population-average dynamics of cells. However, advances in the single-cell measurement techniques have provided new perspectives that are not available from the population-level measurements [14]. Most notably, individual cells even in a genetically homogeneous cell population exhibit substantial cell-to-cell variability [15]. Even though the exact causes of the cellular variability await further studies [15], its implications and importance in medicine as well as biomanufacturing processes are evident; specifically, the cell-to-cell variability has been implicated in the reduced yield and instability of bioreactors [16, 17] as well as the development of drug resistance during the cancer therapy [18, 19].

Because a deterministic model can only simulate the behavior of a "typical" cell in a cell population without considering the heterogeneity among individual cells, the deterministic modeling approach may not be appropriate for modeling and analyzing the heterogeneous cell population. It has been suggested that there are two different sources of such cellular heterogeneity: cell-to-cell difference across the cell population, and reaction stochasticity. Here, the cell-to-cell difference

refers to the fundamental differences between cells that contribute to the cell-to-cell variability. This cellular difference can come from varying levels of gene expression and different cellular micro-environment such as the gradients of morphogens in multicellular development [15, 20]. On the other hand, the reaction stochasticity refers to random fluctuations in reaction kinetics due to low copy numbers of molecules involved in the reactions [15, 21, 22]. Therefore, alternative modeling frameworks are needed to simulate cellular heterogeneity observed in experiments by incorporating both sources of the cellular heterogeneity.

1.2 Literature Review

In order for accounting for the cell-to-cell difference, an individual-based population model (IBPM) framework has been proposed to incorporate the cellular heterogeneity. In the IBPM framework, a model expressed in a system of ordinary differential equations (ODEs) is used to represent the dynamics of an intracellular process, and the cellular heterogeneity is represented by the differences in parameter values used in simulating the ODE. Here, the model parameters are assumed to follow a multivariate probability density function (PDF), which needs to be inferred from experiments. And, the heterogeneous cellular responses can be predicted by solving the ODE multiple times with different model parameters sampled from the PDFs of parameters.

In developing an IBPM, the first step in the IPBM model development process is the construction of an experimentally validated deterministic model. In developing a deterministic model, two aspects need to be considered: model parameter uncertainty and model structural uncertainty. Uncertainty in model parameters attributed to, among other factors, the limited breadth of data used for training the model (e.g., models trained using one dataset with a single stimulus concentration), which makes the models less robust under different conditions. Moreover, the identifiability issue of model parameters [23], which arises due to the model structure as well as the limited availability of experimental data [24], is not always addressed, which may lead to a suboptimal estimation of model parameters. On the other hand, the model structural uncertainty refers to the fact that underlying mechanisms of the system of interest are not fully known. Consequently, constructing an accurate model for such a system can be quite challenging. For example, out of around

100 different stimuli of the NFκB signaling pathway [25], only a handful, such as tumor necrosis factor-α (TNFα) and lipopolysaccharide (LPS), and their reaction mechanisms are well characterized. Consequently, investigating and modeling a signaling pathway induced by a stimulus, which has not been well studied, is non-trivial. Specifically, as the pathway is only partially known beforehand, a number of different model structures need to be formulated and discriminated, which can become very challenging. In the literature, different approaches have been proposed and implemented for solving this problem. First, extensive experiments are performed to characterize as many reactions as possible between intracellular molecules, which in practice is nearly impossible due to the large number of interactions to be studied. Second, the systems biology approach, which iterates between modeling and experiments, can be implemented to improve the model gradually [26, 27, 28]. However, as the initial model is likely to be erroneous, this approach may require a large number of iterations between experiments and model refinement to reach a relatively satisfying model. Third, a number of different initial models, each of which corresponds to a different hypothesis on the signaling pathway structure, are synthesized from the beginning, and the best model structure is selected by solving an optimization problem against experimental observations [29, 30, 31]. Although the optimization-based approach is promising, it has several numerical and algorithmic challenges such as efficiently finding a global optimal solution for a large system [29, 30, 31, 32, 33].

Once an accurate deterministic model is developed, the next step is to calibrate the IBPM against available experimental measurements. Over the last few years, various model calibration methods for IBPMs have been proposed [34, 35, 36, 37, 38, 39, 40]. Koeppl et al. [41] and Hasenauer et al. [10, 42] were among the first ones who proposed methods for estimating the parameter distributions; the former proposed a method combining a Bayesian inference with Markov Chain Monte Carlo method while the later proposed density-matching schemes that parameterize the parameter distributions for the computational efficiency. On the other hand, Stumpf and his coworkers [38, 43] proposed an approximate Bayesian computation scheme for inferring the parameter distributions, which was subsequently used in different biological systems [44, 45, 46]. In

many of these studies, PDFs of all the parameters were estimated simultaneously. Or, the PDFs of only a subset of model parameters from the entire set was selected based on prior knowledge, and only these were estimated. In either approach, the issue of identifiability, an important issue in the calibration of a biological model, is not properly addressed [47, 48]. Here, the identifiability of an IBPM studies the problem whether the PDFs of model parameters can be uniquely estimated from measurement distributions [48, 49, 50]. For an IBPM, the non-identifiability of PDFs of some model parameters arises due to limited quality and quantity of measurements as well as the model structure. Consequently, the PDFs of only a small subset of model parameters are identifiable, and these subsets may not necessarily overlap with the parameter subset selected based on the priori knowledge on the system itself [34, 47, 48].

On the other hand, although the IPBM approach is able to capture the cell-to-cell variability by estimating the PDFs of uncertain parameters, an IPBM has not captured the stochasticity in the reaction kinetics caused by the low copy number of molecules involved. Specifically, in an IPBM setting, once the initial conditions and parameters are sampled from their PDFs and assigned to a cell, the dynamics of a biochemical process occurring within the cell are deterministic. However, in practice, the trajectory is likely to be subject to stochastic fluctuations. Within an intracellular signaling pathway model, one important source of reaction stochasticity comes from binding dynamics between biomolecules expressed on cell membranes and molecules diffusing from extracellular matrix. One way to model such binding processes on cell membranes is the kinetic Monte Carlo (kMC) framework. The kMC is a stochastic modeling approach to simulate detailed surface reactions as well as the evolution of surface micro-configurations. Consequently, it has been widely used to model the microscopic processes of a variety of applications such as protein crystallization [51, 52, 53, 54, 55], thin film deposition [56, 57, 58, 59, 60], catalytic surface reactions [61, 62, 63, 64, 65, 66, 67] and receptor-ligand interactions on a cell membrane [68, 69, 70, 71].

1.3 Research Objectives

Motivated by the above considerations, the followings are the objectives of this dissertations:

1. Implement a systematic approach that consists of first-principle model development, iden-

tifiability analysis, parameter estimation, and model refinement to quantitatively calibrate a model for an intracellular signaling pathway model without overfitting.

2. Develop a semi-mechanistic modeling approach to construct an accurate mathematical model to describe a system that is only partially understood.

3. Develop a systematic approach to estimate the PDFs of model parameters for an IBPM from the PDFs of measurements. This method will address the identifiability of the PDFs as well as the computational efficiency.

4. Develop a kMC modeling framework to explicitly incorporate reaction stochasticity that are present in an intracellular signaling pathway.

1.4 Dissertation Outline

Chapter 2 describes how a systematic approach can be implemented to develop a deterministic model for an intracellular signaling pathway model can be developed. Under this approach, a mathematical model is developed based on the prior knowledge, and *in vitro* measurements are performed to quantitatively calibrate the developed model. Then, in order to avoid potential overfitting issues of the model calibration, a parameter selection and estimation scheme selects key model parameters and estimates their values based on the experimental measurements. Unsatisfactory results from the parameter estimation guides subsequent experiments and appropriate model improvements, and the refined model is calibrated again through parameter estimation. The efficacy of such an approach is illustrated by the NFκB signaling pathway induced by LPS in the presence of brefeldin A (BFA) as a case study.

In Chapter 3, we use the binding kinetics between cholera toxin subunit B (CTB) and GD1b ganglioside on cell membranes as another case study to highlight how the systematic approach described in Chapter 2 can be used to develop and quantitatively calibrate the biological model. Specifically, a first-principle model is developed based on previous experimental studies, and the binding kinetics are experimentally measured to estimate values of the model parameters for the first time.

In Chapter 4, we use the LPS-induced NFκB signaling pathway model in the presence of BFA as a case study to propose a new numerical scheme to develop a time-varying model to efficiently construct a semi-mechanistic model when the system of interest is partially known. This approach consists of global sensitivity analysis, *in vitro* experimental measurements, temporal clustering of experimental measurements, and estimation of values of piecewise constant parameter values.

In Chapter 5, we propose a hybrid modeling approach that improves the prediction accuracy of a first-principle model through estimating correction terms from *in vitro* experimental measurements when the first-principle model is only partially accurate. Specifically, the proposed numerical scheme first identifies a subset of model states whose dynamics impact the measurements most significantly, estimates the values of additional terms necessary for correcting dynamics of the selected states, and develops an artificial neural network to compute correction terms for a given condition so that the resulted hybrid model has an enhanced prediction capability.

In Chapter 6, we propose a new method to construct an IBPM when an accurate deterministic ODE model is developed by any of the modeling methods present in Chapters 2-5. First, a subset of parameters whose PDFs is identifiable are determined through sensitivity analysis, and only these PDFs are estimated. Second, an artificial neural network model is developed to find an empirical relation between these parameter and output PDFs to reduce computational costs of the parameter identification. The proposed approach is validated by estimating the PDFs of parameters of a TNFα signaling model.

In Chapter 7, we propose a kMC modeling framework that can take into account reaction stochasticity, which is another source of the cellular heterogeneity. As a case study, the binding kinetics between CTB and ganglioside on cell membranes are modeled as a case study and compared with their deterministic model to highlight merits of the kMC modeling approach.

2. MATHEMATICAL MODELING AND PARAMETER ESTIMATION OF INTRACELLULAR SIGNALING PATHWAY: APPLICATION TO LPS-INDUCED NFκB ACTIVATION AND TNFα PRODUCTION IN MACROPHAGES *

2.1 Introduction

To integrate multiple signaling pathways, their canonical transcription factors and downstream effector genes are required for cells to respond to various signals they encounter in their micro-environment. Therefore, understanding how information is sensed and processed by cells and the signaling pathways that are engaged by different stimuli can help elucidate cellular behaviors and responses. Typically, cellular signal dynamics and the response to stimuli have been studied using a combination of mathematical modeling and experimental analysis [72, 73]. A majority of these studies has modeled cell signaling at the population level and used population-averaged measurements such as Western blots to infer the dynamics of different proteins in the signaling pathway, as well as the possible network structure of signaling pathways [72]. However, with recent advances in the ability to measure gene and protein expression at the single-cell level (reviewed in [73, 74]), it has become possible to analyze signaling dynamics at the single-cell level. In contrast to the observations from population-average studies, the single-cell studies have demonstrated that individual cells in a clonal population may respond differently to the same stimulus, and the population level measurements could mask the temporal dynamics of individual cells [73]. This variability in the responses of individual cells poses a challenge to their implementation in biology and medicine [19]. Therefore, it is important to understand the stochasticity and heterogeneity in the single-cell responses that might be missed in population-averaged measurements.

Advances in experimental tools for single-cell analysis have led to a significant increase in single-cell studies [73, 74]. Despite these advancements, it is still difficult to study the single-

cell signaling dynamics due to complex interactions at multiple levels between different proteins that are involved in signal transduction [72]. Computational modeling has been proposed as a complementary approach to overcome some of these limitations and gain insights that cannot be obtained solely through experiments [72, 73]. A viable and computationally efficient approach to study the cell-to-cell variability is to use a deterministic model with parameters that have distributions [10, 37, 42, 73]. In this approach, the computational cost is generally reduced by simulating the signaling dynamics through a deterministic modeling approach while the stochasticity is preserved by assigning a set of different parameter values for each simulation based on predetermined parameter distributions.

In order to construct such models, an experimentally validated deterministic model, which can capture average signaling dynamics at the single-cell level, is required. Although various deterministic models have been proposed for several well-studied signal transduction pathways [72, 75], many demonstrate good qualitative, but not quantitative, agreement with the experimental data. This has been attributed to, among other factors, the limited breadth of data used for training the model (e.g, models trained using one dataset with a single stimulus concentration), which makes the models unable to make robust predictions under different conditions. Moreover, the identifiability issue of model parameters [30], which arises due to the model structure as well as the limited availability of experimental data of intracellular proteins [23, 47], is not always addressed, which may lead to a suboptimal estimation of model parameters [23, 76]. Additionally, many models have been constructed and validated based on experimental data obtained from the population-averaged measurements, which mask the signaling dynamics at the single-cell level [73, 77, 78]. Consequently, these models are inadequate to predict the average signaling dynamics of single cells.

Motivated by the above considerations, we developed a deterministic model that can accurately predict the average signaling dynamics of single cells. We chose LPS-induced nuclear factor κB (NFκB) signaling in mouse macrophages for our model system as it is an extensively studied and characterized signaling pathway [3, 75, 79]. In order to address the issues discussed above, both

computational and experimental approaches have been implemented. First, a rigorous numerical scheme is used to identify the most important parameters that are to be estimated in the parameter estimation [80]. Specifically, the sensitivity analysis and the parameter selection method quantitatively assess the significance of each model parameter with respect to experimental measurements under different LPS concentrations and select parameters whose values could be uniquely estimated [23, 81]. Second, flow cytometry with intracellular staining is used to measure the average single-cell dynamics of key molecules involved in the NFκB signaling pathway in response to a broad range of LPS concentrations [82, 83]. In this study, the intracellular concentrations of the inhibitor of κB-α (IκBα) and TNFα were measured. IκBα is an inhibitor of NFκB activity, and therefore, the IκBα dynamics are inversely correlated with the NFκB dynamics. At the same time, the activated NFκB induces the transcription and translation of TNFα upon the stimulation of LPS; hence, the TNFα can also be used to infer the dynamics of the NFκB signaling pathway [79]. The obtained average single-cell kinetics is used to quantitatively calibrate and validate the model. Third, the discrepancy between the experimental measurements and the model predictions reveals important, yet unconsidered mechanisms, which is validated experimentally afterwards and leads to the model refinement. Through this integrated model development methodology, predictions from the resultant model quantitatively agree with the experimental measurements. Therefore, the proposed model represents a first step towards the construction of single-cell semi-stochastic models to investigate the stochasticity of intracellular NFκB signaling in macrophages.

2.2 Material and Methods

2.2.1 Materials and Cell Culture

RAW264.7 cells were obtained from ATCC (Manassas, VA, USA). Dulbecco's Modified Eagle Medium (DMEM) and penicillin/streptomycin were obtained from Invitrogen (Carlsbad, CA, USA). Bovine serum and fetal bovine serum (FBS) were obtained from Atlanta Biologicals (Flowery Branch, GA, USA). Ultrapure LPS derived from *S. minnesota* was obtained from Invivogen

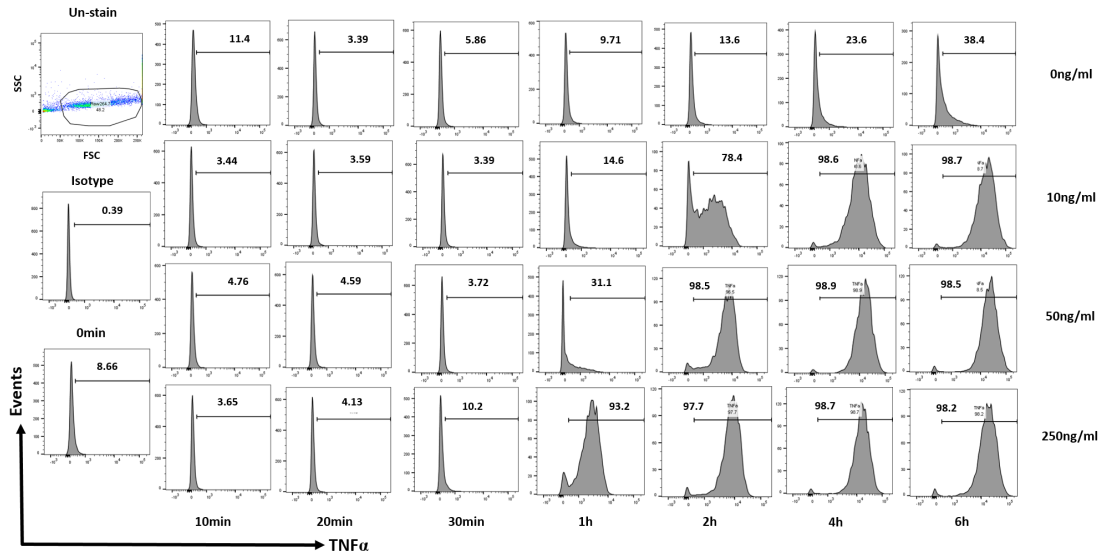


Figure 2.1: Representative histograms of TNFα production. RAW 264.7 cells were stimulated with different concentrations of LPS along with Golgiplug^{TM}, and the production of TNFα were analyzed by flow cytometry.

(San Diego, CA, USA). RAW264.7 macrophages were cultured in DMEM supplemented with 10% FBS, penicillin (200 U/mL) and streptomycin (200 μg/mL) at 37 °C in a 5% CO_2 environment.

2.2.2 Flow Cytometry Analysis

The expression of TNFα and IκBα under different experimental conditions was determined using flow cytometry. RAW264.7 cells were seeded into round-bottomed 96-well plate and stimulated with different concentrations of LPS for the indicated time. Golgiplug^{TM} (BD Biosciences, San Jose, CA, USA) was added along with LPS for TNFα detection experiments to block secretion of TNFα. Cells were then stained with Alexa Flour 700 fluorescence-tagged TNFα antibody (BD Biosciences) and PE-conjugated IκBα antibody (Cell Signaling Technology, Danvers, MA, USA) using the manufacturer's suggested protocol. Stained cells were analyzed using a BD Fortessa flow cytometer (BD Biosciences) The College of Medicine Cell Analysis Facility at the Texas A&M Health Science Center. Ten thousands events per sample were acquired, and the data were analyzed

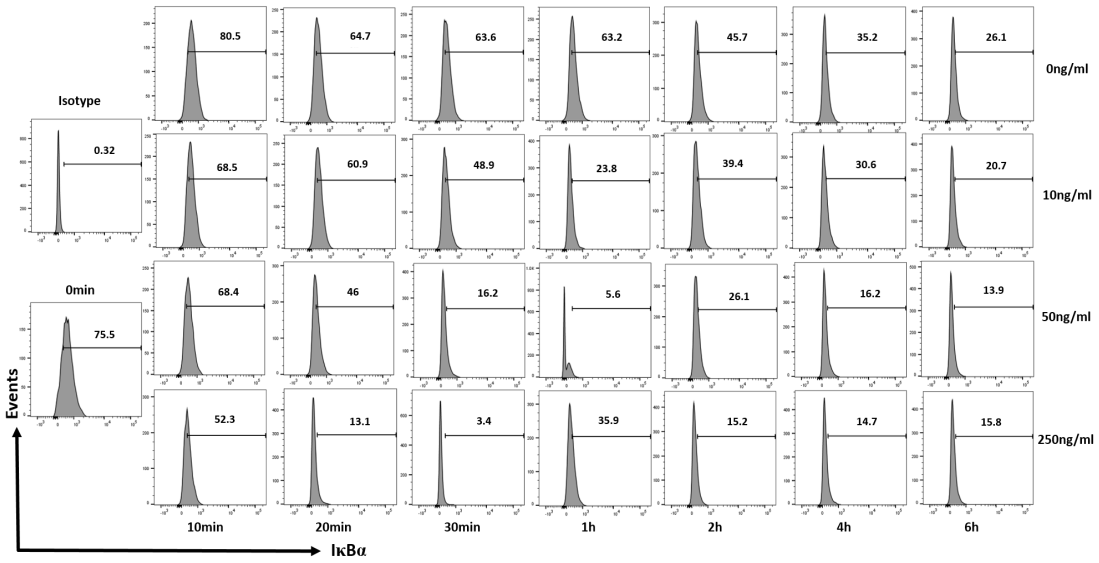


Figure 2.2: Representative histograms of IκBα. RAW 264.7 cells were stimulated with different concentrations of LPS along with Golgiplug$^{\text{TM}}$, and the intracellular IκBα levels were analyzed by flow cytometry.

using FlowJo software (Tree Star, OR, USA). Cells were gated based on side scattered light (SSC) and forward scattered light (FSC) values to eliminate cell debris, and TNFα- and IκBα-positive cells were gated based on the antibody isotype (Figures 2.1-2.3). All experiments were repeated using at least three different cultures.

2.2.3 Model Development

The schematic diagram of the NFκB signaling pathway is illustrated in Figure 2.4. The model used in this study was adopted from Caldwell et al. [1], which takes the extracellular LPS concentration as an input to predict the kinetics of key biomolecules in the NFκB signaling pathway. In this model, by forming a complex with Toll-like receptor 4 (TLR4), LPS activates IκB kinase (IKK) through myeloid differentiation primary response 88 (MyD88)- or TIR (Toll/Interleukin-1 receptor)-domain-containing adaptor-inducing interferon-β (TRIF)-dependent activation of TNF receptor-associated factor 6 (TRAF6). The activated IKK in turn promotes the translocation of NFκB to the nucleus, where the nuclear NFκB induces the transcription of NFκB inhibitors (IκB-

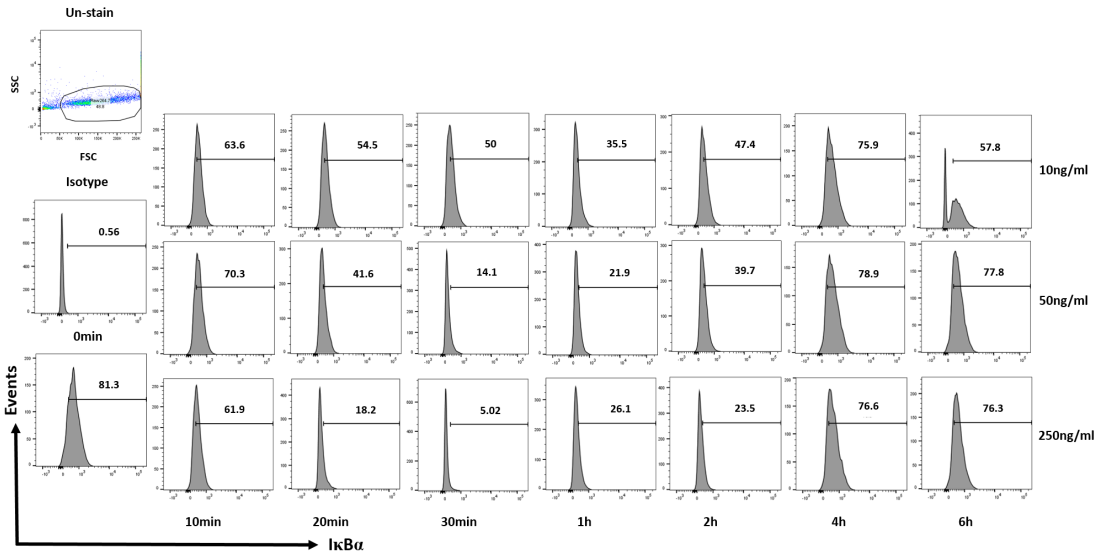


Figure 2.3: Representative histograms of IκBα. RAW 264.7 cells were stimulated with different concentrations of LPS without the addition of Golgiplug$^{\text{TM}}$, and the intracellular IκBα levels were analyzed by flow cytometry.

α, -β, -ϵ and A20), as well as TNFα. Once translated, these inhibitors inhibit the NFκB signaling pathway. In contrast, the translated TNFα is secreted to the extracellular medium, and some of the secreted TNFα proteins will bind with TNFα receptor (TNFR) on the cellular membrane to initiate the TNFα-induced NFκB signaling pathway (see [1, 9, 84] for details of the model).

Additionally, nonlinear functions proposed by Junkin et al. [2] were added to describe how the rates of TNFα production and secretion increase as the amount of activated TRIF complex increases. This model incorporates the TLR4-mediated NFκB dynamics induced by LPS, as well as the production of TNFα in macrophages (see [1, 9] for details). For the purpose of this study, two modifications were made to the model presented by Caldwell et al. [1]. First, transcription delays were ignored to facilitate the simplicity of subsequent calculations for sensitivity analysis and parameter estimation. Second, a new role of A20 protein, which was introduced in the previous model [1, 9] as an inhibitor of the TNFα-induced NFκB signaling [9, 85, 86], was included in the modified model to downregulate the LPS-induced signaling through deubiquitinating

of TRAF6 [87].

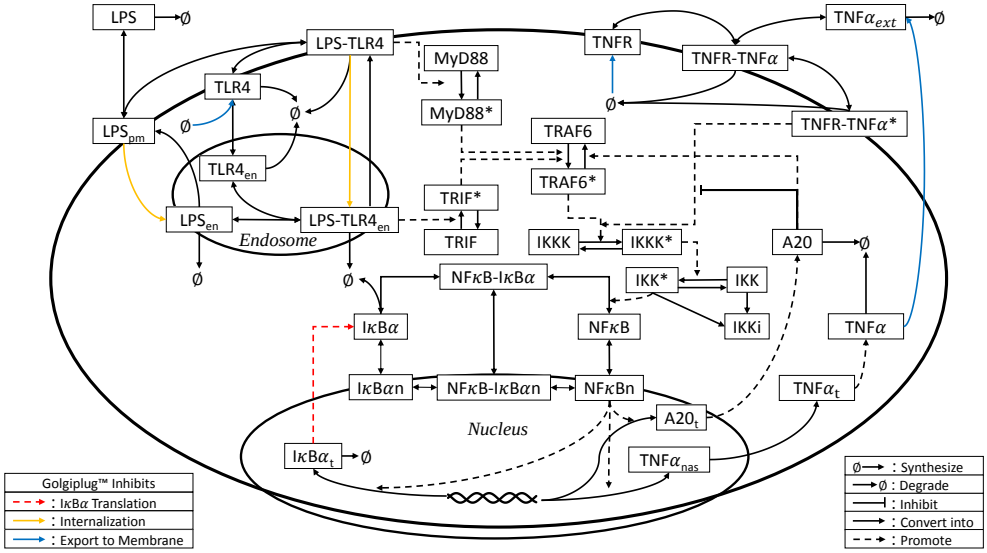


Figure 2.4: Schematic diagram for the LPS-NFκB-TNFα signaling pathway. Due to space limitation, TRIF-dependent regulation of TNFα production, IκBβ and IκBϵ-dependent NFκB deactivation and eIF2α-induced translation inhibition are not illustrated. Furthermore, some states related to TNFα-induced activation of IKK kinase (IKKK) are not shown. Colored arrows indicate the processes affected by the addition of Golgiplug™ (see the text for details).

For this study, the TNFα production at the single-cell level was measured using flow cytometry by adding Golgiplug™ since brefeldin A, the active agent of Golgiplug™, causes the Golgi apparatus to merge with endoplasmic reticulum (ER) and inhibits protein export from the Golgi complex [88, 89]. Hence, the addition of Golgiplug™ enabled us to measure average single-cell production of TNFα. On the other hand, because Golgiplug™ interferes with the normal cellular processes, it inevitably affects the NFκB signaling dynamics. Specifically, Golgiplug™ suppresses the expression of receptors on the cellular membrane, which negatively regulates the LPS-mediated NFκB signaling pathway in different ways. First, the addition of Golgiplug™ can block the translocation of TLR4 and its accessory molecules from the Golgi complex, which leads to the termination of signaling as these receptors are not replenished after turnover [88, 90, 91, 92]. Similarly, TNFR is also depleted from the cellular membrane due to Golgiplug™ [93, 94], which may inhibit

subsequent TNFα autocrine and paracrine signaling [95, 96, 97]. Second, Golgiplug™ can hinder the membrane expression of the cluster of differentiation 14 (CD14), which regulates the endocytosis of LPS or the TLR4-LPS complex [98, 99, 100]. Therefore, the TRIF-dependent pathway, which is initiated only after LPS or LPS-TLR4 is endocytosed into cytoplasm [37, 101], can also be partially impaired. Lastly, the secretion of TNFα proteins translated in response to the NFκB activation will also be inhibited, which helps measure the TNFα production at the single-cell level.

Consequently, the dynamic effects of Golgiplug™ were parameterized and included in the model by the following equations:

$$\begin{aligned} G &= \frac{t}{t+\tau} \\ k_{s_{TNFR},m} &= k_{s_{TNFR}}(1-G) \\ k_{s_{TLR4},m} &= k_{s_{TLR4}}(1-G) \\ k_{en_{LPS},m} &= k_{en_{LPS}}(1-G) \\ k_{en_{cp},m} &= k_{en_{cp}}(1-G) \\ k_{sec,m} &= k_{sec}(1-G) \end{aligned} \tag{2.1}$$

where G is the normalized activity of Golgiplug™, t is the elapsed time from the addition of Golgiplug™, τ is the characteristic time associated with Golgiplug™ activity, $k_{s_{TNFR}}$ and $k_{s_{TLR4}}$ are the constitutive synthesis rates of TNFR and TLR4, respectively, in the absence of Golgiplug™, $k_{en_{LPS}}$ and $k_{en_{cp}}$ are the endocytosis rates of LPS and the LPS-TLR4 complex, respectively, in the absence of Golgiplug™, k_{sec} is the TNFα secretion rate in the absence of Golgiplug™ and $k_{s_{TNFR},m}$, $k_{s_{TLR4},m}$, $k_{en_{LPS},m}$, $k_{en_{cp},m}$ and $k_{sec,m}$ are the corresponding rates in the presence of Golgiplug™. After Golgiplug™ is added to the cells at $t = 0$, G slowly increases from zero to one, where $G = 0$ and $G = 1$ represent no inhibition and complete inhibition, respectively, of protein export from the Golgi apparatus in the presence of Golgiplug™.

Since the signaling kinetics under the stimulation of LPS in the presence of Golgiplug™ were measured experimentally, the dynamic model that consists of the model presented in [1] and Equa-

tion 2.1 was used to simulate the dynamics of LPS-induced NFκB signaling in the presence of Golgiplug™. In general, the dynamic model that simulates the signaling pathway can be represented by a set of nonlinear ODEs as follows:

$$\begin{aligned} \frac{d\boldsymbol{x}}{dt} &= \boldsymbol{f}\left(\boldsymbol{x}, \boldsymbol{\theta}; u\right) \\ y &= g\left(\boldsymbol{x}, \boldsymbol{\theta}; u\right) \end{aligned} \tag{2.2}$$

where $\boldsymbol{x}$ represents the concentration of the biomolecules involved in the signaling pathway (i.e., a vector of states), $\boldsymbol{\theta}$ is a vector of model parameters that describe the biochemical reaction rates in the pro-cess, u is the concentration of LPS added to the cells (i.e., the process input), and y is the model output (i.e., the experimental measurements predicted by the model). When Golgiplug™ is added, Equation 2.1 is included in Equation 2.2, and the overall model consists of 49 states and 146 parameters.

2.2.4 Parameter Estimation

Since we added the Golgiplug™ module to the model developed by Caldwell et al. [1], the integrated dynamic model (the model presented in [1] and Equation 2.1) was quantitatively calibrated by estimating its parameters using experimental measurements in response to different LPS concentrations in the presence of Golgiplug™.

The model parameter values were estimated by minimizing the difference between the experimental measurements and the model predictions of the protein concentration. In this work, we used flow cytometry to measure two key molecules in the LPS-induced NFκB signaling path-way: TNFα and IκBα. Since flow cytometry does not provide direct measurements of protein concentration, the mean fluorescence intensity (MFI), which is a measure of the number of copies of the target molecule per cell, was used to infer the protein concentration by assuming a linear relationship between MFI and protein concentration. The experimental data and model prediction were compared based on fold changes of MFI, which are defined as follows:

$$
\begin{aligned}
y_{I\kappa B\alpha}(t) &= \frac{(x_{I\kappa B\alpha}(t) + x_{I\kappa B\alpha_n}(t) + x_{NF\kappa B\text{-}I\kappa B\alpha}(t) + x_{NF\kappa B\text{-}I\kappa B\alpha_n}(t))}{(x_{I\kappa B\alpha,0} + x_{I\kappa B\alpha_n,0} + x_{NF\kappa B\text{-}I\kappa B\alpha,0} + x_{NF\kappa B\text{-}I\kappa B\alpha_n,0})} \approx \frac{I_{I\kappa B\alpha}(t) - I_{I\kappa B\alpha,c}}{I_{I\kappa B\alpha,0} - I_{I\kappa B\alpha,c}} \\
y_{TNF\alpha}(t) &= \frac{x_{TNF\alpha}(t)}{x_{TNF\alpha,0}} \approx \frac{I_{TNF\alpha}(t) - I_{TNF\alpha,c}}{I_{TNF\alpha,0} - I_{TNF\alpha,c}}
\end{aligned}
\tag{2.3}
$$

where $y_{I\kappa B\alpha}(t)$ and $y_{TNF\alpha}(t)$ are the fold changes of the IκBα and TNFα concentration at time t, $x_{I\kappa B\alpha}$, $x_{I\kappa B\alpha_n}$, $x_{NF\kappa B\text{-}I\kappa B\alpha}$, $x_{NF\kappa B\text{-}I\kappa B\alpha_n}$ and $x_{TNF\alpha}$ are the cytoplasmic IκBα, nuclear IκBα, cytoplasmic IκBα-NFκB complex, nuclear IκBα-NFκB complex and intracellular TNFα concentration, respectively, $x_{i,0}$ is the initial concentration of the corresponding biomolecules, $I_{I\kappa B\alpha}$ and $I_{TNF\alpha}$ are the MFI of IκBα and intracellular TNFα, respectively, and $I_{j,0}$ and $I_{j,c}$, $\forall\ j = \{$IκBα, TNF$\alpha\}$, are the corresponding MFI at $t = 0$ and MFI of negative control, respectively. In each cell, IκBα can be part of four biomolecules ($x_{I\kappa B\alpha}$, $x_{I\kappa B\alpha_n}$, $x_{NF\kappa B\text{-}I\kappa B\alpha}$, $x_{NF\kappa B\text{-}I\kappa B\alpha_n}$); however, flow cytometry measurements can only provide the total IκBα concentration in each cell. Therefore, the simulated concentrations of four IκBα-containing biomolecules were initially summed, and the fold change of the sum (i.e., $y_{I\kappa B\alpha}$) was computed to compare with the measurements in the subsequent parameter estimation procedure.

One of the biggest challenges in estimating parameters of signaling pathways with a large number of parameters is the parameter identifiability issue [23]. That is, the exact values of some model parameters cannot be uniquely determined from experimental measurements even if a large amount of experimental measurements are available [23, 47]. As the proposed model has a large number of parameters, not all the model parameters can be estimated. To this end, a subset of the model parameters, which can be uniquely estimated from the available experimental measurements, was identified through a parameter selection method [23, 81]. Only these parameters were estimated against the experimental data.

First, local sensitivity analysis [23, 102] was performed to compute two different sensitivity matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ to quantify the effect of each model parameter on $y_{I\kappa B\alpha}$ and $y_{TNF\alpha}$ (i.e., the process outputs). $\mathbf{S}_1$ and $\mathbf{S}_2$ represent the sensitivity matrices of the model parameters with respect to $y_{I\kappa B\alpha}$ and $y_{TNF\alpha}$, respectively, when the cells were stimulated with LPS in the presence of

Golgiplug™. Specifically, a sensitivity matrix is defined as:

$$\mathbf{S}_i = \begin{bmatrix} \frac{\partial y_i(t_1)}{\partial \theta_1} & \cdots & \frac{\partial y_i(t_1)}{\partial \theta_{n_p}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_i(t_{N_t})}{\partial \theta_1} & \cdots & \frac{\partial y_i(t_{N_t})}{\partial \theta_{n_p}} \end{bmatrix}, \qquad \forall i = \{\text{I}\kappa\text{B}\alpha, \text{TNF}\alpha\} \tag{2.4}$$

where n_p is the number of parameters in $\boldsymbol{\theta}$ in Equation 2.2, and $\partial y_i(t_l)/\partial \theta_j$ quantifies the effect of a parameter θ_j on an output y_i at $t = t_l, \forall l = 1, \cdots, N_t$, where N_t is the number of measurement instants. $\partial y_i(t_l)/\partial \theta_j$ can be computed by the following equation:

$$\frac{\partial y_i(t_l)}{\partial \theta_j} = \frac{\partial g_i(t_l)}{\partial \boldsymbol{x}^T} \frac{\partial \boldsymbol{x}}{\partial \theta_j} + \frac{\partial g_i(t_l)}{\partial \theta_j} \tag{2.5}$$

Additionally, the term $\partial \boldsymbol{x}/\partial \theta_j$ in Equation 2.5 can be computed by integrating the following equation along with Equation 2.2:

$$\frac{d}{dt} \frac{\partial \boldsymbol{x}(t_l)}{\partial \theta_j} = \frac{\partial \boldsymbol{f}(t_l)}{\partial \boldsymbol{x}^T} \frac{\partial \boldsymbol{x}}{\partial \theta_j} + \frac{\partial \boldsymbol{f}(t_l)}{\partial \theta_j} \tag{2.6}$$

Second, the Gram–Schmidt orthogonalization method [23, 81] was used to identify the p_i most important model parameters to be estimated for each $\mathbf{S}_i, \forall\ i = 1, 2$. Here, p_i is the number of singular values of $\mathbf{S}_i$ whose magnitudes are at least 5% of the largest singular value [80, 81]. As a result, the parameter subset to be estimated, $\boldsymbol{\theta}_s \in \mathbb{R}^{p \times 1}$ where $p \leq p_1 + p_2$, is chosen as the union of the selected parameters from $\mathbf{S}_1$ and $\mathbf{S}_2$. Third, the least-squares problem was solved to estimate the values of $\boldsymbol{\theta}_s$ by minimizing the difference between the model predictions and the experimental data of $y_{TNF\alpha}$ and $y_{I\kappa B\alpha}$ while the values of the unselected parameters were fixed at their nominal values selected from the literature [1, 2, 103, 104] with some modifications.

In this study, three LPS concentrations (10, 50 and 250 ng/mL) were used to stimulate cells, and the MFI of IκBα and TNFα were measured at $t = 0$, 10, 20, 30, 60, 120, 240 and 360 min after the addition of LPS with Golgiplug™ (i.e., $t_l,\ \forall\ l = 1, \cdots, 7$). Specifically, the MFI data from 10 and 250 ng/mL of LPS (i.e., $u_k,\ \forall\ k = 1, 2$) were used to estimate the parameter values,

while the dataset from 50 ng/mL LPS was used to validate the model with the updated parameters. Then, the least-squares problem is formulated as follows:

$$\min_{\boldsymbol{\theta}_s} \sum_{k=1}^{2} \sum_{l=1}^{7} \left[\left(\frac{y_{I\kappa B\alpha,k,1}(t_l) - \hat{y}_{I\kappa B\alpha,k,1}(t_l)}{\hat{y}_{I\kappa B\alpha,k,1}(t_l)} \right)^2 + \left(\frac{y_{TNF\alpha,k,1}(t_l) - \hat{y}_{TNF\alpha,k,1}(t_l)}{\hat{y}_{TNF\alpha,k,1}(t_l)} \right)^2 \right] \tag{2.7}$$

$$\text{s.t. } \frac{d\boldsymbol{x}_{k,i}}{dt} = \boldsymbol{f}_i \left(\boldsymbol{x}_{k,i}, \boldsymbol{\theta}_s; u_k \right), \ \boldsymbol{x}_{k,i}(t=0) = \boldsymbol{x}_0, \ \forall i = 1, 2 \tag{2.8}$$

$$y_{j,k,1} = g_j \left(\boldsymbol{x}_{k,i}, \boldsymbol{\theta}_s; u_k \right), \ j = \{\text{I}\kappa\text{B}\alpha, \text{TNF}\alpha\} \tag{2.9}$$

$$\boldsymbol{x}^{lb} \leq \boldsymbol{x}_{k,i} \leq \boldsymbol{x}^{ub} \tag{2.10}$$

$$\boldsymbol{\theta}_s^{lb} \leq \boldsymbol{\theta}_s \leq \boldsymbol{\theta}_s^{ub} \tag{2.11}$$

where $y_{I\kappa B\alpha,k,1}(t_l)$ and $y_{TNF\alpha,k,1}(t_l)$ are the simulated fold changes of IκBα and TNFα, respectively, through Equation 2.9 at $t = t_l$ under the initial LPS concentration of u_k in the presence of Golgiplug™, $\hat{y}_{I\kappa B\alpha,k,1}$ and $\hat{y}_{TNF\alpha,k,1}$ are the corresponding experimentally measured fold changes and $\boldsymbol{x}_0$ is the vector of the initial conditions of $\boldsymbol{x}$ (see Supplementary Materials Table S1).

In the least-squares problem of Equations 2.7-2.11, the objective function of Equation 2.7 computes the difference between model predictions and the experimental measurements of the proteins in the presence of Golgiplug™. As a whole, the objective function minimizes the difference by varying the values of $\boldsymbol{\theta}_s$. While Equation 2.8 is integrated to compute the predicted protein concentration $\boldsymbol{x}_{k,i}$, $\boldsymbol{f}_1$, which includes Equation 2.1 is integrated if Golgiplug™ is present; otherwise, $\boldsymbol{f}_2$, which does not involve Equation 2.1, is integrated. The initial condition of the model, $\hat{\boldsymbol{x}}_0$, is assumed based on a previous study [1, 9]. Equations 2.10-2.11 impose lower and upper bounds on the states and parameters, respectively, based on previous studies and underlying biological knowledge [1, 9, 37].

It should be noted that we preserved one set of the experimental measurements (one obtained under 50 ng/mL of LPS) to validate the parameter estimation results [105]. As Equations 2.7-2.11 are likely to be non-convex, the choice of the initial guesses is important. In this study, the initial guesses for the above least-squares problem were obtained from Caldwell et al. [1], which were

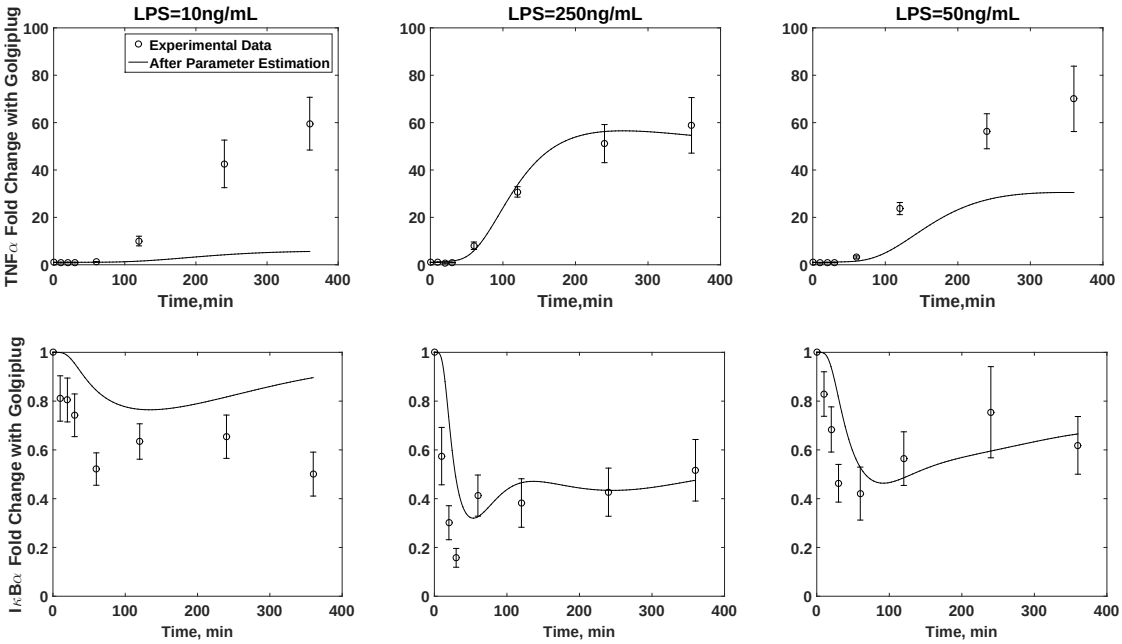
validated experimentally by comparing with the population-level measurements. Therefore, the parameter values estimated by Caldwell et al. [1] were suitable initial guesses. At the same time, Equations 2.7-2.11 were solved multiple times with different initial values to avoid any suboptimal optima. Model simulations and the parameter estimation were performed in MATLAB via its functions *ode15s* and *fmincon*. The absolute and relative tolerance criteria for ode15s were set as 10^{-9}, and *fmincon* was implemented with *multistart* to obtain a better result by solving Equations 2.7-2.11 multiple times with different initial conditions.

2.3 Results

Profiles of *de novo* synthesized intracellular TNFα under the stimulation of LPS in the presence of Golgiplug™ demonstrated that the TNFα production increased around one hour after the stimulation (Figure 2.5). At around the same time, the IκBα concentration reached its minimum, which is consistent with experimental observations in the literature [106, 107, 108]. Subsequently, the IκBα concentration increased due to the induction of IκB transcript (IκB$_t$) by nuclear translation of NFκB, while the TNFα production rate slowed down beyond 4 hours of LPS stimulation (Figure 2.5). It should be noted that no experiments were conducted beyond 6 hours after LPS was added to the cell culture based on the manufacturer's guideline on Golgiplug™ use. This is most likely based on the fact that Golgiplug™ might induce the apoptosis of cells exposed to it for a long time [109, 110]. As a result, the calibrated model is more suitable to describe the early NFκB signaling pathway ($\leq$ 6 hours) upon the LPS stimulation.

2.3.1 Model Validation

Based on the criteria outlined above in the previous section, six parameters (Table 2.1) were selected for parameter estimation. Figure 2.5 shows the simulated profiles of intracellular TNFα and IκBα in macrophages under the stimulation of LPS in the presence of Golgiplug™ after the parameter estimation. While the model predictions agreed well with the experimental data obtained for 250 ng/mL of LPS, less concordance was observed between simulations and experimental data for 10 ng/mL of LPS. Specifically, the simulated concentration of intracellular TNFα was one



(d) (e) (f)

Figure 2.5: Parameter estimation before considering the Golgiplug™-induced ER stress. (**a**–**c**) Measured (empty circle) and simulated (solid line) fold changes of IκBα concentrations over time were plotted under different LPS concentrations in the presence of Golgiplug™. (**d**–**f**) Measured (empty circle) and simulated (solid line) fold changes of intracellular TNFBα concentrations over time were plotted under different LPS concentrations in the presence of Golgiplug™. Indicated amounts of LPS were used for experiments and simulations. Experimental data are given as means $\pm$ SEM (standard error of means) with at least n = 6.

order of magnitude lower than the MFI data, while the simulated IκBα dynamics were qualitatively similar to the measured MFI values. Since the discrepancy between the model prediction and the experimental measurements was pronounced with 10 ng/mL of LPS, we hypothesized that the lack of agreement between the simulations and experimental data was because the effects of Golgiplug™ addition were not adequately represented in the model structure and were more pronounced at the lower LPS concentration.

2.3.2 Golgiplug™-Induced ER Stress

One possible explanation for this discrepancy could be that the addition of Golgiplug™ induced other signaling pathways, which altered NFκB signaling dynamics [111]. As Golgiplug™ prevents protein secretion by causing collapse of the Golgi apparatus into the ER, synthesized

Parameter
TLR4 constitutive generation rate
IKKK-mediated IKK activation (IKK $\rightarrow$ IKKa)
IκBα transcript degradation rate
Hill coefficient of IκBα transcription
Hill coefficient of IκBϵ transcription
Hill coefficient of TNFα transcription

Table 2.1: The selected parameters when Golgiplug™-induced IκB translation inhibition was not considered.

proteins will be redistributed from the Golgi complex into the ER [89]. A direct consequence of Golgiplug™ addition could be accumulation of newly synthesized proteins in ER, which may induce ER stress [111]. It is well established that the ER stress leads to phosphorylation of eukaryotic initiation factor 2 α-subunit (eIF2α), which partially inhibits the translation of IκB in the NFκB signaling pathway [111, 112, 113, 114]. This could lead to a decrease in the overall kinetics of the LPS-induced NFκB signaling as the concentration of IκB proteins would be kept lower, leading to the aforementioned mismatch between the model predictions and experimental data. Since the low LPS concentration induces less IκBα and its isomers (IκBβ and IκBϵ) (Figure 2.5), the entire LPS-induced NFκB pathway dynamics would be affected more significantly by Golgiplug™ at a lower LPS concentration than at a high LPS concentration if the translation of IκBα and its isomers is partially inhibited. If this is true, it can lead to the pronounced disagreement between the model prediction and the experimental measurement under the stimulation of 10 ng/mL LPS as shown in Figure 2.5.

Therefore, we examined whether the Golgiplug™ addition could modulate IκB levels. First, the fold changes in IκBα MFI with the stimulation of LPS alone, Golgiplug ™ alone, and LPS and Golgiplug ™ in macrophages were compared. Figure 2.6a shows that Golgiplug™ alone lowers the concentration of IκBα, and Figures 2.6b–d show that the IκBα kinetics were altered when the cells were stimulated with LPS and Golgiplug™. While IκBα levels initially decreased when cells were exposed to LPS alone, they recovered to pre-stimulation levels after 3 hours of exposure. However, when LPS was added along with Golgiplug™, IκBα levels continued to be lower than

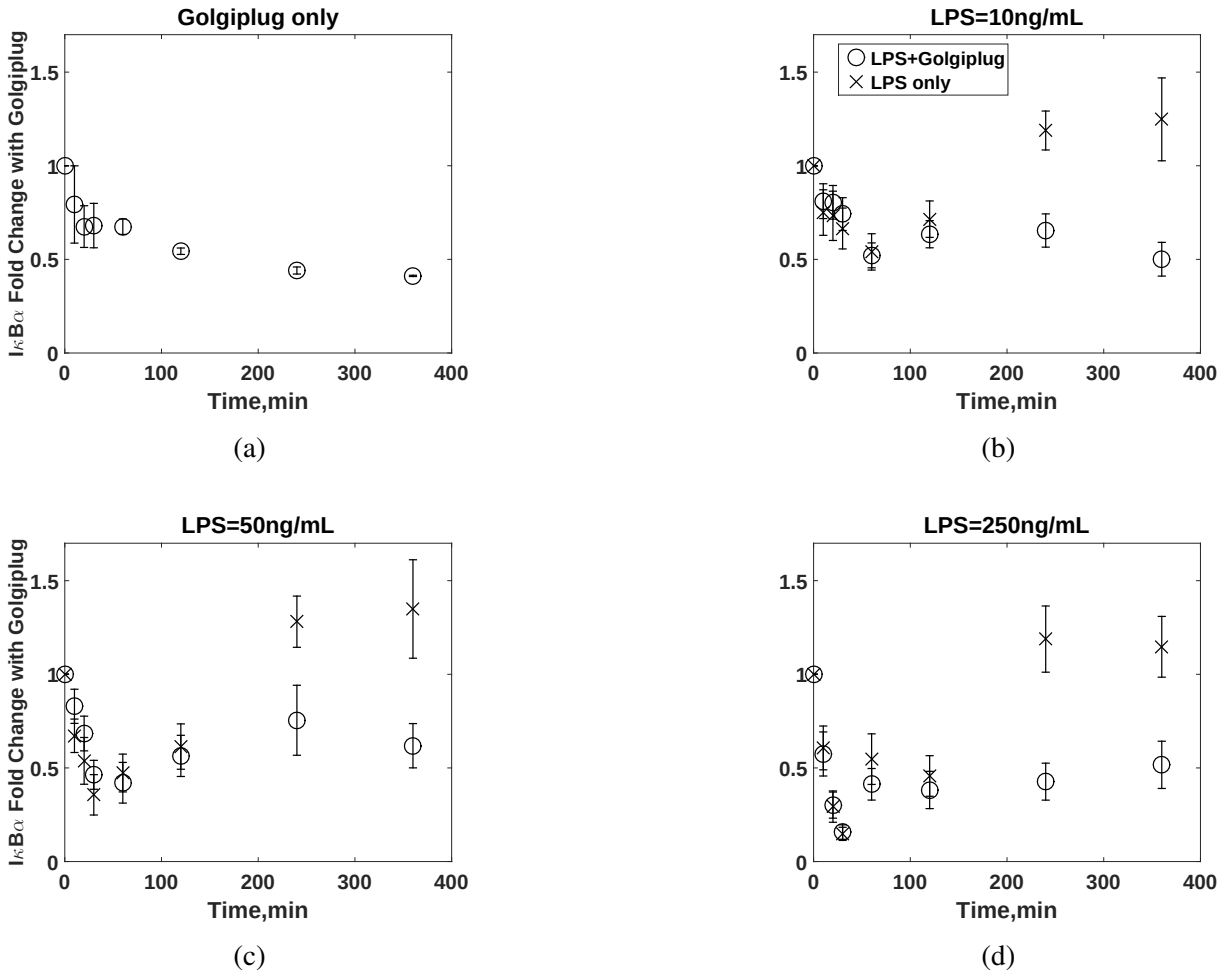


Figure 2.6: Kinetics of IκBα fold changes when the cells were stimulated by (**a**) 0, (**b**) 10, (**c**) 50, and (**d**) 250 ng/mL of LPS in the presence (empty circles) or absence (x marks) of Golgiplug™. Data are given as means $\pm$ SEM with at least n = 3.

pre-stimulation levels (Figures 2.6b–d). These results suggested that Golgiplug™ could affect the IκBα kinetics (presumably through the eIF2α phosphorylation) [112, 114]. This also explains the observations in Figures 2.5a–c, where the intracellular TNFα concentration continued to increase since the Golgiplug™-induced response prolonged the NFκB activation by inhibiting the IκBα synthesis.

Table 2.2: Selected parameters and their newly estimated values for the final model.

Parameter	New Value
Coefficient for eIF2α phosphorylation (ν)	1.00
A20-mediated C1 deactivation	9.04×10^3 (μM min)$^{-1}$
TLR4 constitutive generation rate	3.75×10^{-2} μM min^{-1}
IKKK-mediated IKK activation	4.75×10^3 (μM min)$^{-1}$
Constitutive inactivation of IKK	2.85×10^{-2} min^{-1}
IκBα mRNA degradation rate	5.83×10^{-3} min^{-1}
Hill coefficient of IκBα transcription	4.16
Hill coefficient of IκBϵ transcription	5.00

2.3.3 Model Refinement

In order to account for the Golgiplug™-induced translation inhibition, the following equation was considered in addition to Equation 2.1:

$$k_{tl_i,m} = k_{tl_i}\left(1 - \frac{\nu G}{G + K}\right) \tag{2.12}$$

where $k_{tl_i,m}$ and k_{tl_i} are the IκBi, $\forall\, i = \alpha, \beta, \epsilon$, translation rates in the presence and absence of Golgiplug™, respectively, ν is a coefficient for the maximum translation inhibition and K is the Michaelis constant of the eIF2α phosphorylation. Equation 2.12 was included in Equation 2.2 along with Equation 2.1 for an accurate simulation. The process affected by this translation inhibition is shown in Figure 2.4 via a red arrow.

The proposed dynamic model was calibrated again using the parameter estimation procedure as described above. Since the additional measurements of the IκBα dynamics in the absence of Golgiplug™ were obtained, an extra sensitivity matrix was calculated, and the following was added to the objective function (Equation 2.7):

$$\sum_{k=1}^{2}\sum_{l=1}^{7}\left(\frac{y_{I\kappa B\alpha,k,2}(t_l) - \hat{y}_{I\kappa B\alpha,k,2}(t_l)}{y_{I\kappa B\alpha,k,2}(t_l)}\right)^2$$

where $\hat{y}_{I\kappa B\alpha,k,2}$ and $y_{I\kappa B\alpha,k,2}$ are the simulated and measured IκBα fold changes, respectively, in the

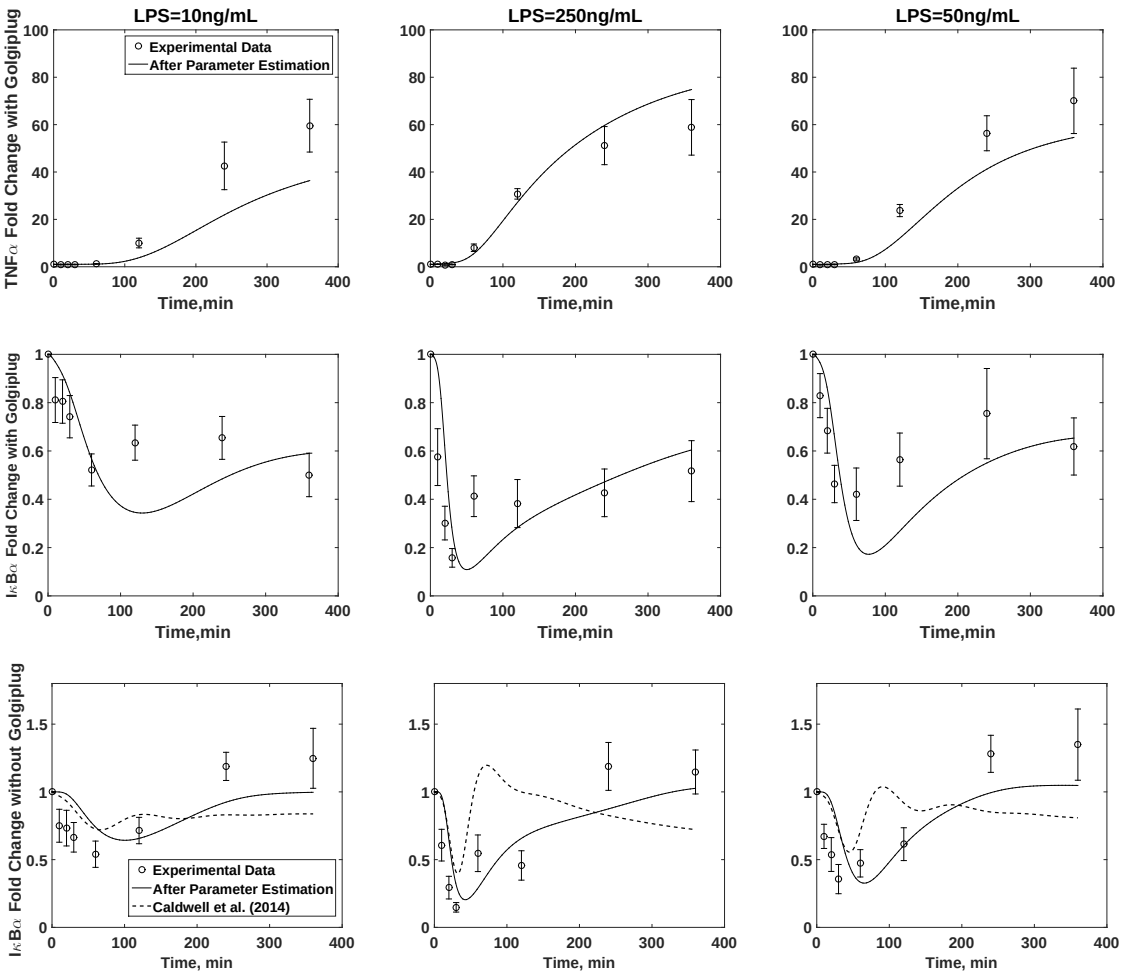
absence of Golgiplug™.

2.3.4 Final Model Validation

Based on the updated model structure and the available experimental data, the aforementioned parameter selection approach determined eight parameters, which could be uniquely estimated (Table 2.2). Most of the parameters selected by the proposed parameter selection procedure were relevant to the core NFκB-IκB feedback system such as Hill coefficients for IκB-α and -ϵ transcription, IKK deactivation, and IκBα transcript degradation rate. The remaining identified parameters are the TLR4 constitutive generation rate, C1 (TNFR complex) deactivation rate and eIF2α phosphorylation coefficient. Hence, all major processes considered in this system, which included the LPS- and TNFα-induced NFκB signaling pathway in the presence of Golgiplug™, were quantitatively validated against the single-cell experimental data.

Figure 2.7 shows simulated fold changes in IκBα and intracellular TNFα after parameter estimation. The simulated profiles were again compared with the experimental data. The normalized root-mean-squares of the parameter estimation results before and after the incorporation of the Golgiplug™ model (Equation 2.12) were 3.8 and 2.5, respectively, which demonstrated the improvement of the model fidelity. Overall, the model predictions were in qualitative and quantitative agreement with both training datasets and validation datasets, as well as the literature data, which validated the prediction capability of the calibrated model, as well as our hypothesis on the effect of Golgiplug™ on the inhibition of IκB translation. The results demonstrated that the calibrated model is capable of predicting input-output responses in the NFκB pathway. Additionally, the predictions from the current model were compared with the model proposed by Caldwell et al. [1] (Figures 2.7g–i). The proposed model was able to predict the observed IκBα dynamics under all LPS concentrations more accurately than the previous model by Caldwell et al. [1], which again demonstrated that the predictive capability of the model was improved in terms of simulating the IκBα dynamics.

In order to further assess the predictive capability of the newly calibrated model, the simulated dynamics of nuclear NFκB levels (i.e., activated NFκB) in the absence of Golgiplug™ were com-



(g) (h) (i)

Figure 2.7: Parameter estimation considering Golgiplug™-induced ER stress. (**a**–**c**) Measured (empty circle) and simulated (solid line) fold changes of IκBα concentrations over time were plotted in the presence of Golgiplug™. (**b**–**f**) Measured (empty circle) and simulated (solid line) fold changes of intracellular TNFα concentrations over time were plotted in the presence of Golgiplug™. (**g**–**i**) Measured (empty circle) and simulated (solid line) fold changes of IκBα concentrations over time were plotted in the absence of Golgiplug™. The IκBα dynamics predicted by the model in [1, 2] were also plotted in (**g**–**i**) for comparison. Indicated amounts of LPS were used for experiments and simulations.

puted and plotted in Figure 2.8a. The maximum NFκB translocation to the nucleus occurred within 2 hours of LPS addition, which was consistent with previous experimental studies [3, 115, 116]. Moreover, as the LPS concentration increased, the nuclear NFκB levels reached their maximum value earlier (i.e., at 50, 60, 75 and 105 min after adding LPS), and the areas under the curves in

Figure 2.8a, which were computed as indicators of the signal strength, were around 20 μM·min for different concentrations of LPS. Interestingly, a 25-fold change in the LPS concentration only resulted in less than a 100% change in the signal strength. This observation was consistent with single-cell studies by Tay et al. [78, 117], where they observed a relatively constant peak intensity and decreasing response time of the NFκB signal in mouse 3T3 cells upon TNFα or LPS stimulation.

Figure 2.8b shows the predicted amount of TNFα secreted under different LPS stimulation conditions in macrophages. As the LPS concentration increased, the concentration of secreted TNFα increased, which was expected since the signal (area under the peak) became stronger. Furthermore, similar to previous studies [1, 3, 118], the TNFα concentration peaked around 5 hours after stimulation and gradually declined thereafter; however, the rate of decline was slower than that reported by Maiti et al. [3] (Figure 2.8c), where they measured the TNFα secretion dynamics from RAW264.7 macrophages in response to LPS stimulation at the population level. This observation was consistent with the observation reported by Xue et al. [77], who observed using human monocyte-derived macrophages the amount of TNFα secreted to the medium from a single cell in a cell population was less than that from an isolated single-cell at 20 hours after the LPS stimulation. This suggested that the simulated dynamics by the proposed model is qualitatively similar to the signaling dynamics of an isolated single-cell instead of population-averaged dynamics, which was expected since the kinetic data obtained under Golgiplug™ were used to train the model.

2.4 Discussion

In this study, we have developed a dynamic model that can accurately simulate the average single-cell dynamics of the NFκB signaling pathway by combining the single-cell measurements and a numerical scheme with sensitivity analysis, parameter selection and parameter estimation. The dynamic model was built based on a previously developed NFκB model [1, 2, 9] and calibrated using the experimental data and the aforementioned numerical scheme. Predictions from the developed dynamic model are in good agreement with the experimental measurements under

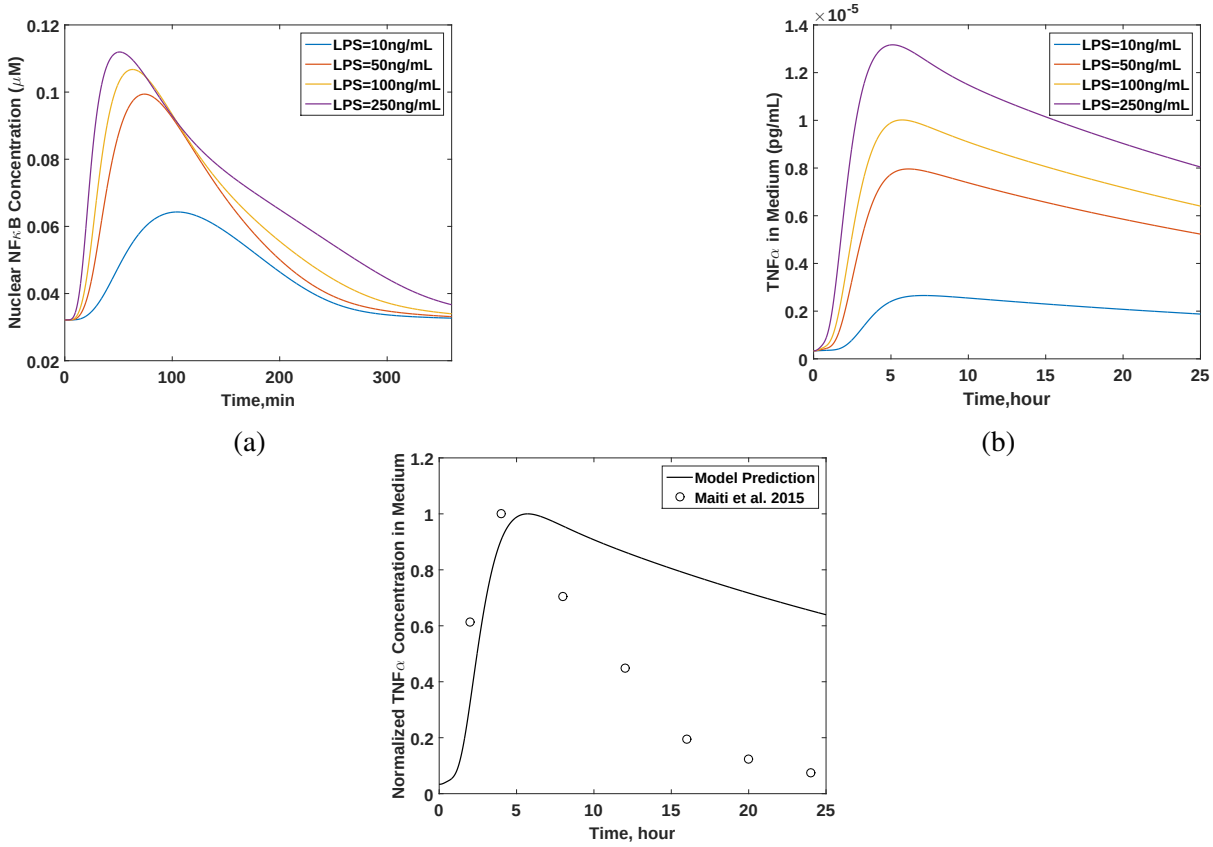


Figure 2.8: Simulated dynamics of NFκB nuclear translation and TNFα secretion. (**a**) Nuclear NFκB concentration and (**b**) the amount of TNFα secreted to the medium upon stimulation by 10, 50, 100 and 250 ng/mL of LPS. (**c**) The simulated dynamics of TNFα concentration in the medium was compared with the measurement by Maiti et al. [3] in response to 100 ng/mL of LPS. The TNFα concentration at each point was normalized to the maximum value obtained.

all LPS concentrations, which demonstrates that the model is capable of simulating the average single-cell dynamics.

Previous studies have used stochastic simulation algorithms such as Gillespie's algorithm [119] and approximate methods of Gillespie's algorithm [51, 120, 121, 122] to study single-cell dynamics and investigate heterogeneity in signaling pathways at the single-cell level [37, 73, 123]. For example, Lipniacki et al. [78, 123] proposed a hybrid stochastic-deterministic model of the TNFα-induced NFκB signaling pathway that was able to reproduce the heterogeneous responses observed in the single-cell measurements [78, 124] and identify possible origins of the heterogeneity. How-

ever, stochastic simulation algorithms are computationally expensive, and they are difficult to fit to experimental measurements for model validation [10, 125, 126]. A more viable method is a semi-stochastic model, which uses deterministic modeling with model parameters that have distributions [10, 37, 42], to reduce the computational cost while still studying the cell-to-cell variability. The dynamic model developed here can accurately simulate average single-cell dynamics and is a first step towards building a semi-stochastic model of the NFκB signaling.

The development of such a deterministic model for building a semi-stochastic model requires accurate parameter estimation, where values of model parameters are estimated by solving an optimization problem (Equations 2.7-2.11). However, parameter estimation is a nontrivial problem due to, but not limited to, ill conditioning, over-fitting and the non-identifiability of model parameters [13, 30, 127]. The ill-conditioning and over-fitting problems during parameter estimation are attributed to the fact that available experimental measurements are usually very limited and noisy, while mathematical models of signaling pathways are often very comprehensive and include a large number of parameters [23, 30]. As a result, the solution to the parameter estimation problem is likely to be non-unique or very sensitive to noise present in the experimental measurements. Furthermore, even if a large number of noise-free experimental measurements are available, the value of a parameter cannot be uniquely determined if the parameter is not identifiable [23, 47]; hence, it is necessary to check the parameter identifiability a priori.

The model developed in this work contains 148 parameters with limited experimental data, and hence, parameter estimation is very likely to suffer from the aforementioned issues in the parameter estimation procedure. Therefore, we implemented an integrated method combining sensitivity analysis and parameter selection before parameter estimation. Specifically, the sensitivity analysis quantified the effects of each parameter on the measurements, and the parameter selection method selected identifiable parameters via Gram–Schmidt orthogonalization. Then, the values of only the selected parameters were estimated in the parameter estimation, while the values of remaining parameters were fixed at their nominal values, which effectively alleviated the ill-conditioning problem by reducing the degrees of freedom in Equations 2.7-2.11 [23, 30, 127].

After parameter estimation, the simulated profiles of intracellular TNFα and IκBα exhibited reasonable agreement between the model predictions and the experimental measurements at all LPS concentrations (Figure 2.7). Furthermore, as shown in Figure 2.8c, model predictions after parameter estimation were distinct from that of a cell population as the simulated profiles were closer to the signaling dynamics of isolated single-cells. This was likely because the use of Golgiplug™ inhibited secretion of cytokines [128] and hence minimized potential autocrine and paracrine signaling from the secreted cytokines. This is important as the autocrine and paracrine signaling has been proposed as a key component in determining the overall signaling dynamics of cells in a population [77, 95, 96, 129]. Therefore, the proposed model, which was trained by the single-cell dynamics from flow cytometry in the presence of Golgiplug™, was able to describe the single-cell NFκB dynamics under minimal cytokine feedback.

It should be noted that the current model simulates the LPS-induced NFκB signaling dynamics in a cell, but it does not consider the initiation of the NFκB signaling pathway by TNFα secreted by neighboring cells. Hence, the flow cytometry measurements obtained in the presence of Golgiplug™ are appropriate to identify realistic parameter values to reproduce average single-cell dynamics. At the same time, as flow cytometry measures cellular responses from thousands of cells simultaneously, flow cytometry can provide distributions of the measurements (Figures 2.1-2.3). Based on this statistical information, one can estimate the distributions of the parameters by different methods such as Bayesian approaches [42, 10] or generalized polynomial chaos [130]. The model with the estimated parameter distributions is then the semi-stochastic model that can be used to study the heterogeneity in cellular responses.

The present study also suggests that cytokine production data acquired using flow cytometry in the presence of Golgiplug™ should be interpreted cautiously. As Golgiplug™ can block cytokine secretion, it is often used to assess the cytokine production at the single-cell level using flow cytometry [131, 132, 133]. The data shown in the work suggest that the dynamics of transcription factors and other signaling intermediates may be altered by the addition of Golgiplug™ (Figure 2.6). Therefore, data from studies using Golgiplug™ need to be interpreted cautiously, and

a model-based approach like the one presented here can be useful in eliminating the effects of Golgiplug™ and extract true signaling dynamics from flow cytometry data.

2.5 Conclusions

We systemically extracted the average single-cell dynamics of the LPS-induced NFκB signaling pathway through the integration of sensitivity analysis and a parameter selection scheme with flow cytometry data of key protein intermediates. Based on the measurements and the model structure, key model parameters were identified and estimated to maximize the prediction accuracy of the calibrated model while avoiding overfitting. The mismatch between the model predictions and experimental observations even after the parameter estimation revealed the existence of a previously unconsidered, yet important, mechanism related to Golgiplug™, which was subsequently validated by experiments and led to the update of the proposed model. Then, the resultant model was validated, and the simulated profiles from the updated model were in good agreement with experimental datasets under three different LPS concentrations. This model can be used as the nominal model to construct a deterministic model that has parameters with distributions and can be used to study the stochasticity in signaling.

3. AN INTEGRATED NUMERICAL AND EXPERIMENTAL FRAMEWORK FOR MODELING OF CTB AND GD1B GANGLIOSIDE BINDING KINETICS *

3.1 Introduction

In the intestinal lumen, *Vibrio Cholerae*, the main culprit of cholera, secrets cholera toxin (CTx). CTx can bind to the host cell membrane, leading to the internalization of CTx into cytoplasm. This process adversely increases the cyclic adenosine monophosphate level and causes severe fluid loss from host cells [134]. Severe dehydration and electrolyte imbalance could threaten the life of the patient if untreated [135]. Identifying the binding mechanisms of CTx to the host cell membrane can assist in understanding the pathogenesis of cholera and developing new CTx inhibitors for treatment.

CTx is one of AB_5 toxins that consists of a catalytic A subunit and five identical B subunits (Figure 3.1) [4, 136, 137]. The pentameric CTB can bind to ganglioside receptors in the host cell membrane to initiate the CTx internalization. Three major characteristics associated with the CTB-ganglioside binding mechanisms were identified by previous studies [11, 138, 139, 140]. First, a CTB can bind up to five ganglioside receptors in the host cell membrane (i.e., pentavalent binding) [5, 11]. Second, the binding affinity between a single CTB binding pocket and a ganglioside receptor can be changed if any adjacent pocket has been occupied by another receptor, which is called cooperative binding effect [5, 6, 134, 138, 141, 142]. Third, ganglioside receptors are found to form aggregates [139, 140, 143, 144, 145], and the receptor aggregates could hinder the association of individual ganglioside receptors with CTB proteins [139, 140, 146]. As a result, the CTB-ganglioside binding mechanism is an inherently complicated dynamic process; therefore, a mathematical model that can accurately predict the dynamics of the binding process is highly desirable to enhance the fundamental understanding and guide the future experimental design.

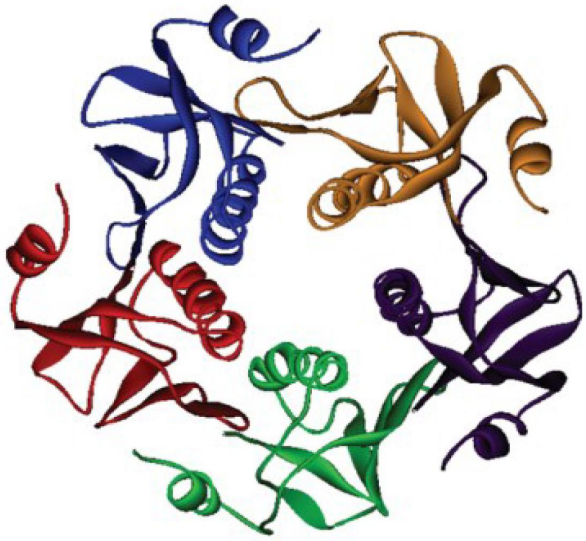


Figure 3.1: Top view of a CTB protein crystal structure (Reprinted from [4] under the permission of Creative Common CC BY license). Five monomers of the CTB are shown in five different color.

Probably due to the complexity of multivalent binding processes, many previous studies [147, 148, 149, 150, 151] still employed a simple single-site binding model, which assumes each CTB could bind with only one ganglioside receptor, to estimate the apparent association and dissociation rate constants. The simplified model misses the inherent multivalent binding process, which is a key characteristic preceding the toxin endocytosis, as well as the receptor aggregation effect [11, 134].

A more detailed mechanistic model was proposed by Lauer et al. [11]. They not only modeled the initial binding step of a free CTB binding to a free ganglioside in a membrane but also considered subsequent binding steps after a CTB gets anchored to the membrane surface. However, the effects of the cooperative binding and the receptor aggregation are still missing in the model, which may lead to inaccurate prediction of the equilibrium binding constant by several folds [5, 11, 139, 141].

Most of prior studies focused on CTB-GM1 interaction as GM1 is known as a primary ganglioside receptor of CTB due to its high avidity [152]. However, a few studies suspected that other

gangliosides also play important roles in the CTB attachment. For example, it was shown that CTB could bind to a cell in the absence of GM1 [152]. In addition, GM1 is of very low abundance in human intestinal epithelial cells (0.0015-0.003 mol% of glycosphingolipids) [153]; thus, a recent publication asked whether GM1 was sufficient to induce the CTx attachment [154]. More recently, it was reported that GM2, GD1b, and fucosyl-GM1 (f-GM1) also bind to CTB, and CTB-GD1b interaction is the strongest among these secondary receptors [6, 155]. Therefore, we focus on modeling CTB-GD1b binding mechanisms.

In this chapter, we developed a systematic framework, which integrates dynamic modeling, experiments, and parameter estimation to acquire more reliable binding rate parameters. Figure 3.2 illustrates the integrated experimental and numerical framework implemented in this study. Specifically, the first-principle modeling and experiments are performed in parallel, which is followed by parameter estimation to calibrate the proposed model. If the model predictions after the parameter estimation disagree with the the experimental measurements considerably, a series of model development, validation, and calibration process is repeated until the satisfactory fit is obtained. A new CTB binding model was established by introducing the cooperative binding effect and the receptor aggregation. The nanocube-based biosensor [6, 7] was implemented to measure the CTB-GD1b binding kinetics under various CTB and GD1b concentrations. We employed a parameter selection method to identify the process model parameters that can be uniquely determined based on the experimental measurements and the model structure to avoid overfitting [156]. The selected parameters were estimated by solving a dynamic optimization problem to minimize the error between model predictions and experimental measurements. The results suggest that our framework can be extended to other multivalent binding systems, enhancing our fundamental understanding of the complex biological processes.

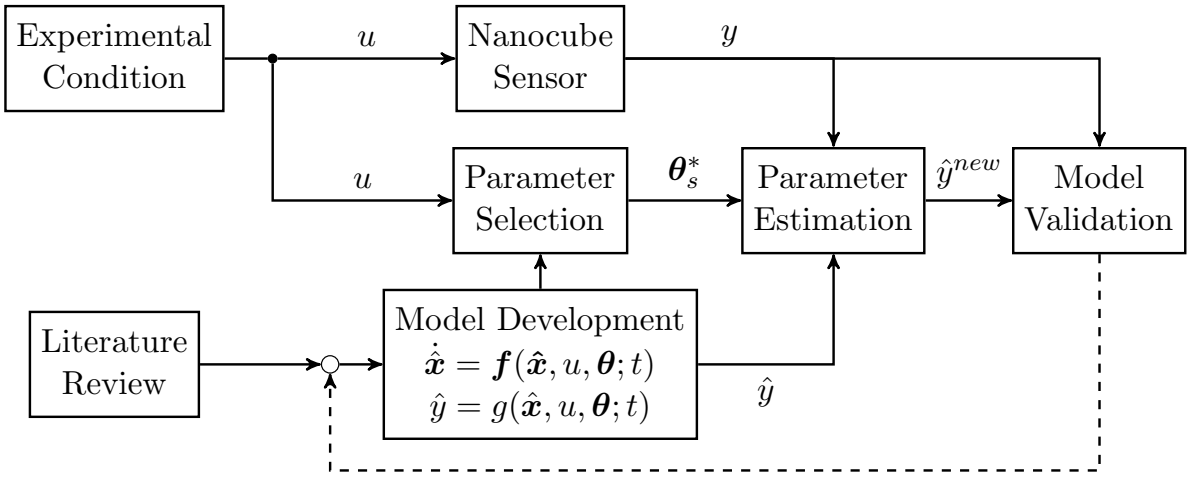


Figure 3.2: A schematic diagram of the model development and calibration procedure via an integrative approach combining experimental and numerical methodologies. u represents the experimental conditions (i.e., GD1b surface density as well as initial CTB concentration), y and $\hat{y}$ are measured and simulated membrane-bound CTB concentration, respectively, $\boldsymbol{\theta}_s^*$ is a set of the most identifiable parameters, and $\hat{y}^{new}$ is the predicted membrane-bound CTB concentration after parameter estimation.

3.2 Mathematical Modeling

The aim of this study is to model the kinetics of CTB proteins binding to GD1b receptors on model membranes. The artificial model membrane contains phospholipids and GD1b receptors, allowing us to tune the membrane compositions for studying the CTB-GD1b binding mechanisms. In addition, the model membrane can form supported lipid bilayers on nanocube sensors [6, 7]. Therefore, we could experimentally monitor CTB binding kinetics and validate the proposed binding model.

Because CTB is a homopentamer (i.e. a protein has five identical binding subunits), Lauer et al. [11] developed a stepwise reaction pathway that considered five different binding intermediates based on the number of bound receptors per CTB. However, it has been shown that the binding affinity can be altered if any adjacent binding pocket has been occupied by another receptor [138, 141, 142]. In order to consider the cooperative effect, we introduced two additional binding states of CTB-GD1b complexes to the CTB binding model developed by Lauer et al. [11]. A CTB-

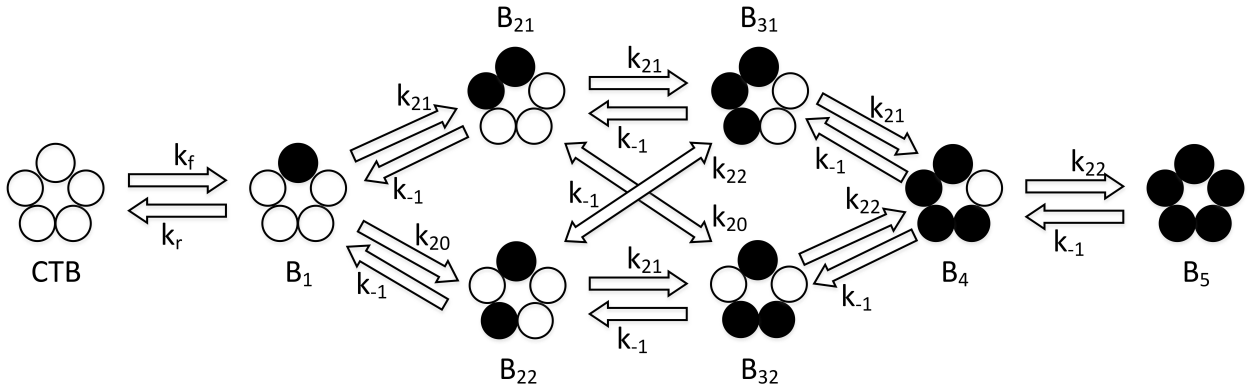


Figure 3.3: A schematic illustration of the dynamic multi-step binding mechanisms between GD1b receptors and a CTB protein (adapted from Lin et al. [5]). Available and filled spheres represent empty and GD1b-bound binding pockets of a CTB protein, respectively.

GD1b complex with two or three bound GD1b receptors has two different configurations based on the relative locations of bound GD1b receptors (Figure 3.3). The binding kinetics of these seven binding complexes are modeled by seven ODEs.

A CTB has to diffuse from the solution phase to a membrane surface before one of its binding sites attaches to a GD1b receptor. At the same time, a free GD1b receptor moves two dimensionally, encounters a CTB protein, and enables subsequent binding events on the membrane. Therefore, the transport process of CTB proteins to the membrane needs to be addressed in order to accurately model the first step of the binding process. In this study, the effective rate constants are used to simplify the modeling process [11, 157, 158]. In addition, it is known that ganglioside receptors can form clusters in the host cell membrane and model membrane [139, 140, 144, 159, 160, 161], and the clusters of GD1b receptors could inhibit the association between CTB and ganglioside receptors due to steric hindrance, increasing the equilibrium dissociation constant [139, 146]. To address this issue, we introduced an empirical factor that accounts for the clustering effect into the association rate constant.

Based on the above considerations, k_f and k_r, which are the effective association and dissociation rate constants between a CTB in solution and a free GD1b in a membrane, are defined as

follows [11, 157, 158, 162]:

$$
\begin{aligned}
k_1 &= k_{1,0} e^{-KR} \\
k_f &= \frac{k_1}{1 + \frac{5 k_1 R}{4\pi D a N_a}} \\
k_r &= \frac{k_{-1}}{1 + \frac{5 k_1 R}{4\pi D a N_a}}
\end{aligned}
\tag{3.1}
$$

where k_1 is the association rate constant between CTB and GD1b that is reduced due to the receptor aggregation, $k_{1,0}$ is the intrinsic association rate constant, K is the binding inhibition coefficient, R is the number of free GD1b receptors per nanocube, D is the effective diffusivity coefficient for a CTB in solution [162], a is the effective radius of a lipid-membrane covered nanocube, N_a is the Avogadro's number, and k_{-1} is the intrinsic dissociation rate constant. A prior study has shown the apparent association constant of the CTB-GM1 binding exponentially decreases when the surface density of GM1 increases from 0.02% to 10% [139]. Thus, we can use an exponential decay function to model the relation of the association rate constant and receptor density. It should be noted that receptor aggregates are not explicitly modeled as a separate state, and R refers to the total number of free GD1b receptors, including both unaggregated and aggregated forms that are not bound to any CTB proteins.

After a CTB attaches to a membrane, the subsequent binding occurs on the two dimensional membrane surface. We introduced a binding rate constant k_2 to describe the binding events between a CTB-GD1b complex and a GD1b receptor on a model membrane [11]. Also, we defined a new rate constant k_{20}, which is k_2 divided by the surface area of the model membrane. As described above, due to the cooperative binding effect, a single k_{20} rate parameter is not sufficient to describe all association events between a CTB-GD1b and a free GD1b receptor. Previous studies [5, 11, 138, 141] reported that the association energy between a CTB binding subunit and a free glycan, which is a sugar headgroup of a ganglioside receptor, could be altered by its adjacent binding subunits. Therefore, we considered three different binding scenarios on model membranes: a GD1b receptor binds to a binding pocket with zero, one or two GD1b-bound neighbors in a

membrane-bound CTB (Figure 3.3). To account for such a cooperative binding effect, two more binding rate constants (k_{21} and k_{22}) were introduced, and the relationships among these rate parameters are defined as follows:

$$\begin{aligned} k_{20} &= k_2/A \\ k_{21} &= \alpha_1 k_{20} \\ k_{22} &= \alpha_2 k_{20} \end{aligned} \tag{3.2}$$

where k_{20} is the effective association rate between a CTB-GD1b complex with a free GD1b receptor in the membrane, A is the surface area of the model membrane, k_{21} and k_{22} are the association rate constants for a free GD1b receptor binding to a binding pocket with one and two GD1b-bound neighbors in a membrane-bound CTB (Figure 3.3), respectively, and α_1 and α_2 are the factors that describe the cooperative binding effect of the CTB-GD1b binding.

The following system of ODEs are proposed to describe the dynamics of CTB-GD1b binding in the model membrane (Figure 3.3):

$$\begin{aligned}
\frac{dB_1}{dt} &= 5k_f CR - k_r B_1 - 2k_{21}B_1R + 2k_{-1}B_{21} - 2k_{20}B_1R + 2k_{-1}B_{22} \\
\frac{dB_{21}}{dt} &= 2k_{21}B_1R - 2k_{-1}B_{21} - 2k_{21}B_{21}R + 2k_{-1}B_{31} - k_{20}B_{21}R + k_{-1}B_{32} \\
\frac{dB_{22}}{dt} &= 2k_{20}B_1R - 2k_{-1}B_{22} - 2k_{21}B_{22}R + 2k_{-1}B_{32} - k_{22}B_{22}R + k_{-1}B_{31} \\
\frac{dB_{31}}{dt} &= 2k_{21}B_{21}R - 2k_{-1}B_{31} + k_{22}B_{22}R - k_{-1}B_{31} - 2k_{21}B_{31}R + 2k_{-1}B_4 \\
\frac{dB_{32}}{dt} &= 2k_{21}B_{22}R - 2k_{-1}B_{32} + k_{20}B_{21}R - k_{-1}B_{32} - 2k_{22}B_{32}R + 2k_{-1}B_4 \\
\frac{dB_4}{dt} &= 2k_{21}B_{31}R + 2k_{22}B_{32}R - 4k_{-1}B_4 - k_{22}B_4R + 5k_{-1}B_5 \\
\frac{dB_5}{dt} &= k_{22}B_4R - 5k_{-1}B_5
\end{aligned} \tag{3.3}$$

which are subject to the following mass balance equation and initial conditions

$$R(t) = R_0(t) - B_1(t) - 2\left(B_{21}(t) + B_{22}(t)\right) - 3\left(B_{31}(t) + B_{32}(t)\right) - 4B_4(t) - 5B_5(t)$$
$$B_{ij}(0) = 0, \quad R(0) = R_0, \quad C(0) = C_0$$

where B_{ij} denotes a CTB-GD1b complex with i number of bound-GD1b receptors in the j^{th} configuration (Figure 3.3), C is the unbound CTB concentration, R_0 is the initial number of GD1b receptors per nanocube, and C_0 is the initial CTB concentration. Since the nanocube biosensor measures the membrane-bound CTB concentration, the process output is the sum of all B_{ij}, which is calculated by:

$$C_B(t) = \frac{c}{N_a}(B_1(t) + B_{21}(t) + B_{22}(t) + B_{31}(t) + B_{32}(t) + B_4(t) + B_5(t)) \tag{3.4}$$

where C_B is the concentration of membrane-bound CTB on all nanocubes in the solution, and c is the number concentration of nanocubes in the solution.

In the following sections, we will use the expressions below to denote Equations 3.3-3.4:

$$\begin{aligned} \dot{\hat{\boldsymbol{x}}}_k &= \boldsymbol{f}\left(\hat{\boldsymbol{x}}_k, \boldsymbol{\theta}, u_k\right) \\ \hat{y}_k &= g\left(\hat{\boldsymbol{x}}_k, \boldsymbol{\theta}, u_k\right) \end{aligned} \tag{3.5}$$

where $\hat{\boldsymbol{x}}_k$ is the predicted state trajectory (i.e., $\hat{\boldsymbol{x}} = [B_1\ B_{21}\ \cdots\ B_5]^T$) under an input u_k (i.e., an initial concentration of CTB and a GD1b surface density), $\boldsymbol{\theta}$ is the vector containing the model parameters, $\hat{y}_k$ is the predicted membrane-bound CTB concentration (i.e., C_B).

3.3 Parameter Estimation

Parameter estimation is essential to make the model predictions more accurate and reliable [23, 163]. It is usually formulated as a least-squares problem; that is, the model parameters are estimated by minimizing the squared deviation between experimental data and model predictions. In this work, among all model parameters (Equation 3.3), the values of a, c, A, D, R_0, and C_0

were determined by the experimental conditions. The remaining unknown model parameters, $\boldsymbol{\theta} = [k_{1,0},\ k_{-1},\ k_2,\ K,\ \alpha_1,\ \alpha_2]^T$, have to be determined via fitting the binding model to experimental data.

It is very tempted to fit all these unknown model parameters simultaneously to a set of available experimental data. An important but often-overlooked first step is to assess the identifiability of the model parameters [23, 156]. If the number of model parameters to be estimated exceeds the number of parameters that could be uniquely determined from the available experimental data, the parameter estimation problem becomes ill-conditioned, which needs to be regularized to compute the optimal solution [76, 81, 164, 165, 166, 167].

3.3.1 Parameter Selection

Among several regularization methods proposed in the literature (reviewed in [23, 163]), the parameter selection method proposed by Chu and Hahn [156] is employed in this work to classify model parameters into groups of identifiable and unidentifiable ones, and only the former is estimated in the subsequent parameter estimation step. In a dynamical system, a set of model parameters $\boldsymbol{\theta}$ is identifiable if

$$\hat{y}(\hat{\boldsymbol{x}}, \boldsymbol{\theta}; u) = \hat{y}(\hat{\boldsymbol{x}}, \boldsymbol{\theta}_1; u) \implies \boldsymbol{\theta} = \boldsymbol{\theta}_1$$

Albeit of its simple definition, it is nontrivial to determine which parameters are identifiable, especially for a nonlinear system like Equation 3.3. In this study, a parameter is selected for estimation if it has a significant impact on the process output while it has a minimum correlation with other parameters [23]. One way to quantify these two criteria at the same time is to perform the local sensitivity analysis of the parameters using a set of nominal values for the model parameters [102]. The local sensitivity $s_{jk}(t_l)$ is defined as a partial derivative of the system output $\hat{y}_k(t_l)$ with respect to the parameter $\theta_j \in \boldsymbol{\theta},\ \forall j = 1, \ldots, 6$, at the sampling time t_l under the input u_k:

$$s_{jk}(t_l) = \left.\frac{\partial \hat{y}_k}{\partial \theta_j}\right|_{t=t_l} = \left.\frac{\partial g}{\partial \hat{\boldsymbol{x}}_k^T}\frac{\partial \hat{\boldsymbol{x}}_k}{\partial \theta_j}\right|_{t=t_l} + \left.\frac{\partial g}{\partial \theta_j}\right|_{t=t_l} \tag{3.6}$$

In order to compute the time evolution of s_{jk}, the values of $\partial\hat{\boldsymbol{x}}_k/\partial\theta_j$ at every sampling time are calculated by integrating Equation 3.7, which is derived by differentiating Equation 3.5 with respect to θ_j as follows:

$$\frac{d}{dt}\left(\frac{\partial\hat{\boldsymbol{x}}_k}{\partial\theta_j}\right)_{t=t_l} = \frac{\partial\boldsymbol{f}}{\partial\hat{\boldsymbol{x}}_k^T}\frac{\partial\hat{\boldsymbol{x}}_k}{\partial\theta_j}\bigg|_{t=t_l} + \frac{\partial\boldsymbol{f}}{\partial\theta_j}\bigg|_{t=t_l} \tag{3.7}$$

which is subject to the following initial conditions

$$\frac{\partial\hat{\boldsymbol{x}}_k}{\partial\theta_j}\bigg|_{t=0} = \mathbf{0} \tag{3.8}$$

where $\partial\hat{\boldsymbol{x}}_k/\partial\theta_j$ is the sensitivity of state $\hat{\boldsymbol{x}}_k$ with respect to the parameter θ_j under input u_k.

Suppose there are N_1 different initial CTB concentrations tested under each of N_2 different GD1b surface densities (i.e., the process input $\mathbf{u} \in \mathbb{R}^{N\times 1}$, where $N = N_1 \times N_2$), and the membrane-bound CTB concentration (i.e., the process output y) is measured at M_k different time instants for each u_k, $\forall k = 1, \ldots, N$. Then, Equations 3.6-3.8 are integrated to compute sensitivities at M_k sampling times under N process inputs. The obtained sensitivities can form a sensitivity matrix $\mathbf{S} \in \mathbb{R}^{d\times 6}$, where $d = \sum_{k=1}^{N} M_k$, as follows:

$$\mathbf{S} = \begin{bmatrix} \frac{\hat{\theta}_1}{\hat{y}_{1,ss}}s_{11}(t_1) & \cdots & \frac{\hat{\theta}_p}{\hat{y}_{1,ss}}s_{p1}(t_1) \\ \vdots & \ddots & \vdots \\ \frac{\hat{\theta}_1}{\hat{y}_{1,ss}}s_{11}(t_{M_1}) & & \frac{\hat{\theta}_p}{\hat{y}_{1,ss}}s_{p1}(t_{M_1}) \\ \vdots & \ddots & \vdots \\ \frac{\hat{\theta}_1}{\hat{y}_{N,ss}}s_{1N}(t_1) & & \frac{\hat{\theta}_p}{\hat{y}_{N,ss}}s_{pN}(t_1) \\ \vdots & \ddots & \vdots \\ \frac{\hat{\theta}_1}{\hat{y}_{N,ss}}s_{1N}(t_{M_N}) & \cdots & \frac{\hat{\theta}_p}{\hat{y}_{N,ss}}s_{pN}(t_{M_N}) \end{bmatrix} \tag{3.9}$$

where $\hat{y}_{k,ss}$ is the predicted steady-state value of output $\hat{y}$ under input u_k, and $\hat{\theta}_j$ is the nominal value of θ_j. In order to eliminate possible scaling effects of different parameters and the outputs on the sensitivity, each column is multiplied by the corresponding parameter $\hat{\theta}_j$, and each row is di-

vided by the corresponding $\hat{y}_{k,ss}$. Therefore, each column of the sensitivity matrix $\mathbf{S}$ represents the sensitivity vector of a parameter at M_k time instants under N process inputs when the measurement is available.

Utilizing the sensitivity matrix $\mathbf{S}$, the aforementioned two criteria for the parameter selection can be quantified by computing D-optimality criterion, which is defined as follows [156]:

$$\phi_D(\mathbf{S}_s^T\mathbf{S}_s) = \det(\mathbf{S}_s^T\mathbf{S}_s) \tag{3.10}$$

where $\mathbf{S}_s \in \mathbb{R}^{d\times p_s}$ is the sensitivity matrix that consists of a set of columns in $\mathbf{S}$ corresponding to the selected parameters, and p_s is the number of selected parameters, which is obtained by counting the number of singular values of $\mathbf{S}$ greater than 0.1% of the largest singular value [81]. A parameter subset with the largest ϕ_D value contains the most identifiable parameters that will be estimated in parameter estimation.

It should be noted that the result of the local sensitivity analysis highly depends on the nominal values of the model parameters. As a result, if exact values of model parameters are not known *a priori*, the selected parameter solely based on the local sensitivity analysis might not be robust [13, 156, 168]. In this work, the nominal values of the kinetic parameters are unreliable due to the limited number of relevant studies in the past. And, depending on the experimental techniques, the nominal values of $k_{1,0}$, k_{-1}, α_1, and α_2 reported in different studies vary in a wide range [147, 148, 155]. Moreover, the values of k_2 and K have not been quantified. Thus, we implemented an approach proposed by Chu and Hahn [156] to determine the most identifiable parameters by combining the local sensitivity analysis with Latin hypercube sampling (LHS). The procedure can be summarized as follows [156]:

1. Enumerate every possible combination of choosing p_s out of the six model parameters
2. Determine the ranges of values for all the model parameters
3. Generate q parameter samples by LHS with a uniform distribution

4. Compute D-optimality criterion (ϕ_D) for each parameter combination generated in Step 1 with each parameter sample from Step 3.

5. Compute the mean D-optimality criterion ($\bar{\phi}_D$) for each parameter combination

The most identifiable parameter subset is the one with the largest $\bar{\phi}_D$. We selected LHS because it is an efficient and unbiased sampling methodology [156, 169]. More details of the methodology can be found in [156].

3.3.2 Least-squares problem for parameter estimation

In the parameter estimation step, only the parameters identified in the parameter selection step, $\boldsymbol{\theta}_s$, are estimated by the least-squares problem while the remaining parameters are fixed at their nominal values (Table 3.1). The least-squares problem is formulated as follows:

$$\min_{\boldsymbol{\theta}_s} \quad \sum_{k=1}^{N}\sum_{l=1}^{M_k}\left(\frac{y_k(t_l)-\hat{y}_k(t_l)}{E_k(t_l)}\right)^2 \tag{3.11a}$$

$$\text{s.t.} \quad \hat{y}_k(t_l) = g\left(\hat{\boldsymbol{x}}_k(t_l), \boldsymbol{\theta}_s, u_k\right) \tag{3.11b}$$

$$\dot{\hat{\boldsymbol{x}}}_k(t_l) = \boldsymbol{f}\left(\hat{\boldsymbol{x}}_k(t_l), \boldsymbol{\theta}_s, u_k\right), \quad \hat{\boldsymbol{x}}_k(t=0) = \boldsymbol{x}_0 \tag{3.11c}$$

$$\boldsymbol{x}^{lb} \leq \hat{\boldsymbol{x}}_k \leq \boldsymbol{x}^{ub} \tag{3.11d}$$

$$\boldsymbol{\theta}_s^{lb} \leq \boldsymbol{\theta}_s \leq \boldsymbol{\theta}_s^{ub} \tag{3.11e}$$

where $\boldsymbol{\theta}_s \in \mathbb{R}^{p_s \times 1}$ is a set of the selected parameters, $E_k(t_l)$ is the standard error of $y_k(t_l)$, and $\boldsymbol{x}_0$ is the vector containing the initial conditions of $\boldsymbol{x}$.

$k_{1,0}\ (M^{-1}s^{-1})$	$k_{-1}\ (s^{-1})$	$k_2\ (nm^2s^{-1})$	K	α_1	α_2
1.6×10^4	0.032	0.021	0.01	0.5	0.5

Table 3.1: The model parameters and their nominal values.

In the least-squares problem of Equation 3.11, the objective function of Equation 3.11a represents the squared deviation between the model predictions and the experimental measurements, which is to be minimized by adjusting the values of $\boldsymbol{\theta}_s$. Equation 3.11d imposes the lower and upper bounds on states based on experimental observations and the knowledge on this biological system, and the constraint of Equation 3.11e ensures the parameter values are consistent witthe previous studies on the CTB-ganglioside binding process [5, 147, 148].

3.3.3 Experimental methods

The kinetic datasets of CTB binding to the model membrane containing 1, 2 and 4 mol% GD1b at 21, 32, 40 and 52 nM CTB concentrations were acquired by the nanocube-based sensor. The binding kinetics were used for the parameter estimation and the model validation. Specifically, the collected experimental measurements were divided into a training dataset for the parameter estimation process and a validation dataset for the model validation process. All the data considered was in one batch.

3.3.3.1 Materials

Disialoganglioside GD1b(NH_4^+salt)(Galβ1-3GalNAcβ 1-4(Neu5Acα2-8)(Neu5Acα2-3)Gal β1-4Glc-Ceramide) were purchased from Matreya LLC (State College, PA). 1,2-dioleoyl-sn-glycero-3-phosphocholine (DOPC), 1,2-dioleoyl-sn-glycero-3-phospho-L-serine - sodium salt (DOPS), and 1,2-dioleoyl-sn-glycero-3-phosphoethanolamine-N-(biotinyl) (biotin-PE) were obtained from Avanti Polar Lipids (Alabaster, AL). CTB (lyophilized powder) from *Vibrio Cholerae*, streptavidin from Streptomyces avidinii (Stp) and casein from bovine milk were purchased from Sigma-Aldrich. Chloroform and methanol were purchased from Fisher Scientific. All the CTB binding experiments were performed in Tris-buffered saline (TBS) from Sigma Aldrich while the experiments involving Stp and biotin-PE were performed in phosphate buffered saline (PBS) from CulGeneX.

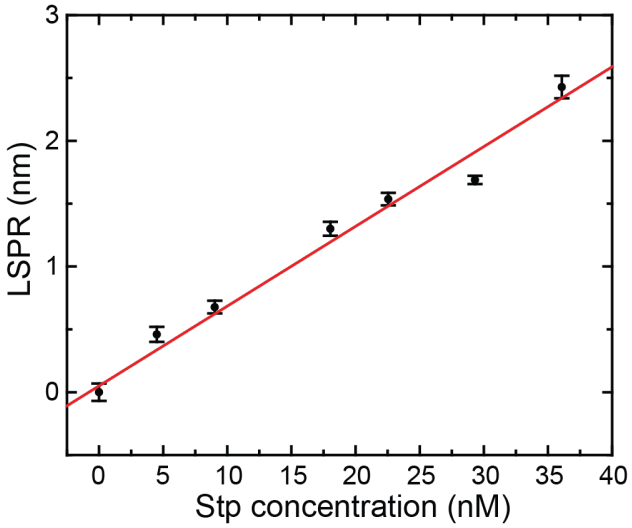


Figure 3.4: The binding curve of Stp-biotin used to calibrate the nanocube-sensor for this study (see [6, 7] for the detailed procedures).

3.3.3.2 Synthesis & calibration of the nanocube sensor

The synthesis of silica coated silver nanocubes was followed as reported in Worstell et al. [6, 155]. The silver nanocubes were prepared based on polyol method, and silica coating on the silver nanocubes was done using Stöber's™ process. The quality control of each nanocube synthesis batch was done by a set of characterization procedures, including (1) figure-of-merit measurements, (2) film thickness measurements using transmission electron microscopy (TEM), (3) reconstituted cell membrane confirmation using cryo-TEM, and (4) the calibrations of bound protein density using the standard protein-membrane binding systems [6, 7]. In Step 4, Stp-biotinyl lipid binding was used as the positive control. In brief, the lipid bilayer (89% DOPC/10% DOPS/1% biotin-PE) coated nanocubes were titrated with Stp, and the average Stp surface density on nanocubes was evaluated (Figure 3.4). All the CTB densities shown in the kinetic binding curves have been calibrated by this positive control.

3.3.3.3 Supported lipid membrane preparation

Small unilamellar vesicles (SUVs) were prepared at 3 g/L concentration. 1 mol%, 2 mol%, and 4 mol% of GD1b stored in chloroform/methanol/water mixture was mixed with 10 mol% of DOPS and 89 mol%, 88 mol%, and 86 mol% of DOPC stored in chloroform. The mixture was dried using a rotary evaporator (Heidolph Hei-VAP Value®) for five minutes. The dried lipids were then rehydrated with Milli-Q® water, followed by extrusion through 100 nm polycarbonate filters (Whatman) using a Mini-extruder (Avanti Polar Lipids). A previously established protocol of sequential addition of silica coated silver nanocubes and buffer (TBS or PBS) solution to SUVs was used to form supported lipid membranes [6, 155]. There was a small modification of using vortex mixer instead of bath sonicator for mixing after each sequential addition. After membrane coating, the nanocubes were incubated with 0.5 g/L casein in 1X TBS or 1X PBS solution to prevent non-specific binding.

3.3.3.4 CTB binding measurement

17 μL of the lipid membrane coated nanocubes was added to two wells in the 384-well plate, followed by addition of 3 μL of 1X TBS solution to one well (control) and 3 μL of the required CTB concentration to the other well. Kinetic measurements of extinction spectra of the solutions were started immediately after CTB addition in a UV/Vis microplate spectrophotometer equipped with a CCD (FLUOstar Omega®, BMG-Labtech) and continued for three hours. Each binding measurement was repeated in four wells. The quadrupolar LSPR peak location was estimated by fitting the measured absorption spectra to a fifth order polynomial. All binding experiments were carried out in triplicate and repeated at least twice.

3.4 Results

The datasets obtained with the CTB concentration of 21, 32 and 52 nM ($N_1 = 3$) under the GD1b receptor density of 1, 2 and 4 mol% ($N_2 = 3$) were used for parameter estimation; therefore, N = 9 in Equation 3.11. The binding kinetics measured at 40 nM CTB concentration were used to validate the model [105].

3.4.1 Parameter selection results

Although the nominal values of model parameters are needed for computing the sensitivity matrix, the information of rate parameters are limited. MacKenzie et al. [147] and Kuziemko et al. [148] measured values of $k_{1,0}$, and k_{-1} for the CTB-GD1b and CTB-GM1 binding kinetics; however, the binding kinetics reported in these two studies were quite different. This is because they used a single-site binding model to describe the inherent multivalent binding process. Hence, these values were not applicable to our model.

We estimated the nominal values of the model parameters by scaling down the rate parameters of the CTB-GM1 binding reported by Lauer et al. [11]. Prior studies [147, 148] also reported $k_{1,0}$ values in the CTB-GD1b binding system were around 60% in the CTB-GM1 binding system, and k_{-1} value in the former was one order of magnitude higher than that in the latter. Hence, the kinetic parameters ($k_{1,0}$, k_{-1}, k_2) adopted from Lauer et al. [11] were scaled accordingly as the nominal values of the CTB-GD1b binding model. The nominal values of the cooperative binding factors α_1 and α_2 were adopted from a recent study reported by Krishnan et al. [155] (Table 3.1).

	Parameter subset	$\log_{10}\bar{\phi}_D$	$\log_{10}\phi_D$ (at nominal values)	Probability to be the optimum
$p_s = 3$	$k_{1,0}$, k_{-1}, K	5.5	8.6	0.76
	$k_{1,0}$, k_2, K	4.1	7.9	0.04
	k_{-1},k_2, K	3.3	7.9	0.01
$p_s = 2$	$k_{1,0}$, K	6.0	7.9	0.26
	k_{-1}, K	5.8	8.0	0.51
	$k_{1,0}$, k_{-1}	2.7	4.6	0.08
$p_s = 4$	$k_{1,0}$, k_{-1}, k_2, K	2.0	7.8	0.72
	$k_{1,0}$, k_{-1}, K, α_1	1.2	7.1	0.02
	$k_{1,0}$, k_{-1}, K, α_2	-0.04	6.0	0.09

Table 3.2: The result of the parameter selection procedure.

With the given nominal parameter values, the sensitivity analysis and parameter selection were performed as described in the previous section. The number of identifiable parameters is deter-

mined by the magnitude of singular values of **S**; the threshold used in this study was the number of singular values of **S** greater than 0.01% of the largest singular value. Based on this criterion, three out of the six model parameters are determined to be identifiable from the experimental data obtained from the nanocube-based biosensor ($p_s = 3$). In order to carry out the sampling-based sensitivity analysis [156], the parameter values were allowed to vary between 1% to 10000 $\%$ of their nominal values listed in Table 3.1, and 5000 different sets of the parameter values in these ranges were sampled and used to perform the local sensitivity analysis 5000 times. For $p_s = 3$, the best parameter subset was $\{k_{1,0}, k_{-1}, K\}$ based on the average D-optimality criterion value, $\bar{\phi}_D$ (Table 3.2). Furthermore, its probability to be chosen as the optimal set was highest (Table 3.2), where the probability was calculated by dividing the number of samples when each parameter subset had the largest value of $\bar{\phi}_D$ by the total number of samples.

As a comparison, the parameter selection procedure was repeated with $p_s = 2$ and 4. Their ϕ_D values at the nominal parameter values and their probabilities to be the optimal set were presented in Table 3.2. When the value of p_s increased from two to three, the $\bar{\phi}_D$ value of the most optimal parameter set decreased slightly while the $\bar{\phi}_D$ value decreased by more than three orders of magnitude when the value of p_s changed from three to four. This indicates at most three model parameters are identifiable, and estimating more than three parameters is likely to result in a non-unique solution [23, 156].

It should be noted that under all conditions the $\bar{\phi}_D$ value of a subset deviated significantly from that at the nominal parameter values (Table 3.2), which suggested that selecting parameters solely based on the nominal parameter values could be misleading [156]. For example, the D-optimality criterion value of the set $\{k_{1,0}, K\}$ was slightly smaller than that of the set $\{k_{-1}, K\}$ at the nominal values, but their relationship was reversed when their $\bar{\phi}_D$ values were compared (Table 3.2). This analysis suggests that the local analysis along with LHS is a viable option to determine the set of identifiable model parameters if their nominal values were unreliable and only a wide range of their values was known *a priori*.

3.4.2 Parameter estimation results

In order to estimate the selected parameters, the least-squares problem (Equation 3.11) was solved by varying the values of $k_{1,0}$, k_{-1} and K while the values of the other parameters were fixed at their nominal parameter values (Table 3.1). Equation 3.11 is a dynamic optimization problem with ODEs as one of its constraints [168]. Here, we implemented a sequential approach to solve the dynamic optimization problem using Matlab software suite [168, 170, 171]. Specifically, at the inner loop of the dynamic optimization problem, numerical integration of Equation 3.11b was performed using *ode15s* function in Matlab to compute the objective function as well as its gradient numerically. Then, the optimization method was applied at the outer loop to minimize the objective function via *fmincon* with *MultiStart* in Matlab to obtain an improved result by solving Equation 3.11 with 50 different different initial conditions [170, 171]. Running in a Windows desktop with 3.60 GHz Quad-Core Intel i7 Processors, the total CPU time for solving the optimization problem was around 900 seconds. In average, it took about 0.6 seconds to evaluate Equation 3.11a once, and it was evaluated 5500 times, which took around 830 seconds (90% of the total CPU time) through a parallel computation scheme utilizing four cores. For a large-scale problem with numerous constraints, a simultaneous approach such as the collocation problem method will improve the computational costs [172]. The estimated parameter values are listed in Table 3.3.

$k_{1,0}$ $(M^{-1}\ s^{-1})$	k_{-1} (s^{-1})	K
2.5×10^{3}	6.6×10^{-3}	9.3×10^{-4}

Table 3.3: Estimated values of the selected parameters.

The estimated CTB-GD1b association rate constant is one order of magnitude lower than that of the CTB-GM1 binding estimated by Lauer et al. [11]. This result is consistent with the previous studies of the CTB-ganglioside binding kinetics [147, 148, 149]. The estimated value of dissoci-

ation constant (k_{-1}) was about the same order of magnitude ($\sim 10^{-3}s^{-1}$) compared to that of the CTB-GM1 binding kinetics estimated [11] while previous studies [147, 148] reported k_{-1} of the CTB-GD1b binding kinetics was one order of magnitude larger.

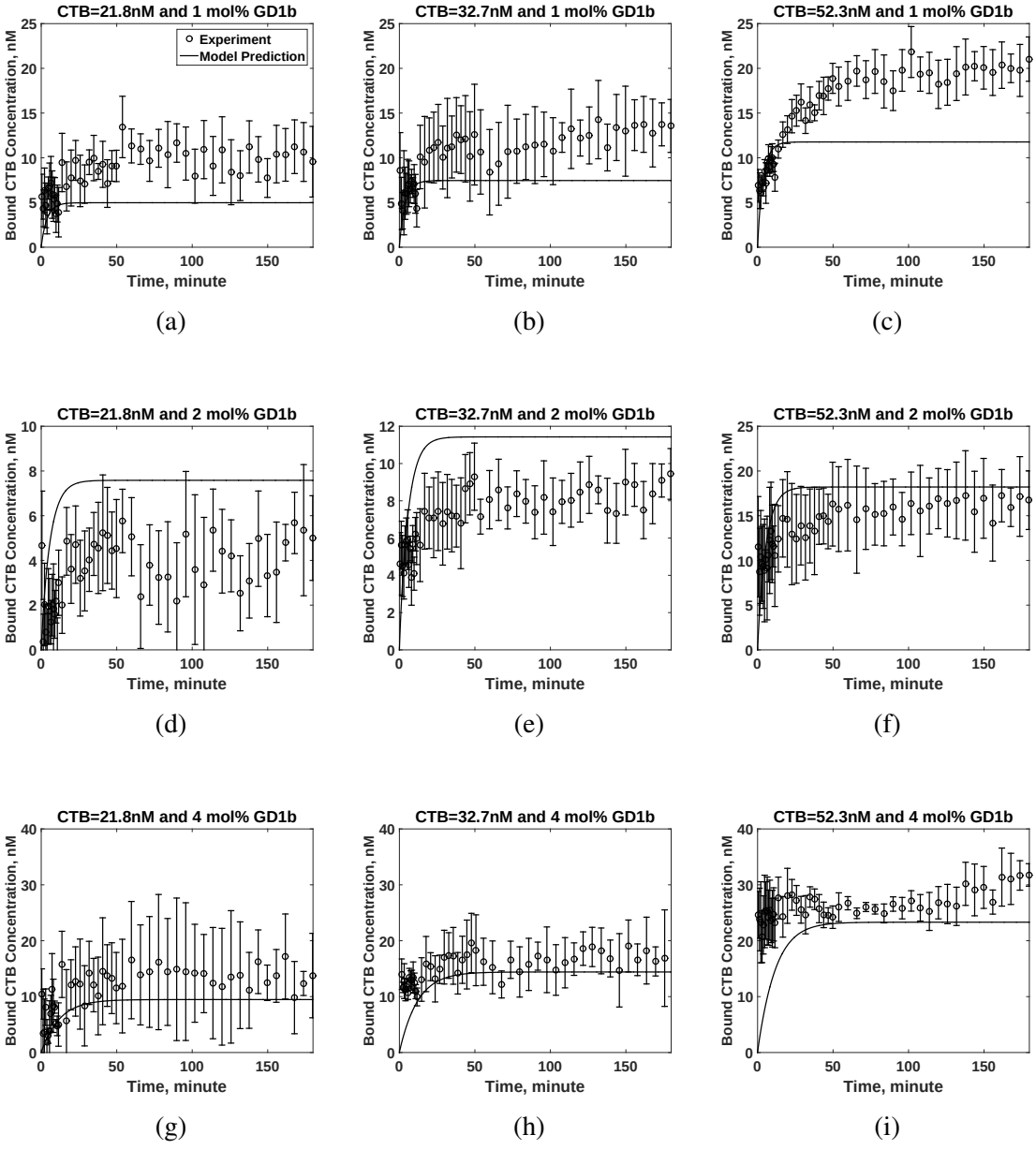


Figure 3.5: The results of the parameter estimation. Model predictions were compared with the experimental datasets used in the parameter estimation. Experimental data are given as means $\pm$ SEM.

The experimental and the predicted kinetics of the CTB-GD1b binding were shown in Figure 3.5. Comparing the simulated profiles with the experimental data used for model fitting, the proposed model agreed reasonably well with the experimental observations except CTB binding to 1 mol% of GD1b at 50nM CTB concentration. To further validate the proposed model, we compared the model predictions with the experimental measurements that were not used in the parameter estimation (Figure 3.6). The results demonstrated our parameter selection and estimation procedures could improve the predictive capability of the proposed dynamic model without overfitting.

The same parameter selection and estimation procedure was implemented to calibrate the model proposed by Lauer et al. [11]. The k_{10} and k_{-1} were determined as identifiable parameters, so they were estimated using the same experimental measurements. Then, we compared the normalized root-mean-square error (RMSE) of the two models. The RMSE values were 150 and 76 for the model proposed by Lauer et al. [11] and for the one proposed in this work, respectively, which demonstrated the proposed model was better than the one proposed by Lauer et al. [11].

3.5 Discussion

CTB binding to membrane receptors can be significantly influenced by the cell membrane environment and cooperative actions between CTB and its receptors. It requires a comprehensive analysis to understand this dynamic binding process. This study developed a systematic tool to estimate the binding parameters used in the CTB-GD1b recognition. We successfully integrated the first-principle modeling, the biosensing platform, and the algorithm of parameter selection and estimation (Figure 3.2). Specifically, we established a multi-step dynamic model, which considered the multivalency of CTB, the cooperative binding effect, and the receptor aggregation process to simulate the CTB-GD1b binding on model membranes. To acquire more reliable binding parameters, the nanocube sensing platform was employed to measure the CTB binding kinetics in the cell-membrane-mimicking environment. The proposed binding model was fit to the experimental data by estimating the identifiable parameters. This systematic approach could ensure that the values of estimated parameters were accurate and robust.

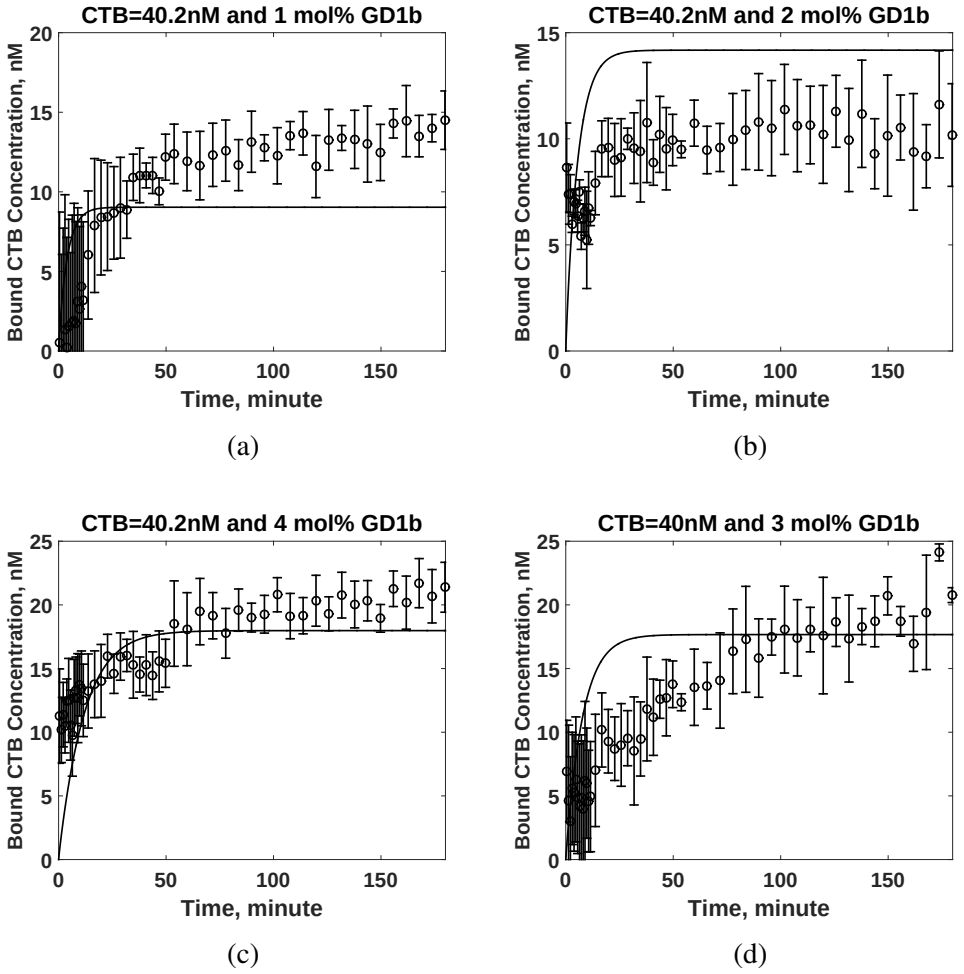


Figure 3.6: Validation of the parameter estimation results. Datasets obtained under 40.2 nM CTB concentration were used to validate the predictive capability of the proposed model since these datasets were not used in the parameter estimation. Experimental data are given as means $\pm$ SEM (standard error of means).

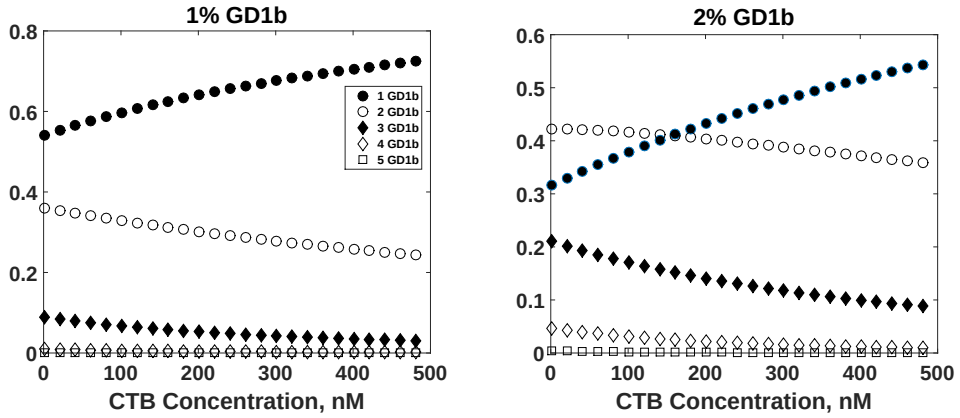
The values of model parameters were estimated by the model fitting and then validated by conducting a new set of experiments (Figures 3.5 and 3.6, respectively). While the model predictions agreed well with the experimental measurements obtained under 4% GD1b surface density, there was a small degree of discrepancy between the model predictions and the experimental data under 1% and 2% GD1b surface densities (Figure 3.5). It is probably due to the higher experimental errors at these experimental conditions. It is worth noting that the error bar in Figure 3.5 represents the standard error of the measurements, and the model predictions still fall within the

95% confidence intervals of the experimental measurements. To further validate the reliability of the fitted parameters, additional experiments were conducted (Figure 3.6). The model predictions agree well with the new set of experimental measurement. This result demonstrates that our modeling procedures could improve the predictive capability even though the experimental errors were considerable.

A distinct advantage of the mathematical model is allowing us to estimate the average valency of membrane-bound CTB proteins (i.e. the average number of receptors binding to CTB proteins). Although a single ganglioside binding to CTx is sufficient for intoxication, prior studies have shown multivalent binding can significantly increase CTx toxicity [173, 174]. In order to describe the toxicity, it is essential to estimate the average valency of membrane-bound CTB proteins. It is difficult to experimentally quantify the valency; hence, mathematical modeling becomes an attractive tool.

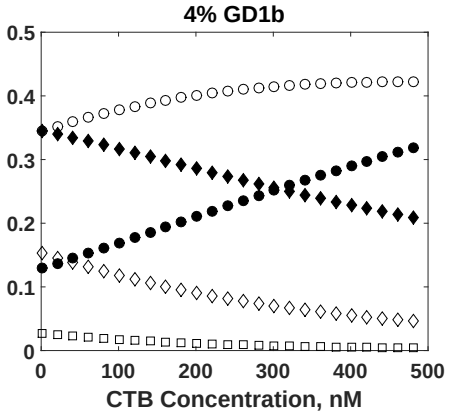
We calculated the average valency of membrane-bound CTB proteins at various GD1b surface densities and CTB concentrations using the resultant model (Figure 3.7). The percentage of CTB proteins bound to only one GD1b increased with CTB concentrations as more CTB proteins competed to bind with relatively scarce GD1b receptors. At the same time, the percentage of membrane-bound CTB proteins that had more than two bound GD1b increased when the CTB concentration decreased. Nevertheless, the percentage of the membrane-bound CTB proteins with more than four bound GD1b receptors remained low at the conditions we tested. Considering the fact that physiological concentrations of ganglioside and CTB are low [153, 175], we expect most of membrane-bound CTB proteins receive two GD1b receptors in a homogeneous membrane containing only GD1b receptors. Although CTx binding to membrane is not the only mechanism involving in the pathogenesis, this study offers a mathematical approach to assist the biologist in estimating the influences of membrane receptors in toxin entry.

It should be noted that we reported the average and standard errors of individual experimental results conducted on different days (i.e., inter-day validation), which was one of the factors causing the data fluctuations shown in Figures 3.5-3.6. Furthermore, the low GD1b surface densities



(a)

(b)



(c)

Figure 3.7: Predicted changes in the average valency of CTB proteins under various conditions by the proposed model. The amounts of membrane-bound CTB proteins with the different number of bound GD1b receptors under different CTB and GD1b concentrations were computed and plotted against the CTB concentration after being normalized by the total number of membrane-bound CTB proteins after 3 hours.

considered in this work rendered the CTB-GD1b binding process to be stochastic as the number of GD1b receptors on a model membrane would be far less than Avogadro's number, which would inevitably introduce some fluctuations in the experimental measurements. Our previous study [7] has shown that the fluctuations in the measurements were largely due to the process itself. Specifically, when binding kinetics of fluorescence-labeled CTB binding to GM1 receptors were measured by the nanocube-sensor and fluorescence correlation spectroscopy [176], a similar degree of fluctuation was observed through two different methods [7]. To partially handle these

undesired uncertainties in the measurement data, the objective function of the parameter estimation (Equation 3.11a) was normalized by the standard errors so that the measurement data with larger uncertainties would be less important during the parameter estimation process. Additionally, we have developed in kMC model for the CTB-GD1b binding kinetics in order to analyze to what extent the measurement uncertainties are originated from the stochasticity present in the CTB-GD1b binding kinetics, which will be beneficial for the design of future experiments to minimize the measurement uncertainties (see Chapter 7).

3.6 Conclusion

Numerous pathogens and their virulence factors (e.g. viruses, bacteria, and toxins) recognize the host cell membrane receptors via multivalent bindings. So far, most of biological studies still rely on the classic monovalent binding model to describe this essential process. In this work, we developed a systematic approach to model the dynamic process of the multi-step CTB-GD1b binding. We introduced several key features, including the stepwise reaction, the cooperative binding effect, and the membrane receptor aggregation, into the CTB-binding model. The parameter estimation and the parameter selection strategies were implemented for the model fitting. In addition, the nanocube-based biosensor was employed to measure the CTB-GD1b binding kinetics. The identifiable parameters were estimated by minimizing the difference between the experimental measurements and the model predictions. The calibrated model was able to make an accurate and robust prediction of the membrane-bound CTB concentration. Because many recent biological studies suggested that multivalency is a fundamental principle in lectin-glycan recognition [177, 178], there is an urgent need of a reliable multivalent binding analysis. Our study shows the potential of mathematical modeling in the analysis of complex multivalent binding on cell membrane surfaces. The future work involves analyzing a more complicated hetero-multivalent binding process where CTB proteins can bind to different types of membrane receptors simultaneously.

4. IDENTIFICATION OF A TIME-VARYING INTRACELLULAR SIGNALING MODEL THROUGH DATA CLUSTERING AND PARAMETER SELECTION: APPLICATION TO NFκB SIGNALING PATHWAY INDUCED BY LPS IN THE PRESENCE OF BFA *

This chapter proposes a method to construct a time-varying model for a partially known system. Specifically, when an accurate model for describing a signaling pathway under one stimulus is available, we can modify that model to describe the same signaling pathway under a lesser-known stimulus [179, 180]. Hereafter, we refer to the model constructed for the well-studied stimulus as the nominal model. The rationale for using the nominal model is two-fold. First, the nominal model already contains a number of important pathway components as well as their interactions, which are likely to be important under the lesser-known stimuli as well. Second, this approach avoids the lengthy model selection procedure, which requires a number of different candidate models to be synthesized, calibrated, and compared [179]. On the other hand, the structure of the nominal model is likely to be insufficient to describe the signaling dynamics under the lesser-known stimulus due to unincorporated and unknown reactions and components specific to this stimulus [179]. Therefore, a poorly characterized signaling pathway induced by a lesser-known stimulus needs to be described by a data-driven approach to complement the inaccuracy of the nominal model. Here, we choose to introduce time-varying parameters to the nominal model, which is usually time-invariant, based on the available experimental measurements [180, 181]. Through this data-driven approach, a more accurate model for the lesser-known stimulus can be derived based on the nominal model and the available data.

Motivated by the above considerations, we propose a numerical scheme to construct a time-varying model to simulate an intracellular signaling pathway with a lesser-known stimulus based on a nominal model. First, global sensitivity analysis is performed on the nominal model to identify

*Reprinted with the permission from "Identification of a time-varying intracellular signalling model through data clustering and parameter selection: application to NF-κB signalling pathway induced by LPS in the presence of BFA," Lee et al., 2019, *IET Systems Biology*, 13, 169-179. Copyright 2019 by Dongheon Lee, Arul Jayaraman and Joseph S. Kwon distributed under Creative Commons Attribution-NonCommercial-NoDerivs License

a set of parameters that are identifiable given the model structure and experimental observations, and only these parameters are assumed to vary with time. Next, the temporal profiles of the model parameters are partitioned into several temporal subdomains whose boundaries are determined by clustering the experimental observations. And the parameters determined by the sensitivity analysis have fixed values in each temporal subdomain. Finally, a least-squares problem is solved to estimate the values of the parameters in each temporal subdomain by minimizing the difference between the model predictions and the experimental data.

The chapter is organized as follows: first, the motivation for formulating a time-varying signaling model is presented. Next, the proposed methodology that consists of the optimal temporal clustering and the global sensitivity analysis to construct a time-varying model is presented in details. Lastly, the proposed methodology is implemented to develop a time-varying model for the NFκB signaling pathway induced by LPS in the presence of BFA to assess the efficiency and accuracy of the proposed scheme.

4.1 Background

4.1.1 System Description

Consider an intracellular signaling pathway initiated by an external stimulus, u, which has been well characterized by the following model:

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{f}\left(\mathbf{x}, \boldsymbol{\theta}, u; t\right); \quad \mathbf{x}(0) = \mathbf{x}_0 \\ \mathbf{y} &= \mathbf{g}\left(\mathbf{x}, \boldsymbol{\theta}, u; t\right) \end{aligned} \tag{4.1}$$

where $\mathbf{x} \in \mathbb{R}^{n_x}$ is the state vector, $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$ is the parameter vector, $\mathbf{x}_0$ is the initial value of the state vector $\mathbf{x}$, and $\mathbf{y} \in \mathbb{R}^{n_y}$ is the output vector.

When a lesser-known stimulus, u_a, is added to a cell, the signaling dynamics deviate significantly from those predicted by Equation 4.1. Due to the disparity in our understanding of the roles of u_a, u, and their interplay in the signaling dynamics, little information on the signaling dynamics is available *a priori*. Consequently, the construction of a high-fidelity model, which faithfully sim-

ulates the signaling dynamics initiated by u_a and u, requires iterative experimentation and model refinement, which can be an arduous and lengthy process [8, 27, 182].

A more viable alternative is to approximate the dynamics induced by u_a through introducing a time-varying model where $\boldsymbol{\theta}$ in Equation 4.1 changes with time so that the well-defined model (Equation 4.1) can be used to describe the signaling dynamics under the two stimuli [181]. To this end, the temporal profile of $\boldsymbol{\theta}$ is described as piece-wise constant functions. Under this representation, the entire temporal domain is divided into several temporal subdomains, each of which has its own parameter values. Consequently, the following modified form of Equation 4.1 is used to describe the signaling dynamics under two stimuli:

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{f}\left(\mathbf{x}, \boldsymbol{\theta}_{\sigma(t)}, u; t\right); \quad \mathbf{x}(0) = \mathbf{x}_0 \\ \mathbf{y} &= \mathbf{g}\left(\mathbf{x}, \boldsymbol{\theta}_{\sigma(t)}, u; t\right) \\ \sigma(t) &= i \quad \text{if } t \in \mathbb{T}_i, \quad i = \{1, \ldots, n_\sigma\} \end{aligned} \tag{4.2}$$

where $\boldsymbol{\theta}_i \in \mathbb{R}^{n_\theta}$, where $i = 1, \ldots, n_\sigma$, is the vector of parameter values used when time t belongs to temporal subdomain, $\mathbb{T}_i$, n_σ is the number of temporal subdomains, and $\sigma(t)$ is the discrete variable to denote which $\boldsymbol{\theta}_i$ is used at time t.

Under this formulation, the overall temporal domain is partitioned into n_σ subdomains, where different values of $\boldsymbol{\theta}$ are used. From here on, u_a is neglected since the use of $\boldsymbol{\theta}_{\sigma(t)}$ implies the presence of u_a.

4.1.2 Experimental Measurements

In order to train and validate Equation 4.2, $\mathbf{y}$ is measured experimentally under different conditions. Here, n_u different values of u with a fixed value of u_a were used. Due to the technical and economical constraints in a biological experiment, $\mathbf{y}$ can be measured at only a few sampling time instants, t_l, $l = 1, \ldots, N_t$, where N_t is the number of sampling instants [183]. Also, it should be noted that commonly used biochemical assays such as Western blots, flow cytometry, or microarrays typically give qualitative or semi-quantitative datasets, which measure relative but not

absolute concentrations of biomolecules [42]. In other words, the measured output is defined as:

$$z_i^s(t_l) = c_i \cdot \hat{y}_i(u^s; t_l) + \nu_i \tag{4.3}$$

where $z_i^s,\ i = 1, \ldots, n_y$, is the relative measured output that is corrupted with measurement noise under the input $u^s,\ s = 1, \ldots, n_u$, $\hat{y}_i$ is the output in absolute concentration that is not directly measurable in the experiments, c_i is the proportional constant relating z_i^s and $\hat{y}_i$, and ν_i is measurement noise. Here, it is assumed that the mean value of ν_i can be inferred during the equipment calibration procedure [184].

Since the values of $\mathbf{c}$ are not usually known beforehand, an alternative quantity is computed to facilitate the comparison between the model and the experimental measurements. Specifically, fold changes of the measurements are calculated as follows [8]:

$$\bar{y}_i^s(t_l) = \frac{z_i^s(t_l) - \bar{\nu}_i}{z_i^s(t_1) - \bar{\nu}_i} = \frac{\hat{y}_i(u^s; t_l)}{\hat{y}_i(u^s; t_1)} \tag{4.4}$$

where $\bar{y}_i^s(t_l)$ is the fold change of z_i^s at t_l, t_1 is the first sampling instant (usually $t_1 = 0$), and $\bar{\nu}_i$ is the average measurement noise that can be obtained by performing a negative control measurement without reagents.

4.1.3 Problem Statement

In this study, we seek to construct a time-varying model (Equation 4.2) by estimating the temporal dynamics of $\boldsymbol{\theta}$, and this can be achieved by addressing the following two problems:

Problem. *Given the model (Equation 4.2) and the experimental measurements (Equations 4.3-4.4), determine the number of temporal subdomains, n_σ, as well as the temporal subdomains, $\mathbb{T}_i,\ \forall i = 1, \ldots, n_\sigma$.*

Since n_σ and $\mathbb{T}_i$ are not known *a priori*, the experimental measurements are clustered to estimate the value of n_σ and the temporal subdomains, $\mathbb{T}_i$.

Problem. *Given the model (Equation 4.2), the experimental measurements (Equation 4.3), and the*

temporal subdomains of the parameters ($\mathbb{T}_1, \ldots, \mathbb{T}_{n_\sigma}$), estimate the values of parameters, $\boldsymbol{\theta}_i$, $i = 1, \ldots, n_\sigma$, in each temporal subdomain.

For many intracellular signaling pathways, only a small subset of $\boldsymbol{\theta}$ is identifiable from the experimental measurements [47, 183]. As a result, the additional parameters introduced in the time-varying model (Equation 4.2), which increases the size of the parameter space by n_σ-fold, are likely to be even more unidentifiable. Hence, a sequential parameter selection methodology is implemented to identify the most important parameters in $\boldsymbol{\theta}$, and the values of these parameters in each temporal subdomain are estimated. The resultant model then can be used to investigate the system dynamics and design the optimal experiments for future studies to advance our understanding of systems.

4.2 Temporal Clustering

Since the intracellular signaling dynamics are described by the time-varying model Equation (4.2) with the piecewise constant $\boldsymbol{\theta}$, the value of n_σ and all the temporal subdomains, $\mathbb{T}_i$, need to be determined. In this work, they are inferred by clustering the experimental measurements into several temporal subdomains in a way that the data points contained in each subdomain exhibit similar temporal behaviors [185]. This inference assumes that the time-invariant parameters in one temporal subdomain, $\mathbb{T}_i$, result in the relatively uniform dynamics in $\mathbf{y}$.

For the given experimental measurements $\mathbf{D} \in \mathbb{R}^{N_c \times N_t}$, where $N_c = n_u \cdot n_y$,

$$\mathbf{D} = \begin{bmatrix} \bar{y}_1^1(t_1) & \cdots & \bar{y}_1^1(t_{N_t}) \\ \vdots & \ddots & \vdots \\ \bar{y}_1^{n_u}(t_1) & \cdots & \bar{y}_1^{n_u}(t_{N_t}) \\ \bar{y}_2^1(t_1) & \cdots & \bar{y}_2^1(t_{N_t}) \\ \vdots & \ddots & \vdots \\ \bar{y}_2^{n_u}(t_1) & \cdots & \bar{y}_2^{n_u}(t_{N_t}) \\ \vdots & \ddots & \vdots \\ \bar{y}_{n_y}^{n_u}(t_1) & \cdots & \bar{y}_{n_y}^{n_u}(t_{N_t}) \end{bmatrix} \tag{4.5}$$

a clustering algorithm will assign N_t column vectors of $\mathbf{D}$ into n_σ different temporal subdomains by minimizing the distance between vectors in a subdomain and the center of the subdomain, which is measured by the following intracluster error sum [185]:

$$\Lambda = \sum_{i=1}^{N_t} \sum_{k=1}^{n_\sigma} z_{ik} \left\| \mathbf{D}_i - \mathbf{c}_k \right\|_2^2 \tag{4.6}$$

where $\mathbf{D}_i$ is the i^{th} column of $\mathbf{D}$, z_{ik} is a binary variable indicating whether $\mathbf{D}_i$ is in the k^{th} subdomain, and $\mathbf{c}_k \in \mathbb{R}^{N_c}$ is the center of the k^{th} cluster.

Since a value of n_σ is not known *a priori*, a clustering method is implemented with all possible number of subdomains (1,. . .,N_t) to find an optimal n_σ by computing and comparing the values of Λ as well as the inter-cluster error sum, Γ, which is defined as follows [185]:

$$\Gamma = \sum_{k=1}^{n_\sigma} \left\| \mathbf{c}^\circ - \mathbf{c}_k \right\|_2^2 \tag{4.7}$$

where $\mathbf{c}^\circ \in \mathbb{R}^{N_c}$ is the global cluster center, which is defined as

$$c_j^\circ = \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbf{D}_{ji} \tag{4.8}$$

where c_j° is the j^{th} element of $\mathbf{c}^\circ$. When an optimal clustering configuration is achieved, the value of Λ is minimized while the value of Γ is maximized to achieve the maximum intracluster similarity and intercluster dissimilarity [185, 186]. Mathematically, this is quantified by the clustering balance, ϵ, which was proposed in [186], as follows:

$$\epsilon = 0.5\Gamma + 0.5\Lambda \tag{4.9}$$

where 0.5 in front of Γ and Λ is a weight coefficient, which can be adjusted based on the problem [187]. As Λ and Γ are expected to decrease and increase, respectively, with the number of subdomains, a turning point in the value of ϵ determines the optimal value of n_σ [185]. Once the value of

n_σ is determined, all the $\mathbb{T}_i$ can also be determined by clustering $\mathbf{D}$ into n_σ temporal subdomains.

4.3 Parameter Estimation

The aim of the parameter estimation is to quantitatively calibrate a model so that it can make an accurate and robust prediction of the system, which then can be used to analyze underlying mechanisms and design optimal experiments [105, 182]. We can formulate a parameter estimation for Equation 4.2 as a least-squares problem to minimize the difference between model predictions and measurements as follows:

$$\min_{\mathbf{c},\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_{n_\sigma}} \quad \sum_{s=1}^{n_u}\sum_{i=1}^{n_y}\sum_{l=1}^{N_t}\left(y_i(u^s;t_l)-\bar{y}_i^s(t_l)\right)^2 \tag{4.10a}$$

$$\text{s.t.} \quad \dot{\mathbf{x}} = \mathbf{f}\left(\mathbf{x}, \boldsymbol{\theta}_{\sigma(t_l)}, u^s; t_l\right); \quad \mathbf{x}(0) = \mathbf{x}_0 \tag{4.10b}$$

$$\mathbf{y} = \mathbf{g}\left(\mathbf{x}, \boldsymbol{\theta}_{\sigma(t_l)}, u^s; t_l\right) \tag{4.10c}$$

$$\sigma(t_l) = k \quad \text{if } t_l \in \mathbb{T}_k; \;\; k = \{1,\ldots,n_\sigma\} \tag{4.10d}$$

$$\boldsymbol{\theta}^{lb} \leq \boldsymbol{\theta}_k \leq \boldsymbol{\theta}^{ub} \tag{4.10e}$$

where $\boldsymbol{\theta}^{lb}$ and $\boldsymbol{\theta}^{ub}$ are lower and upper bounds for the values of the model parameters, respectively.

It should be noted that the parameter estimation (Equation 4.10) is often ill-conditioned and results in a non-unique solution [23]. This is especially problematic for calibrating biological models since biological systems are often partially observable and over-parameterized (i.e., $n_y \ll n_\theta$) [47]. As the time-dependency of the model parameters are introduced, the issue of the non-uniqueness in the parameter estimation exacerbates since the number of parameters increases by n_σ-fold. In order to handle this issue, we assume that only identifiable parameters, which is a subset of $\boldsymbol{\theta}$, vary with time while the remaining parameters are time-invariant and fixed at their nominal values. Consequently, this study carries out the parameter selection methodology before the parameter estimation to determine the identifiable parameters and estimate their values in each $\mathbb{T}_i$ by solving the least-squares problem.

4.3.1 Parameter Selection

The objective of the parameter selection procedure is to determine the identifiable parameters that will be estimated in the parameter estimation step. In this study, two global sensitivity analysis techniques are implemented to determine which parameters are identifiable.

4.3.1.1 Sensitivity Analysis

In the literature, several analytical methods have been proposed to determine parameter identifiability, including Taylor series expansion [188], differential algebra [189], or similarity transformation [190]. But these methods require symbolic manipulation and thus only applicable to a relatively small system (for $n_\theta + n_y \leq 10$) due to the computational requirements of these methods [13].

Alternatively, the parameter identifiability can be assessed by sensitivity analysis (SA), which evaluates the importance of the model parameters by quantifying changes in model outputs due to changes in model parameters. A common method is the local SA method that is based on the direct differentiation of a system model with respect to its parameters. However, the evaluation of the system model as well as its derivatives with respect to its parameters depends on the values of the model parameters, which are unknown before the parameter estimation. Therefore, a result of the local SA method is local in nature and likely to be unreliable, particularly when the parameter values are largely uncertain [156, 191].

In this study, two global SA methods, Morris method [192] and Sobol' method [193], are implemented sequentially to determine the most important parameters. Even though the parameters take different values in each temporal subdomain, the model structure remains the same. Hence, the results of global sensitivity analysis on the time-invariant model will be valid for the time-varying one because the global sensitivity analysis computes the importance of model parameters over the entire parametric domain. Therefore, all the analysis in the following sections is conducted based on the time-invariant model (Equation 4.1).

4.3.1.2 *Morris Method*

The Morris method computes the average sensitivity of a model parameter by calculating the average change in model outputs due to changes in its value. Specifically, the value of a parameter $\theta_j \in \boldsymbol{\theta},\ j = 1, \ldots, n_\theta$, is perturbed by Δ_j to compute its effect on an output y_i, which is quantified as follows [102]:

$$d_{ij}(u^s; t_l) = \frac{y_i(\mathbf{x}, \theta_1, \ldots, \theta_j + \Delta_j, \ldots, \theta_{n_\theta}, u^s; t_l) - y_i(\mathbf{x}, \boldsymbol{\theta}, u^s; t_l)}{\Delta_j} \tag{4.11}$$

where $d_{ij}(u^s; t_l)$ is called the elementary effect of θ_j on y_i at time $t_l,\ l = 1, \ldots, N_t$. By calculating N_m different d_{ij} with N_m different values of Δ_j, the average sensitivity measure of the parameter θ_j, which is denoted as s_{ij}, is computed as follows:

$$\begin{aligned} s_{ij}(u^s; t_l) &= \frac{1}{N_m} \frac{\theta_j}{y_i(\mathbf{x}, \boldsymbol{\theta}, u^s; t_l)} \sum_{k=1}^{N_m} \left| d_{ij}^{(k)}(u^s; t_l) \right| \\ &= \frac{1}{N_m} \frac{\theta_j}{y_i(\mathbf{x}, \boldsymbol{\theta}, u^s; t_l)} \times \sum_{k=1}^{N_m} \left| \frac{y_i(\mathbf{x}, \theta_1, \ldots, \theta_j + \Delta_j^{(k)}, \ldots, \theta_{n_\theta}, u^s; t_l) - y_i(\mathbf{x}, \boldsymbol{\theta}, u^s; t_l)}{\Delta_j^{(k)}} \right| \end{aligned} \tag{4.12}$$

where $\Delta_j^{(k)}$ is the k^{th} perturbation applied to parameter θ_j. Here, $d_{ij}(u^s; t_l)$ is normalized by $\theta_j / y_i(\mathbf{x}, \boldsymbol{\theta}, u^s; t_l)$ to eliminate possible scaling effects [102]. And, the suggested value for N_m is $r(n_\theta + 1)$, where r is usually around six [194].

Then, the final scaled sensitivity of all the model outputs with respect to a parameter across all the time instants is defined as follows:

$$S_j = \frac{1}{n_y} \sum_{i=1}^{n_y} \left\| \begin{bmatrix} s_{ij}(u^1; t_1) & \cdots & s_{ij}(u^1; t_{N_t}) & \cdots & s_{ij}(u^{n_u}; t_{N_t}) \end{bmatrix} \right\|_2 \tag{4.13}$$

where $s_{ij}(u^s; t_l)$ is the average sensitivity computed under input $u_s,\ s = 1, \ldots, n_u$, at time t_l.

Although the Morris method is conceptually simple and easy to be implemented, it has a limited capability in capturing the nonlinear output behavior and the dependency among parameters [102]. Therefore, this study utilizes the Morris method as a screening tool to reduce the number

of parameters to be analyzed by the Sobol' method, which overcomes the problems of the Morris method but is computationally more expensive.

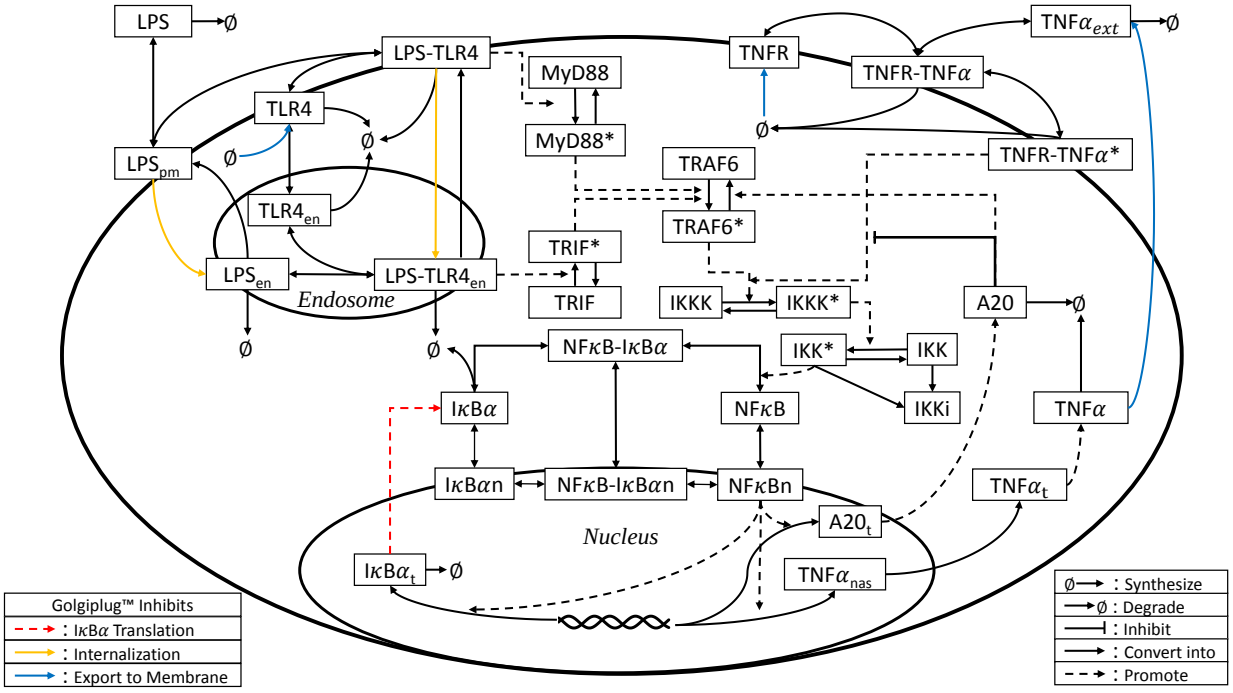


Figure 4.1: Schematic diagram for the LPS-induced NFκB signaling pathway (adapted from [8]). Due to space limitation, TRIF-dependent regulation of TNFα production [2] and the feedback regulation between NFκB and IκB-β and -ϵ are not illustrated. Also, the NFκB activation induced by TNFα-TNFR is not shown in details due to the limited space (see [9] for details). Colored arrows indicate the processes affected by the addition of BFA (see text for details).

4.3.1.3 Sobol' Method

Once the Morris method screens out less important parameters from $\boldsymbol{\theta}$, the importance of the remaining parameters, which are denoted as $\bar{\boldsymbol{\theta}} \in \mathbb{R}^{n_p}$, $n_p < n_\theta$, is analyzed via the Sobol' method. Different from the local SA method or the Morris method, the Sobol' method is a variance-based method. Specifically, the sensitivity of a parameter is computed by quantifying how much each parameter contributes to the output variance. A brief overview on the Sobol' SA method is presented below, and further details can be found in [193, 194, 195].

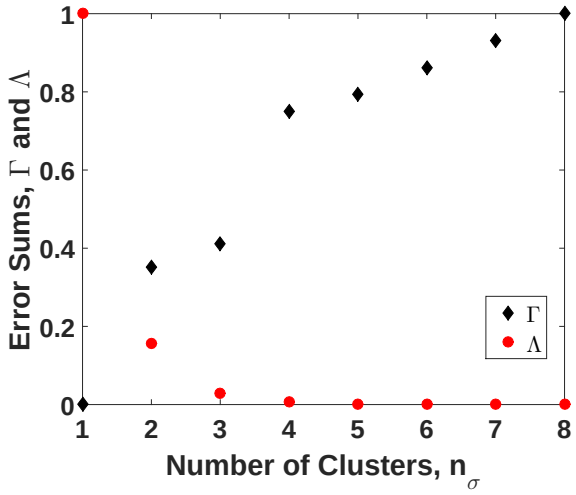


Figure 4.2: Values of the intracluster error sum (Λ) and intercluster error sum (Γ) with the number of subdomains (n_σ). Each error sum is normalized by its maximum value.

The main idea of the Sobol' method is the decomposition of the model output into summands of increasing dimensionality. Specifically, a model output y can be decomposed as follows:

$$y(\bar{\boldsymbol{\theta}}) = y_0 + \sum_{i=1}^{n_p} y_i(\bar{\theta}_i) + \sum_{i=1}^{n_p-1}\sum_{j>i}^{n_p} y_{i,j}(\bar{\theta}_i,\bar{\theta}_j) + \sum_{i=1}^{n_p-s+1} \cdots \sum_{j>i+s-2}^{n_p} y_{i,\ldots,j}\left(\bar{\theta}_i,\ldots,\bar{\theta}_j\right) + y_{1,\ldots,n_p}(\bar{\boldsymbol{\theta}}) \tag{4.14}$$

where s, where $3 \leq s \leq n_p$, is the number of parameters involved in a summand $y_{i,\ldots,j}\left(\bar{\theta}_i,\ldots,\bar{\theta}_j\right)$, and y_0 is defined as follows:

$$y_0 = \int y(\bar{\boldsymbol{\theta}}) d\bar{\boldsymbol{\theta}} \tag{4.15}$$

Here, we assume y_0 is a constant and the integrals of every summand over any of its variables are zero, i.e.,

$$\int y_{i_1,\ldots,i_r}(\bar{\theta}_{i_1},\ldots,\bar{\theta}_{i_r}) d\bar{\theta}_k = 0 \forall k = i_1,\ldots,i_r,\ 1 \leq i_1 < \cdots < i_r \leq n_p \tag{4.16}$$

in order for the decomposition of y as above (Equation 4.14) to hold [183, 193].

Then, by assuming that y is square integrable, its variance can be expressed as follows:

$$\begin{aligned} V &= \int \left(y^2(\bar{\boldsymbol{\theta}}) - y_0^2\right) d\bar{\boldsymbol{\theta}} = \int \left(\sum_{i=1}^{n_p} y_i^2(\bar{\theta}_i) + \sum_{i=1}^{n_p-1} \sum_{j>i}^{n_p} y_{i,j}^2(\bar{\theta}_i, \bar{\theta}_j) + \ldots + y_{1,\ldots,n_p}^2(\bar{\boldsymbol{\theta}}) \right) d\bar{\boldsymbol{\theta}} \\ &= \sum_{i=1}^{n_p} V_i + \sum_{i=1}^{n_p-1} \sum_{j>i}^{n_p} V_{i,j} + \sum_{i=1}^{n_p-s+1} \cdots \sum_{j>i+s-2}^{n_p} V_{i,\ldots,j} + V_{1,\ldots,n_p} \end{aligned} \tag{4.17}$$

where V is the total variance of a model output y, and $V_{i,\ldots,j}$ is the partial variance of the output due to the parameters $\bar{\theta}_i, \ldots, \bar{\theta}_j$.

Based on V and $V_{i,\ldots,j}$, the importance of a parameter, $\bar{\theta}_j$, can be quantified by the first order and total sensitivity indices, which are defined as follows [196]:

$$\begin{aligned} S_{S_j} &= \frac{V_j}{V} \\ S_{T_j} &= \frac{1}{V}\left(V_j + \sum_{k\neq j}^{n_p} V_{j,k} + \ldots + V_{1,\ldots,n_p} \right) = \frac{1}{V}(V - V_{\sim j}) \end{aligned} \tag{4.18}$$

where S_{S_j} and S_{T_j} are the first-order and total sensitivity indices, respectively, of the model parameter $\bar{\theta}_j$, V_j is the partial variance of a model output due to $\bar{\theta}_j$, and $V_{\sim j}$ is the partial variance of a model output due to joint effects of the model parameters $\bar{\boldsymbol{\theta}}$ except $\bar{\theta}_j$. Here, S_{S_j} refers to the main effect of the parameter $\bar{\theta}_j$, and S_{T_j} measures the importance of a parameter, $\bar{\theta}_j$, by taking into account the direct effect (S_{S_j}) as well as its joint effects with other parameters. It should be noted that the difference between S_{T_j} and S_{S_j} indicates how much $\bar{\theta}_j$ is involved in interactions with other parameters in terms of changing the model output [195].

In this study, a Monte Carlo method proposed by Homma and Saltelli [196] is implemented to estimate the total sensitivity indices. First of all, two matrices ($\mathbf{A}$ and $\mathbf{B} \in \mathbb{R}^{N_{int}\times n_p}$) are generated randomly from the parameter space via a Sobol' sequence to produce parameter samples without overlapping [193, 197]. Here, N_{int} is the sample size for the Monte Carlo estimation, which is typically around a few hundreds to thousands [194]. Then, another set of matrices $\mathbf{C}_j \in \mathbb{R}^{N_{int}\times n_p}$, $\forall j = 1, \ldots, n_p$, can be defined for every parameter, $\bar{\theta}_j$, by replacing the j-th

column of $\mathbf{B}$ with the j-th column of $\mathbf{A}$. Next, the model outputs can be computed for all the sampled parameter values in the matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}_j$. Finally, the first-order and total sensitivity indices in Equation 4.18 can be approximated as follows:

$$\begin{aligned} S_{S_j}(u^s;t_l) &\approx \frac{1/N_{int}\sum_{i=1}^{N_{int}} y(u^s,\mathbf{a}^{(i)};t_l)\cdot y(u^s,\mathbf{c}_j^{(i)};t_l) - f_0^2(u^s;t_l)}{1/N_{int}\sum_{i=1}^{N_{int}}(y(u^s,\mathbf{a}^{(i)};t_l)^2 - f_0^2(u^s,t_l)} \\ S_{T_j}(u^s;t_l) &\approx 1 - \frac{1/N_{int}\sum_{i=1}^{N_{int}} y(u^s,\mathbf{b}^{(i)};t_l)\cdot y(u^s,\mathbf{c}_j^{(i)};t_l) - f_0^2(u^s;t_l)}{1/N_{int}\sum_{i=1}^{N_{int}}(y(u^s,\mathbf{a}^{(i)};t_l)^2 - f_0^2(u^s,t_l)} \end{aligned} \tag{4.19}$$

where $\mathbf{a}^{(i)}$, $\mathbf{b}^{(i)}$, and $\mathbf{c}_j^{(i)}$ are i-th rows of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}_j$, respectively, and $f_0^2(u^s,t_l)$ is defined as follows:

$$f_0^2(u^s,t_l) = \left(\frac{1}{N_{int}}\sum_{i=1}^{N_{int}} y_i(u^s,\mathbf{a}^{(i)};t_l)\right)^2 \tag{4.20}$$

Since the proposed model (Equation 4.2) has n_y outputs obtained under n_u values for u sampled at N_t time instants, a lumped sensitivity metric for the total sensitivity index is defined to ease the parameter selection process as follows:

$$ST_{ij} = \frac{1}{N_c}\sum_{s=1}^{n_u}\sum_{l=1}^{N_t} S_{i,T_j}(u^s;t_l) \tag{4.21}$$

where ST_{ij} is the average total sensitivity of y_i with respect to $\bar{\theta}_j$, which will be used to select the most influential parameters, and S_{i,T_j} is the total sensitivity index, S_{T_j}, computed for an output, $y_i,\ i = 1,\ldots,n_y$. Similarly, a lumped sensitivity metric for the first-order sensitivity index is defined as follows:

$$SS_{ij} = \frac{1}{N_c}\sum_{s=1}^{n_u}\sum_{l=1}^{N_t} S_{i,S_j}(u^s;t_l) \tag{4.22}$$

where SS_{ij} is the average first-order sensitivity of y_i with respect to $\bar{\theta}_j$, which will be used to select the most influential parameters, and S_{i,S_j} is the first-order sensitivity index, S_{S_j}, computed for an output, $y_i,\ i = 1,\ldots,n_y$.

Based on the values of ST_{ij} and SS_{ij}, a set of identifiable parameters, $\boldsymbol{\Theta} \in \mathbb{R}^{n_{p_s}}$, where $n_{p_s} \leq n_p$, can be identified from $\bar{\boldsymbol{\theta}}$. And the final parameter estimation problem (Equation 4.10) is

reformulated as follows:

$$\min_{\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_{n_\sigma}} \quad \sum_{s=1}^{n_u}\sum_{i=1}^{n_y}\sum_{l=1}^{N_t}\left(y_i(u^s;t_l)-\bar{y}_i^s(t_l)\right)^2 \tag{4.23a}$$

$$\text{s.t.} \quad \dot{\mathbf{x}} = \mathbf{f}\left(\mathbf{x}, \boldsymbol{\Theta}_{\sigma(t_l)}, u^s; t_l\right); \quad \mathbf{x}(0) = \mathbf{x}_0 \tag{4.23b}$$

$$\mathbf{y} = \mathbf{g}\left(\mathbf{x}, \boldsymbol{\Theta}_{\sigma(t_l)}, u^s; t_l\right) \tag{4.23c}$$

$$\sigma(t_l) = k \quad \text{if } t_l \in \mathbb{T}_k; \;\; k = \{1,\ldots,n_\sigma\} \tag{4.23d}$$

$$\boldsymbol{\theta}^{lb} \leq \boldsymbol{\Theta}_{\sigma(t_l)} \leq \boldsymbol{\theta}^{ub} \tag{4.23e}$$

The computational time required for the sensitivity analysis depends on the sample size, the number of parameters, and the time for running a model. For the Morris and Sobol' methods, the numbers of simulations required to compute the sensitivity indices are $n_\theta \times N_m \times n_u$ and $(n_p + 2)\, N_{int} \times n_u$, respectively, which shows that the computational cost will increase linearly. Moreover, the computational cost of solving Equation 4.23 depends on the time required for running a model, the number of the model parameters, the number of the temporal subdomains, and the number of different initial guesses to solve Equation 4.23.

4.4 Application to NFκB Signaling

In this section, we applied the proposed methodology to model the NFκB signaling dynamics in RAW murine macrophages induced by LPS in the presence of BFA.

4.4.1 NFκB Signaling Pathway

NFκB is an important regulator of inflammation and immune responses in various immune cells such as macrophages [198]. Under homeostatic conditions, the activity of NFκB is minimal because it is sequestered by isomers of IκB (inhibitors of κB) proteins such as IκB-α, -β and -ϵ [198]. In the classical NFκB activation pathway, an external stimulus (e.g., LPS) activates IKK, which leads to degradation of IκB and thus activates NFκB [198]. Then, the derepressed NFκB protein translocates to the nucleus and upregulates the expression of various target genes such as

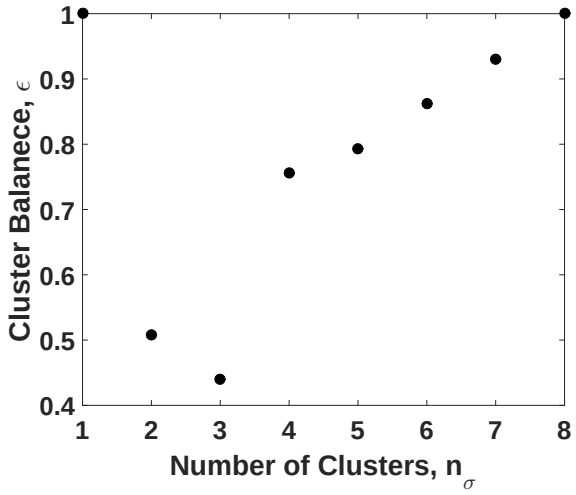


Figure 4.3: Change in the cluster balance value (ϵ) with the change in the number of subdomains (n_σ). The cluster balance value is normalized by its maximum value.

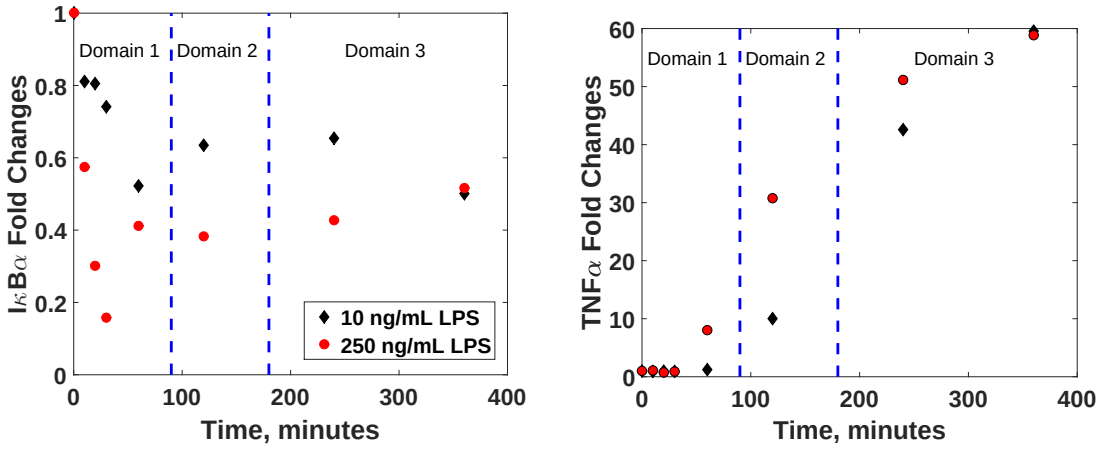
IκB, and pro-inflammatory cytokines such as TNFα, which propagates the inflammatory signals to adjacent cells and tissues [3, 199].

As a component in gram-negative bacteria's outer membranes, LPS is a potent activator of the NFκB signaling pathway in macrophages through Toll-like receptor 4 (TLR4) [200]. By forming a complex with LPS, TLR4 and its accessory molecules activate NFκB signaling through the classical activation pathway as described earlier. In contrast, BFA activates NFκB through an alternate signaling pathway [201, 202]. Since exposure to BFA leads to the Golgi apparatus fusing with the endoplasmic reticulum (ER), normal intracellular trafficking is disrupted, which leads to accumulation of proteins in ER. This, in turn, initiates the ER-stress pathway and leads to the activation of NFκB [8, 88, 201].

Although several mechanisms have been proposed to explain how NFκB activity is induced by the ER-stress pathway, mechanistic details have not been fully elucidated yet due to the complexity of the ER-stress signaling pathway [203, 204]. Furthermore, recent studies demonstrated that interactions between the ER-stress and NFκB signaling pathways are actually bidirectional, which further complicates the system analysis (see [202] and references therein). To unravel the complexity of the ER-stress signaling pathway, several computational models [203, 204, 205] have

been proposed; however, they have not been validated thoroughly under various physiological conditions, whereas the NFκB signaling pathway model has been continuously tested and improved since the early 2000s [75, 206, 207]. Furthermore, few studies have attempted to model the crosstalk between the ER-stress and NFκB signaling pathways. Consequently, this study chose to use the time-varying model to represent the LPS-induced NFκB signaling dynamics in the presence of BFA because the detailed model structure is still not known fully. The proposed model can be used to design future experiments that can help elucidate the underlying molecular interactions in future studies.

Motivated by the above considerations, we considered the LPS-induced NFκB signaling model as the well-characterized model (Equation 4.1) while the model for the NFκB signaling dynamics induced by LPS in the presence of BFA is considered as the unknown high-fidelity model, which would be approximated by the LPS-induced signaling model with time-varying parameters.



(a) (b)

Figure 4.4: Temporal subdomains of the measured (a) IκBα and (b) TNFα dynamics from [8]. The data points in black diamonds and red circles were measured under 10 and 250 ng/mL of LPS, respectively, in the presence of 1μg/mL of BFA, and three different temporal subdomains are separated by blue dash lines.

Rank	Parameter	Normalized S_j
1	Hill coefficient for TNFα transcription	1.00
2	Constant for TRIF*-induced TNFα production enhancement (K_{a0})	0.99
3	Constant for TRIF*-induced TNFα production enhancement (K_a)	0.29
4	TNFα protein synthesis rate constant	0.28
5	TNFα nascent mRNA processing rate constant	0.16
6	TNFα protein degradation rate constant	0.12
7	Maximum degradation rate constant for TNFα transcript	0.089
8	IKK*-mediated degradation rate constant for IκBα in NFκ-IκBα	0.081
9	IKKK*-mediated IKK activation rate constant	0.053
10	Constitutive IKKK activation rate constant	0.032
11	IκBα transcript degradation rate constant	0.032
12	IκBα translation rate constant	0.025
13	IκBα degradation rate in nucleus	0.023
14	IκBα degradation rate in cytoplasm	0.023
15	EC50 constant for TNFα transcription	0.022
16	Constitutive IκBα transcription rate constant	0.021
17	Hill coefficient for IκBα transcription	0.021
18	NFκB-induced TNFα transcription rate constant	0.015
19	Constitutive IKKK deactivation rate constant	0.012
20	Rate constant for IκBα and NFκB association in nucleus	0.011
21	Constitutive rate constant for IKK inactivation (IKK $\rightarrow$ IKKi)	0.011
22	Constitutive rate constant for IKK activation	0.010

Table 4.1: Result of Morris Sensitivity Analysis

4.4.2 Dynamic Model of LPS-induced NFκB Signaling

The schematic diagram for the NFκB signaling pathway and the TNFα production induced by LPS in the presence of BFA is shown in Figure 4.1. The starting point of the model is the LPS-induced NFκB signaling model developed by Hoffmann et al. [1, 2], where the LPS-NFκB signaling pathway model was adopted from Caldwell et al. [1], and a model describing the regulation of the TNFα production by internalized LPS-TLR4 complexes was adopted from Junkin et

al. [2]. Lee et al. [8] further updated the model by incorporating a new role for A20 protein as an inhibitor of LPS-induced signaling. Also, the well-known effect of BFA addition on the collapse of Golgi apparatus was taken into account by introducing time-dependent decays in rate constants associated with protein secretion and protein translocation to the membrane [8]. The model outputs are the dynamics of IκBα protein and intracellular TNFα protein (i.e., $n_y = 2$), and the updated model contains 49 states and 146 parameters (i.e., $n_x = 49$ & $n_\theta = 146$, and see [8] for the details on the model).

The datasets obtained through flow cytometry in our previous study [8] were used to perform the temporal clustering as well as the parameter estimation required in the proposed methodology. As discussed earlier, the datasets obtained through flow cytometry are relative data, which will not give the measurements in absolute concentrations, so the fold changes were computed based on Equation 4.4. Therefore, the model output functions $\mathbf{y} = \mathbf{g}\left(\mathbf{x}, \boldsymbol{\theta}_{\sigma(t)}, u; t\right)$ are also defined as the fold change of the two states with respect to their initial conditions as follows:

$$\begin{aligned} y_1(t_l) &= \frac{\text{I}\kappa\text{B}\alpha_{\text{total}}(t_l)}{\text{I}\kappa\text{B}\alpha_{\text{total}}(t_1)} \\ y_2(t_l) &= \frac{\text{TNF}\alpha(t_l)}{\text{TNF}\alpha(t_1)} \end{aligned} \tag{4.24}$$

$y_1(t_l)$ and $y_2(t_l)$ are the predicted fold changes of IκBα and intracellular TNFα concentrations, respectively, at time t_l, and TNF$\alpha(t_l)$ and IκB$\alpha_{\text{total}}(t_l)$ are the predicted IκBα and intracellular TNFα concentrations, respectively, by the model.

4.4.3 Temporal Clustering

In the flow cytometry experiments described in our previous study [8], the sampling time instants were 0, 10, 20, 30, 60, 120, 240, 360 minutes (i.e., $N_t = 8$) after LPS were added to the cell culture in the presence of BFA. Two LPS concentrations (10 and 250 ng/mL) and one concentration of BFA (1μg/mL) were used to obtain the experimental datasets (i.e., $n_u = 2$). Then, the temporal clustering methodology described in the preceding section was implemented to partition the measurement datasets to determine the optimal value of n_σ as well as the corresponding tem-

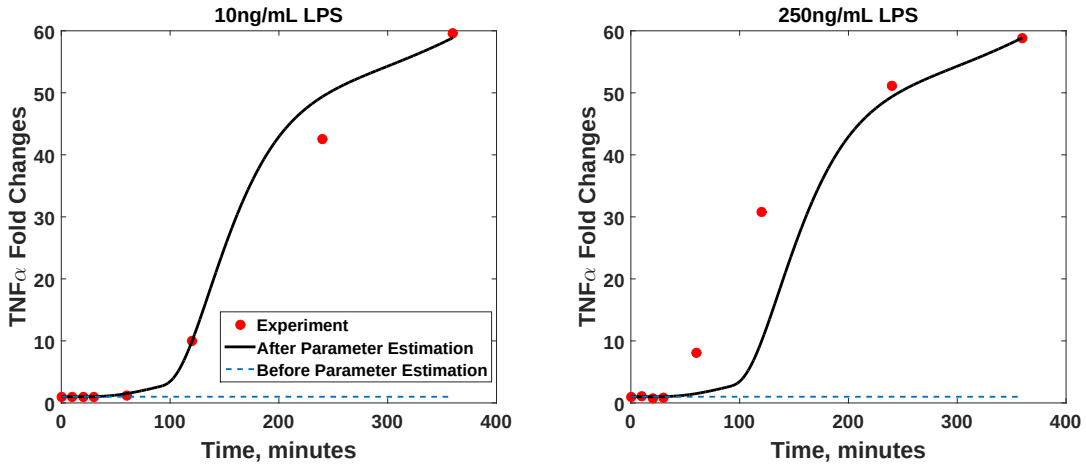
poral subdomains, $\mathbb{T}_i$, $\forall i = 1, \ldots, n_\sigma$. In this work, the k-means clustering algorithm was used via *kmeans* function available in MATLAB, and multiple initial conditions were used to initialize the k-means clustering for each number of subdomain.

Since the value of N_t is eight, the maximum number of possible subdomains is eight in this work. Figure 4.2 shows the changes in the intra- and inter-cluster error sums (Λ and Γ, respectively) with the number of subdomains. As expected, the value of Λ decreases with the number of subdomains while the value of Γ increases. Based on these two values, the cluster balance (ϵ) defined in Equation 4.9 can be computed for each number of subdomains and plotted in Figure 4.3. As described earlier, a turning point in Figure 4.3 is used to determine the optimal value of n_σ, which is found to be three.

Based on $n_\sigma = 3$, each temporal subdomain can be determined by clustering the experimental datasets into three temporal subdomains, which are shown in Figure 4.4. Specifically, the first, second and third subdomains contain the data points spanning from 0 to 60 minutes, 120 minutes, and 240 to 360 minutes, respectively. Each temporal subdomain can be interpreted to represent a different phase of the NFκB signaling pathway induced by LPS in the presence of BFA. The first subdomain shows the early phase of the NFκB signaling, where IκBα is quickly degraded while TNFα has not been synthesized. The second subdomain corresponds to the transition from the late phase of the LPS-induced NFκB signaling pathway, where the rate of TNFα synthesis accelerates and the IκBα is being re-synthesized, to the BFA-dominated signaling. The last subdomain can be seen as BFA-induced NFκB dynamics, where the IκBα concentration is sustained at a low level due to the inhibition of its translation by BFA [8].

4.4.4 Sensitivity Analysis Result

The Morris and Sobol' sensitivity methods were implemented as described above, and all the sensitivity computation was performed in parallel in ADA supercomputing cluster at Texas A&M University. The result of the sensitivity analysis via the Morris method is shown in Table 4.1. For each parameter, six different values were randomly sampled from its parameter domain ranging from 10% to 1000% of its nominal value, and the average sensitivity of each parameter with



(a) (b)

Figure 4.5: The result of parameter estimation. The predicted dynamics of TNFα before (dash line) and after (solid line) the parameter estimation were compared with the experimental observations under (a) 10 ng/mL and (b) 250 ng/mL of LPS in the presence of BFA.

respect to the two outputs was computed (Equation 4.13). Table 4.1 only lists the parameters whose sensitivity measures were at least 1% of that of the most important parameter. Interestingly enough, parameters whose normalized S_j values are at least 0.1 are the ones directly involved in the TNFα dynamics such as synthesis rate constants and degradation rate constants of TNFα transcripts and proteins. On the other hand, the parameters involved in the IκBα dynamics appear to be less important. This is probably because of the intrinsic property of the NFκB reaction network. Specifically, the existence of the feedback loop formed between NFκB and IκB proteins, where the activation of one will be inhibited by the other [84], ensures that an abrupt change in a reaction rate relevant to IκBα is less likely to result in an abrupt change in the overall IκBα dynamics.

Out of 22 parameters selected from the Morris method, the first eleven parameters were further analyzed by the subsequent sensitivity analysis through the Sobol' method [102]. Here, the parameters after the eleventh parameters were not further analyzed as their significance became negligible since the cumulative sum of the normalized S_j value does not increase more than 1% after the eleventh parameter [156]. Same as the Morris method, the values of these eleven parameters

Rank	Parameter	ST_{ij}	SS_{ij}
	With Respect to IκBα		
1	IKK*-mediated degradation rate constant for IκBα in NFκ-IκBα	0.49	0.30
2	IKKK*-mediated IKK activation rate constant	0.36	0.19
3	Constitutive IKKK activation rate constant	0.31	0.10
4	IκBα transcript degradation rate constant	0.30	0.15
5	K_{a0}	0.075	0.02
6	TNFα protein degradation rate constant	0.075	0.02
7	Maximum degradation rate constant for TNFα transcript	0.075	0.02
8	TNFα nascent mRNA processing rate constant	0.075	0.02
9	TNFα protein synthesis rate constant	0.075	0.02
10	Hill coefficient for TNFα transcription	0.075	0.02
11	K_a	0.075	0.02
	With Respect to TNFα		
1	Hill coefficient for TNFα transcription	0.95	0.08
2	K_a	0.85	0.03
3	TNFα nascent mRNA processing rate constant	0.79	0.00
4	IKK*-mediated degradation rate for IκBα in NFκ-IκBα complexes	0.78	0.00
5	K_{a0}	0.76	0.01
6	Maximum degradation rate constant for TNFα transcript	0.46	0.00
7	TNFα protein degradation rate constant	0.43	0.01
8	Constitutive IKKK activation rate constant	0.41	0.00
9	IκBα transcript degradation rate constant	0.25	0.00
10	IKKK*-mediated IKK activation rate constant	0.22	0.00
11	TNFα protein synthesis rate constant	0.21	0.00

Table 4.2: Result of Sensitivity Analysis by the Sobol' Method

were varied from 10% to 1000% of their nominal values, and the sample size for the integration approximation (N_{int}) was 5000. It should be noted that two different sensitivity indices were computed for each parameter with respect to each output separately through the Sobol' method, and parameters that are important to at least one output were selected for the subsequent parameter estimation.

Based on the values of ST_{ij} and SS_{ij} computed with respect to the IκBα dynamics (Equations 4.21-4.22), the top four parameters were selected for the subsequent estimation since the values of both sensitivity measures were one order of magnitude larger than the remaining ones (Table 4.2). With respect to the TNFα dynamics, the ST_{ij} values are relatively large while the

SS_{ij} values are quite small for all the parameters. This means that the parameters are highly dependent on each other in terms of making changes in the TNFα dynamics while one parameter has a little effect in changing the TNFα dynamics. Therefore, only one parameter, the Hill coefficient for TNFα transcription, was selected for the parameter estimation as it has the highest sensitivity measures. In summary, five parameters were selected to vary with time, and their values in each temporal subdomain were determined in the following parameter estimation step (Table 4.3).

Parameter	Parameter values in each temporal subdomain		
	$\mathbb{T}_1$	$\mathbb{T}_2$	$\mathbb{T}_3$
IKK*-mediated degradation rate constant for IκBα in NFκ-IκBα ($(\mu$M·min$)^{-1}$)	2.59	0.23	0.04
IKKK*-mediated IKK activation rate constant ($(\mu$M·min$)^{-1}$)	5200	52	4230
Constitutive IKKK activation rate constant (min^{-1})	5×10^{-6}	1.3×10^{-7}	4.9×10^{-6}
IκBα transcript degradation rate constant (min^{-1})	0.33	0.18	0.12
Hill coefficient for TNFα transcription	3.73	1.96	2.02

Table 4.3: Result of the Parameter Estimation

4.4.5 Parameter Estimation

With the results from the temporal clustering and sensitivity analysis, the parameter estimation problem (Equation 4.23) was solved to obtain the values of these parameters in each temporal subdomain (Table 4.3). Here, the model evaluation and the parameter estimation were performed via MATLAB built-in functions, *ode15s* and *fmincon*, and the *multistart* function available in MATLAB was used to solve the optimization problem multiple times with different initial values.

Figures 4.5-4.6 showed the predicted dynamics of TNFα and IκBα after the parameter estimation. In order to show the improvement of the prediction accuracy, the predicted dynamics after the parameter estimation were compared with experimental measurements [8] and those predicted before the estimation. The prediction accuracy for the dynamics of the proteins was significantly

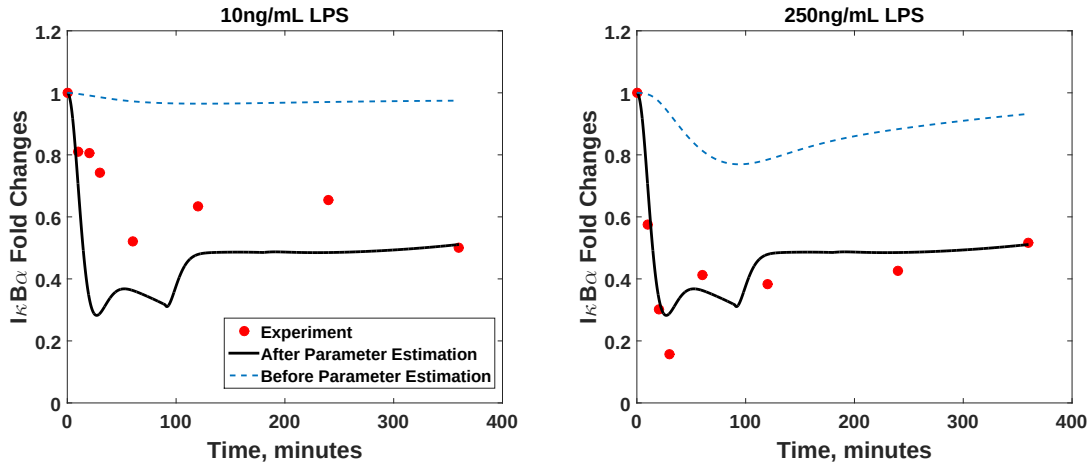
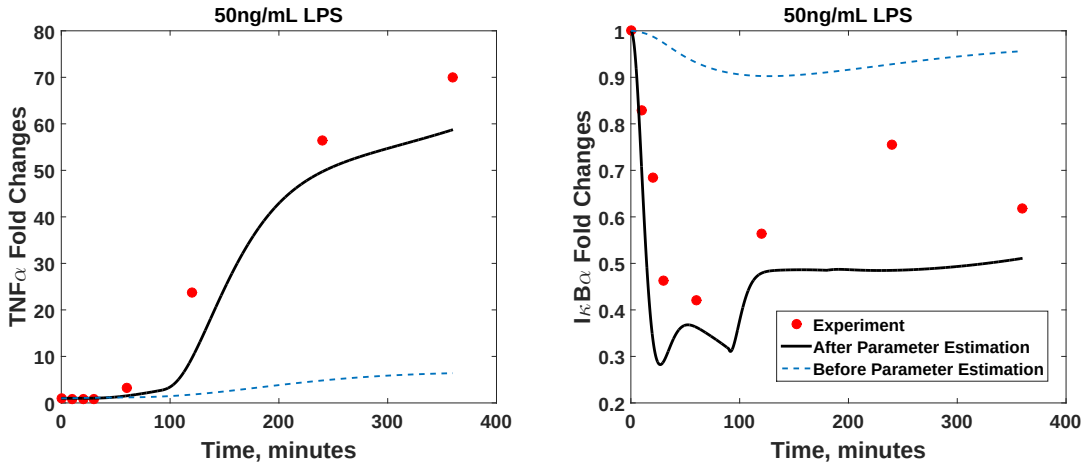


Figure 4.6: The result of parameter estimation. The predicted dynamics of IκBα before (dash line) and after (solid line) the parameter estimation were compared with the experimental observations under (a) 10ng/mL and (b) 250ng/mL of LPS in the presence of BFA.

improved. In particular, the model was able to track the TNFα dynamics very accurately under both conditions (Figure 4.5). Although there was some discrepancy between the model prediction and the experimental measurements for the IκBα dynamics under 10ng/mL LPS, the overall prediction was improved.

In order to further validate the resulted model, the prediction accuracy of the resulted model was assessed with the experimental dataset, which was not used to train the model. In Figure 4.7, the TNFα and IκBα dynamics predicted by the model under 50ng/mL LPS in the presence of BFA were plotted and compared with the corresponding experimental dataset. As shown in Figure 4.7, the resultant model was able to accurately predict the TNFα and IκBα dynamics reasonably well even though the 50ng/mL dataset was not used in the model calibration, which demonstrated the robustness of the calibrated model and thus validate the proposed methodology.

Lastly, the resulted model was compared with our previous model, which partially incorporated the ER-stress signaling pathway through further experimentation and literature survey. Figures 4.8-4.9 compare the model performance of these two models by comparing their temporal dynamics under two different LPS concentrations. In general, both models were similar in terms of re-

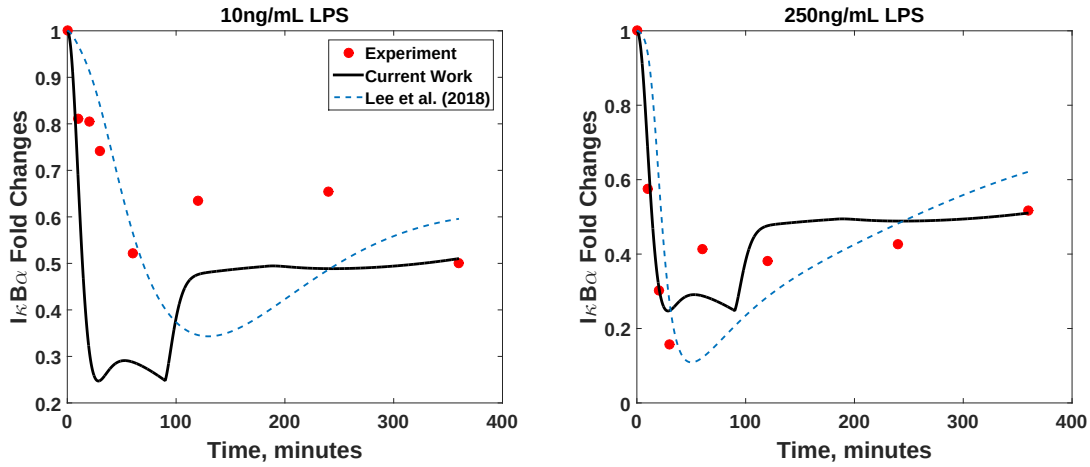


(a) (b)

Figure 4.7: The validation of the parameter estimation results with an independent dataset, which was not used in the model calibration. The predicted dynamics of (a) TNFα and (b) IκBα before (dash line) and after (solid line) the parameter estimation were compared with the experimental observations under 50ng/mL of LPS in the presence of BFA.

producing the IκBα dynamics while the model developed in the current work predicted slightly more accurately in terms of root-mean-squares (RMS) of the parameter estimation: the normalized RMS of the parameter estimation for the presented model is 1.75 while the RMS value for the previous model is 2.29, which showed the improved model accuracy by implementing the proposed approach. It should be noted that the development of the previous model went through the iterative implementation of experiments and modeling, which can be time-consuming. However, through the proposed approach, one can get a model with a reasonable prediction accuracy in a short amount of time.

Since intracellular signaling pathways regulate various cellular behaviors, their dynamics and outcomes bear great importance for studying and predicting the tissue-level responses *in vivo*. One important factor dictating the signaling pathways is different stimuli that initiate the pathways. As discussed in the manuscript, there can be multiple stimuli for one signaling pathway, and the number of stimuli is likely to be high for those with highly complex network structures. For example, it has been found that there are around 100 stimuli that can trigger the NFκB signaling

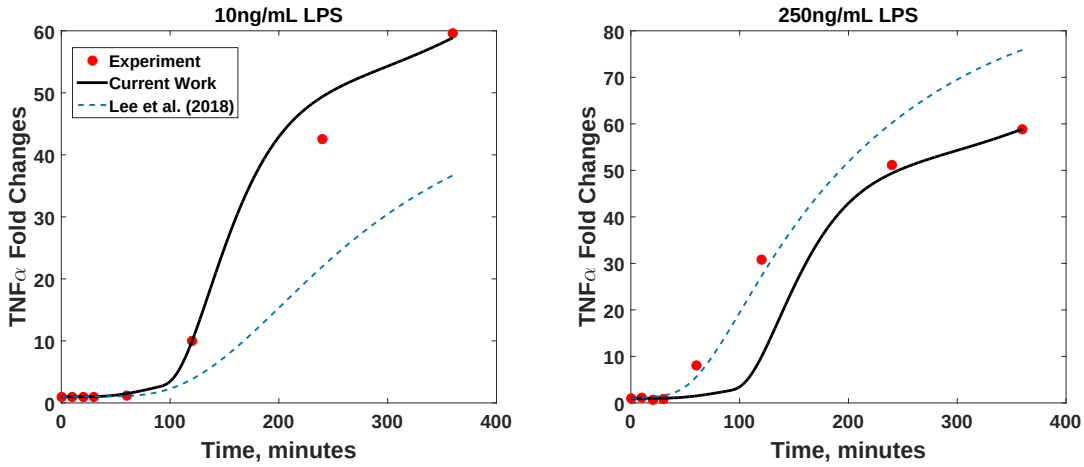


(a) (b)

Figure 4.8: The comparison between the model developed in this study and the model developed by [8]. The predicted dynamics of IκBα were compared with the experimental observations under (a) 10 ng/mL and (b) 250 ng/mL (b) LPS concentration in the presence of BFA.

pathway. Moreover, the dynamics of one signaling pathway induced by different stimuli can be very different since these stimuli activate the intracellular signaling pathway through different mechanisms. Again, with the NFκB signaling pathway as an example, TNFα and LPS, two well-known stimuli of the NFκB signaling pathway, activate the signaling pathway through two different molecules (TNFα receptor and TLR4, respectively), resulting the distinctive signaling dynamics.

Therefore, the comprehensive characterization of an intracellular signaling pathway is nontrivial since each stimulus of the signaling pathway has its own distinct activation mechanism and corresponding dynamics. Under this circumstance, a model-based approach can be implemented to facilitate the study. However, this model-based approach is often practical only for a handful of well-characterized stimuli such as TNFα and LPS for the NFκB signaling pathway since the underlying signaling mechanisms induced by these stimuli are relatively well studied. Motivated by the above considerations, the current study proposes a methodology to construct a data-driven mechanistic model for those less-studied stimuli, whose corresponding signaling dynamics are less characterized. This is feasible since the mechanisms of the signaling pathway induced by different stimuli overlap with each other. For example, the NFκB signaling pathway network induced by



(a) (b)

Figure 4.9: The comparison between the model developed in this study and the model developed by [8]. The predicted dynamics of TNFα were compared with the experimental observations under (a) 10 ng/mL and (b) 250 ng/mL LPS concentration in the presence of BFA.

TNFα and LPS will change the level of IKK activation, whose activation level will determine how fast NFκB proteins will be freed from their inhibitors. Therefore, for a new or less-studied stimulus of an intracellular signaling pathway, its corresponding signaling dynamics can be described by modifying the nominal model into a time-varying one as discussed in the manuscript. Then, the constructed model can be used for the optimal experimental design to enhance our understanding.

It should be noted that the proposed methodology is a semi-data-driven approach, where the model construction is guided by both the available experimental data and the mechanistic model. Specifically, based on the experimental data, the temporal profiles of the model parameters are inferred to complement the model mismatch due to the use of a nominal model. As a result, the resultant model can provide relatively accurate predictions in spite of the incomplete knowledge of the underlying system. At the same time, the use of the mechanistic model allows the resultant model to be used in the detailed analysis of the underlying mechanisms, which is difficult to be performed through a data-driven model.

Additionally, the proposed time-varying model was able to robustly predict the dynamics of IκBα and TNFα proteins, which are the core components in the NFκB signaling pathway, under

the various conditions although the detailed ER-stress signaling mechanisms were not incorporated into the model. Due to the accuracy and robustness of the model, it can be used in the future studies to design optimal experiments to enhance our understandings on how BFA can activate the NFκB signaling pathway.

Although the proposed methodology can be used to obtain a more accurate and predictive model as described above, it has the following limitations. First, the increase in the number of parameters need to be estimated due to the temporal partitioning the parameters may exacerbate the unidentifiability issue in the model calibration. This can be a severe issue since a signaling pathway model is often over-parameterized while the available experimental measurements are limited. Second, the identified model may not reflect the true mechanisms associated with the less-studied stimulus. Specifically, the proposed method relies on the global sensitivity analysis to identify which parameters are time-varying, but it does not consider any biological significance while selecting the parameters. Therefore, the identified temporal profiles of the parameters may not have the biological relevance, which will constrain the process analysis based on the resultant model. It should be noted that this limitation can be mitigated by adding additional constraints into the minimization problem (Equation 4.23) so that the resultant parameters retain their biological significance.

4.5 Conclusion

In this work, we presented a methodology for constructing a time-varying model for an intracellular signaling pathway when its reaction network is not fully known *a priori*. First, experimental data were clustered through the k-mean clustering algorithm to determine the temporal subdomains for the model parameters, where the parameters have different values in each temporal subdomain. Next, the global sensitivity analysis, which uses the Morris and Sobol' methods in sequence, was carried out to identify the most important parameters with respect to the model outputs. And only these parameters were determined to be time-varying while the remaining parameters were fixed at their nominal values. Finally, the least-squares problem was solved to estimate the values of five parameters in each temporal subdomain to construct an accurate time-varying model. The

proposed methodology was implemented to model the NFκB signaling pathway induced by LPS in the presence of BFA to predict the dynamics of IκBα and TNFα proteins. The prediction accuracy of the resulted model was comparable to that of a more detailed model proposed by Lee et al. [8], which demonstrated the performance of the proposed methodology. In summary, the proposed methodology speeds up the overall model development process without losing the prediction accuracy by avoiding the time-consuming procedure of experimentation and literature survey for developing a high-fidelity model.

5. HYBRID MODELING APPROACH TO CONSTRUCT A SEMI-MECHANISTIC MODEL OF PARTIALLY KNOWN INTRACELLULAR SIGNALING PATHWAY

5.1 Introduction

When our prior understanding of a dynamic system is limited, first-principle mathematical models may not be able to generate accurate predictions. Alternatively, a data-driven model can be developed from available experimental measurements [208]. The resulted data-driven model can describe the system adequately, even when mechanistic understanding is limited [209, 210, 211]. However, a data-driven model has narrow applicability as it is tailored to describe input-output relationships contained in the training datasets [212, 213, 214].

As an alternative, a hybrid modeling approach that combines first-principles and data-driven modeling techniques has been proposed to describe a partially known process [208]. Hybrid models have better prediction capabilities than first-principle models, and they have better generalizability and interpretability than data-driven models [212, 213, 215]. One classical example of the hybrid modeling approach is to model a fedbatch bioreactor, where the biomass growth rate is inferred from process data and coupled with mass conversation laws [213, 216]. In this model, the mass conservation laws represent our prior knowledge about the system (i.e., the first-principle model), where its growth rate is uncertain and inferred from experiment measurements to improve the model's prediction accuracy. Due to its merits, the hybrid modeling approach has been implemented in various research areas such as bioprocess development and optimization [212, 213, 216, 217], modeling propagation of fractures during the hydraulic fracturing process [215], transcription factor dynamics [218], tumor development and treatment [219], and flour beetles population dynamics [220, 221].

5.2 Preliminaries

In this study, the following hybrid modeling formulation is adopted [179, 222]:

$$\begin{aligned} \dot{\boldsymbol{x}} &= \boldsymbol{f}\left(\boldsymbol{x}, \boldsymbol{\theta}, u; t\right) + \boldsymbol{w}(\boldsymbol{x}(u;t);t); \quad \boldsymbol{x}(0) = \boldsymbol{x}_0 \\ \boldsymbol{y} &= \boldsymbol{g}\left(\boldsymbol{x}, \boldsymbol{\theta}, u; t\right) \end{aligned} \tag{5.1}$$

where $\boldsymbol{x} \in \mathbb{R}^{n_x}$ is the state vector whose dynamics are partially correct, u is the external perturbation given to the system, $\boldsymbol{w}(\boldsymbol{x}(u;t);t) \in \mathbb{R}^{n_x}$ is the vector of correction terms necessary for the model to compensate for model-system mismatch due to incomplete understanding of the system, and $\boldsymbol{y} \in \mathbb{R}^{n_y}$ is the output vector of the system. Here, it is assumed that the correction terms depend on the values of states and time [212, 223].

In order for Equation 5.1 to properly predict true dynamics of the system, the values of $\boldsymbol{w}$ need to be inferred from experimental measurements at every time instant along the temporal trajectory of the system. Suppose that experimental datasets are obtained under n_u different concentrations of u, and $\boldsymbol{y}$ are measured at N_t discrete time instants from $t = 0$ to $t = t_{N_t}$. Then, the following minimization can be solved:

$$\min_{\mathcal{H}} \quad \sum_{s=1}^{n_u} \sum_{i=1}^{n_y} \sum_{l=1}^{N_t} \left(\frac{y_i(u_s;t_l) - \bar{y}_i(u_s;t_l)}{\bar{y}_i(u_s;t_l)} \right)^2 \tag{5.2a}$$

$$\text{s.t.} \quad \dot{\boldsymbol{x}} = \boldsymbol{f}\left(\boldsymbol{x}, u_s; t\right) + \boldsymbol{w}(\boldsymbol{x}(u_s;t);t); \quad \boldsymbol{x}(0) = \boldsymbol{x}_0 \tag{5.2b}$$

$$\boldsymbol{y} = \boldsymbol{g}\left(\boldsymbol{x}, u_s; t\right) \tag{5.2c}$$

$$\boldsymbol{w}(\boldsymbol{x}(u_s;t);t) = \mathcal{H}(\boldsymbol{x}(u_s;t);t) \tag{5.2d}$$

where $y_i(u_s;t_l)$ and $\bar{y}_i(u_s;t_l)$ are simulated and experimentally measured outputs, respectively, at time $t = t_l$, where $l = 1, \ldots, N_t$ in response to the s^{th} input, and $\mathcal{H}(\boldsymbol{x}(u_s;t);t)$ is a function that computes the values of $\boldsymbol{w}$ for given values of the model states and the current time.

In the above least-squares problem, we seek to estimate $\mathcal{H}$ so that values of $\boldsymbol{w}$ can be computed once the value of $\boldsymbol{x}$ and current time are known. However, the functional forms of $\boldsymbol{w}$ (i.e., $\mathcal{H}$) are

usually unknown a priori. Although there are some methods proposed in the literature to identify functional forms from the data, inferring functional forms usually requires a large amount of data and can be computationally expensive [214, 224, 225, 226]. Consequently, we assume $\mathcal{H}$ to be a shallow artificial neural network (ANN) [212, 213, 215]. Specifically, this study adopts a shallow ANN that has one input layer, one hidden layer, and one output layer, where each layer contains multiple neurons as shown in Figure 5.1. Here, an ANN is chosen here due to its proven ability to represent any arbitrary input-output relations with sufficient accuracy [227, 228].

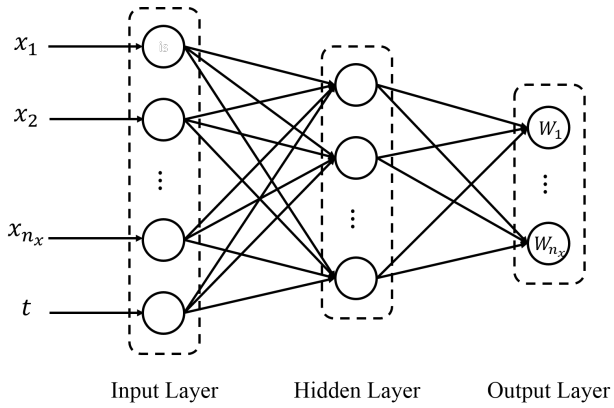


Figure 5.1: A shallow ANN

For each neuron in the hidden layer of a shallow ANN, the following hyperbolic tangent sigmoid transfer function is used:

$$o_i = \frac{2}{1 + e^{-2\hat{u}_i}} - 1, \quad \hat{u}_i = \sum_{j=1}^{n_u} (\alpha_{ij} \cdot z_j + b_j), \; \forall i = 1, \ldots, N_n \tag{5.3}$$

where o_i is the output of the i^{th} neuron in a hidden layer, $\hat{u}_i$ is the weighted sum of inputs given to the i^{th} neuron in the hidden layer, n_u is the number of inputs given to the shallow ANN, α_{ij} is a weight term for the input z_j to the ith neuron, z_j is the j^{th} input of the ANN, b_j is a bias given to the i^{th} neuron, and N_n is the number of neurons in the hidden layer [229]. For the purpose of this study, the inputs to the ANN are the states and time (i.e., $\boldsymbol{z} = [x_1 \; x_2 \; \cdots \; x_{n_x} \; t]$), so the number of

inputs to the ANN (i.e., n_u) will be equal to $n_x + 1$.

On the other hand, the outputs of the ANN will be computed as follows:

$$w_k = \sum_{i=1}^{N_n} \beta_{ki} o_i + c_k,\ k = 1, \ldots, n_x \tag{5.4}$$

where w_k is the k^{th} element of $\boldsymbol{w}$, β_{ki} is the weight given to o_i for w_k, and c_k is a bias term given to w_k.

After assuming the computed ANN as $\mathcal{H}$, a solution to Equation 5.2 can be obtained by two subsequent steps: first, the following optimization problem is solved to obtain $\boldsymbol{w}(t)$ under the n_u different conditions:

$$\min_{\boldsymbol{w}(u_s;t)} \quad \sum_{s=1}^{n_u} \sum_{i=1}^{n_y} \sum_{l=1}^{N_t} \left(\frac{y_i(u_s;t_l) - \bar{y}_i(u_s;t_l)}{\bar{y}_i(u_s;t_l)} \right)^2 \tag{5.5a}$$

$$\text{s.t.} \quad \dot{\boldsymbol{x}} = \boldsymbol{f}\left(\boldsymbol{x}, u_s; t\right) + \boldsymbol{w}(\boldsymbol{x}(u_s;t);t); \quad \boldsymbol{x}(0) = \boldsymbol{x}_0 \tag{5.5b}$$

$$\boldsymbol{y} = \boldsymbol{g}\left(\boldsymbol{x}, u_s; t\right) \tag{5.5c}$$

and, second, from the obtained values of $\boldsymbol{w}(t)$, a shallow ANN is developed for predicting the values of $\boldsymbol{w}$ for given $\boldsymbol{x}$ and current time. While the development of an ANN is easier as multiple software packages such as MATLAB Neural Network Toolbox are available to use, the first sub-problem (Equation 5.5) is more difficult to solve. Specifically, the inference of $\boldsymbol{w}(t)$ is likely to be ill-conditioned since a mathematical model for a biological system is often over-parameterized with a large number of states (i.e., large n_x) with a limited number of measurements (i.e., low n_y and N_t) to calibrate the model [23, 47]. Moreover, as how $\boldsymbol{w}$ evolves with time is unknown, the values of $\boldsymbol{w}$ need to be estimated at a large number of time points, which will significantly exceed N_t (i.e., the number of time points where the measurements are taken). As a result, the overall $\boldsymbol{w}$ inference problem is very likely to be ill-posed [76, 185]. Hence, the estimated values of $\boldsymbol{w}$ become unreliable, and the resulted hybrid model constructed based on the estimated $\boldsymbol{w}$ will be difficult to be generalized for future predictions. Therefore, this study will focus on how to solve

the first subproblem, the inference of $\boldsymbol{w}$, in a systematic way.

5.3 Problem Statement

In summary, this section seeks to solve the following problems

Problem. *Given the first-principle model (Equation 4.1) and the experimental datasets measured at N_t time instants under n_u different inputs, estimate the values of $\boldsymbol{w}$ at multiple time points by addressing the ill-posedness of the $\boldsymbol{w}$ inference problem.*

Problem. *Given the estimated the values of $\boldsymbol{w}$, develop a shallow ANN with the states and time as inputs and $\boldsymbol{w}$ as outputs of the ANN.*

5.4 Proposed Methodology

As described above, the inference of $\boldsymbol{w}$ is likely to be ill-conditioned. In this study, several assumptions are made to reduce the dimension of decision variables. First, the values of $\boldsymbol{w}$ will be estimated only when the measurements are taken (i.e., from $t = 0, \ldots, t_{N_t}$). Second, instead of adding $\boldsymbol{w}$ to all model states as in Equation 5.1, a subset of states is selected a priori, and only these states are given the correction terms. Hence, the resulted hybrid model can be formulated as follows:

$$\begin{aligned}
&\dot{\boldsymbol{x}} = \boldsymbol{f}\left(\boldsymbol{x}, u_s; t\right) + \boldsymbol{H}\begin{bmatrix} w_{s_1}(t) \\ \vdots \\ w_{s_{n_s}}(t) \end{bmatrix} \quad \boldsymbol{x}(0) = \boldsymbol{x}_0 \\
&\boldsymbol{y} = \boldsymbol{g}\left(\boldsymbol{x}, u_s; t\right) \\
&H_{ij} = \begin{cases} 1 & \text{if} \quad x_i = x_{s_j}, \quad i = 1, \ldots, n_x,\ j = 1, \ldots, n_s \\ 0 & \text{otherwise.} \end{cases}
\end{aligned} \tag{5.6}$$

where $\boldsymbol{H}$ is a $n_x \times n_s$ matrix, $\boldsymbol{x}_s = [x_{s_1}, \cdots, x_{s_{n_s}}]^T \in \mathbb{R}^{n_s}$ is a subset of $\boldsymbol{x}$ whose dynamics are corrected by $\boldsymbol{w}_s = [w_{s_1}, \cdots, w_{s_{n_s}}]^T \in \mathbb{R}^{n_s}$, H_{ij} is the entry in $\boldsymbol{H}$ at the i^{th} row and the j^{th} column.

It should be noted that Equation 5.6 assumes that each correction term adjust the dynamics of the corresponding state.

With the above two assumptions, this study formulates the $\boldsymbol{w}_s$ inference problem as a least-squares problem with a L_2 regularization. With these approaches, the inference of the correction terms can be formulated as follows:

$$\min_{\boldsymbol{W}} \quad \sum_{s=1}^{n_u}\sum_{l=1}^{N_t}\sum_{i=1}^{n_y}\left(\frac{y_i(u_s;t_l)-\bar{y}_i(u_s;t_l)}{\bar{y}_i(u_s;t_l)}\right)^2+\frac{\alpha}{2}\cdot R(\boldsymbol{W}) \tag{5.7a}$$

$$\text{s.t.} \quad \dot{\boldsymbol{x}}=\boldsymbol{f}\left(\boldsymbol{x},u_s;t\right)+\boldsymbol{H}\boldsymbol{w}_s(u_s;t) \quad \boldsymbol{x}(0)=\boldsymbol{x}_0 \tag{5.7b}$$

$$\boldsymbol{y}=\boldsymbol{g}\left(\boldsymbol{x},u_s;t\right) \tag{5.7c}$$

$$R(\boldsymbol{W})=\sum_{s=1}^{n_u}\sum_{l=1}^{N_t}\|\boldsymbol{w}_s(u_s;t_l)\|_2^2 \tag{5.7d}$$

$$\boldsymbol{W}=\begin{bmatrix}\boldsymbol{w}_s(u_1;t_1) & \cdots & \boldsymbol{w}_s(u_1;t_{N_t}) & \boldsymbol{w}_s(u_2;t_1) & \cdots & \boldsymbol{w}_s(u_{n_u};t_{N_t})\end{bmatrix} \tag{5.7e}$$

$$\hat{\boldsymbol{w}}_s(u_s;t)=\begin{cases}\boldsymbol{w}_s(u_s;t_l), & \text{if } t=t_l,\ l=1,\ldots,N_t\\ \boldsymbol{w}_s(u_s;t_l)+(t-t_l)\frac{\boldsymbol{w}_s(u_s;t_{l+1})-\boldsymbol{w}_s(u_s;t_l)}{t_{l+1}-t_l}, & \forall t\notin\begin{bmatrix}0, & \ldots, & t_{N_t}\end{bmatrix}\end{cases} \tag{5.7f}$$

where $\boldsymbol{W}\in\mathbb{R}^{n_s\times(n_u\cdot N_t)}$ is the correction term matrix to be estimated, $R(\boldsymbol{W})$ is the L_2 regularization term, and α is the tuning parameter for regularization. Since the values of $\boldsymbol{w}_s$ are estimated only at the time points when the measurements are taken, a linear interpolation is employed to compute values of $\boldsymbol{w}_s$ at time $t_k\notin[0,\ \ldots,\ t_{N_t}]$, as described in Equation 5.7f. Also, it should be noted that the L_2 regularization is chosen so that Equation 5.7 is more robust to experimental noise, and its objective function is continuously differentiable [76, 230, 231, 232].

5.4.1 Selection of $\boldsymbol{x}_s$

Before solving Equation 5.7, we need to determine which states should be classified into $\boldsymbol{x}_s$. Specifically, two questions need to be answered: first, what the dimension of $\boldsymbol{x}_s$ (i.e., n_s) is, and second, when the value of n_s is known, which states in $\boldsymbol{x}$ should be selected to form $\boldsymbol{x}_s$. In this study, we employ a graph-theoretical approach to determine a state subset, to which the correction

terms are given.

First, this study aims to make the hybrid model (Equation 5.6) to be invertible by letting n_s, the dimension of $\boldsymbol{w}_s$ to be equal to n_y. If a system is invertible, for a given value of $\boldsymbol{x}_0$, unique values of $\boldsymbol{y}$ will correspond to unique values of inputs [233, 234]. Therefore, one could reconstruct the values of inputs from available output measurements. Hence, if we view $\boldsymbol{w}_s$ in the hybrid model (Equation 5.6) as inputs to the system and the hybrid model is invertible, the values of $\boldsymbol{w}_s$ can be uniquely characterized from given measurements [233, 235]. Hence, it is our best interest to select the dimension of $\boldsymbol{w}_s$ as well as its placement so that the resulted hybrid model is invertible, which will attenuate the ill-posedness of the inverse problem. Daoutidis and Kravaris [233] have shown that a dynamic system is invertible when the following matrix is nonsingular:

$$C(\boldsymbol{x}) = \begin{bmatrix} \mathcal{L}_{h_1}\mathcal{L}_f^{r_1-1}g_1(\boldsymbol{x}) & \cdots & \mathcal{L}_{h_{n_s}}\mathcal{L}_f^{r_1-1}g_1(\boldsymbol{x}) \\ \vdots & \ddots & \vdots \\ \mathcal{L}_{h_1}\mathcal{L}_f^{r_{n_y}-1}g_{n_y}(\boldsymbol{x}) & \cdots & \mathcal{L}_{h_{n_s}}\mathcal{L}_f^{r_{n_y}-1}g_{n_y}(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} c_{11} & \cdots & c_{1n_s} \\ \vdots & \ddots & \vdots \\ c_{n_y1} & \cdots & c_{n_yn_s} \end{bmatrix} \tag{5.8}$$

where $C(\boldsymbol{x})$ is the characteristic matrix of the system, $\mathcal{L}$ represents Lie derivative defined as $\mathcal{L}_f g_i(\boldsymbol{x}) = \sum_{j=1}^{n_x}(\partial g_i/\partial x_j)f_j(\boldsymbol{x})$, h_k, where $k = 1, \ldots, n_s$, is a k^{th} column vector of the matrix $\boldsymbol{H}$ in Equation 5.6, and r_i is the relative order of the output y_i with respect to $\boldsymbol{w}_s$, which is defined as the smallest integer for which

$$\begin{bmatrix} \mathcal{L}_{h_1}\mathcal{L}_f^{r_i-1}g_i(\boldsymbol{x}) & \cdots & \mathcal{L}_{h_{n_s}}\mathcal{L}_f^{r_i-1}g_i(\boldsymbol{x}) \end{bmatrix} \neq \begin{bmatrix} 0 & \cdots & 0 \end{bmatrix} \tag{5.9}$$

or $r_i = \infty$, if such integer does not exist [235].

Therefore, one prerequisite for $C(\boldsymbol{x})$ to be nonsingular is to ensure this matrix is square by setting $n_s = n_y$. Once the value of n_s is determined, the second step is to examine to which states the n_y number of correction terms should be placed.

In this study, the selection of $\boldsymbol{x}_s$ is achieved by considering the following two criteria:

1. For each output y_i, $i = 1, \ldots, n_y$, only one state is chosen so that the placement of a correc-

tion term on this state will have the maximum 'physical closeness' to this output.

2. While an output has the maximum 'physical closeness' with one correction term, this output will have the minimum correlation with other correction terms in the system.

By meeting the above two criteria, $\boldsymbol{w}_s$ will have the maximal correlation with only one correction term, which will minimize the likelihood of the ill-posedness of Equation 5.7 [191].

In this study, the correlation between an output and a correction term is assessed by a relative order matrix, which is defined as follows:

$$r = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n_s} \\ r_{21} & r_{22} & \cdots & r_{2n_s} \\ \vdots & & \vdots & \vdots \\ r_{n_y1} & r_{n_y2} & \cdots & r_{n_yn_y} \end{bmatrix} \tag{5.10}$$

where r_{ij} is the relative order of the output y_i with respect to $w_{s_j},\ j = 1, \ldots, n_s$, which is defined as the smallest integer for which $\mathcal{L}_{h_j}\mathcal{L}_f^{r_{ij}-1}g_i(\boldsymbol{x}) \neq 0$ or $r_{ij} = \infty$ if such integer does not exist. Additionally, the relationship between r_i and r_{ij} is as follows:

$$r_i = \min(r_{i1}, r_{i2}, \cdots, r_{in_s}) \tag{5.11}$$

With this relative matrix, the above two criteria can be redefined in terms of relative orders [236]:

- The value of $\sum_{i=1}^{n_y} r_i$ needs to be minimal.
- The value of $\sum_i \sum_{j \neq i} r_i / r_{ij}$ needs to be minimal.

Previous studies have demonstrated that the relative order measures 'physical closeness' between a correction term and an output [235, 236]: a lower value of r_i represents a stronger connection between $\boldsymbol{w}_s$ and y_i. So, the first criterion renders $\boldsymbol{w}_s$ to have the maximum impacts on the outputs.

On the other hand, the second criterion is to render each w_{s_j} will have a maximum impact on only one output while having minimal impacts on the remaining outputs [236, 237].

In order to select a combination based on the above criteria, the following steps are taken:

1. Enumerate all n_y permutations of $\boldsymbol{x}$.
2. For each candidate, construct the corresponding relative order matrix, and compute its $\sum_{i=1}^{n_y} r_i$.
3. Find a candidate that has the lowest value of $\sum_{i=1}^{n_y} r_i$.
4. In case multiple candidates have equal minimal $\sum_{i=1}^{n_y} r_i$ values, compute their $\sum_i \sum_{j \neq i} r_i / r_{ij}$ values, and select a candidate with the lowest value.

For implementing the above procedure, a relative order matrix has to be constructed for each candidate. Instead of performing iterative Lie differentiation that can be computationally expensive, Daoutidis and Kravaris [235] have demonstrated that a graph-based approach can be used to easily evaluate relative orders of the system, which will be briefly discussed below.

5.4.2 Graph-theoretical Approach

A state-space model of a process (Equation 5.6) can be represented by a digraph, which is defined by a set of vertices and a set of edges by the following rules [235]:

- States ($\boldsymbol{x} \in \mathbb{R}^{n_x}$), outputs ($\boldsymbol{y} \in \mathbb{R}^{n_y}$), and manipulated inputs ($\boldsymbol{w}_s \in \mathbb{R}_{n_s}$) are represented by a set of vertices in a digraph.
- If $\partial f_i(\boldsymbol{x}) / \partial x_j \neq 0,\ i, j = 1, \ldots, n_x$,, there is a unidirectional edge pointing from the vertex of x_j to that of x_i.
- If $\partial f_i(\boldsymbol{x}) / \partial w_{s_k} \neq 0,\ k = 1, \ldots, n_s$, there is a unidirectional edge pointing from the vertex of w_{s_k} to that of x_i.
- If $\partial y_l / \partial x_j \neq 0,\ l = 1, \ldots, n_y$, there is a unidirectional edge pointing from the vertex of x_j to that of y_l.

In a digraph, a path from one vertex to another is a sequence of edges without repeating vertices, and the path length is the number of edges included in one particular path [235, 238]. Figure 5.2 provides an example on how a state-space model is translated into a digraph. Specifically, Figure 5.2 is a representation of the following dynamic system:

$$\begin{aligned} \frac{dx_1}{dt} &= f_1(x_2) + u \\ \frac{dx_2}{dt} &= f_2(x_3) \\ \frac{dx_3}{dt} &= f_3(x_1) \\ y &= g(x_3) \end{aligned} \tag{5.12}$$

What is important for this study is the connection between a digraph and a relative order matrix: Daoutidis and Kravaris [235] demonstrated that r_{ij} can be calculated by computing the shortest path length from an input w_{s_j} to an output y_i as follows:

$$r_{ij} = l_{ij} - 1 \tag{5.13}$$

where l_{ij} is the shortest path length from an input w_{s_j} to an output y_i. Therefore, the relative order matrix can be easily computed once a digraph of a state-space model is constructed.

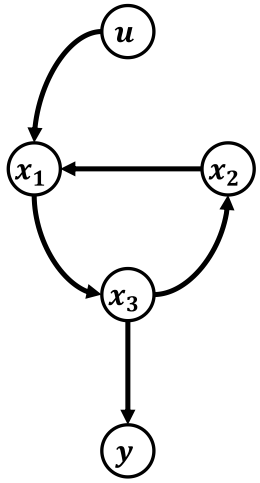


Figure 5.2: An illustration of a diagrph.

In summary, the procedures for selecting $\boldsymbol{x}_s$, whose dynamics will be corrected by $\boldsymbol{w}_s$, are illustrated with the following steps:

1. Set n_s, the size of $\boldsymbol{x}_s$, to be equal to the dimension of outputs (i.e., n_y).

2. Enumerate all n_s permutations of $\boldsymbol{x}$ as candidates of $\boldsymbol{x}_s$.

3. Construct a digraph by adding correction terms to each $\boldsymbol{x}_s$ candidate enumerated in the previous step.

4. Construct the corresponding relative order matrix based on Equation 5.13.

5. Find a configuration that has the lowest $\sum_{i=1}^{n_y} r_i$ value, and the configuration with the lowest value is the best selection for $\boldsymbol{x}_s$.

6. If there are multiple candidates of $\boldsymbol{x}_s$ with the equal lowest $\sum_{i=1}^{n_y} r_i$ value, compute their $\sum_i \sum_{j \neq i} r_i / r_{ij}$ values, and select the digraph with the lowest value.

Once the optimal $\boldsymbol{x}_s$ is chosen, Equation 5.7 is solved to infer the values of $\boldsymbol{W}$. With the estimated $\boldsymbol{W}$, an ANN model will be developed to predict the values of $\boldsymbol{w}_s$ for any given values of $\boldsymbol{x}$ and time so that the developed ANN can be integrated with the original state-space model (Equation 4.1) to predict the dynamics of the system accurately even with partial knowledge of the system.

5.5 Results

The above hybrid modeling approach is implemented to construct a semi-mechanistic model for the NFκB signaling pathway dynamics stimulated by LPS and BFA together. As described earlier, the dynamics under these two stimuli are not well studied; therefore, the system is a suitable example for constructing a hybrid model. Here, the previously developed LPS-induced NFκB signaling pathway model is used as the first-principle model [8], and correction terms are estimated from the available experimental measurements.

5.5.1 Selection of Locations of Correction Terms

As outlined in the previous section, the first step in developing a hybrid model is to optimize the number of correction terms needed as well as to which states these correction terms should be added. As the number of outputs for the system of interest is two (i.e., TNFα and IκBα), we assume that the dimension of $\boldsymbol{x}_s$ is also two to ensure the resulted hybrid model is invertible.

Second, all two-permutations from 49 model states are enumerated, and their corresponding digraphs are constructed to compute their corresponding relative order matrices. Based on the constructed relative order matrices, the minimum value of $\sum_{i=1}^{n_y} r_i$ is found to be two, and Table 5.1 lists all the configurations whose $\sum_{i=1}^{n_y} r_i$ values equal to two.

x_{s1}	x_{s2}	r_{11}	r_{12}	r_{21}	r_{22}	$\sum_{i=1}^{n_y} r_i$	$\sum_i \sum_{j \neq i} r_i / r_{ij}$
5	37	1	6	6	1	2	0.33
1	37	1	6	5	1	2	0.37
3	37	1	6	5	1	2	0.37
34	37	1	6	5	1	2	0.37
2	37	1	6	4	1	2	0.417
4	37	1	6	4	1	2	0.417
5	28	1	4	6	1	2	0.417
1	28	1	4	5	1	2	0.45
3	28	1	4	5	1	2	0.45
34	28	1	4	5	1	2	0.45
2	28	1	4	4	1	2	0.50
4	28	1	4	4	1	2	0.5

Table 5.1: All the configurations with the minimal $\sum_{i=1}^{n_y} r_i$ value.

It should be note that r_{1j} and r_{2j} compute the relative order with respect to IκBα and TNFα measurements, respectively, in Table 5.1.

Based on the result presented in Table 5.1, x_5 and x_{37} are chosen as the best candidate to add $\boldsymbol{w}_s$ since this pair has the lowest $\sum_i \sum_{j \neq i} r_i / r_{ij}$ value. It should be noted that both x_5 and x_{37} represent the concentrations of IκBα and TNFα transcripts, respectively. Therefore, adding correction terms to these states is a reasonable choice since one of the important factors affecting the dynamics of a protein concentration is its transcript dynamics. Hence, by correcting the dynamics of these two proteins' transcript through inferring $\boldsymbol{w}_s$, the resulted hybrid model will have a higher chance to predict the output dynamics more accurately.

With $\boldsymbol{x}_s = [x_1 \;\; x_{37}]$, the regularized least-squares problem (Equation 5.7) is solved to estimate $\boldsymbol{W}$ that contains the values of $\boldsymbol{w}_s$ at eight time points under three input concentrations. Since the value of regularization parameter α in Equation 5.7 is unknown beforehand, its optimal value is determined by five-fold cross-validation. For this study, there are three experimental datasets, each of which contains fold changes of TNFα and IκBα concentrations measured at eight time instants under a specific LPS concentration. For the purpose of the cross-validation, the experimental datasets are divided into training and validation dataset in five different ways, and the regularized least-square problem (Equation 5.7) is solved with one particular value of α with respect to each five training dataset. Then, the optimal value of α is chosen by examining average model errors with respect to both the training and validation datasets. Particularly, for each α value tested, the following average error criterion is computed:

$$\Lambda = \frac{1}{5} \sum_{i=1}^{5} \left[\frac{1}{N_{tl}} \sum_{j} \left(y_{ij}^{tl} - \bar{y}_{ij}^{tl} \right)^2 + \frac{1}{N_{vl}} \sum_{l} \left(y_{il}^{vl} - \bar{y}_{il}^{vl} \right)^2 \right] \tag{5.14}$$

where Λ is the average error for a given α value, N_{tl} and N_{vl} are the number of data points included in the training and validation datasets, respectively, y_{ij}^{tl} and $\bar{y}_{ij}^{tl}$ are the j^{th} predicted and measured output in the i^{th} training dataset, respectively, and y_{ij}^{vl} and $\bar{y}_{ij}^{vl}$ are l^{th} predicted and measured output in the i^{th} training dataset, respectively.

Figure 5.3 plots changes in the value of Λ with respect to the value of α. From this result, the optimal value of α is determined to be 0.001, and the estimated values of $\boldsymbol{W}$ obtained with $\alpha = 0.001$ are considered as optimal.

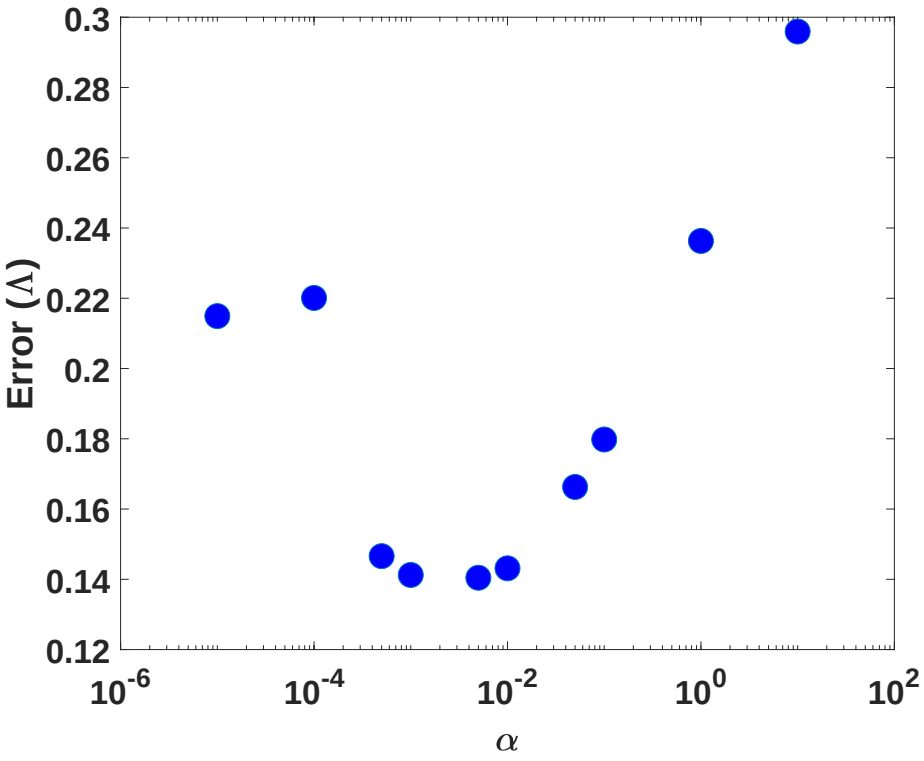


Figure 5.3: A plot showing how the value of average error (Λ) changes with the value of α. The error value equals to 0.14, which is the minimum, when $\alpha = 0.001$.

Before constructing an ANN model, the accuracy of the inferred $\boldsymbol{W}$ is assessed by comparing the experimental measurements and the predictions from the model with the inferred $\boldsymbol{W}$. The root mean squared error is 28.75. Figures 5.4-5.5 compare the predicted and measured TNFα and IκBα dynamics under three different LPS concentrations. Additionally, the predictions of the model coupled with the inferred $\boldsymbol{W}$ are compared with those of the model without $\boldsymbol{W}$. Figures 5.4-5.5 show the addition of $\boldsymbol{W}$ significantly improves the model accuracy across all three LPS concentrations. Specifically, the addition of $\boldsymbol{W}$ renders the model prediction to match with the experimental trend,

which shows the sustained low concentration of IκBα. At the same time, the addition of $\boldsymbol{W}$ helps the model predictions agree much better with the measured TNFα dynamics. Overall, these comparisons have demonstrated that the integration of $\boldsymbol{W}$ greatly improves the predictive capability of the hybrid model (Equation 5.6).

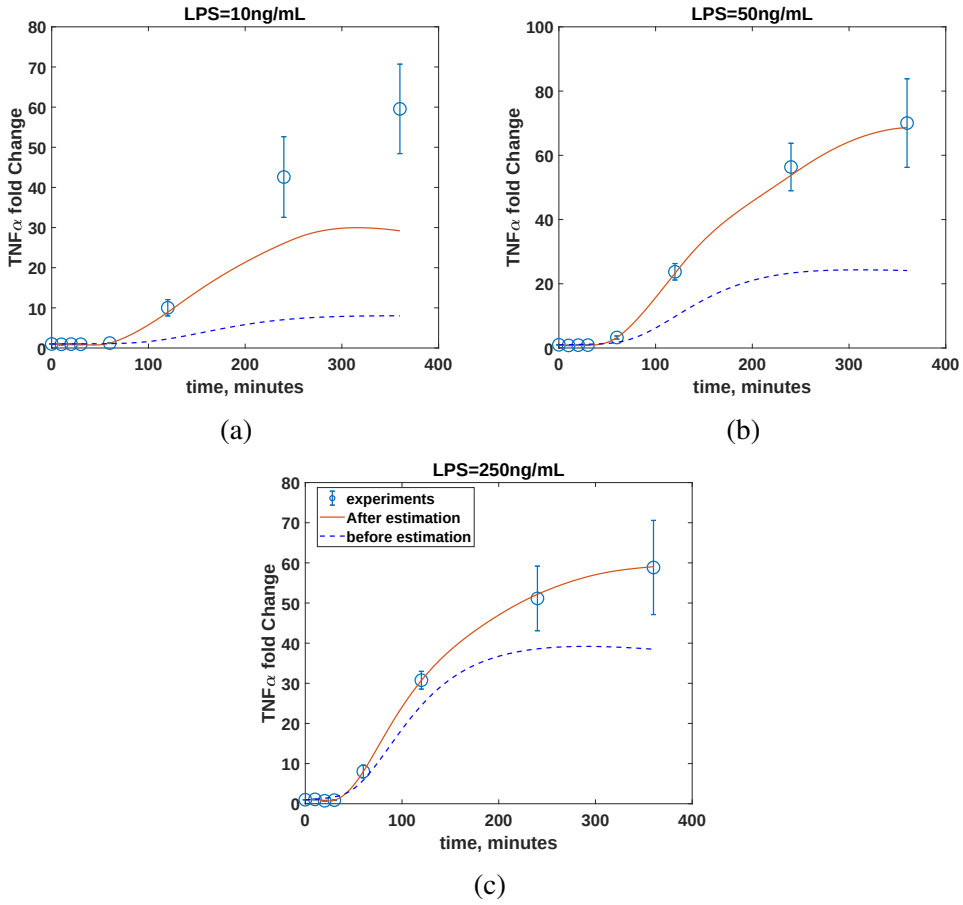


Figure 5.4: Comparison between predicted (red solid line) and measured (blue empty circle) TNFα dynamics in the presence of BFA under the LPS concentrations of (a) 10ng/mL, (b) 50ng/mL, and (c) 250ng/mL. Blue dash lines represent the model predictions without the correction terms ($\boldsymbol{W}$).

Lastly, with the inferred $\boldsymbol{W}$, an ANN is developed with the values of $\boldsymbol{x}$ and time as inputs to the ANN and $\boldsymbol{w}_s$ as outputs to the ANN. The ANN development is necessary so that the model coupled with the developed ANN can be used to predict the system dynamics under a new LPS concentration and to generalize the hybrid model performance. In order to increase accuracy of

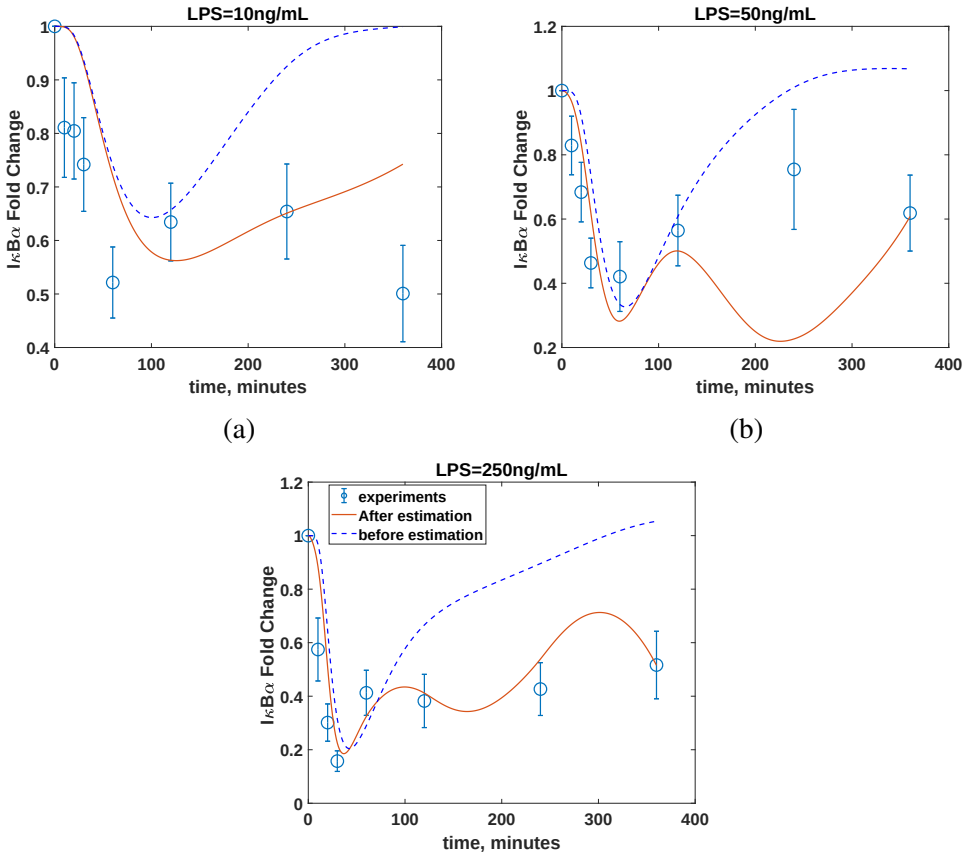


Figure 5.5: Comparison between predicted (red solid line) and measured (blue empty circle) IκBα dynamics in the presence of BFA under the LPS concentrations of (a) 10ng/mL, (b) 50ng/mL, and (c) 250ng/mL. Blue dash lines represent the model predictions without the correction terms ($\boldsymbol{W}$).

the ANN to be developed, the values of $\boldsymbol{w}_s$ are linearly interpolated from $t = 0$ to $t = 360$ minutes with one minute as the sampling interval based on the inferred $\boldsymbol{w}_s$ values. Accordingly, the values of $\boldsymbol{x}$ at the same time instants are also computed with the interpolated $\boldsymbol{w}_s$ values and the hybrid model (Equation 5.6). This approach will ensure that the size of the datasets to be used for developing ANN is large enough.

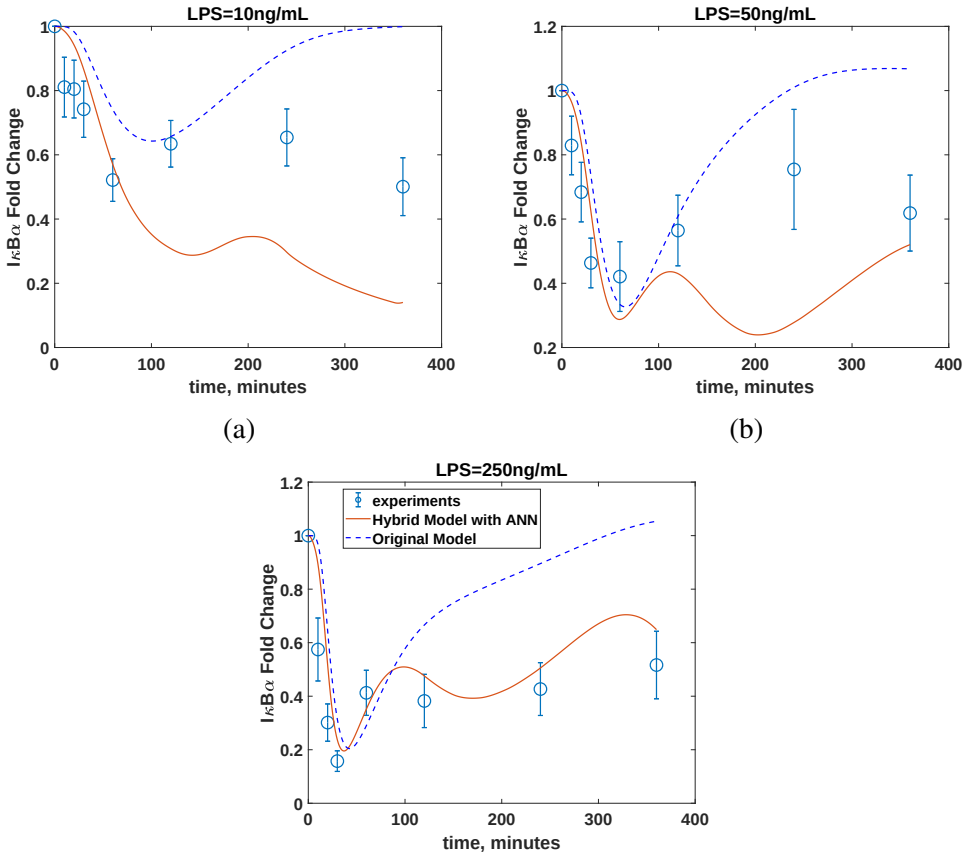


Figure 5.6: Comparison between IκBα dynamics predicted from the hybrid model (red solid line) and measurements (blue empty circle) in the presence of BFA under the LPS concentrations of (a) 10ng/mL, (b) 50ng/mL, and (c) 250ng/mL. Blue dash lines represent the model predictions without the ANN.

In this study, a shallow ANN with one hidden layer with ten neurons are used for describing an empirical mapping from $\boldsymbol{x}$ and t to $\boldsymbol{w}_s$. All calculations related to the ANN construction are implemented in the MATLAB Neural Network Toolbox. The R^2 statistics for the developed ANN model are 0.999 and 0.998 with respect to the training and validation datasets, respectively, which show the accuracy of the developed ANN. Then, the developed ANN is integrated with the first-principle model to construct the hybrid model. The root mean squared error of the constructed the hybrid model with respect to the experimental measurements is 63.9. Figures 5.6-5.7 show the prediction capabilities of the resultant hybrid model by comparing its predictions with the available

experimental results.

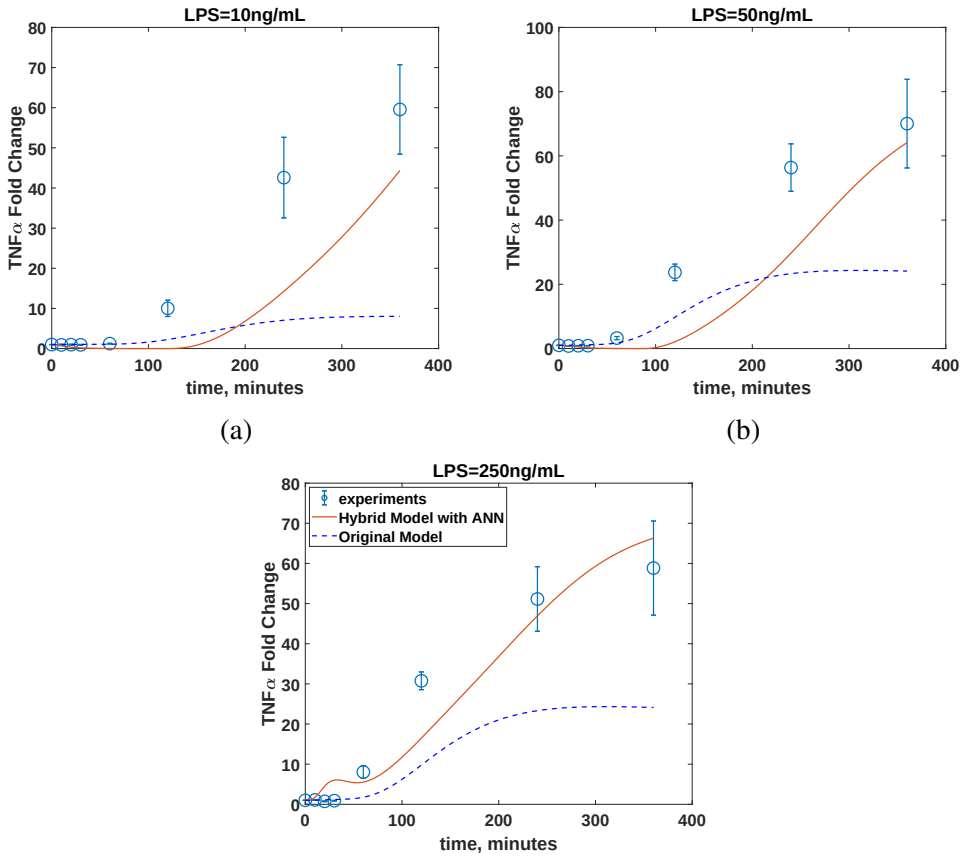


Figure 5.7: Comparison between TNFα dynamics predicted from the hybrid model (red solid line) and measurements (blue empty circle) in the presence of BFA under the LPS concentrations of (a) 10ng/mL, (b) 50ng/mL, and (c) 250ng/mL. Blue dash lines represent the model predictions without the ANN.

It should be noted that the most significant model-system mismatch occurs under the 10 ng/mL LPS concentration throughout Figures 5.4-5.7. It is hypothesized that this is due to the low LPS concentration. Since the concentration of BFA is constant, its effects are more pronounced at a lowe LPS concentration. Specifically, at 10 ng/mL LPS concentration, the first-principle model becomes least accurate as the effects of the BFA-induced signaling pathways dominate the effect by the low LPS concentration, which may contribute to the largest discrepancy. Increasing the

dimension of $\boldsymbol{w}_s$ may further improve the prediction accuracy due the increase in the degree of freedom. Alternatively, the first-principle model can be modified further by incorporating known mechanisms of the BFA-induced signaling pathways (see [8] for an example) to improve the first-principle model before estimating the values of $\boldsymbol{w}_s$.

5.6 Conclusion

In this work, a systematic method has been proposed to construct a hybrid model to describe the dynamics of an intracellular signaling pathway when our knowledge about the system is limited. Specifically, by combining merits of the first-principle and data-driven modeling approaches, a hybrid model is more suitable under such circumstances. This study employs one particular formulation (Equation 5.6), where its first-principle part is correct by additional correction terms inferred from experimental observations. One key step in constructing such a hybrid model is to infer the values of correction terms ($\boldsymbol{w}$) to improve the model predictions. As this inference problem is likely to suffer from overfitting issues, several measures are implemented to reduce the dimension of $\boldsymbol{w}$ to be estimated. Specifically, a graph-theoretical approach has been implemented to estimate only a subset of $\boldsymbol{w}$ (i.e., $\boldsymbol{w}_s$) from the given experimental measurements. Once the $\boldsymbol{w}_s$ values are estimated, an ANN is developed so the first-principle model integrated with the developed ANN can be used for predicting the dynamics of the signaling pathway even under new conditions. In the future, the hybrid model will be used for solving a model-based optimal experimental design problem to explore unknown mechanisms for further improving the first-principle model.

6. IDENTIFICATION OF CELL-TO-CELL HETEROGENEITY THROUGH SYSTEMS ENGINEERING APPROACHES *

6.1 Background

Once an accurate deterministic ODE model is developed by any of the methods implemented or proposed in previous chapters, an IBPM can be developed by estimating model parameters' PDFs. In this regard, we propose a new methodology to estimate the PDFs of parameters from single-cell measurements by handling the identifiability issue without increasing the computational cost. First, a normal-distribution assumption is made to ensure the overall inference problem is finite-dimensional. Second, a parameter selection method that consists of the extended Sobol' method is implemented to assess the practical identifiability of each parameter's PDF to determine which parameter's PDF should be estimated. Third, an ANN is developed to predict the measurements for given PDFs of parameters, and it is evaluated iteratively to solve the overall PDF inference problem in a computationally efficient manner. The proposed methodology is implemented to infer the PDFs of model parameters in TNFα signaling pathway by utilizing *in silico* measurements.

6.2 Preliminaries

6.2.1 Model descriptions

In the IBPM framework, the dynamics of individual cells within a cell population are represented by the following ODE [10, 34, 42]:

$$\mathcal{M} = \begin{cases} \dot{\mathbf{x}}^{(i)}(t) & = \mathbf{f}\left(\mathbf{x}^{(i)}, \boldsymbol{\theta}^{(i)}, u; t\right); \quad \mathbf{x}^{(i)}(0) = \mathbf{x}_0^{(i)}, \;\; \boldsymbol{\theta}^{(i)} \sim p_\theta(\boldsymbol{\theta}) \\ \mathbf{y}^{(i)}(t) & = \mathbf{g}\left(\mathbf{x}^{(i)}, \boldsymbol{\theta}^{(i)}, u; t\right); \quad \forall i = 1, \ldots, M \end{cases} \tag{6.1}$$

where $\mathbf{x} \in \mathbb{R}^{n_x}$ is the state vector, $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$ is the parameter vector, u is the external stimulus as an input to the system, x_0 is the initial condition of $\mathbf{x}$, $p_\theta(\boldsymbol{\theta}) : \mathbb{R}_+^{n_\theta} \to \mathbb{R}_+$ is the multivariate

*Reprinted with the permission from "Identification of cell-to-cell heterogeneity through systems engineering approaches," by Lee et al., 2020, *AIChE Journal*, 66, e16925. Copyright 2020 by John Wiley & Sons, Inc.

PDF of $\boldsymbol{\theta}$, $\mathbf{y} \in \mathbb{R}^{n_y}$ is the output vector, M represents the number of cells, and index i represents the i^{th} cell in the population. In this study, $\mathbf{x}_0$ is obtained by running Equation 6.1 with $u = 0$ until the system reaches an equilibrium and setting the values of $\mathbf{x}_0$ at the equilibrium. It should be noted that the current IBPM modeling framework assumes the cell-to-cell communication is minimal [10]. By simulating Equation 6.1) M different times with M different values for $\boldsymbol{\theta}$, the heterogeneous responses within the cell population can be simulated. As discussed earlier, Equation 6.1 assumes the extrinsic source is a dominating factor for the heterogeneous dynamics of an intracellular biochemical reaction pathway. This assumption is generally valid for many systems as the intrinsic source of the heterogeneity may have only transient effects on the overall reaction dynamics.

6.2.2 Experimental measurements

In this work, we consider the output measurements are provided by flow cytometry, which is widely used to infer the intracellular protein concentrations since 1970s [239]. Flow cytometry is suitable for studying the heterogeneous single-cell dynamics because it measures the concentration of a biomolecule of interest from thousands of cells at a time. Specifically, at each sampling time $t_k,\ k = 1, \ldots, N_t$, where N_t is the number of sampling times, flow cytometry provides experimental measurements in the form of population snapshots, $\hat{D}_k$, which is expressed as follows [10, 42]:

$$\hat{D}_k = \{\hat{\mathbf{y}}^{(1)}(t_k), \hat{\mathbf{y}}^{(2)}(t_k), \cdots, \hat{\mathbf{y}}^{(N_k)}(t_k)\} \tag{6.2}$$

where $\hat{\mathbf{y}}^{(i)}(t_k)$ is the measurement from the i^{th} cell in the cell population, and N_k is the number of cells analyzed at the k^{th} sampling time, which is usually around a few thousands.

It should be noted that a cell has to be removed from the cell culture to be analyzed by flow cytometry, so the flow cytometry does not provide the time-series measurement from the same cell [10, 42]. That is, two measurements, $\hat{\mathbf{y}}^{(i)}(t_k)$ and $\hat{\mathbf{y}}^{(i)}(t_j)$, where $k \neq j$, are not from the same cell. Consequently, the measurement snapshots, $\hat{D}_k, \forall k = 1, \ldots, N_t$, are assumed to be independent from each other [42].

Similar to other biochemical assays, flow cytometry provides semi-quantitative datasets; in other words, it only measures the relative concentrations of intracellular biomolecules [240]. In this study, we assume the relative concentration measured by the flow cytometry can be expressed in terms of the absolute concentration as follows:

$$\hat{y}_j(t_k) = c_j \cdot \hat{x}_j(t_k) + \nu_j, \quad j = 1, \ldots, n_y \tag{6.3}$$

where $\hat{x}_j(t_k)$ is the concentration of a biomolecule, x_j, in the absolute concentration at time t_k, $\hat{y}_j(t_k)$ is the relative concentration of $\hat{x}_j$ measured by the flow cytometry, c_j is the proportional constant relating $\hat{y}_j$ and $\hat{x}_j$, and ν_j is the measurement noise. Here, we assume both c_j and ν_j are mutually independent and log-normally distributed (i.e., $\ln c_j \sim \mathcal{N}\left(\mu_{c_j}, \sigma_{c_j}\right)$ and $\ln \nu_j \sim \mathcal{N}\left(\mu_{\nu_j}, \sigma_{\nu_j}\right)$, respectively), [184]. For the purpose of this study, we assume $\mu_{c_j}, \mu_{\nu_j}, \sigma_{c_j}$, and σ_{ν_j} are known *a priori* through the equipment calibration [184, 241].

Based on this description of the experimental measurement (Equation 6.3), the model output, y_j, in Equation 6.1 is defined accordingly:

$$y_j(t_k) = \bar{c}_j(t_k) \cdot x_j(t_k) + \bar{\nu}_j(t_k) \tag{6.4}$$

Here, $\bar{c}_j(t_k)$ and $\bar{\nu}_j(t_k)$ are randomly sampled from the known distributions of c_j and ν_j, respectively.

6.2.3 Problem statement

In this study, we seek to infer the PDFs of IBPM model parameters, $p_\theta(\boldsymbol{\theta})$ (Equation 6.1), utilizing the available population snapshot measurements, $\{\hat{D}_k\}_{k=1}^{N_t}$. Mathematically, we seek to

solve the following minimization problem [10]:

$$\min_{p_\theta(\boldsymbol{\theta})} \quad \sum_{k=1}^{N_t} \left\| \hat{p}_{\hat{y}(t_k)}(\hat{\boldsymbol{y}}(t_k)) - \hat{p}_{y(t_k)}(\boldsymbol{y}(t_k)) \right\|_2 \tag{6.5a}$$

$$\text{s.t.} \quad \dot{\mathbf{x}}^{(i)}(t_k) = \boldsymbol{f}\left(\mathbf{x}^{(i)}, \boldsymbol{\theta}^{(i)}, u; t_k\right) \tag{6.5b}$$

$$\mathbf{y}^{(i)}(t_k) = \mathbf{g}\left(\mathbf{x}^{(i)}, \boldsymbol{\theta}^{(i)}, u; t_k\right) \tag{6.5c}$$

$$\mathbf{x}^{(i)}(0) = \mathbf{x}_0^{(i)},\ \boldsymbol{\theta}^{(i)} \sim p_\theta(\boldsymbol{\theta}), \quad \forall i = 1, \ldots, M \tag{6.5d}$$

$$D_k = \{\mathbf{y}^{(1)}(t_k), \mathbf{y}^{(2)}(t_k), \cdots, \mathbf{y}^{(M)}(t_k)\} \tag{6.5e}$$

$$\int p_\theta(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1, \quad p_\theta(\boldsymbol{\theta}) \geq 0 \tag{6.5f}$$

where $\hat{p}_{\hat{y}(t_k)}(\hat{\boldsymbol{y}}(t_k)) : \mathbb{R}_+^{n_y} \to \mathbb{R}_+$ is the output PDF at time t_k based on the snapshot measurement $\hat{D}_k$, $\hat{p}_{y(t_k)}(\boldsymbol{y}(t_k)) : \mathbb{R}_+^{n_y} \to \mathbb{R}_+$ is the simulated output PDF obtained by solving Equations 6.5b-6.5c M times for the given $p_\theta(\boldsymbol{\theta})$, and D_k is the simulated snapshot measurement. The constraints (Equation 6.5f) ensure the estimated $p_\theta(\boldsymbol{\theta})$ is a PDF. It should be noted that $\hat{p}_{\hat{y}(t_k)}(\hat{\boldsymbol{y}}(t_k))$ and $\hat{p}_{y(t_k)}(\boldsymbol{y}(t_k))$ are estimated from the population snapshot datasets, $\hat{D}_k$ and D_k, respectively, through a PDF estimation method such as Gaussian mixture modeling.

Unfortunately, this minimization problem (Equation 6.5) is nontrivial to solve. First of all, the problem is infinite-dimensional since its decision variable is a function instead of a finite-dimensional vector [10, 242]. Usually, one has to rely on an approximation method to solve such an optimization problem. Second, it is expected that the PDFs of the entire parameter set are difficult to be estimated accurately from the measurements due to the unidentifiability of the PDFs of the model parameters [23, 47, 76]. Specifically, the PDFs of only a small subset of the model parameters, $\boldsymbol{\theta}_s \in \mathbb{R}^{n_{\theta_s}}$ where $n_{\theta_s} < n_\theta$, are often identifiable from the available experiments [23, 47, 243]. Therefore, even if a solution to Equation 6.5 is obtained through an approximation scheme, the solution to the minimization problem above is likely to be unreliable or subject to a large uncertainty. Lastly, solving the IBPM (Equations 6.5b-6.5e) is computationally demanding, so the overall minimization problem (Equation 6.5) requires an immense computational effort as

well.

In summary, this study proposes a new numerical scheme to address the following problem:

Problem. *Given the IBPM (Equation 6.1) and the measured population snapshot data,* $\hat{D}_k,\ k = 1, \ldots, N_t$*, estimate the underlying parameter probability density function,* $p_\theta(\boldsymbol{\theta})$*, accurately in a numerically efficient manner.*

6.3 Proposed methodology

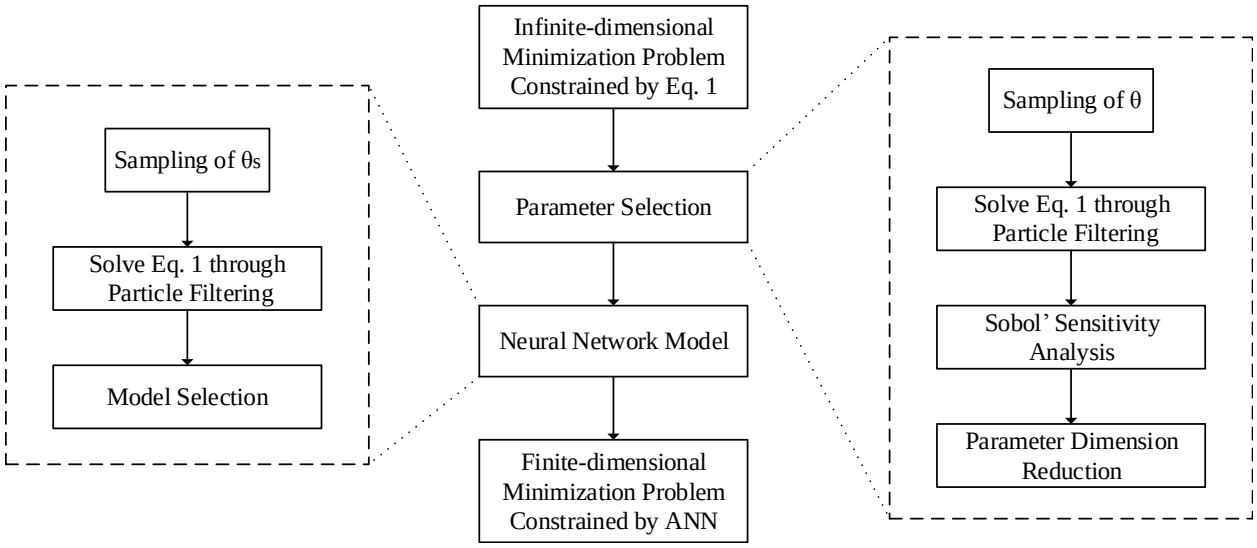


Figure 6.1: An overall flow diagram for describing the proposed methodology. Equation 6.1 refers to the IBPM.

In this study, we develop a systemic framework to efficiently solve the proposed PDF estimation problem (Equation 6.5) while considering the identifiability of $p_\theta(\boldsymbol{\theta})$. Specifically, this approach consists of dimension reduction of $p_\theta(\boldsymbol{\theta})$, which is the decision variable of Equation 6.5, and development of an ANN to replace the computationally expensive IBPM (Equations 6.5b-6.5e). The overall procedure is illustrated in Figure 6.1, and each step is described below.

6.3.1 Assumptions

For the purpose of this study, we make one assumption: the PDF of $\boldsymbol{\theta}$ is assumed to be a multivariate normal distribution [36]. The assumption is motivated by the two considerations. First, since a normal distribution is fully characterized by its mean and covariance matrix, the normal-distribution assumption allows the above minimization problem (Equation 6.5) to be reduced to the one with a finite number of variables (i.e., means, and standard deviations (SDs), and covariances of $\boldsymbol{\theta}$). Second, previous studies [42, 10] attempted to infer $p_\theta(\boldsymbol{\theta})$ through the PDF parameterization. Although these approaches resolved the infinite dimensionality of the problem, the implementation of such methodologies generally results in a large number of decision variables to be estimated, which is likely to exacerbate the unidentifiability problem. In contrast, the normal-distribution assumption renders the total number of decision variables limited to $M_s = 2n_\theta + \frac{n_\theta!}{2!(n_\theta-2)!}$. In the end, the normal-distribution assumption reduces the decision variables of Equation 6.5 to finite-dimensional vectors, $\boldsymbol{\mu} \in \mathbb{R}^{n_\theta}$ and $\boldsymbol{\sigma} \in \mathbb{R}^{n_\theta + \frac{n_\theta!}{2!(n_\theta-2)!}}$, from an infinite-dimensional function, $p_\theta(\boldsymbol{\theta})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the vector of means and the one with SDs and covariances of $\boldsymbol{\theta}$, respectively.

6.3.2 Parameter selection

The first step of the proposed methodology is the parameter selection [156], where the dimension of $\boldsymbol{\theta}$ is reduced through identifiability analysis. Specifically, its goal is to select a subset of $\boldsymbol{\theta}$ whose PDFs can be accurately estimated from the population snapshot data. For deterministic modeling approaches that utilize ODEs, this procedure is fairly well established with various methods proposed in the past [23, 34, 47]. However, the identifiability of an IBPM has not been examined as thoroughly as that of a deterministic model. Only available method for the IBPM is proposed by Zeng et al. [48, 50], but their method is only applicable for a linear IBPM while the most of the models for intracellular processes are nonlinear. Therefore, their method is not implemented in this study.

Under this circumstance, this study performs sensitivity analysis to determine the most important parameters and assumes that these parameters' PDFs are identifiable. Specifically, the Sobol'

method [193], a global sensitivity analysis method, is implemented. Through the sensitivity analysis, a parameter is chosen to be estimated if 1) it has a significant impact on ouputs and 2) it has minimum correlations with other parameters. Theses two criteria will be evaluated through the Sobol' sensitivity index [23, 191].

For an ODE-based deterministic model, the parameter identifiability is often evaluated numerically via a sensitivity analysis method, which assesses the significance of each model parameter by computing the change in the model output due to the change in the parameter value [195]. The Sobol' method is appropriate for the purpose of this study since results from the Sobol' method are global and, thus, more reliable when the parameter values are largely uncertain or subject to PDFs [197, 243].

In an ODE modeling setting, the Sobol sensitivity index of a parameter θ_i is computed based on the variance of model outputs $\mathbf{y}$, whose dimension is usually finite (i.e., n_y). However, since the output and parameters of an IBPM are both PDFs (i.e., $\hat{p}_{y(t_k)}(\boldsymbol{y}(t_k)|p_\theta(\boldsymbol{\theta}))$ and $p_\theta(\boldsymbol{\theta})$, respectively), the Sobol' sensitivity index of an IBPM should be computed based on the variance of $\hat{p}_{y(t_k)}(\boldsymbol{y}(t_k)|p_\theta(\boldsymbol{\theta}))$ with respect to the change in $p_\theta(\boldsymbol{\theta})$. Specifically, the Monte Carlo method proposed by Homma and Saltelli [196] is implemented along with the discretization of the output dimension. The detailed procedures are explained as follows:

1. Two matrices ($\mathbf{A}$ and $\mathbf{B} \in \mathbb{R}^{N_s \times M_s}$) are generated randomly via a Sobol' sequence to sample $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ [193, 195], where N_s is the Monte Carlo sample size and M_s is the total number of statistical parameters (i.e., means, SDs, and covariances) needed to be considered for a system with n_θ model parameters ($M_s = 2n_\theta + \frac{n_\theta!}{2!(n_\theta-2)!}$).

2. Another set of matrices $\mathbf{C}_j$, $\forall j = 1, \ldots, M_s$, is constructed by substituting the j^{th} column of $\mathbf{B}$ by the j^{th} column of $\mathbf{A}$.

3. Determine $p_\theta(\boldsymbol{\theta})$ and the corresponding $\hat{p}_{y(t_k)}(\boldsymbol{y}(t_k)|p_\theta(\boldsymbol{\theta}))$, $\forall k = 1, \ldots, N_t$, based on each row vector of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}_j$ through the particle filtering and Gaussian mixture model approximation (GMM) (see Supplementary Materials for the descriptions on these methods).

4. Discretize the dimension of each output, $y_m, m = 1, \ldots, n_y$, into M_1 uniformly spaced points, $\mathcal{Y}_m = \{y_m^1, \ldots, y_m^{M_1}\}$, and create a meshgrid matrix, $\boldsymbol{\mathcal{Y}} \in \mathbb{R}^{M_1^{n_y} \times n_y}$. Here, each row vector of $\boldsymbol{\mathcal{Y}}$, which is denoted as $\boldsymbol{\mathcal{Y}}_i$, contains discrete points in the n_y-dimensional output space.

5. Evaluate each $\hat{p}_{y(t_k)}(\boldsymbol{y}(t_k)|p_\theta(\boldsymbol{\theta}))$ obtained in Step 3, and compute the sensitivity index at the discretized points, $\boldsymbol{\mathcal{Y}}_i = \{y_1^{i_1}, y_2^{i_2}, \ldots, y_{n_y}^{i_{n_y}}\}$, $\forall\, i_1, \cdots, i_{n_y} = 1, \ldots, M_1$, at each sampling time as follows:

$$S_{ij}(\boldsymbol{\mathcal{Y}}_i(t_k)) \approx 1 - \frac{1/N_s \sum_{l=1}^{N_s} p\left(\boldsymbol{\mathcal{Y}}_i(t_k)|\boldsymbol{\mathcal{N}}(\mathbf{b}^{(l)})\right) p\left(\boldsymbol{\mathcal{Y}}_i(t_k)|\boldsymbol{\mathcal{N}}(\mathbf{c}_j^{(l)})\right) - \bar{p}_i^2\left(\boldsymbol{\mathcal{Y}}_i(t_k)|\boldsymbol{\mathcal{N}}(\mathbf{a}^{(l)})\right)}{1/N_s \sum_{l=1}^{N_s} \left(p\left(\boldsymbol{\mathcal{Y}}_i(t_k)|\boldsymbol{\mathcal{N}}(\mathbf{a}^{(l)})\right)\right)^2 - \bar{p}_i^2\left(\boldsymbol{\mathcal{Y}}_i(t_k)|\boldsymbol{\mathcal{N}}(\mathbf{a}^{(l)})\right)} \tag{6.6}$$

where

$$\bar{p}_i\left(\boldsymbol{\mathcal{Y}}_i(t_k)|\boldsymbol{\mathcal{N}}(\mathbf{a}^{(l)})\right) = \frac{1}{N_s}\sum_{l=1}^{N_s} p\left(\boldsymbol{\mathcal{Y}}_i(t_k); \boldsymbol{\mathcal{N}}(\mathbf{a}^{(l)})\right),\ \forall i = 1 \ldots, M_1^{n_y} \tag{6.7}$$

and $\boldsymbol{\mathcal{N}}(\mathbf{a}^{(l)}), \boldsymbol{\mathcal{N}}(\mathbf{b}^{(l)})$, and $\boldsymbol{\mathcal{N}}(\mathbf{c}_j^{(l)})$ are the multivariate normal PDFs defined by the l^{th} rows of the matrices $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}_j$, respectively.

6. An effective sensitivity index, ST_j of the j^{th} statistical parameters of $p_\theta(\boldsymbol{\theta})$ now can be defined over all the sampling times and all the points in $\boldsymbol{\mathcal{Y}}$:

$$ST_j = \frac{1}{N_t M_1^{n_y}} \sum_{i=1}^{M_1^{n_y}} \sum_{k=1}^{N_t} S_{ij}(\boldsymbol{\mathcal{Y}}_i(t_k)) \tag{6.8}$$

Here, the value of M_s equals to $2n_\theta + \frac{n_\theta!}{2!(n_\theta-2)!}$, since a system with n_θ number of parameters has n_θ number of means and SDs, and $\frac{n_\theta!}{2!(n_\theta-2)!}$ number of covariances between the model parameters.

In this procedure, $p_\theta(\boldsymbol{\theta})$ can be fully characterized by each row vector of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}_j$ since it contains sampled means and covariance matrix of $\boldsymbol{\theta}$. Based on the Sobol' sensitivity indices, a subset of important statistical parameters, $\boldsymbol{\phi} \in \mathbb{R}^{n_s},\ n_s < M_s$, can be identified. In this work, we assume only the model parameters, whose at least one of their statistical parameters are included in $\boldsymbol{\phi}$, are normally distributed, which will be denoted as $\boldsymbol{\theta}_s \in \mathbb{R}^{n_{\theta_s}},\ n_{\theta_s} \leq n_\theta$, hereafter. In contrast,

the remaining parameters are fixed at their nominal values (i.e., they are not subject to PDFs). Consequently, the decision variables of the minimization problem (Equation 6.5)) are reduced to a vector whose dimension is n_s.

It should be noted that this study employs the total sensitivity index as a measure to assess the two criteria of identifiability. The total sensitivity index, ST, can be decomposed into two terms: the first-order and higher-order sensitivities. The first-order sensitivity quantifies the direct effect of a statistical parameter, which computes its significance. On the other hand, the higher-order sensitivity computes its correlations with other parameters. Therefore, a parameter should have a high magnitude of the first-order sensitivity with a low higher-order sensitivity.

6.3.3 Neural network model

Through the normal-distribution assumption and the parameter selection, the minimization problem becomes finite dimensional. However, it is still numerically challenging to solve it directly. The difficulty mainly stems from the fact that it has the IBPM (Equations 6.5b-6.5e) as one of its constraints. Motivated by the above considerations, an ANN (i.e., surrogate model) is developed to approximate the IBPM. In this study, the ANN approach is chosen over other surrogate modeling approaches because ANNs can approximate any arbitrary input-output relations with sufficient accuracy as long as a dataset with a sufficient size is provided to train the ANN [227, 228].

In order to ensure the developed ANN can accurately represent the IBPM (Equations 6.5b-6.5e), a sufficient number of datasets are needed for training the ANN. The procedures for generating the training datasets are listed below:

1. A matrix $\bar{A} \in \mathbb{R}^{N_n \times n_s}$ is generated via Sobol' sequence to obtain samples of $\boldsymbol{\phi}$, where N_n is the sample size.

2. Discretize the parameter space of each parameter in $\boldsymbol{\theta}_s$ into M_2 uniformly spaced points, $\mathcal{P}_j = \{\theta_j^1, \ldots, \theta_j^{M_2}\},\ \forall j = 1, \ldots, n_{\theta_s}$, and create a meshgrid matrix, $\boldsymbol{\mathcal{P}} \in \mathbb{R}^{M_2^{n_{\theta_s}} \times n_{\theta_s}}$. Here, each row vector of $\boldsymbol{\mathcal{P}}$, which is denoted as $\boldsymbol{\mathcal{P}}_j$, contains a discrete point in the n_{θ_s}-

dimensional parameter space.

3. Evaluate each $p_{\theta_s}(\boldsymbol{\theta}_s)$ characterized by each row of $\bar{\mathbf{A}}$ at their discretized grid points, $\boldsymbol{\mathcal{P}}_j$, and stack the evaluated values side-by-side to form the input dataset matrix, $\bar{\mathbf{B}} \in \mathbb{R}^{N_n \times (M_2^{n_{\theta_s}} \times n_{\theta_s})}$.

4. The particle filtering method and GMM (see Supplementary Materials) are implemented sequentially to estimate N_n different $\hat{p}_{y(t_k)}(\boldsymbol{y}(t_k)|p_{\theta_s}(\boldsymbol{\theta}_s)))$, $\forall k = 1, \ldots, N_t$, corresponding to N_n row vectors of $\bar{A}$.

5. Every set of $\{\hat{p}_{y(t_k)}(\boldsymbol{y}(t_k)|p_{\theta_s}(\boldsymbol{\theta}_s))\}_{k=1}^{N_t}$ corresponding to each row of $\bar{\mathbf{A}}$ is evaluated at $\boldsymbol{\mathcal{Y}}$ to construct a matrix $\bar{\mathbf{C}} \in \mathbb{R}^{N_n \times (N_t \times M_1^{n_y})}$ as the output dataset.

Here, it should be noted that all the remaining parameters, which do not belong to the set $\boldsymbol{\theta}_s$, are fixed at their nominal values while generating the training datasets. During the model training, the training datasets (both $\bar{\mathbf{B}}$ and $\bar{\mathbf{C}}$) are divided into training, validation, and testing sets [244].

In this study, the structure of ANN is restricted to that with only one hidden layer, whose neuron has the hyperbolic tangent sigmoid transfer function, and one linear output layer. For each neuron in the hidden layer, the following function is used:

$$o_i = \frac{2}{1 + e^{-2\hat{u}_i}} - 1, \quad \hat{u}_i = \sum_{j=1}^{N_u} \left(w_{ij} \cdot u_j + b_j \right), \quad \forall i = 1, \ldots, N$$

where o_i is the output of the i^{th} neuron in the hidden layer, $\hat{u}_i$ is the weighted sum of inputs applied to the i^{th} neuron in the hidden layer of the ANN, N_u is the number of inputs given to the ANN, w_{ij} is the weight given to the input u_j at the i^{th} neuron, u_j is the j^{th} input of the ANN, b_j is a bias term given to the i^{th} neuron, and N is the number of neurons in the hidden layer.

On the other hand, the output of the ANN is calculated as follows:

$$O_k = \sum_{i=1}^{N} \left(W_{ki} o_i + B_k \right), \forall k = 1, \ldots, N_o$$

where O_k is the k^{th} output of the ANN, N is the number of neurons in the hidden layer, W_{ki} is the

weight term given to o_i for O_k, B_k is a bias term given to O_k, and N_o is the number of outputs of the ANN. As a result, the total number of hyper-parameters used in the ANN is $N_u \cdot N + N + N \cdot N_o + N_o = N(N_u + 1) + N_o(N + 1)$.

In order to determine the optimal number of neurons in the hidden layer, ANNs with the different number of hidden-layer neurons are trained, and their expected AIC_c ($E(\text{AIC})_c$) values (see Supplementary Materials) are computed. Specifically, an ANN for a given number of hidden-layer neurons is trained N_{tr} times with partitioning the training dataset differently to compute its $E\left(\text{AIC}_c\right)$ values. Then, the ANN with the smallest $E\left(\text{AIC}_c\right)$ is chosen as the optimal one.

6.3.4 Re-formulation of PDF estimation problem

In summary, the proposed methodology reduces the original infinite-dimensional and computationally expensive problem to a finite-dimensional and numerically tractable one through the parameter selection and the ANN development. Through this proposed methodology, the original minimization problem (Equation 6.5) is reformulated as follows:

$$\min_{\boldsymbol{\phi}} \quad \sum_{k=1}^{N_t} \left\| \hat{p}_{\hat{y}(t_k)}(\hat{\boldsymbol{y}}(t_k)) - \hat{p}_{y(t_k)}(\boldsymbol{y}(t_k)) \right\|_2 \tag{6.9a}$$

$$\text{s.t.} \quad \hat{p}_{y(t_k)}\left(\boldsymbol{y}(t_k)\right)\right) \approx h\left(\boldsymbol{\phi}\right), \ \ \forall k = 1, \ldots, N_t \tag{6.9b}$$

$$\boldsymbol{\phi}^{lb} \leq \boldsymbol{\phi} \leq \boldsymbol{\phi}^{ub} \tag{6.9c}$$

where $h(\cdot)$ is the developed ANN that predicts the output distribution for the given PDF of parameters characterized by $\boldsymbol{\phi}$. It should be noted that the l-2 norm in the objective function (Equation 6.9a) is evaluated via the standard trapezoid rule [10]. In this study, Equation 6.9a is solved 50 times with 50 different initial guesses via *fmincon* along with Multistart in Matlab to minimize the chance of obtaining a suboptimal local solution. The allowed number of iteration and the tolerance on the constraint violation are set to be 10^4 and 10^{-12}, respectively, to achieve a high accuracy.

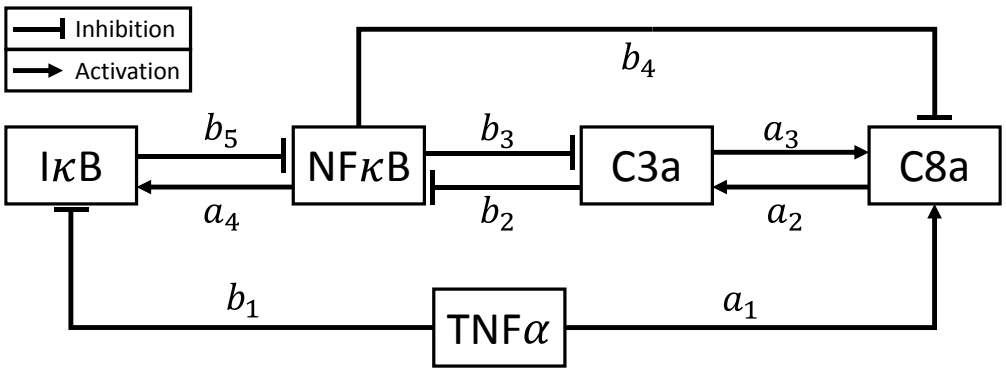


Figure 6.2: Schematic diagram for the TNFα signaling pathway. Adapted from Hasenauer et al. [10].

6.4 Results: Application to TNFα Signaling Pathway

Here, our proposed methodology was validated by implementing it in two different scenarios. First, the PDFs of model parameters were inferred when only one measurement was available and the correlation parameters were not considered. Second, the PDFs of the model parameters were estimated when two measurements under the consideration of the parameter correlation.

6.4.1 TNFα Signaling Pathway

We implemented and validated the proposed methodology by estimating the PDFs of model parameters in the TNFα signaling pathway model [245] based on *in silico* measurement data. The mechanisms of co-activation of pro- and anti-apoptotic pathways by TNFα were modeled by Chaves et al. [245], which is illustrated in Figure 6.2. This model describes the dynamic interaction among four key proteins, which are NFκB, active caspase 8 (C8a), active caspase 3 (C3a), and IκB, with TNFα as an external stimulus to the system. Through modeling the dynamics and interactions among these proteins, the model represents a simplified version of three regulatory networks present in the overall TNFα signaling pathway: the anti-apopotic NFκB network, the pro-apoptotic caspase activation network, and the crosstalks between these two pathways [246]. A detailed discussion on this system as well as its dynamic model is presented in [245, 246].

The corresponding ODE model is formulated as follows [42]:

$$\begin{aligned}
\dot{x}_1 &= -x_1 + \frac{1}{2}\left(inh_4(x_3) \cdot act_1(u) + act_3(x_2)\right) \\
\dot{x}_2 &= -x_2 + act_2(x_1) \cdot inh_3(x_3) \\
\dot{x}_3 &= -x_3 + inh_2(x_2) \cdot inh_5(x_4) \\
\dot{x}_4 &= -x_4 + \frac{1}{2}\left(inh_1(u) + act_4(x_3)\right)
\end{aligned} \tag{6.10}$$

where $x_i,\ i = 1, \ldots, 4$ represent the relative activities of C8a, C3a, NFκB, and IκB, respectively, and u represents the TNFα. The initial concentrations $\mathbf{x}_0 = [x_1(0)\ x_2(0)\ x_3(0)\ x_4(0)]^T$ are the steady states of Equation 6.10 with $u = 0$. Also, the functions, inh_i and act_i, in the model are rational functions given by:

$$\begin{aligned}
inh_i(x_j) &= \frac{x_j^2}{a_i^2 + x_j^2} \\
act_i(x_j) &= \frac{b_i^2}{b_i^2 + x_j^2}
\end{aligned} \tag{6.11}$$

where $a_i,\ i = 1, \ldots, 5$, and $b_i,\ , i = 1, \ldots, 4$, are the model parameters ranging from 0 and 1 (Table 6.1) [42]. So, the parameter vector of this model is $\boldsymbol{\theta} = \begin{bmatrix} a_1 & \ldots & a_5 & b_1 & \ldots & b_4 \end{bmatrix}^T$, and their nominal values are listed in Table 6.1.

Parameter	Value	Parameter	Value
a_1	0.6	b_1	0.4
a_2	0.2	b_2	0.7
a_3	0.2	b_3	0.3
a_4	0.5	b_4	0.5
		b_5	0.4

Table 6.1: Nominal parameter values of the TNFα signaling model

6.4.2 Parameter Selection Results

We started with the case where only one measurement was available (i.e., $n_y = 1$) and parameter correlations were not considered (i.e, $M_s = 2n_\theta$). In order to assess the identifiability of $p_\theta(\boldsymbol{\theta})$ in the TNFα signaling pathway (Equation 6.10), the extended Sobol' method was implemented as described above. Specifically, since the dimension of $\boldsymbol{\theta}$ was nine, significance of their eighteen statistical moments (i.e., first and second moments of each model parameter) was evaluated by computing their sensitivity indices. In this study, the sample size of 1000 was used to perform the Monte Carlo evaluation of the sensitivity indices (i.e., $N_s = 1000$), and $p(y)$ was evaluated at 100 uniformly distributed grid points in the y dimension (i.e., $M_1 = 100$).

Figure 6.3 shows the results of the Sobol' sensitivity analysis. Here, a_2 represents the rate constant of the C8a-induced C3a activation. It is evident that values of sensitivity indices were quite close to each other except those of a_2's mean and SD. This suggests that their ST values were dominated by their higher-order sensitivities, which indicates that these parameters were highly correlated and not identifiable. Based on this result, the statistical moments of a_2 were determined to be most influencing, and it was assumed that only a_2 follows the normal distribution while the remaining model parameters (i.e., $a_1, a_3, a_4, b_1, \ldots, b_4$) were fixed at their values shown in Table 6.1. Consequently, the decision variables of the PDF estimation problem were reduced to the mean and SD of a_2 from entire eighteen means and SDs of the model parameters.

6.4.3 Neural Network Model Development

For developing an ANN, 2000 pairs of $\{\mu_{a_2}, \sigma_{a_2}\}$ values were randomly sampled (i.e., $N_n = 2000$) to obtain $\bar{\mathbf{A}}$, and the corresponding output PDFs were approximated through the particle filtering. Then, these inferred output PDFs were evaluated at 100 uniformly distributed grid points in the y dimension to construct the output dataset matrix, $\bar{\mathbf{C}} \in \mathbb{R}^{2000\times(100\times N_t)}$. At the same time, the parameter space of a_2 was discretized into 200 uniformly spaced grid points (i.e., $M_2 = 200$) to construct the input dataset matrix, $\bar{\mathbf{B}} \in \mathbb{R}^{2000\times 200}$.

ANNs with various numbers of neurons in the hidden layer were trained 100 times by the train-

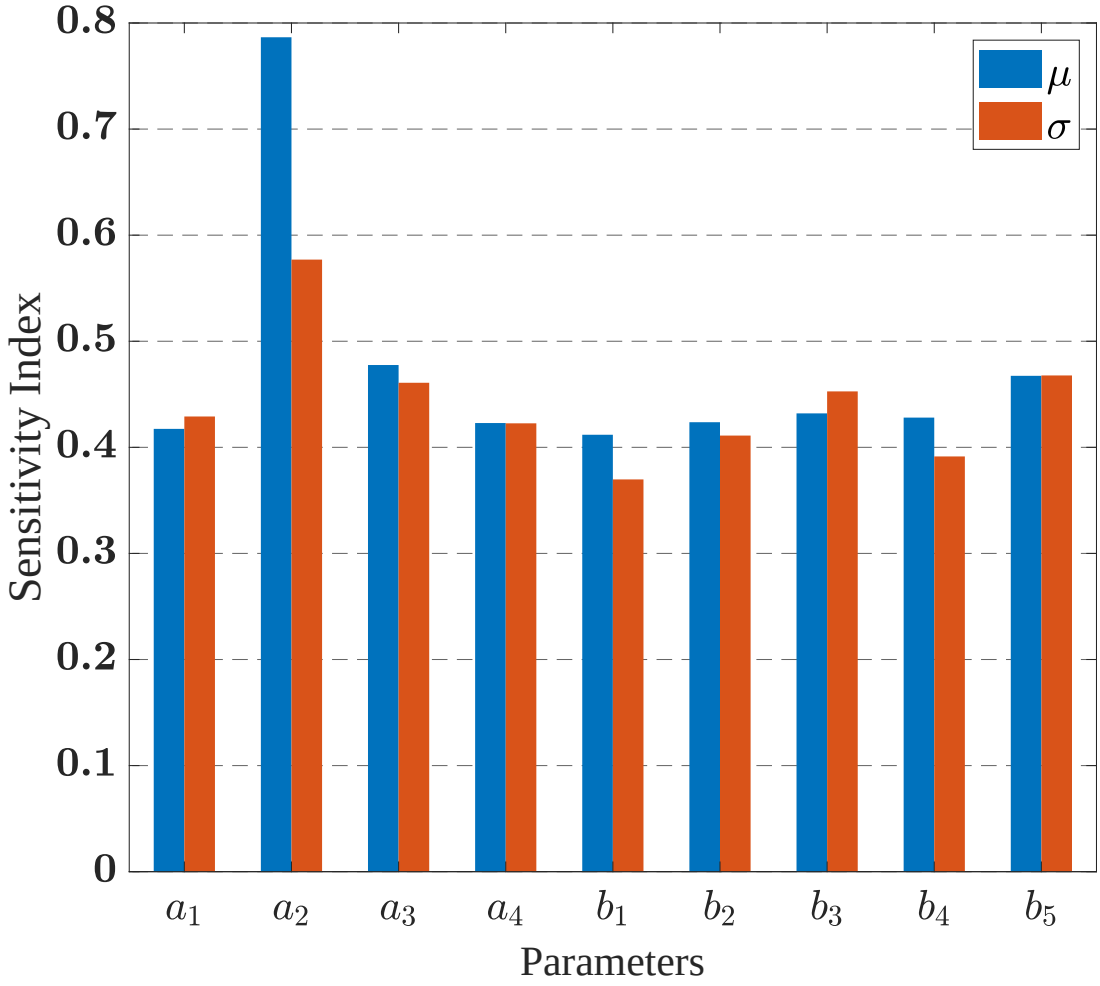


Figure 6.3: Sensitivity Analysis Results. Sensitivity indices of the first and second moments (μ and σ, respectively) of the model parameters in the TNFα signaling pathway.)

ing dataset with different partitions to compute $E(\text{AIC}_c)$ [244]. Here, all computations relevant to the ANN development were performed in the MATLAB Neural Network Toolbox. Figure 6.4 shows the $E(\text{AIC}_c)$ values of ANNs with different sizes. Based on Figure 6.4, the ANN with 11 neurons in the hidden layer was chosen since the $E(\text{AIC}_c)$ value reached the minimum value and remained unchanged after it. Based on the current structure of the ANN, the number of hyper-parameters used in the ANN is $N(N_u+1)+N_o(N+1) = 11\times(200+1)+1000\times(11+1) = 14211$. The R^2 statistics for the final ANN model were 0.984 with respect to the training dataset and 0.978

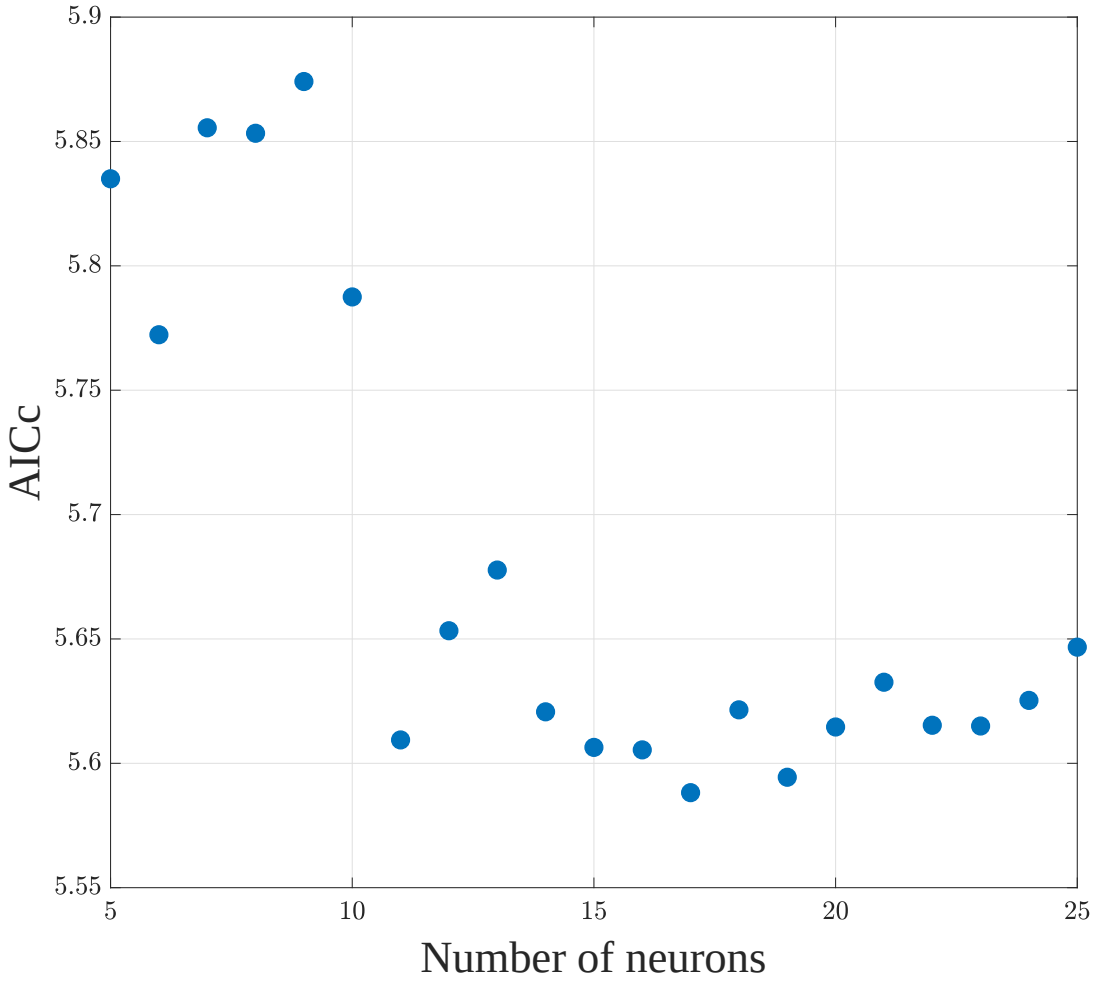


Figure 6.4: Change of $E(\text{AIC})_c$ values with the change in the number of neurons in the hidden layer.

with respect to the validation dataset.

6.4.4 *In Silico* Experimental Measurement

From the previous studies, it has been shown that the TNFα signaling exhibits a high degree of cell-to-cell variability [37, 246, 247]. To generate artificial experimental measurements, we assumed a_4 and b_3 to be log-normally distributed and mutually independent. Then, the activity of

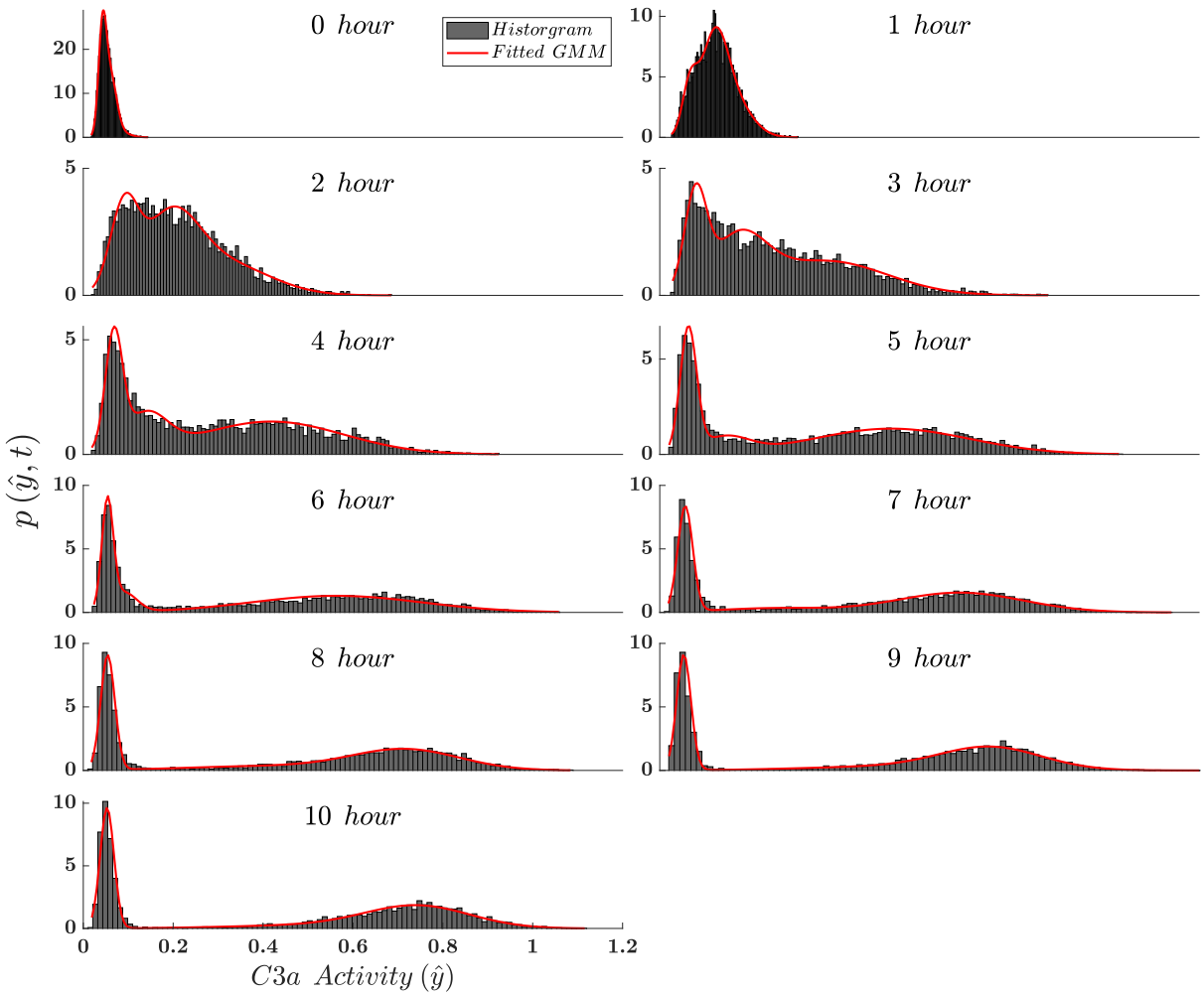


Figure 6.5: Artificial flow cytometry measurements and the Gaussian Mixture model (GMM) fitted to the flow cytometry measurements.

C3a (x_2) was measured every hour from 0 to 10 hours subject to the following TNFα stimulation:

$$u(t) = \begin{cases} 1, & \forall\, t \in [0,\ 2] \\ 0, & \forall\, t > 2 \end{cases} \tag{6.12}$$

At each sampling point, t_k, x_2 was measured from 5000 cells [8, 10], and the measurements were subsequently corrupted with noise (Equation 6.3). Specifically, the proportional factor (c) and measurement noise (ν) were log-normally distributed, and their statistics were known *a priori*

from the equipment calibration, where μ_c, σ_c, μ_ν, and σ_ν were 0, 0.1, -3, and 0.3, respectively.

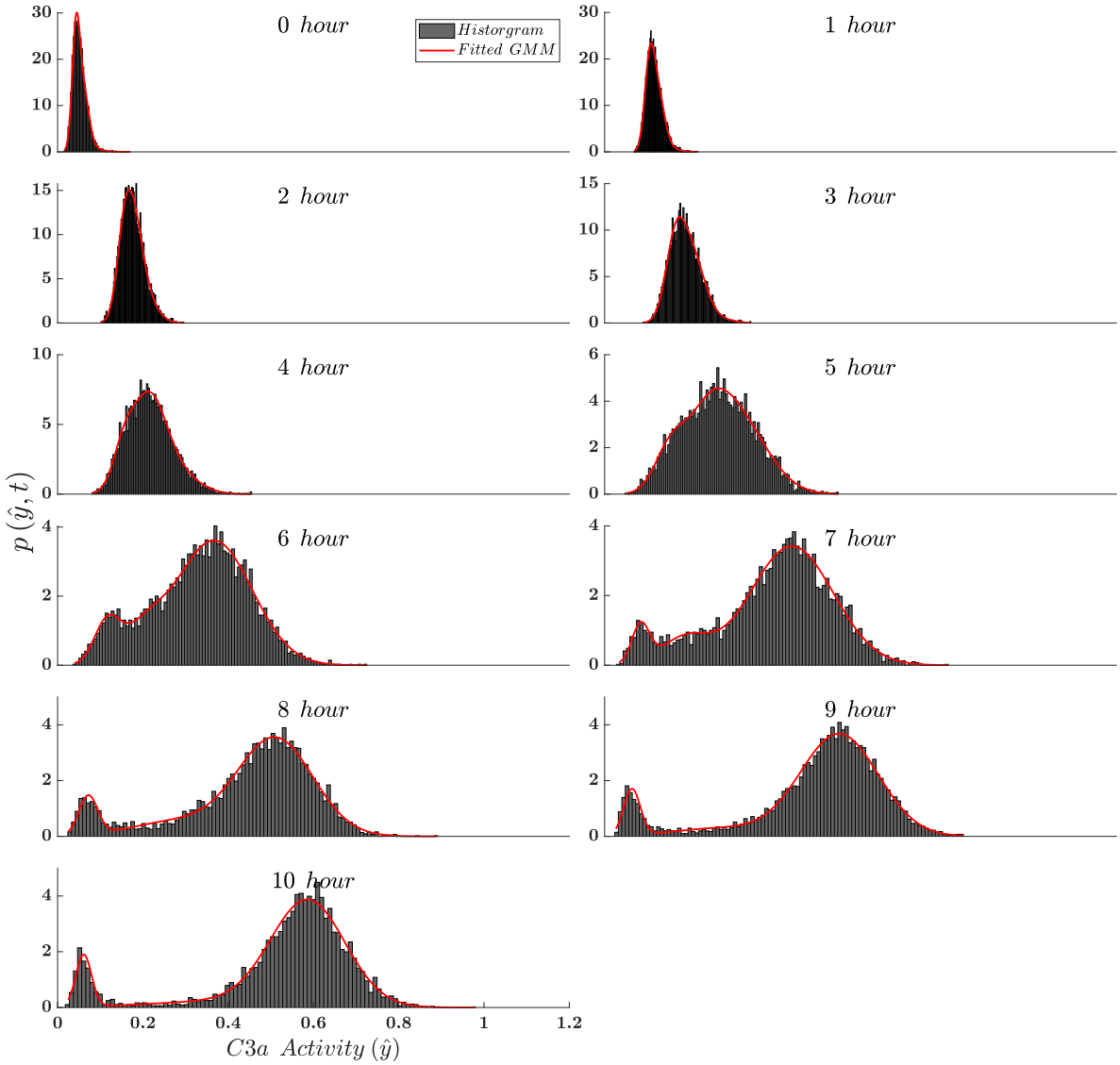


Figure 6.6: The second set of an artificial flow cytometry measurements and GMM fitted to the flow cytometry measurements.

Figure 6.5 illustrates an exemplary experimental measurement when the natural logarithms of a_4 and b_3 were distributed according to $\mathcal{N}(0.5, 0.2)$ and $\mathcal{N}(0.3, 0.15)$, respectively. Figure 6.5 indicates that the originally homogeneous cell population (i.e., $t = 0$) evolved into a heterogeneous one with two distinct subsets characterized by the low and high levels of C3a activities.

Additionally, it also shows that the GMM smoother performed well to estimate the underlying PDF of the obtained population snapshot data, which justified the use of the GMM smoother for approximating the underlying PDFs.

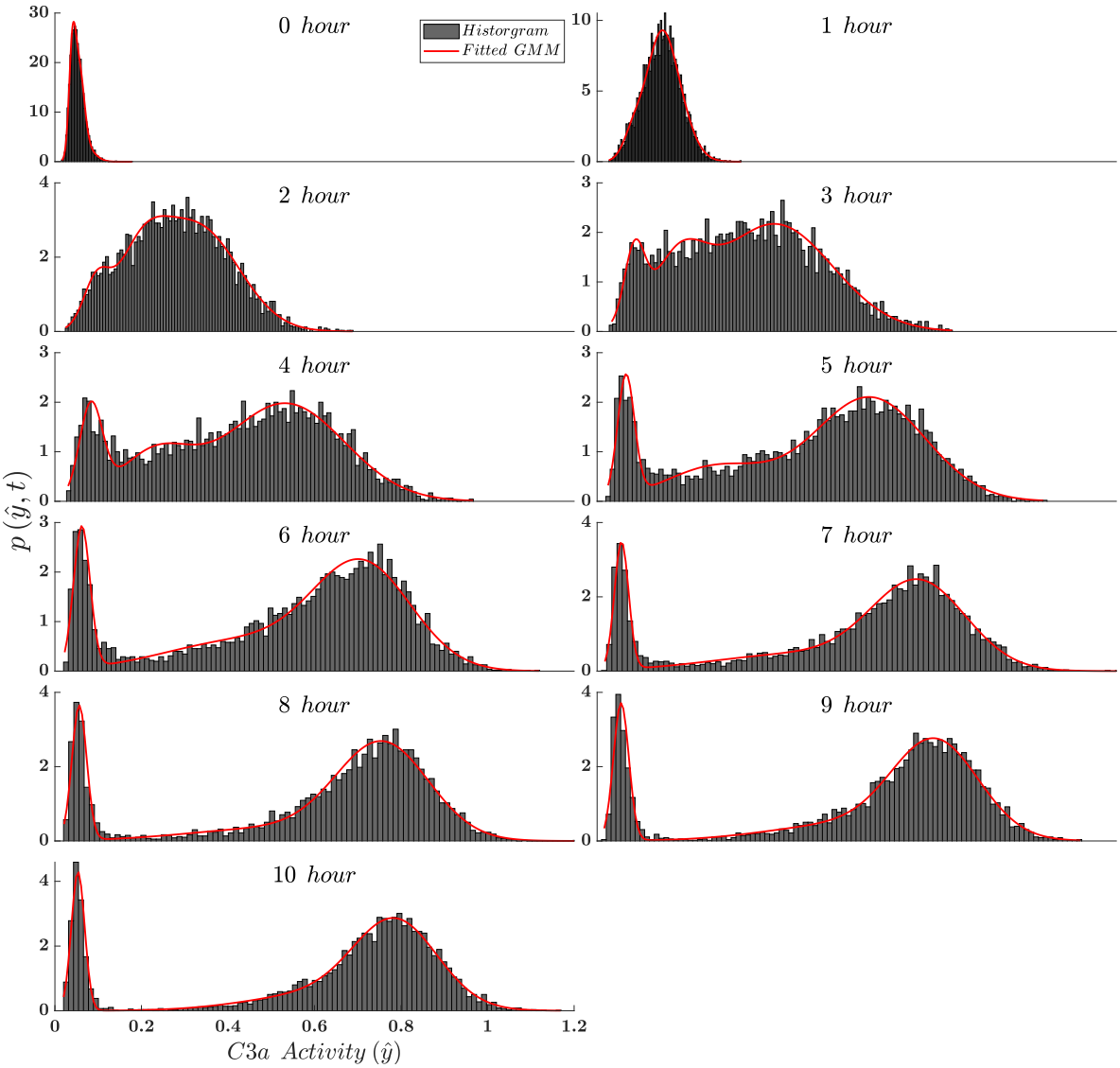


Figure 6.7: The third set of an artificial flow cytometry measurements and GMM fitted to the flow cytometry measurements.

To test the robustness of the proposed approach, two additional datasets were generated with different PDFs for a_4 and b_3 (Figures 6.7-6.7). The proposed scheme was implemented to infer the

PDFs of a_2 from these three datasets (i.e., Figures 6.5-6.7).

6.4.5 Parameter Estimation Result

The minimization problem (Equation 6.9) was solved to estimate the PDF of a_2 ($p(a_2)$) for each of the three datasets generated in the previous section (Figures 6.5-6.7), and the means and SDs of the estimated PDFs are listed in Table 6.2. All the calculations were performed in MATLAB via functions *fmincon* with *multistart* to solve Equation 6.9 with 50 different initial conditions [191]. To assess the validity of the estimated PDFs, the corresponding $\hat{p}_{y(t_k)}(y(t_k))$ was computed and plotted in Figures 6.8-6.10. Overall, the predicted output PDFs were in qualitative and quantitative agreement with the *in silico* experimental measurements; especially, the IBPM with the inferred $p(a_2)$ was able to predict the temporal evolution from the unimodal distribution at $t = 0$ to the bimodal distribution at later time points.

Dataset used for the estimation	μ	σ
Figure 6.5	0	0.26
Figure 6.6	0.19	0.03
Figure 6.7	0	0.17

Table 6.2: Estimated mean and SD of a_2

6.4.6 When Two Measurements are Available

Lastly, we implemented the proposed methodology to a system with two measurements, where parameters were possibly correlated with each other. Specifically, it was assumed that the activities of C3a and IκB (i.e., x_2 and x_4) were measured every hour from 0 to 10 hours subject to the same TNFα profile (Equation 6.12). To this end, a_4 and b_3 were log-normally distributed, and the noise for both both of two measurements were also log-normally distributed, whose statistics were shown

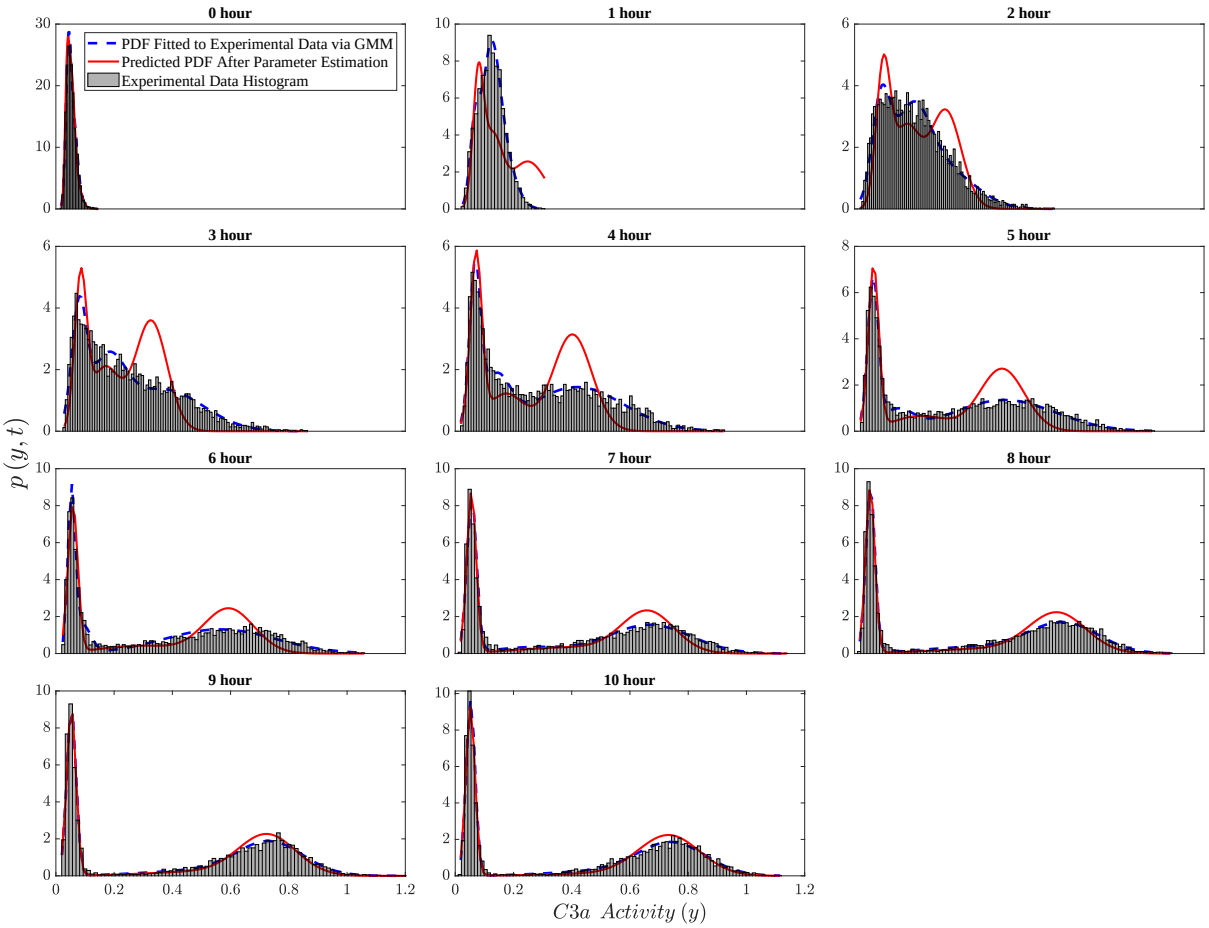


Figure 6.8: The PDF of a_2 was inferred from the artificial experimental data shown in Figure 6.5, and the predicted output PDFs with the inferred PDF of a_2 were plotted with the original histograms and the fitted PDFs.

as follows:

$$\begin{bmatrix} \mu_c, & \sigma_c, & \mu_\nu, & \sigma_\nu \end{bmatrix} = \begin{cases} \begin{bmatrix} 0 & 0.1 & -3 & 0.3 \end{bmatrix} \text{ for } & x_2 \\ \begin{bmatrix} 0 & 0.05 & -1.9 & 0.25 \end{bmatrix} \text{ for } & x_4 \end{cases} \tag{6.13}$$

Figure 6.11 illustrates the approximated PDFs of the *in silico* experimental measurements when the natural logarithms of a_4 and b_3 were distributed according to $\mathcal{N}(0.5, 0.2)$ and $\mathcal{N}(0.3, 0.15)$, respectively. As before, the cell population eventually evolves into a heterogeneous population with two distinct subsets in terms of C3a activity while the cell population is relatively more homogeneous in terms of IκB activity.

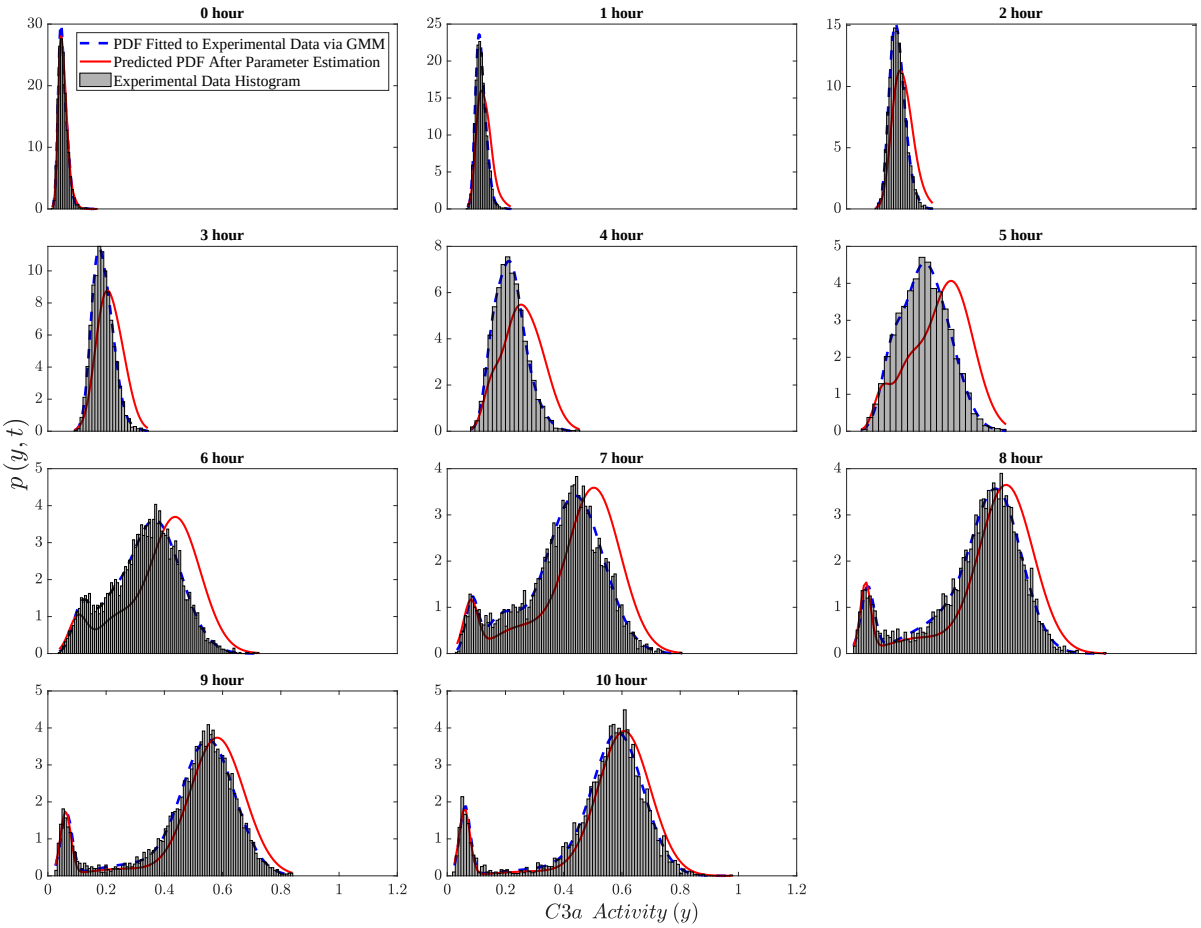


Figure 6.9: The PDF of a_2 was inferred from the artificial experimental data shown in Figure 6.6, and the predicted output PDFs with the inferred PDF of a_2 were plotted with the original histograms and the fitted PDFs.

For assessing the identifiability of $p_\theta(\boldsymbol{\theta})$ with respect to the two available measurements (i.e., $n_y = 2$), the sensitivity analysis was implemented as described earlier with $N_s = 500$ and $M_1 = 50$, and its results are shown in Figure 6.12. From Figure 6.12, it is clear that the means of a_2 and a_4 are the most important statistical parameters. Therefore, we assumed that only a_2 and a_4 were subject to PDFs while the remaining parameters were fixed at their nominal values. Also, since none of the covariances between the model parameters was selected through the sensitivity analysis, it was assumed that the PDFs of a_2 and a_4 are two mutually independent normal distributions since the correlation between a_2 and a_4 will not influence the PDFs of the two outputs significantly. It should be noted that the variances of a_2 and a_4 were fixed at 0.5 for the subsequent

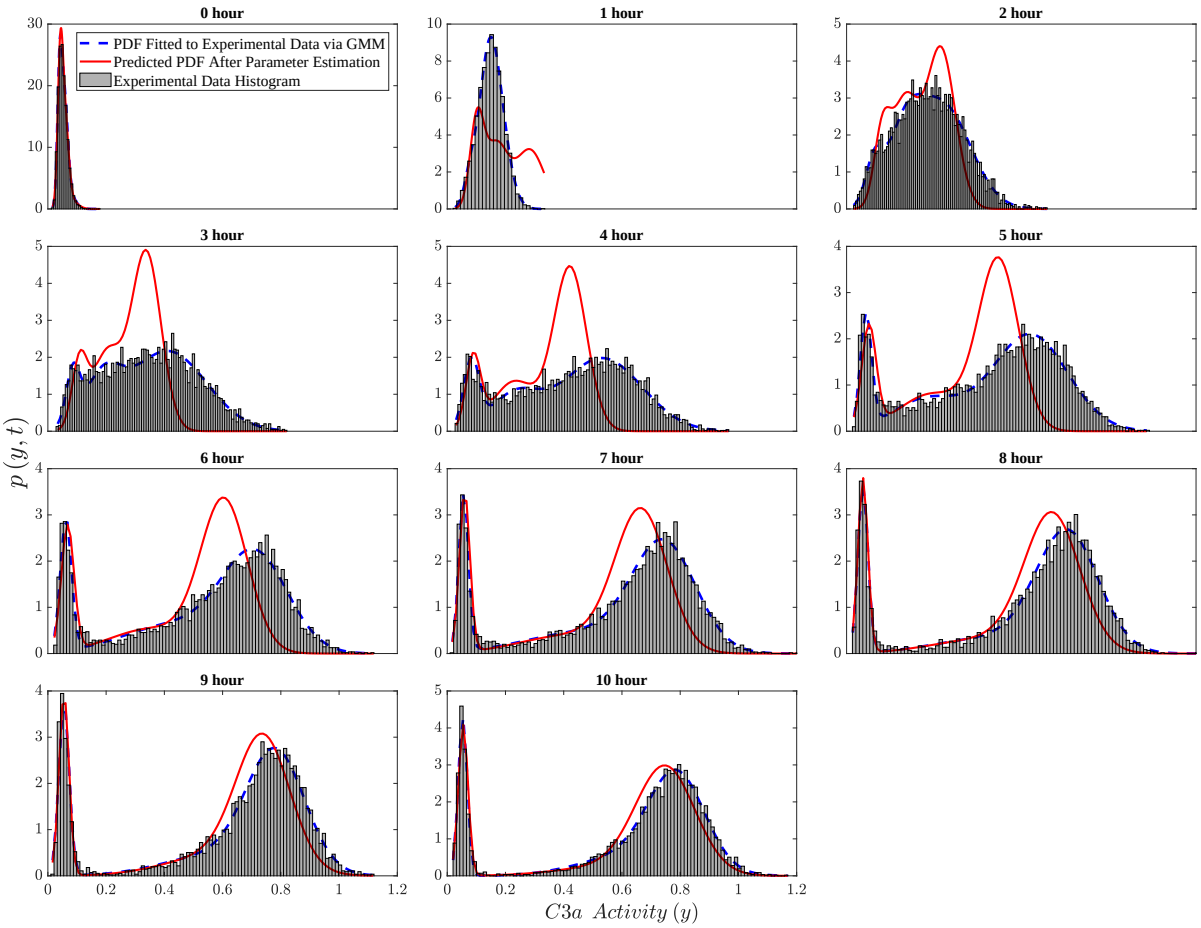


Figure 6.10: The PDF of a_2 was inferred from the artificial experimental data shown in Figure 6.7, and the predicted output PDFs with the inferred PDF of a_2 were plotted with the original histograms and the fitted PDFs.

ANN development and parameter estimation.

As the means of a_2 and a_4 were identified as the most influencing factors, the corresponding ANN was developed accordingly. To this end, 5000 pairs of $\{\mu_{a_2},\ \mu_{a_4}\}$ values were randomly sampled (i.e., $N_n = 5000$) to obtain $\bar{\mathbf{A}}$, and the corresponding output dataset matrix $\bar{\mathbf{C}}$ was obtained with $M_1 = 10$. With the obtained training datasets, the optimal number of neurons in the hidden layer was determined by finding the minimum $E(\text{AIC}_c)$ values. Figure 6.13 plots the values of $E(\text{AIC}_c)$ of ANNs with different number of neurons in the hidden layer, and the ANN with five neurons in the hidden layer was chosen to be optimal based on Figure 6.13. Based on the current structure of the ANN, the number of hyper-parameters used in the ANN is $N(N_u + 1) + N_o(N +$

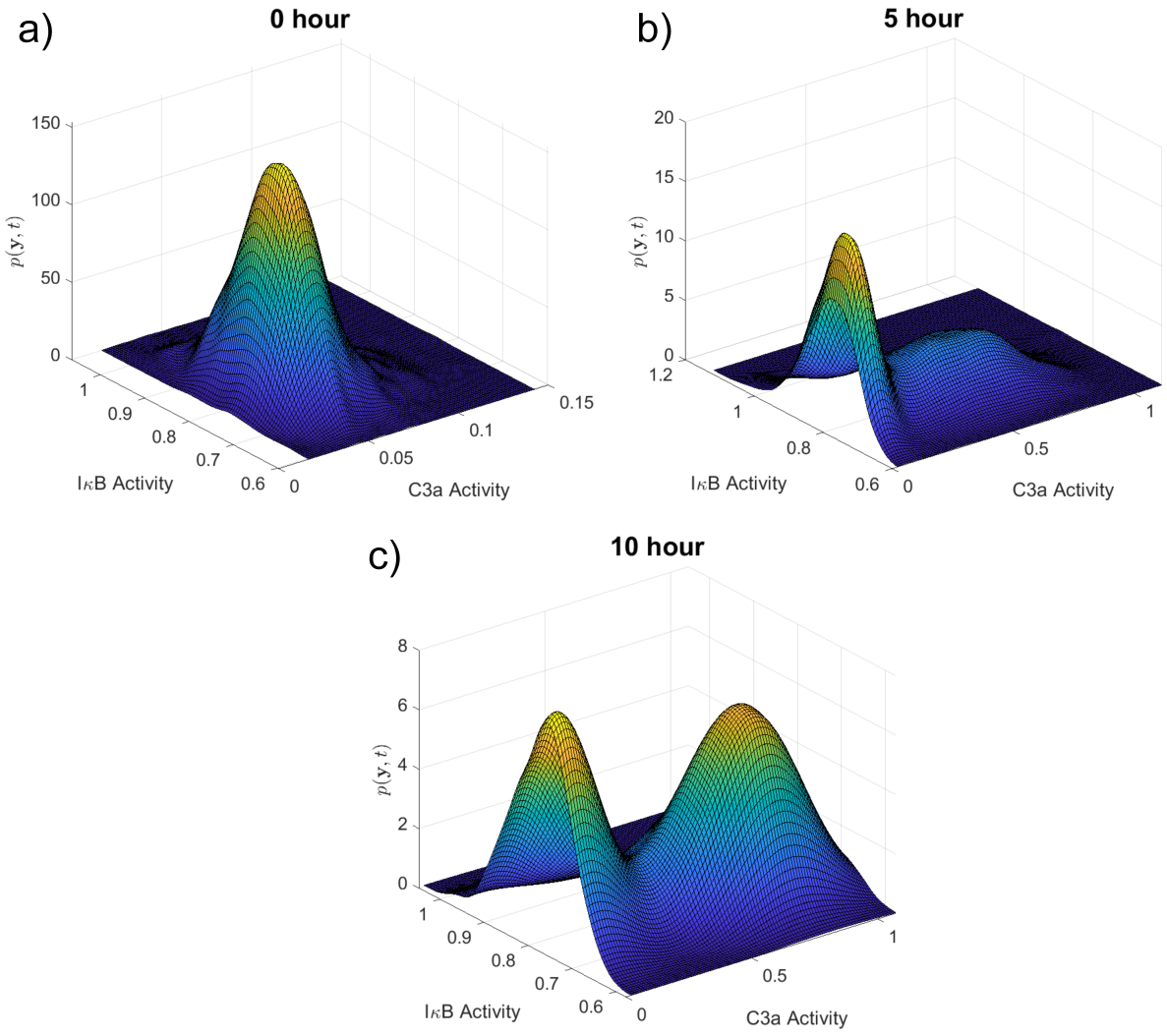


Figure 6.11: Artificial flow cytometry measurements when both C3a and IκB are measured. PDFs of the output based on the estimated PDF of b_1. The approximated PDFs of the artificial flow cytometry measurements at $t = 0, 5, 10$ hours when the C3a and IκB activities were measured.

$1) = 5 \times (80000 + 1) + 10000 \times (5 + 1) = 460005$. The R^2 statistics for the final ANN model were 0.772 with respect to the training dataset and 0.748 with respect to the validation dataset.

Based on the trained ANN and the *in silico* experimental measurements, the means of a_2 and a_4 were inferred by solving Equation 6.9. The inferred values of μ_{a_2} and μ_{a_4} are 1.5 and 1.3, respectively. In order to validate the estimation results, the corresponding $\hat{p}_{y(t_k)}(\boldsymbol{y}(t_k))$ were computed and plotted in Figure 6.14 along with the *in silico* experimental measurements. Overall, the pre-

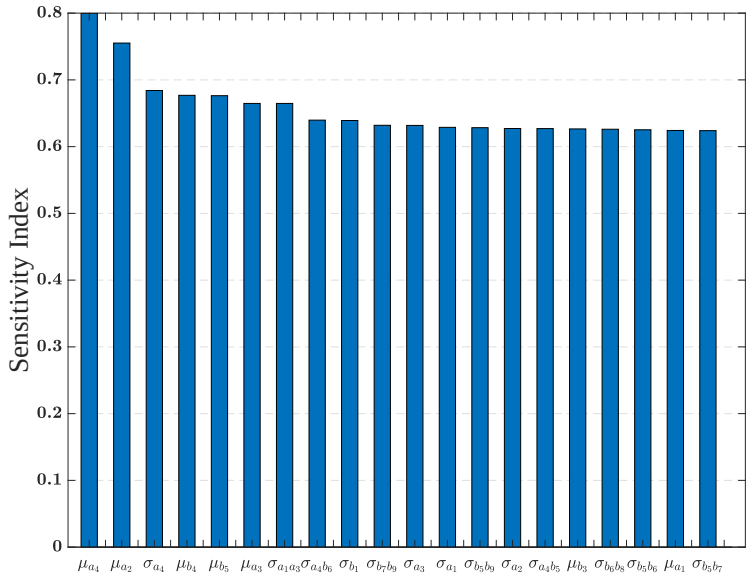


Figure 6.12: The results of the Sobol' method of the parameters in the TNFα signaling pathway when the C3a and IκB activities were measured. For clarity, the sensitivity indices of only twenty statistical parameters of $\boldsymbol{\theta}$ are shown.

dicted output PDFs agreed with the *in silico* measurements, which demonstrated that the proposed methodology can be applicable even when there are more than one measurement.

6.5 Discussion

A central problem in systems biology is the construction of predictive mathematical models for biological systems, and its key part is the inference of accurate model parameters [47]. Due to the limitations in experimental techniques as well as the inherent complexity of biological systems, the parameter estimation problem is usually ill-posed with many parameters that are not able to be identified from experimental measurements [23, 36]. Therefore, the parameter identifiability analysis has been extensively investigated by the systems biology community [47]. In previous studies employing the IBPM approach, the parameter identifiability was not explicitly analyzed. Instead, the PDFs of all parameters were estimated [44], or PDFs of a subset of parameters, which were selected based on prior knowledge, were estimated; therefore, the resultant models were mathematically suboptimal [10, 37, 42]. In contrast, this study implemented the Sobol' sensitivity

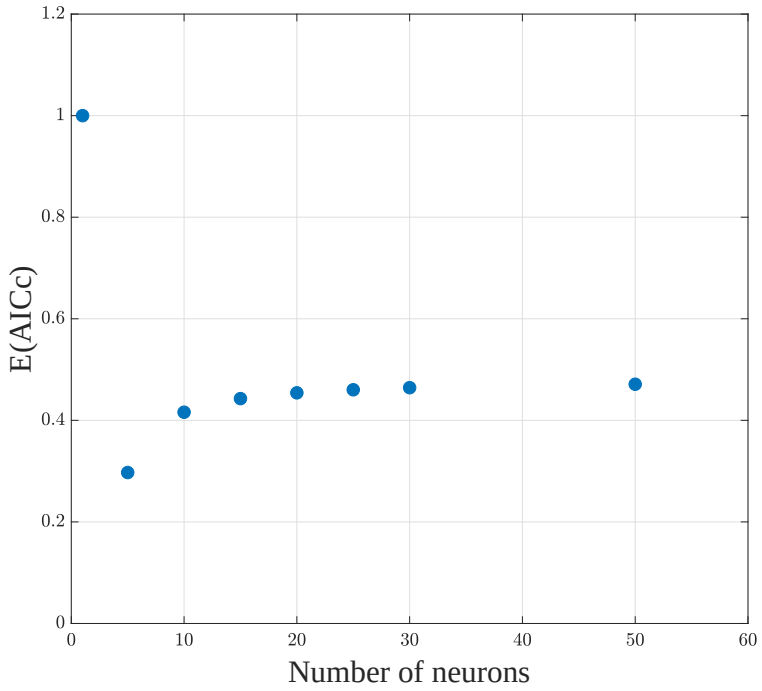


Figure 6.13: Changes in $E(\text{AIC}_c)$ values with the change in the number of neurons in the hidden layer when the C3a and IκB activities were measured.

analysis method to assess the identifiability of an IBPM numerically. Based on the presented TNFα signaling pathway case study, the importance of the identifiability analysis is evident. Specifically, when the PDF of a_2, which was identified as the most important, was estimated, the model with the inferred PDF of a_2 was able to predict the experimental measurements. Although the estimation results presented in Figures 6.8-6.14 agree reasonable well with the measurements, there were noticeable mismatches present. Particularly, the secondary peaks present in measured PDFs of C3a activities at 5 and 10 hours in Figure 6.14 were not well reproduced in the predicted PDFs. Such mismatch may be reduced by improving the accuracy of an ANN through adaptive sampling and/or iterations between surrogate modeling and optimization [248, 249, 250]. Such approaches would detect the inaccuracy of the ANN due to the insufficient sampling and re-train the ANN to achieve a better estimation result.

Previously, Hasenauer et al. [10, 42] proposed efficient density-matching schemes to infer

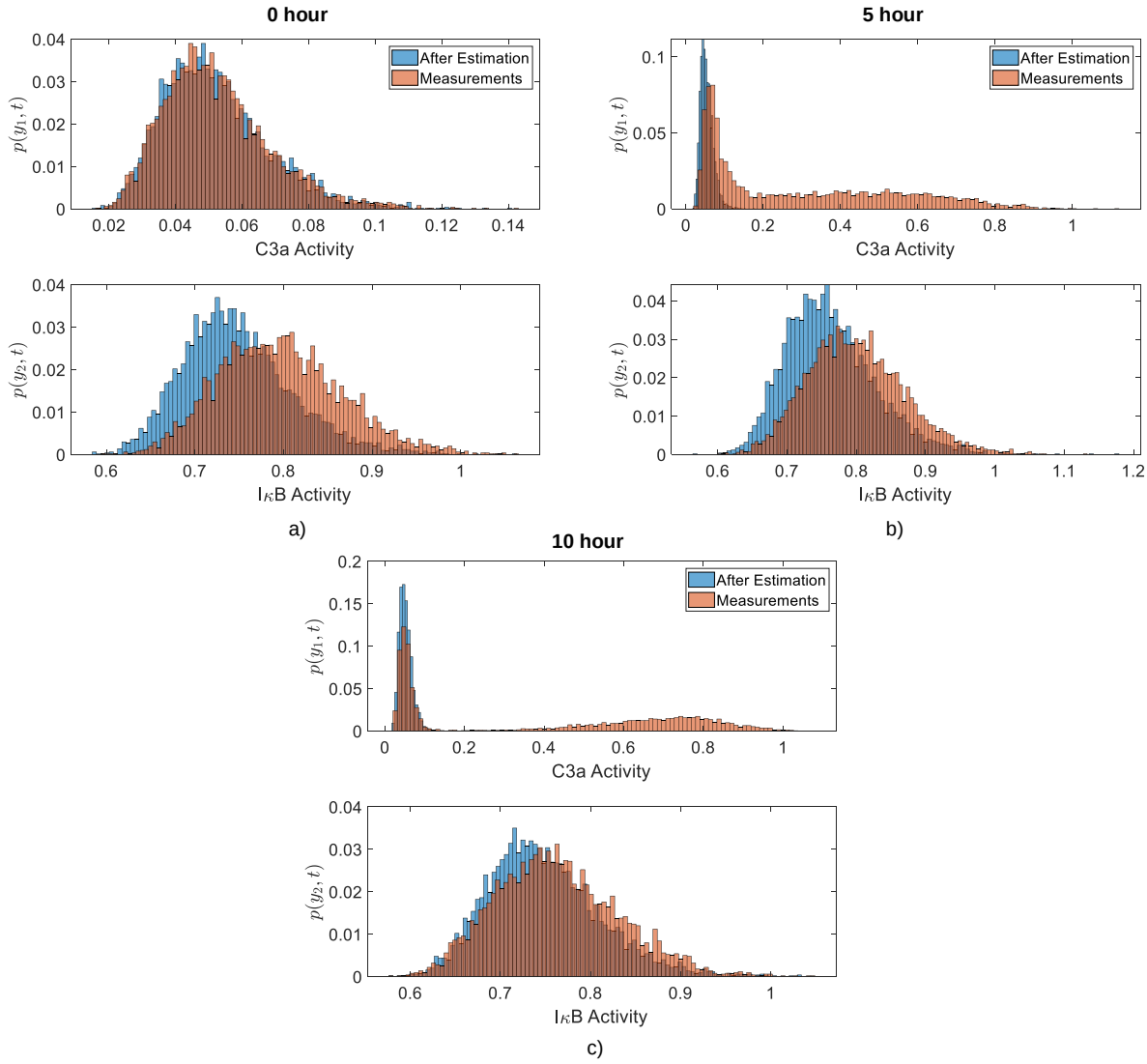


Figure 6.14: PDFs of the outputs based on the estimated PDFs of a_2 and a_4. Histograms of *in silico* measurements are compared with those predicted with the estimated means of a_2 and a_4 at t =0, 5, and 10 hours.

PDFs of model parameters, where they parameterized the PDFs of parameters into a linear combination of multiple known PDFs whose corresponding output PDFs are computed *a priori*. Through this parameterization, the PDF inference problem becomes much easier to be solved since it becomes a convex optimization problem and does not require to solve the forward model (Equation 6.1) repetitively, which is computationally expensive. Similarly, one of the main advantages of the proposed methodology in our work is to avoid simulating the forward model iteratively

while solving parameter inference problem (Equation 6.9) since the empirical mapping between PDFs of model parameters and measurements is constructed beforehand. The key difference lies in whether identifiability of PDFs is considered or not. Specifically, Hasenauer et al. [10, 42], at least in their case studies, identified a subset of important parameters based on the prior information and estimated their PDFs only. However, for many systems, it is unknown *a priori* which model parameters exhibit cell-to-cell variabilities. Also, even if such information is available, it may not be possible to uniquely estimate the PDFs of those parameters if those PDFs are not identifiable. In comparison, the proposed methodology in this work tries to address the identifiability issue by performing the global sensitivity analysis of the PDFs of model parameters to determine the identifiable PDFs, which are estimated in the subsequent inference problem. Consequently, the possibility of obtaining a non-unique solution will be minimized.

The other contribution of this study is to propose the use of a surrogate model (i.e., ANN). Since the ANN is developed offline with its validity spanning the entire parameteric space, the minimization problem (Equation 6.9) could be solved in a very efficient manner. Without the ANN development, the minimization problem (Equation 6.5) is computationally expensive because $1)$ each iteration requires to solve the ODE model (Equation 6.1) a large number of times to estimate $p(y)$ and 2) the problem has to be solved multiple times with multiple initial conditions to avoid locally optimal solutions. Therefore, the introduction of ANN reduces the computational cost. The proposed methodology can be implemented in a biomanufacturing process for estimation and control, as cell-to-cell variability can result in reduced yields and a bioreactor instability [16, 17, 34]. Although it is now feasible to measure the dynamics of biomolecules at the single-cell level [251, 252], it is still difficult to measure the dynamics of every biomolecule of interest at the single-cell level [14]. Therefore, the proposed methodology can be used to estimate the PDFs of unmeasurable biomolecules by estimating the underlying PDFs of parameters so that the biological system can be monitored more effectively. Furthermore, it is known that the phenotypes of cells can evolve in a long term due to various reasons such as genetic mutations [253]. Consequently, the PDFs of the model parameters need to be updated continuously in a real time, which requires

a numerically efficient method. Therefore, the proposed methodology can be an attractive option to achieve this goal. Also, the proposed approach can be used as a method to develop a grey-box model to improve the prediction accuracy, when there is a considerable amount of model uncertainty. In many circumstances, a first-principle model is partially accurate due to our limited understandings of a process [8, 243]. As a result, a model training via parameter estimation is unlikely to be successful due to the model mismatch. On the other hand, our proposed methodology can be used for modeling and estimating a partially known system by introducing PDFs of model parameters, which will take into account the uncertainty in a model structure.

6.6 Conclusion

In this work, we presented a new methodology for estimating the PDFs of parameters in an intracellular biochemical reaction from the population snapshot measurements. Since this was formulated as a density matching problem (Equation 6.5) with its decision variables to be infinite-dimensional, the normal-distribution assumption and the Sobol' sensitivity analysis method were implemented sequentially to reduce the dimension and transformed the decision variables to be finite-dimensional (i.e., requiring the first and second moments of the parameters only). Next, in order to avoid the excessive computational cost associated with solving the minimization problem, an ANN was developed beforehand to find an empirical mapping from the PDFs of parameters into the output PDF. Finally, the re-formulated minimization problem constrained by the developed ANN was solved to estimate the PDFs of parameters. To validate the proposed methodology, it was implemented to infer the PDFs of the selected parameters in the TNFα signaling pathway based on the experimental measurements generated *in silico*. The model with the estimated PDFs of the selected parameters was able to predict the output PDFs that agreed reasonably well with the measured output PDFs used in the estimation, which validated the proposed methodology.

7. KINETIC MONTE CARLO MODELING OF MULTIVALENT BINDING OF CTB PROTEINS WITH GM1 RECEPTORS*

7.1 Background

In Chapter 3, we have developed a new multi-step binding model describing interactions between CTB and ganglioside. Although the newly developed model performs better than existing models in the literature [11, 147, 148, 149, 151], there are three limitations in this modeling approach. First, in order for a deterministic model to account for the surface micro-configuration (e.g., receptor clustering and lipid rafts), which has been reported as one of the key factors determining the CTB-ganglioside binding kinetics [11, 139, 143, 144], a large number of states and parameters are required [56]. Second, the surface densities of GM1 receptors are usually low in the gastrointestinal tract, which will increase stochastic fluctuations in the binding kinetics [153, 254]; however, such fluctuations are not easy to be captured by deterministic models [68, 254]. Lastly, implementing the deterministic models for a cell membrane containing more than one type of ganglioside receptors becomes increasingly difficult since all the possible binding configurations between CTB proteins and ganglioside receptors need to be tracked during the simulation [255].

Motivated by these limitations, in Chapter 7, a kMC framework is proposed to model the CTB-GM1 binding kinetics on a cell membrane. The kMC is a stochastic modeling approach to simulate detailed surface reactions as well as the evolution of surface micro-configurations. Consequently, it has been widely used to model the microscopic processes of a variety of applications such as protein crystallization [51, 52, 53, 54, 55], thin film deposition [56, 57, 58, 59, 60], catalytic surface reactions [61, 62, 63, 64, 65, 66, 67], and receptor-ligand interactions on a cell membrane [68, 69, 70, 71]. The proposed kMC modeling approach can handle the aforementioned shortcomings of the deterministic models for the CTB-ganglioside binding system. Specifically, first, a kMC model can explicitly describe the surface micro-configuration of how membrane-bound CTB proteins

*Reprinted with the permission from "Kinetic Monte Carlo modeling of multivalent binding of CTB proteins with GM1 receptors," by Lee et al., 2018, *Computers & Chemical Engineering*, 118, 283-295. Copyright 2018 by Elsevier Ltd.

and free GM1 receptors are distributed on a cell membrane as well as its influence on the binding kinetics, which is difficult to be captured by deterministic models [59, 69, 70]. Second, the kMC framework computes the binding kinetics in a probabilistic manner that is suitable for describing a stochastic nature of the CTB-ganglioside binding process such as the fluctuation in the binding kinetics when the concentration of CTB proteins in the solution and the density of GM1 receptors on the surface of a cell membrane are low [56, 256]. Third, since the kMC framework executes one event at every time step (i.e., a discrete-time model), the process dynamics are described by a series of jumps from one surface micro-configuration to another in a discrete manner, while the deterministic model updates simultaneously all of the possible binding configurations in a continuous manner. The kMC model becomes extremely advantageous when the number of states to be considered is huge, as in the case of a cell membrane containing multiple types of receptors.

In this chapter, a kMC modeling framework is proposed to predict the binding kinetics between CTB proteins in the solution phase and GM1 receptors localized on a cell membrane. Furthermore, the subsequent binding of membrane-bound CTB proteins with additional GM1 receptors on a cell membrane is considered (Figure 7.1). It is important to note that GM1 is the primary ganglioside receptor for CTB proteins [6]. Since the CTB-GM1 binding process takes place at both solution and membrane-bound phases, the proposed kMC model needs to consider the binding kinetics in both phases and their interdependence. Specifically, the CTB-GM1 binding process in the solution phase is modeled by the steady-state diffusion equation to capture the CTB proteins diffusing towards or away from the cell membrane [11, 157]. On the other hand, the surface binding process is modeled by a standard kMC scheme to describe the surface reactions of membrane-bound CTB proteins and free GM1 receptors on a cell membrane. For a manageable computational requirement, the migration of GM1 receptors is decoupled from other surface reaction processes since it is an overwhelmingly frequent event than others [59, 60]. The simulation results indicated that the proposed kMC model is comparable with the experimentally verified deterministic model proposed by [11], which demonstrates the viability of the proposed stochastic modeling framework for a CTB-ganglioside binding system.

The chapter is organized as follows: first, we present the dynamic model that describes the CTB-GM1 binding kinetics in the solution phase and on a cell membrane. Next, a simulation lattice that represents a cell membrane is introduced to mimic the cell membrane environment, and the execution procedure of kMC simulations is provided with details. Lastly, the simulation results of binding kinetics predicted by the proposed kMC model are compared with those of an experimentally verified deterministic model.

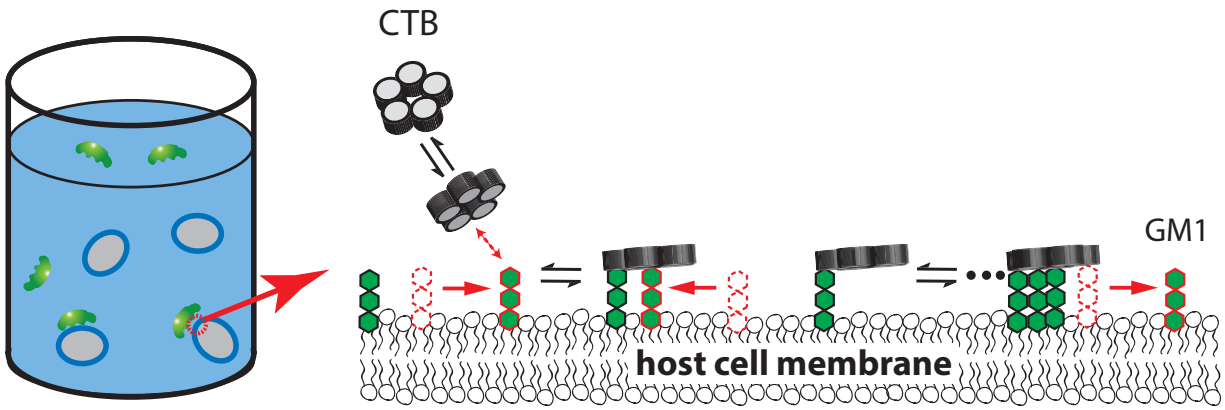


Figure 7.1: Macroscopic (left) and microscopic (right) illustrations of the CTB-GM1 binding process. CTB proteins are in the solution phase while GM1 receptors are localized on cell membranes.

7.2 CTB-GM1 binding process description and modeling

We consider CTB proteins binding to a cell membrane containing only GM1 ganglioside receptors (Figure 7.1). Initially, CTB proteins are in the solution phase, and GM1 receptors are localized on the cell membrane. The CTB-GM1 binding process only occurs on the cell membrane after CTB proteins are transported to the cell membrane via diffusion, which requires an integrated model capable of capturing both the solution-phase diffusion of CTB proteins as well as multiple microscopic reactions and transport phenomena (Figure 7.2).

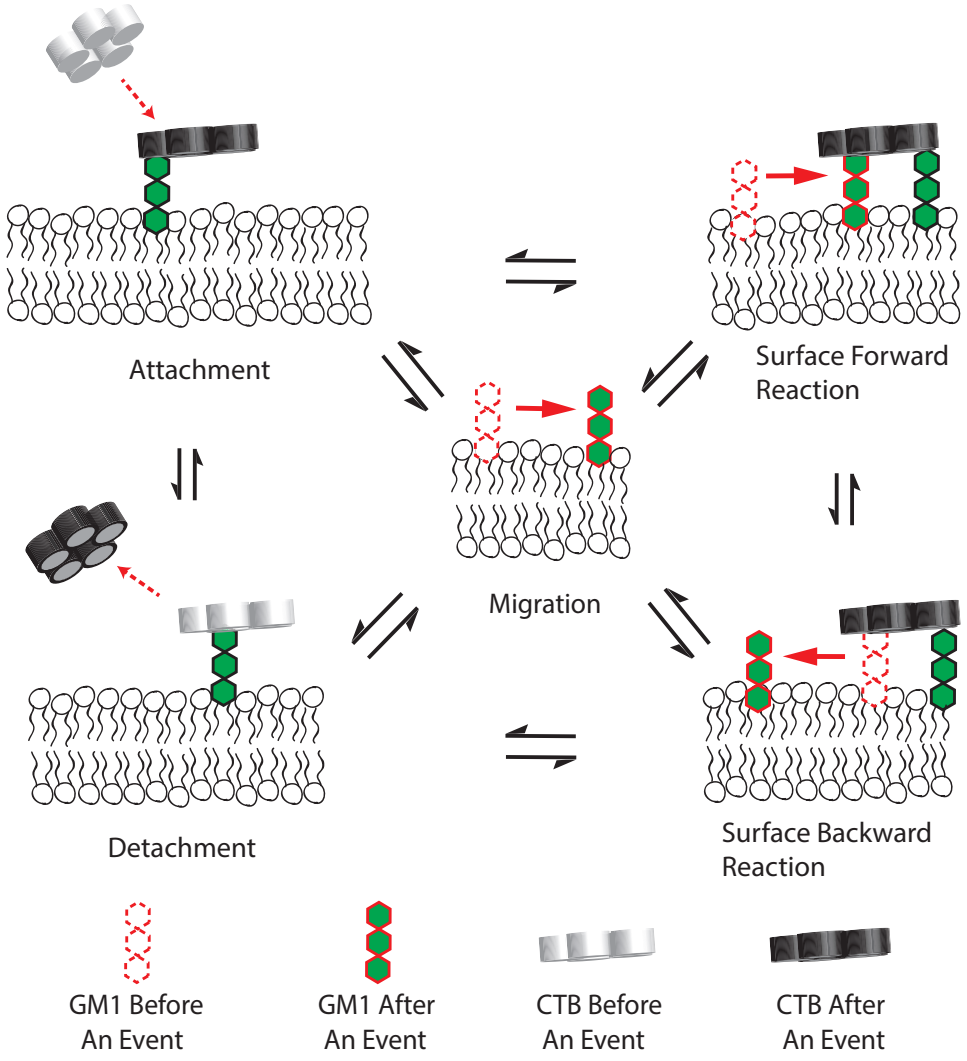


Figure 7.2: Schematic illustrations for microscopic processes considered to describe the CTB-GM1 binding on a cell membrane.

7.2.1 Solution phase binding model

7.2.1.1 CTB diffusion and reactions

The solution phase binding model consists of two processes: CTB attachment to a cell membrane and CTB detachment from a cell membrane. The former refers to a two-step process where a protein diffuses to the vicinity of a cell membrane and attaches to the cell membrane by binding

with a free GM1 receptor, which can be illustrated by the following reaction:

$$\text{CTB(s)} + \text{GM1(m)} \rightarrow \text{CTB-}[\text{GM1}]_1 \tag{7.1}$$

where CTB(s) represents a CTB protein in the solution phase, GM1(m) represents a GM1 receptor on a cell membrane, and CTB-$[\text{GM1}]_1$ represents a membrane-bound CTB protein with one bound GM1 receptor. Similarly, the CTB detachment refers to a two-step process where a membrane-bound CTB protein (i.e., CTB-$[\text{GM1}]_1$ in Equation 7.1) dissociates with its bound GM1 receptor and diffuses away from the cell membrane as follows:

$$\text{CTB-}[\text{GM1}]_1 \rightarrow \text{CTB(s)} + \text{GM1(m)} \tag{7.2}$$

7.2.1.2 Rates of CTB attachment and detachment

In this study, we introduce effective rate constants to treat the two-step CTB attachment as a one-step process that takes into account simultaneously the diffusion and binding processes of CTB proteins. The lumped rate constant will be functions of the intrinsic CTB-GM1 association rate constant, the number of free GM1 receptors on a cell membrane, the diffusion coefficients of a CTB protein in the solution, and the radius of a cell [157].

Suppose a cell with a radius of a is placed at the origin of a spherical coordinate system, the steady-state diffusion equation for a CTB protein in the solution phase is as follows:

$$D\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{dC_s}{dr}\right) = 0 \tag{7.3}$$

where D is the diffusivity coefficient of a CTB protein in the solution phase, and $C_s(r)$ is the spatially varying CTB concentration. The first boundary condition is as follows:

$$r \rightarrow \infty \qquad C_s \rightarrow C \tag{7.4}$$

where C is the bulk CTB concentration. The second boundary condition can be obtained by applying a mass balance on a cell membrane. More specifically, at the surface of a cell membrane where $r = a$, it is assumed that the overall CTB diffusion rate equals to the rate at which CTB proteins bind with GM1 receptors on the cell membrane, which can be defined as [257]:

$$4\pi a^2 D N_A \left.\frac{dC_s}{dr}\right|_{r=a} = 5k_0 \cdot C_s(r=a) \cdot R \tag{7.5}$$

where N_A is the Avogadro's number, k_0 is the intrinsic association rate constant for the CTB-GM1 binding, and R is the number of free GM1 receptors on the cell membrane. Here, a coefficient of 5 is included because there are five different ways for a CTB protein to associate with a free GM1 receptor to result in one configuration for a membrane-bound CTB protein with one bound GM1 receptor (C_1 in Figure 7.3) as a CTB protein is a homogeneous pentamer. Applying these two boundary conditions, the analytical solution to Equation 7.3 is obtained as follows:

$$C_s(r) = \frac{-5k_0 aCR}{4\pi DaN_A + 5k_0 R}\frac{1}{r} + C \tag{7.6}$$

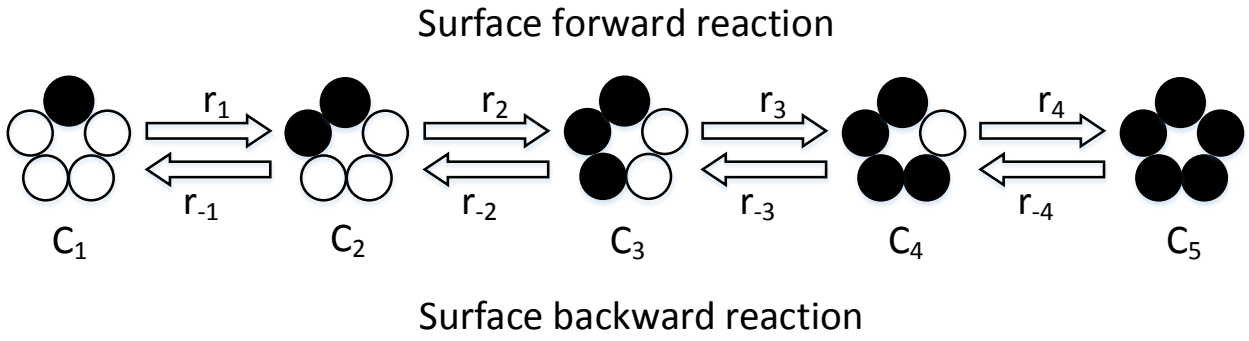


Figure 7.3: A schematic diagram for the step-wise CTB-GM1 binding kinetics on a cell membrane adopted from [11]. C_i represents the membrane-bound CTB protein with i bound GM1 receptors. Empty and filled circles represent the binding pockets of a membrane-bound CTB protein without and with bound GM1 receptors, respectively. r_i and r_{-i}, $\forall i = 1, \ldots, 4$, are the corresponding reaction rates.

By introducing the effective association rate constant, k_f, the overall CTB flux to a cell membrane can be defined as $5k_fCR$, which should be equal to the left hand side of Equation 7.5. Consequently, k_f can be defined as follows [11]:

$$k_f = \frac{k_0}{1 + \frac{5k_0R}{4\pi DaN_A}} \tag{7.7}$$

Similarly, the effective dissociation rate constant, k_r, that describes both the CTB detachment and diffusion away from a cell membrane is defined as follows [157]:

$$k_r = \left(1 - \frac{5k_0RN_A}{5k_0RN_A + 4\pi Da}\right) k_{-1} = \frac{k_{-1}}{1 + \frac{5k_0R}{4\pi DaN_A}} \tag{7.8}$$

where k_{-1} is the intrinsic CTB-GM1 dissociation constant.

Now, the CTB attachment and detachment rates are computed by

$$\begin{aligned} r_a &= 5k_fCR \\ r_d &= k_rC_1 \end{aligned} \tag{7.9}$$

where r_a and r_d are the CTB attachment and detachment rates, respectively, C is the CTB concentration in the solution phase, R is the number of free GM1 receptors on the cell membrane, and C_i is the number of membrane-bound CTB proteins with i bound GM1 receptors.

7.2.1.3 CTB mass balance

In this work, the change in the total CTB concentration determined by the interplay between the CTB detachment and attachment processes can be described by the following mass balance equation:

$$\frac{dC}{dt} = (r_d - r_a)\frac{N_{cell}}{N_A \cdot V} \tag{7.10}$$

where N_{cell} is the number of cells suspended in the solution phase, and V is the volume of solution.

7.2.2 Surface binding model

As this is the first time applying the kMC methodology to model the dynamic CTB-GM1 binding process, the microscopic surface binding interactions between CTB proteins and GM1 receptors are described in great detail (Figure 7.2). The proposed surface binding model consists of the multi-step binding process and the migration of GM1 receptors on a cell membrane.

7.2.2.1 CTB and GM1 surface reactions

The multi-step binding mechanisms proposed by [11] are adopted to describe the interactions between CTB proteins and GM1 receptors on a cell membrane. Specifically, they considered four surface forward reactions and four surface backward reactions (Figure 7.3). Here, the surface forward reaction refers to a binding process where a membrane-bound CTB protein binds with an additional GM1 receptor on a cell membrane, which can be illustrated by the following reaction:

$$\text{CTB-}[\text{GM1}]_i + \text{GM1} \rightarrow \text{CTB-}[\text{GM1}]_{i+1}, \quad \forall i = 1, \cdots, 4 \tag{7.11}$$

where i represents the number of bound GM1 receptors. Since a CTB protein has five binding pockets, the surface forward reaction can take place until a membrane-bound CTB protein becomes fully occupied with five bound GM1 receptors. On the other hand, the surface backward reaction refers to a process where a membrane-bound CTB protein dissociates one of its bound GM1 receptors, which can be illustrated by the following reaction:

$$\text{CTB-}[\text{GM1}]_i \rightarrow \text{CTB-}[\text{GM1}]_{i-1} + \text{GM1}, \quad \forall i = 2, \cdots, 5 \tag{7.12}$$

During the surface forward and backward reactions, the intermolecular distances between free GM1 receptors and membrane-bound CTB proteins are constantly changing due to the migration of those molecules on a cell membrane. Specifically, a free GM1 receptor cannot bind with a binding pocket of a membrane-bound CTB protein if their intermolecular distance exceeds a threshold value, which will hereafter be denoted as l_c. In this work, all free GM1 receptors are allowed to

migrate on a cell membrane while membrane-bound CTB proteins are assumed to be immobile: since considering the migration of both membrane-bound CTB proteins and free GM1 receptors will bring similar effects on the intermolecular distances, it is decided to consider only the migration of free GM1 receptors for computational efficiency.

7.2.2.2 Surface binding rates

The rate equations for four surface forward and four surface backward reactions are presented as follows:

$$\begin{aligned} r_i &= (5-i)\frac{k_1}{4\pi a^2}C_i R \\ r_{-i} &= (i+1)k_{-1}C_{i+1}, \qquad \forall i = 1, \cdots, 4 \end{aligned} \tag{7.13}$$

where r_i is the forward reaction rate from C_i to C_{i+1}, k_1 is the surface forward association rate constant, a is the radius of a cell, and r_{-i} is the backward reaction rate from C_{i+1} to C_i. Here, the factors $(5-i)$ and $(i+1)$ are included to show that there are $(5-i)$ and $(i+1)$ ways in which the corresponding reactions can occur with identical results due to the symmetric nature of a CTB protein.

As a GM1 receptor is allowed to migrate to any adjacent sites including diagonally neighboring sites, the migration rate can be computed via the following equation [68, 258]:

$$r_m = N_{emp} k_D / l^2 \tag{7.14}$$

where r_m is the migration rate for GM1 receptors, N_{emp} is the total number of empty neighboring sites available for the migration of free GM1 receptors, k_D is the diffusion coefficient of GM1 receptors on a cell membrane, and l is the average distance between two neighboring lattice sites (details on lattice sites will be presented in the following section). It should be noted that Equation 7.14 computes the mean rate of the receptor migration to any neighbors, including 1^{st} and 2^{nd} nearest (i.e., diagonal) neighbors. This is an acceptable practice to simulate the CTB-GM1 binding process as the rate of migration is overwhelmingly dominating compared to the rates of the

remaining events, which will be discussed further in a subsequent section. However, if the value of k_m decreases very significantly, two separate migration rates should be computed to represent the migration rates to 1st and 2nd nearest neighbors differently with different values of l and N_{emp}.

7.3 Kinetic Monte Carlo Model

7.3.1 Simulation Lattice

A two-dimensional solid-on-solid lattice with a periodic boundary condition is implemented to model the foregoing surface interactions among GM1 receptors and CTB proteins via a kMC methodology, which considers the CTB attachment and detachment, GM1 migration, and surface forward and backward reactions (Figure 7.2) [51, 53, 68]. The dimension of the lattice sites is $(N \cdot l_s) \times (N \cdot l_s)$, where N is the number of lattice sites on each side, and l_s is the lattice spacing. In this work, we considered a simplified cell membrane, composing of two molecules: GM1 receptor and a phospholipid (DOPC), which does not interact with CTB proteins. Under this assumption, the value of l_s is set to be 0.96 nm, which is the diameter of DOPC head group estimated based on its surface area (0.72 nm^2) [259]. We could introduce different numbers of GM1 receptors on the two-dimensional lattice sites to reach the desired surface densities of GM1 receptors. Prior to the simulation, GM1 receptors are randomly distributed over the simulation lattice. Each lattice site can be either empty or occupied by one GM1 receptor, and GM1 receptors can migrate to eight neighboring sites, including the four diagonal directions, unless they are occupied (Figure 7.4). In this case, l in Equation 7.14 equals to $((1 + \sqrt{2})/2) \cdot l_s$, which is the average value of the lattice spacing, l_s, and the distance between two diagonally neighboring sites, $\sqrt{2} \cdot l_s$

A CTB protein consists of five identical monomers to form a pentemeric ring whose diameter is 6 nm [260]. In this work, a CTB protein on the lattice sites is approximated by a pixelated pentagon that covers a total of 24 lattice sites (Figure 7.5). Hence, the covered lattice sites that represent a membrane-bound CTB protein is around 22 nm^2, which is very close to the surface area of a regular pentagon that inscribes a circle of diameter of 6 nm. In this work, we assumed a CTB protein attaches to the lattice sites with the following configuration: the CTB will be aligned

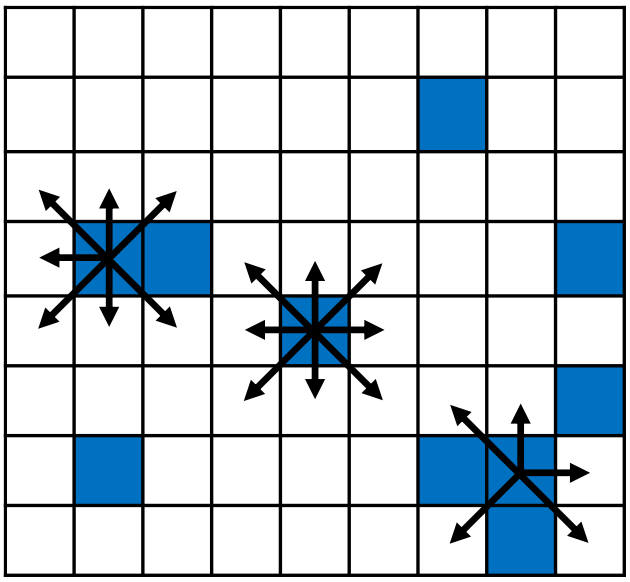


Figure 7.4: A schematic illustration of GM1 receptors on the lattice sites for the kMC simulation. Squares in blue represent lattice sites occupied by GM1 receptors, and arrows indicate possible migration directions.

with the lattice sites so that all of its five binding pockets belong to a plane that is in parallel to the two-dimensional lattice sites. As membrane-bound CTB proteins are assumed to be immobile, their bound GM1 receptors are also immobile and only available for a surface backward reaction or a CTB detachment event.

7.3.2 Kinetic Monte Carlo Implementation

In this work, a hybrid n-fold kMC algorithm is applied to the CTB-GM1 binding process occurring on a cell membrane [261], and the GM1 migration event is handled separately by another approach for computational efficiency [59, 60], which will be discussed in the following subsection. The total surface reaction rate, r_t, is defined as follows:

$$r_t = r_a + r_d + \sum_{i=1}^{4} (r_i + r_{-i}) \tag{7.15}$$

Then, a random number, $\xi_1 \in (0, 1]$, is generated. If $\xi_1 < r_a/r_t$, the CTB attachment event is selected; if $r_a/r_t \leq \xi_1 < (r_a + r_d)/r_t$, the CTB detachment event is selected; if $(r_a + r_d +$

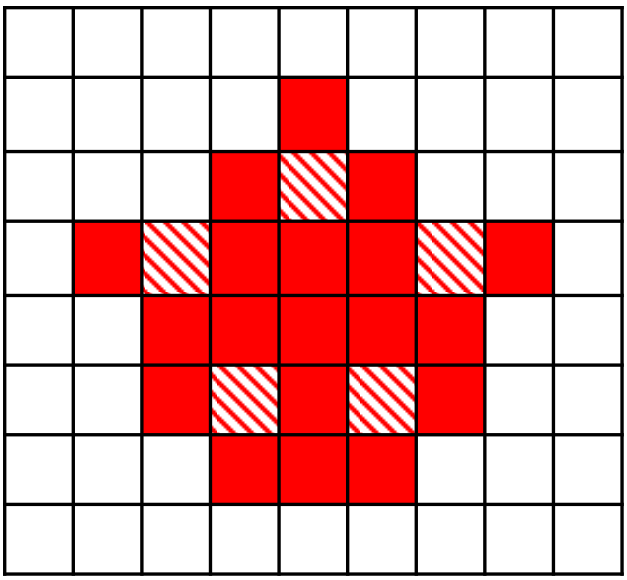


Figure 7.5: A schematic illustration of a pixelated CTB protein. Each square box represents a lattice site with its side length of 0.96 nm, and filled squares represent the lattice sites covered by a membrane-bound CTB protein. Five shaded boxes represent the five binding pockets of the membrane-bound CTB protein.

$\sum_{i=1}^{j-1} r_i)/r_t \leq \xi_1 < (r_a + r_d + \sum_{i=1}^{j} r_i)/r_t,\ \forall j = 1, \ldots, 4$, the surface forward reaction event from C_j to C_{j+1} is selected; if $(r_a + r_d + \sum_{i=1}^{4} r_i + \sum_{i=1}^{j-1} r_{-j})/r_t \leq \xi_1 \leq (r_a + r_d + \sum_{i=1}^{4} r_i + \sum_{i=1}^{j} r_{-j})/r_t,\ \forall j = 1, \ldots, 4$, the surface backward reaction event from C_{j+1} to C_j is selected.

Since the kMC model simulates only a part of a cell membrane, the mass balance equation of Equation 7.10 should be modified to account for the difference in the surface areas of an actual cell membrane and the two-dimensional lattice as follows:

$$C(t + \Delta t) = C(t) \pm \left(A_r \cdot \frac{N_{cell}}{N_A \cdot V} \right) \tag{7.16}$$

where Δt is the elapsed time of one CTB attachment or detachment event, and A_r is the ratio between the surface area of a cell membrane ($4\pi a^2$) and the lattice surface area ($N^2 l_s^2$). It is important to note that Equation 7.16 is a discrete-time model, which is readily implementable in the kMC modeling approach. A flow diagram of the overall kMC simulation procedures is shown in Figure 7.6.

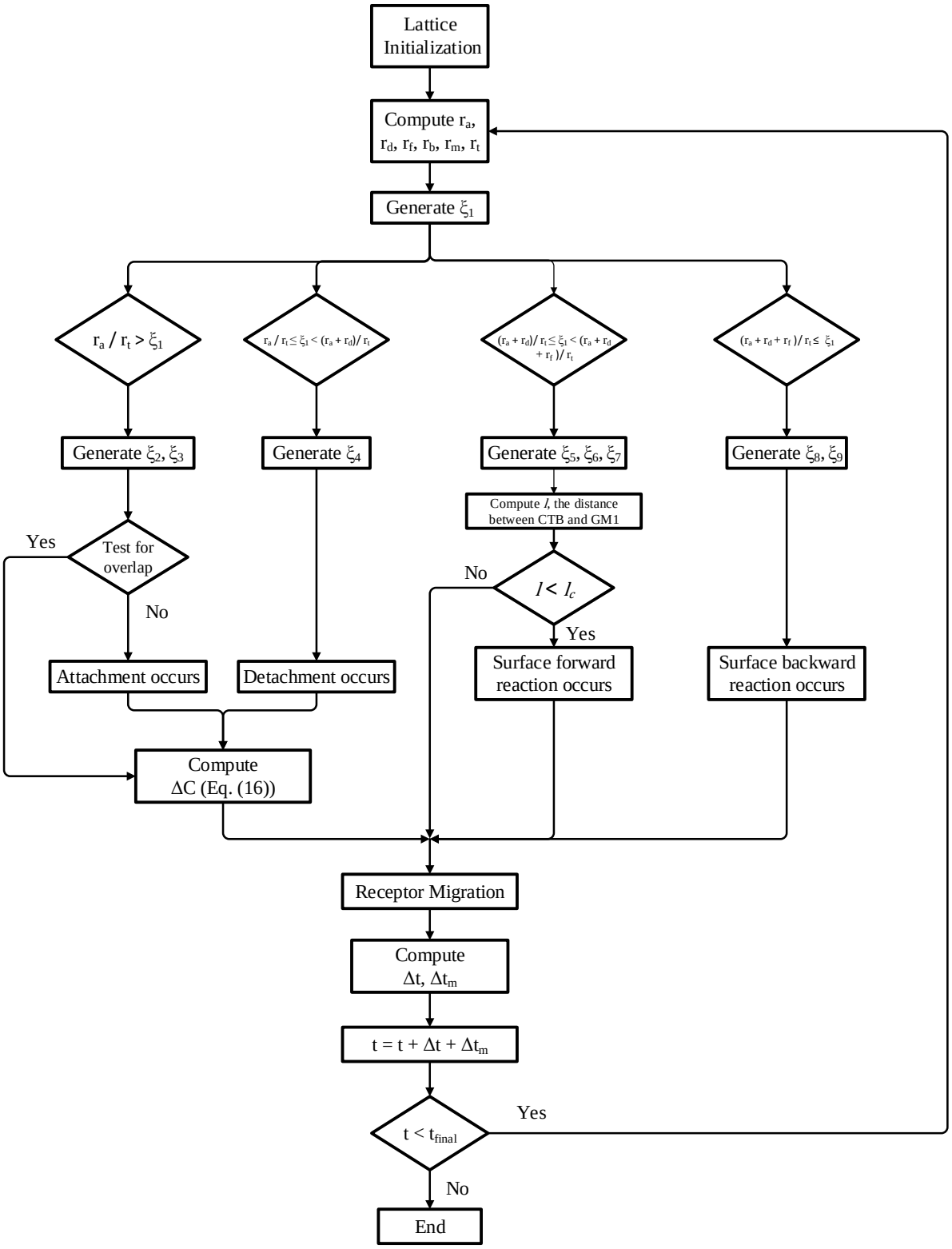


Figure 7.6: The kMC flow chart. Due to the space limitation, only one surface forward reaction (r_f) and one surface backward reaction (r_b) are shown.

When the CTB attachment event, r_a, is selected, a random integer $\xi_2 \in (0, R]$ is generated to select which free GM1 receptor on the lattice sites will be bound to the incoming CTB protein, and another random integer $\xi_3 \in [1, 5]$ is generated to select one out of the five available binding sites of the incoming CTB protein to which the selected free GM1 receptor will be bound. It should be noted that the incoming CTB protein needs a group of available lattice sites around the selected GM1 receptor for an attachment event to occur (Figure 7.5). If an enough area is not reserved due to overlapping with any of existing membrane-bound CTB proteins on the lattice sites, the attachment event is rejected, and the kMC simulation proceeds to the next step.

When the CTB detachment event, r_d, is selected, an integer $\xi_4 \in (0, C_1]$ is randomly generated to select a membrane-bound CTB protein with only one bound GM1 receptor, and the selected CTB protein will leave the cell membrane while the dissociated GM1 receptor is now available for binding with other membrane-bound or incoming CTB proteins.

When the surface forward reaction rate, r_i, is selected, a random integer $\xi_5 \in (0, C_i]$ is generated to select a membrane-bound CTB protein with i bound GM1 receptors to bind with an additional GM1 receptor, which is then selected by generating another random integer $\xi_6 \in (0, R]$. Additionally, another random integer $\xi_7 \in [1, 5 - i]$ is generated to select a binding pocket of the selected CTB protein to which the selected free GM1 receptor will be bound. It should be noted that the intermolecular distance between the selected binding pocket and the selected GM1 receptor is important. Specifically, if the intermolecular distance is larger than the value of l_c, the surface forward reaction between the selected CTB protein and GM1 receptor will be rejected, and the kMC simulation proceeds to the next step.

When the surface backward event, r_{-i}, is selected, a random integer $\xi_8 \in (0, C_{i+1}]$ is generated to determine a membrane-bound CTB protein that currently has $i + 1$ bound GM1 receptors. Then, another random integer $\xi_9 \in [1, i + 1]$ is generated to determine a GM1 receptor bound to the selected membrane-bound CTB protein. The selected GM1 receptor dissociates from the membrane-bound CTB protein and becomes free.

Once an event is selected and executed, an additional random number $\xi_t \in (0, 1]$ is drawn to

determine the elapsed time associated with the executed event [51]. Specifically, a time increment, Δt_t, is computed based on ξ_t and r_t as follows:

$$\Delta t_t = \frac{-\ln \xi_t}{r_t} \tag{7.17}$$

7.3.2.1 Decoupling GM1 migration

Although the GM1 surface density is usually low, the rate of GM1 receptor migration is still much larger than the rates of all other events combined (Figure 7.7). Specifically, if the GM1 surface density is 1 mol% and $N = 100$, there are 100 free GM1 receptors on the lattice sites; therefore, the migration rate is around 10^9/s via Equation 7.14 since the diffusivity of GM1 receptors is estimated to be 8.25×10^6 nm^2/s [140, 262]. In contrast, the CTB attachment rate is 0.1/s via Equation 7.13 when the CTB concentration is 100 nM and the k_f value is about 10^4/(M·s). The rates of other events are even smaller than the GM1 attachment rate. From this comparison, it is clear that the migration rate is much greater than the rate of other reactions by several orders of magnitude. Under this condition, the effect of the receptor migration on the membrane reactions becomes negligible, which allows to approximate the receptor migration [263]. Consequently, the GM1 migration is decoupled from the event selection for computational efficiency. More specifically, after executing one non-migration event, the GM1 migration event is carried out by randomly reassigning locations of all free GM1 receptors: two random integers from $(0, N]$ are generated to determine a new location for every free GM1 receptor. If a new location is occupied by another receptor, this procedure is repeated until the GM1 receptor finds an empty site. The time increment associated with the decoupled migration process is calculated as follows [59, 60]:

$$\Delta t_m = \frac{-\ln \xi_m}{r_t} \tag{7.18}$$

where ξ_m is another random number sampled from (0,1].

Consequently, the updated time instant after the event execution is $t_{k+1} = t_k + \Delta t_t + \Delta t_m$, where

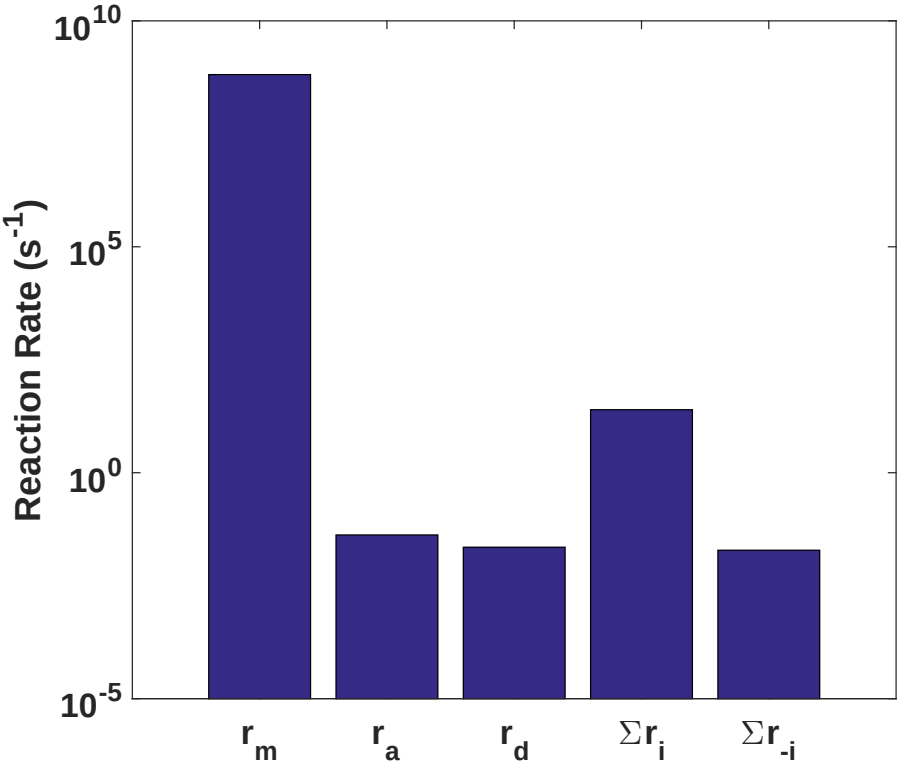


Figure 7.7: Reaction rates of microscopic events considered in the proposed kMC model with the CTB concentration of 25 nM and the GM1 surface density of 0.01 mol%.

t_{k+1} and t_k are the next and current time instants, respectively. It should be noted the time will elapse by $(\Delta t_t + \Delta t_m)$ even if the selected event is rejected due to the surface micro-configuration. The simulation will continue to proceed until the pre-specified end time. The kMC simulation is implemented in C# with Visual Studio 2017, and it is executed on a Windows desktop with 3.60 GHz Quad-Core Intel i7 Processors.

7.4 Results

The values of the model parameters used for the kMC simulation are listed in Table 7.1. Please note that these values were adopted from [11] except k_1. In the model proposed by [11], only surface forward and backward reactions are considered without the GM1 migration. Therefore, the k_1 in their model in fact is an effective surface forward association rate constant, which is a function

of the intrinsic surface forward association rate constant, the average distance between membrane-bound CTB proteins, the l_c value, and the GM1 diffusivity on a cell membrane [162]. On the other hand, k_1 in the proposed kMC model refers to the intrinsic surface forward association rate constant as the GM1 migration and l_c are explicitly incorporated in the proposed kMC framework. Therefore, the k_1 value proposed by [11] was not able to be readily used in the proposed kMC model.

Parameter	Value	Source
k_0 (/(M·s))	2.8×10^{4}	[11]
k_{-1} (/s)	3.2×10^{-3}	[11]
k_1 (μm^2/s)	1.3×10^{-2}	Estimated
k_D (m^2/s)	8.25×10^{-12}	[264]
D (m^2/s)	10^{-10}	[155]
l_c (nm)	1.45	[265]
a (μm)	2.5	Fixed
N_{cell}/V (/L)	5×10^{8}	Fixed
l_s (nm)	0.96	[259]
N	500	Fixed
d_{mem}(nm)	3.6	[266]

Table 7.1: Values of the model parameters used in the kMC simulation

Unfortunately, there was no previous study reported the k_1 value of the CTB-GM1 binding kinetics in this context. Therefore, we estimated the value of k_1 based on the following correlation with k_0, which is the intrinsic rate constant describing the CTB-GM1 binding in the solution phase (Equation 7.7) [162]:

$$k_1 = k_0 \cdot \frac{1}{d_{mem} N_A} \tag{7.19}$$

where d_{mem} is the thickness of a cell membrane, which is set to be 3.6 nm [266].

Additionally, the values of k_D and D were adopted from previous studies [155, 259, 264], and the value of a was fixed at 2.5 μm (Table 4.1). The value of N_{cell}/V in Equation 7.16 was fixed at

5×10^8 cell/L, which was taken from [11]. Lastly, the threshold distance of $l_c = 1.45$ mm was used for the surface forward reactions, which is one half of the distance between two adjacent binding pockets of a CTB protein [265].

In order to validate the receptor migration decoupling strategy, separate simulations without the receptor migration were executed to validate the accuracy of the proposed kMC model. In these simulations, the migration event was explicitly considered in the event selection by considering Equation 7.14, the migration event rate calculation, in calculating the total surface reaction rate, r_t. The obtained predictions were compared with those with the migration decoupling. As shown in Figure 7.8, the simulated binding kinetics with and without kinetics showed a reasonable consistency, which demonstrated the validity of the migration decoupling strategy implemented in this study.

In order to validate the proposed kMC model, the total number of membrane-bound CTB proteins (i.e., $\sum_{i=1}^{5} C_i$) was compared with that predicted by the deterministic model proposed by [11]. Since the deterministic model was experimentally validated, a reasonable match between the two models would validate predictive capability of the proposed kMC model. Figures 7.9-7.10 present the total number of membrane-bound CTB proteins predicted by the proposed kMC model when the GM1 surface densities are 0.01 and 0.03 mol% with 3, 5, 10 and 15 nM CTB concentrations. Although some discrepancy between the predictions from these two models existed due to the lack of efforts to carry out a sophisticated parameter estimation procedure, the total number of membrane-bound CTB proteins predicted by the proposed kMC model and the deterministic model by [11] were reasonably close to each other, which suggests that the proposed kMC model is a valid mathematical tool to simulate the CTB-GM1 binding kinetics on a cell membrane.

Each realization of the kMC model represents the CTB-GM1 binding process on a cell membrane. Therefore, we can simulate the cell-to-cell variability in the CTB-GM1 binding kinetics by running the kMC model under the same condition multiple times; the simulation results are different because the generated random numbers that determine a series of the surface reaction rates of CTB proteins and GM1 receptors will not be identical. In Figures 7.9-7.10, ten realizations of

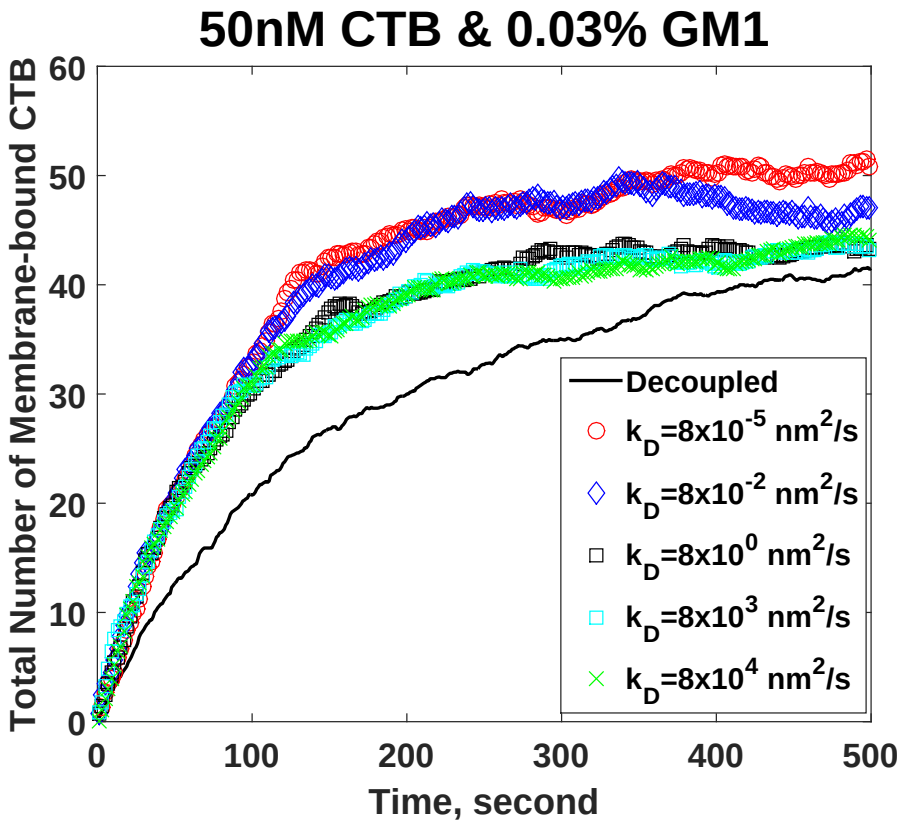


Figure 7.8: Validation of the migration decoupling strategy. The solid line was the predicted kinetics with decoupling the migration while dotted points were the predicted kinetics without decoupling under the particular migration rate constant. For each condition, ten realizations were obtained, and their mean values were plotted.

the kMC model were plotted together to compare how the total number of membrane-bound CTB proteins changes with time under the same condition of CTB concentrations and GM1 surface densities: Specifically, in Figure 7.9, the maximum of the total number of membrane-bound CTB proteins after 1000 seconds was almost four times the minimum of the total number of membrane-bound CTB proteins among the realizations. This fluctuation was attributed to the low copy number of CTB proteins and GM1 receptors. Here, we simulated a cell membrane with a low GM1 surface density (< 0.1 mol%), and the corresponding number of GM1 receptors on a cell membrane was around 10^4. In the physiological condition, GM1 is approximately 0.001 mol% of total glycolipids in human intestinal epithelia [153], and the typical total glycolipid fraction of the entire cell

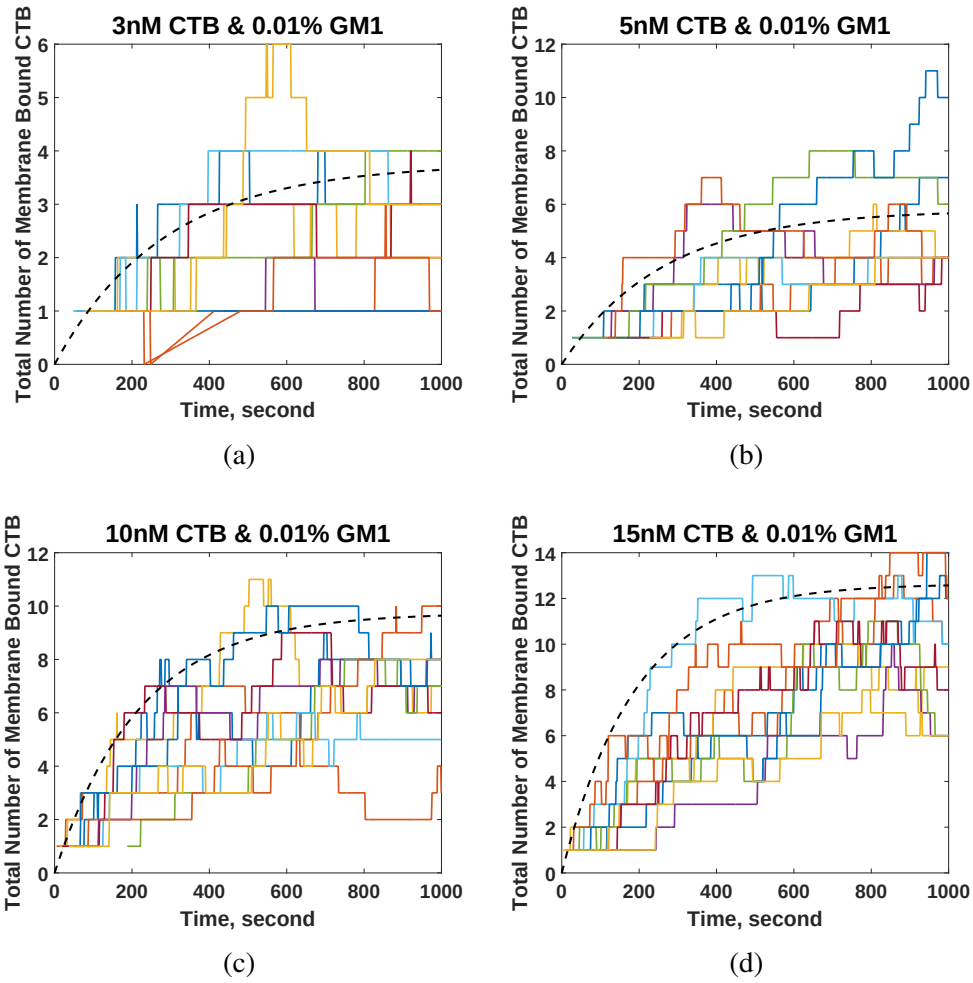


Figure 7.9: Comparing the CTB-GM1 binding kinetics predicted by the proposed kMC model and the deterministic model proposed by [11] for the GM1 surface density of 0.01 mol%. Ten realizations were plotted for each CTB concentration.

membrane for a mammalian cell is around 5 mol% [267]. Therefore, the continuum assumption, which is necessary for the use of deterministic models, may not be valid because the number of receptors on a cell membrane is far smaller than Avogadro's number, resulting in the rate of the CTB-GM1 binding process to fluctuate [126]. Therefore, the proposed kMC model is more suitable than deterministic models to capture the fluctuation in the binding kinetics attributed to low CTB concentrations and GM1 surface densities.

By increasing the CTB concentration from 3 nM (Figure 7.10a) to 15 nM (Figure 7.10d), the difference in the maximum and minimum total number of membrane-bound CTB proteins among

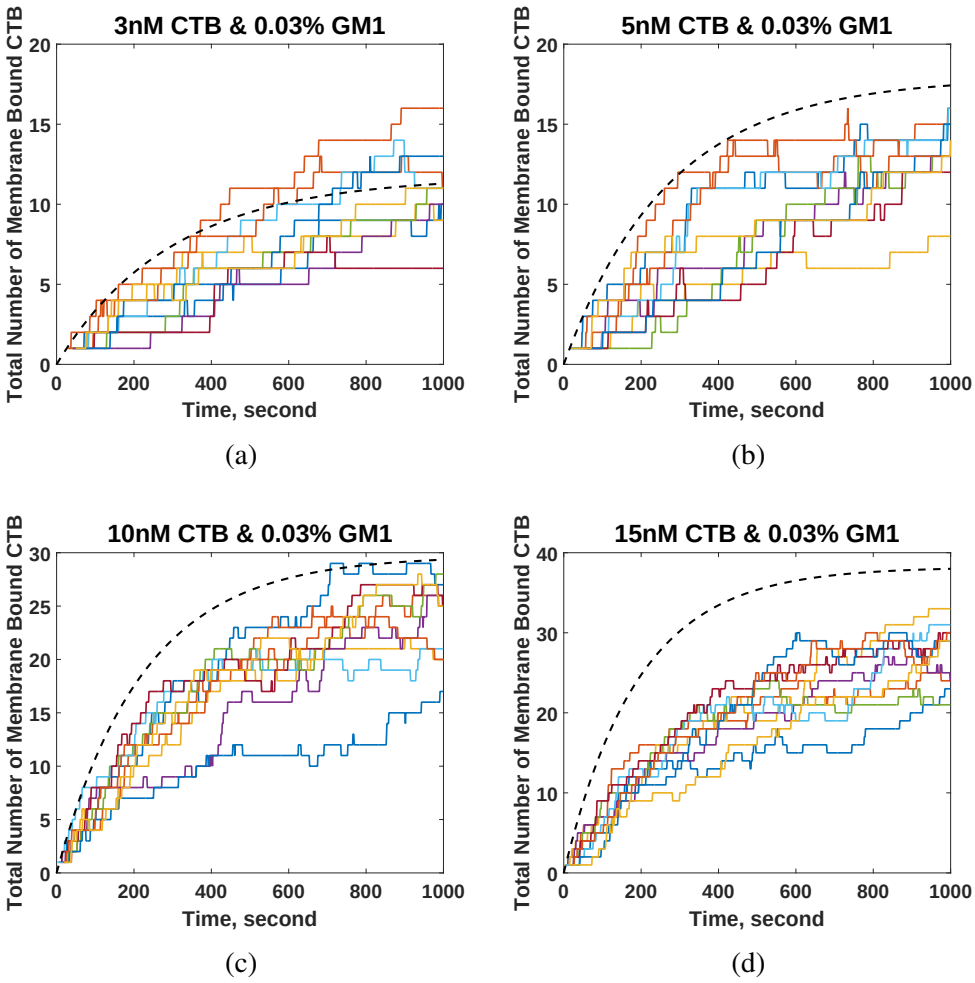


Figure 7.10: Comparing the CTB-GM1 binding kinetics predicted by the proposed kMC model and the deterministic model proposed by [11] for the GM1 surface density of 0.03 mol%. Ten realizations were plotted for each CTB concentration.

ten realizations at 1000 seconds reduced by 30% at the GM1 surface density of 0.03 mol%. Also, the same difference decreased by 20% when the GM1 surface density increased from 0.01 mol% (Figure 7.9d) to 0.03 mol% (Figure 7.10d) at the CTB concentration of 15 nM. These results demonstrated that the fluctuation in the CTB-GM1 binding rates decreased with the increase in the number of CTB proteins and GM1 receptors involved in the binding process; this observation justifies the use of deterministic models at high CTB concentrations and GM1 surface densities.

In addition to the total number of membrane-bound CTB proteins, it is also important to check the number of bound GM1 receptors per a membrane-bound CTB protein on a cell membrane,

which will be denoted as valency. A recent study elaborated that the CTx toxicity increases with the valency [173]. In this work, the average valency over 1000 seconds was simulated using the proposed kMC approach. Figure 7.11 presents the average valency increased with the GM1 surface density, while the average valency decreased as the CTB concentration increased. It should be noted that 100 realizations were computed to calculate the average valency for each condition to obtain a better statistics, and the simulation lattice with $N = 1000$ was used for the GM1 surface density of 0.01 mol% and CTB concentration of 1 nM to minimize the potential finite size effect.

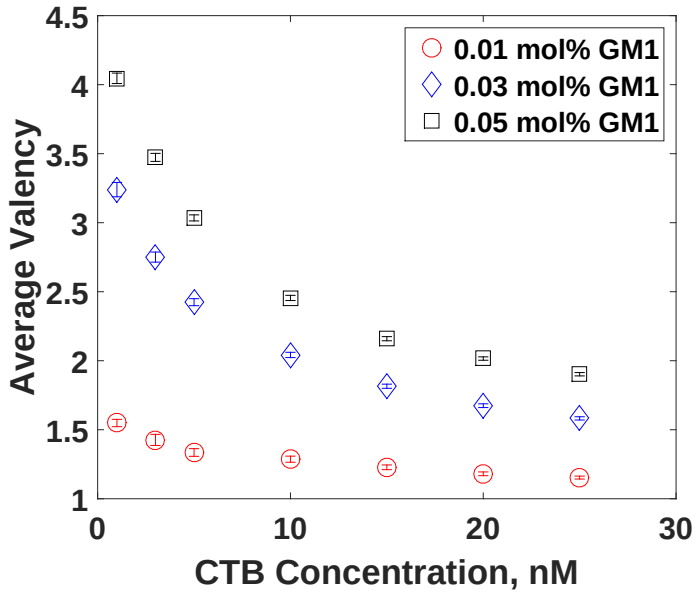


Figure 7.11: Average valency under various GM1 surface densities and CTB concentrations at $t = 1000$ seconds. Data points are given as means $\pm$ standard error of means.

7.5 Discussion

In this work, a kMC model was proposed to describe the CTB-GM1 binding kinetics by considering various microscopic processes in the solution phase as well as on a cell membrane. By taking into account the significance of each microscopic process depending on its reaction rate, the proposed kMC model was able to describe the stochastic nature of the CTB-GM1 binding ki-

netics, particularly when the number of participating CTB proteins and GM1 receptors is low. At the same time, the simulated profiles were in reasonable agreement with those predicted by the experimentally validated deterministic model proposed by [11]. It should be noted a stochastic model and a deterministic model are likely to make different predictions when the number of reactants are low [69, 126]. However, Mayalawa et al. [69] demonstrated that a deterministic model can accurately describe a biomolecular interaction on a membrane when the migration rate on the membrane is much faster than those of reactions. Based on the parameter values available from the literature (Table 4.1), the CTB-GM1 binding process is expected to be a migration-dominating process as described above. Consequently, the deterministic model by [11] can describe the dynamics accurately enough to be used in assessing the accuracy of the proposed kMC model.

It should be noted that the parameters used in the proposed kMC model were taken from the deterministic model proposed by [11] after a minor modification. However, these values are suboptimal for the kMC model since these values were originally obtained for a deterministic model, resulting in a discrepancy shown in Figures 7.9 and 7.10. The optimal parameter values for the kMC model can be acquired by employing rigorous parameter estimation methods. Specifically, various parameter estimation methods for stochastic dynamic models were proposed such as maximum likelihood methods and Bayesian inference methods [126, 268], but these methods are computationally inefficient because of the high computational requirements for stochastic models to obtain accurate simulation results. One way to handle this computational issue is to derive a reduced model that approximates the dominant dynamics of the original stochastic model during the parameter estimation step [269, 270].

In order to obtain high-throughput data of the CTB-GM1 binding kinetics under different GM1 concentrations, we should be able to manipulate the composition of receptors on a cell membrane. A novel nanocube biosensor, which was recently developed by Wu and his colleagues [7, 6, 155], is a viable option to collect high-throughput data for the parameter estimation as the CTB-GM1 binding kinetics can be measured quantitatively under a cell-membrane-mimicking environment while the composition of receptors on a model membrane can be manipulated.

Lastly, it should be noted that the proposed kMC model is a flexible framework, which can be easily extended to incorporate other characteristics of multivalent binding dynamics on a cell membrane. Recent studies [6, 152, 155, 271] have demonstrated that the presence of weak receptors along with strong receptors leads to an increase in the number of membrane-bound proteins, which is mainly caused by a mechanism called reduction of dimensionality (RD) [155, 271, 272]: once a multivalent protein is localized on a membrane by strong receptors, effective association rates between membrane-bound proteins and both strong and weak receptors are enhanced by several folds due to the reduction of the reaction domain from three to two dimensions [155]. As a cell membrane consists of multiple types of receptors, the RD mechanism is expected to play a key role in the binding process *in vivo*. Therefore, it is necessary for a model to be capable of describing the multivalent membrane reactions accurately to predict a binding process on an actual cell membrane more realistically. And the proposed kMC model can be easily extended to study effects of the presence of other receptors along with GM1 receptors by considering a few more events in the event selection procedure.

Furthermore, previous studies have revealed that the receptors are not uniformly distributed on a cell membrane and its spatial organization is an important factor in determining kinetics of membrane reactions and their subsequent downstream signaling pathways [68, 70, 139, 273]. For example, multiple types of CTB receptors, including GM1 receptors, are found to form microclusters on a cell membrane, which mainly consist of glycolipids, cholesterol and membrane proteins [274]. In [139], the authors showed that the formation of the GM1 clusters on a membrane inhibited the CTB-GM1 binding. On the other hand, the inhibition of the cluster formation by removing cholesterol from a membrane decreased the amount of cholera toxin internalized into the cytoplasm and attenuated the cholera toxin toxicity [275, 276]. As the proposed kMC model explicitly accounts for the temporal evolution of the surface configuration, it can be modified to take into account the clustering of the receptors, which will provide valuable insights on the effects of the clustering on a more realistic binding process.

7.6 Conclusion

The present work focuses on the development of a kMC modeling framework to describe the CTB-GM1 binding process on a cell membrane. Since CTB proteins are initially in the solution phase while GM1 receptors are localized on the cell membrane, the proposed modeling framework considered the binding processes over the two spatial domains. Specifically, the solution phase binding model was developed by taking into account the diffusion, attachment and detachment processes of CTB proteins utilizing the steady-state diffusion equation. On the other hand, a lattice kMC model was developed to account for the GM1 receptor migration and the multi-step surface association and dissociation reactions between membrane-bound CTB proteins and GM1 receptors on a cell membrane. Along with the surface binding reactions, the intermolecular distances between membrane-bound CTB proteins and free GM1 receptors were considered to account for the effect of migration on the surface micro-configuration, which was not considered by [11]. For computational efficiency, the GM1 migration process was decoupled from the kMC event selection procedure to reduce the high computational cost introduced by considering the GM1 migration. Then, a mass balance equation was implemented to integrate the lattice kMC model with the solution phase model, and the integrated model was compared with the experimentally validated deterministic model proposed by [11]. The reasonable agreement between predictions by two models demonstrated the prediction accuracy of the proposed kMC model. Furthermore, the proposed model was able to capture the stochastic nature of the CTB-GM1 binding kinetics on a cell membrane, which became more evident with a low copy number of CTB proteins and GM1 receptors.

8. SUMMARY & FUTURE WORK

8.1 Summary

Due to the intrinsic stochasticity as well as the cell-to-cell differences, the signaling dynamics in a clonal population of cells exhibit cell-to-cell variability at the single-cell level, which is distinct from the population-average dynamics. Even the bacteria, which are the simplest form of life, also show population heterogeneity, which suggests the cellular heterogeneity is a ubiquitous feature of every population of living organisms [277]. Therefore, it is of great interest for research communities to explore onsets and roles of such cellular heterogeneity. For example, Stamatakis [277] suggested two reasons why one needs to study the cellular heterogeneity: first, in order for utilizing mammalian or bacterial cell bioreactors to produce commercial products, it is important to explicitly consider the cell-to-cell variability since it will directly impact the production yield as well as the process design and control of bioreactors; and second, it has been suggested that the cellular heterogeneity is important for survival of cells even under sudden changes in cellular environment. Hence, this study aims to capture and investigate these important phenomena through computational modeling approaches. Specifically, this study proposed systematic methods to address two main sources of the cellular heterogeneity, that is, cell-to-cell differences and reaction stochasticity.

For developing a model that can address the first source of the cellular heterogeneity, an IBPM modeling approach is implemented. Development of an IBPM consists of two steps, where its ODE model needs to be constructed first and its model parameters' PDFs need to be inferred from the PDFs of the measurements. Hence, this study presented how to achieve each goal in a systematic manner. For developing an ODE model, two scenarios are considered. First, when the underlying mechanisms are well understood, an integrative approach that consists of *in vitro* experimentations, parameter selection and estimation, model validation, and model refinement is implemented so that model predictions match well with experimental observations. In the second

scenario, we presented two different numerical approaches to construct an accurate model even when underlying mechanisms are only partially known. The first method aims to construct a time-varying model by assuming the most influential model parameters to be piecewise constant whose temporal dynamics are determined by temporal clustering of measurements and parameter estimation. On the other hand, the second method constructs a hybrid model, where its state dynamics are constantly adjusted by additional correction terms inferred from experimental datasets. While estimating values of these necessary correction terms, the dimension of the correction terms is minimized by the graph-theoretical approach to avoid ill-posedness as well as the computational costs of the inference problem. Once an accurate ODE model is developed by any of three methods mentioned earlier, parameters' PDFs are inferred from the measurement PDFs. This study proposed a three-step procedure to estimate parameters' PDFs at a reasonable computational requirement.

Lastly, we propose a kMC modeling framework to incorporate reaction stochasticity in a receptor-ligand binding system, which is an initial event in an intracellular signaling pathway. This is necessary since a receptor-ligand binding process is another source of the cellular heterogeneity. While the IBPM focuses on capturing the cell-to-cell differences, a kMC model is designed to capture stochasticity that is intrinsic to a biochemical reaction due to low copy numbers of reactants involved in the process. Furthermore, the proposed kMC model offers a way to take into account the multivalent binding, which is common for many receptor-ligand interactions on a cell membrane, migration of receptors on a membrane, and non-uniform distributions of receptors on a membrane. Since these features along with the reaction stochasticity are not easily captured by an ODE or IBPM, such a kMC model offers a flexible modeling framework to study how the stochasticity and heterogeneous spatial configurations contribute to the cellular heterogeneity in an intracellular signaling pathway.

8.2 Future Work

This study laid the foundations for formulating a multiscale mathematical model for describing cellular heterogeneity in response to perturbation in their environment. In what follows, a list of potential improvements currently under consideration in the lab is provided:

1. Construction of an IBPM requires an accurate ODE model before estimating PDFs of its model parameters. Constructing an accurate ODE model requires us to address both parametric and structural uncertainties. In this regard, Chapters 2 and 3 addressed how we can estimate the values of model parameters so that we can improve the model accuracy. Such parameter estimation scheme assumes that the underlying model structure is accurate. On the other hand, Chapter 5 discussed how we can construct a hybrid model to infer correct state trajectories for addressing the underlying structural model uncertainty. Since it is likely to have both parametric and structural model uncertainty while developing a model, Dr. Kwon's group aims to propose a simultaneous parameter and state estimation scheme. Since the numbers of model parameters and states are usually large for many intracellular signaling pathway models, a dimension reduction scheme combining sensitivity analysis and observability analysis will be implemented to determine a subset of states and parameters that need to be estimated from experiments. Once this subset is identified, a regularized least-squares problem can be solved to determine parameter values as well as trajectories of states.

2. In this study, optimization-based approaches have been implemented to infer parameter values. Such approaches aim to obtain the best estimates of parameter values by minimizing discrepancy between the experimental measurements and model's predictions. Alternatively, Bayesian approaches can be used to estimate unknown parameter values [38, 43, 46]. On top of estimating parameter values, a Bayesian approach also computes uncertainty associated with the estimated values of model parameters. The computed uncertainty will help assess the robustness of the estimation results. In this regard, Dr. Kwon's lab intends to implement a Bayesian approach such as approximate Bayesian computation methodology [38] to estimate the parameter values as well as their uncertainties to assess how reliable they are.

3. Constructing a first-principle model for an intracellular signaling pathway is to translate our prior understanding of the pathway into a model that can be used for *in silico* experiments to validate hypotheses regarding the system and generate new ones to be tested in a new

set of experiments. For this purpose, the model needs to be trained and validated by experimental measurements. Chapter 2 discussed how to improve the NFκB model accuracy by iterations between model training, validation, and model updates. But, the final model presented in the Chapter still did not agree perfectly with the experimental results, which suggests the model deficiency. In this regard, Dr. Kwon's lab plans to further improve the accuracy of the model by further experimentation and model update. Specifically, we plan to incorporate BFA-induced ER-stress response mechanisms much more in details so that the model accuracy can be improved. The improved model will be further trained and validated by new experimental measurements. To obtain a good model training result, it is imperative to obtain new reliable experimental measurements. To this end, we will aim to measure expression levels of TNFα and other proteins, which are secreted to extracellular matrix. Specifically, we want to measure such proteins instead of IκBα, since expression levels of these secreted proteins are expected to be higher than proteins that remain inside a cell; as a result, the noise-to-signal ratio in measurements will be smaller and thus increase the estimation accuracy.

4. In this study, two sources of the cellular heterogeneity are modeled separately with different modeling approaches. Specifically, the cell-to-cell differences are modeled by IBPM by estimating the PDFs of model parameters while the reaction stochasticity is considered through the kMC simulation framework. Since both sources are important and biologically significant, a multiscale model is needed to integrate these two approaches to consider both sources of the cellular heterogeneity. Specifically, in this multiscale model, the on-lattice kMC model will explicitly consider spatial distributions of biomolecules on cell membranes and stochastic binding dynamics between these molecules and external molecules from a cellular environment. As such receptor-ligand dynamics are usually the initiating events for many intracellular signaling pathways, it is important to investigate how the stochastic dynamics on cell membranes will propagate along intracellular signaling pathways and how these integrated dynamics will contribute to the cell-to-cell heterogeneity. In this regard, in

the future, Dr. Kwon's lab aims to develop a multiscale modeling framework incorporating IBPM and kMC so that both sources of the cellular heterogeneity are incorporated to study the origins and impacts of such heterogeneity.

5. Although the parameter estimation has been performed for an ODE or an IBPM model to enhance their predictive power, the parameters for a kMC model have not be estimated from experimental measurements. In Chapter 7, by using the same parameter values, the accuracy of the proposed kMC model was validated by comparing its predictions with its deterministic counterpart. However, the proposed kMC model is intrinsically different from the deterministic one since the kMC model explicitly considers the surface migration as well as the effects of membrane configurations. It can be erroneous for the kMC model to use the parameter values derived from the deterministic model. Hence, the parameter estimation is necessary for the kMC model to increase its accuracy and interpretability. In this regard, Dr. Kwon's lab aims to develop a computationally efficient parameter estimation scheme for a kMC model by developing a surrogate model for the kMC model as well as the adaptive sampling to reduce the computational costs.

REFERENCES

[1] A. B. Caldwell, Z. Cheng, J. D. Vargas, H. A. Birnbaum, and A. Hoffmann, "Network dynamics determine the autocrine and paracrine signaling fucntions of TNF," *Genes & Development*, vol. 28, pp. 2120–2133, 2014.

[2] M. Junkin, A. J. Kaestli, Z. Cheng, C. Jordi, C. Albayrak, A. Hoffmann, and S. Tay, "High-content quantification of single-cell immune dynamics," *Cell Reports*, vol. 15, pp. 411–422, 2016.

[3] S. Maiti, W. Dai, R. C. Alaniz, J. Hahn, and A. Jayaraman, "Mathematical modeling of pro- and anti-inflammatory signaling in macrophages," *Processes*, vol. 3, pp. 1–18, 2015.

[4] K. J. Baldauf, J. M. Royal, K. T. Hamorsky, and N. Matoba, "Cholera toxin B: One subunit with many pharmaceutical applications," *Toxins*, vol. 7, no. 3, pp. 974–996, 2015.

[5] H. Lin, E. N. Kitova, and J. S. Klassen, "Measuring positive cooperativity using the direct ESI-MS assay. cholera toxin B subunit homopentamer binding to GM1 pentasaccharide," *American Society for Mass Spectrometry*, vol. 25, no. 1, pp. 104–110, 2013.

[6] N. C. Worstell, P. Krishnan, J. D. Weatherston, and H.-J. Wu, "Binding cooperativity matters: A GM1-like ganglioside-cholera toxin B subunit binding study using a nanocube-based lipid bilayer array," *PLoS ONE*, vol. 11, no. 4, p. e0153265, 2016.

[7] H.-J. Wu, J. Henzie, W.-C. Lin, C. Rhodes, Z. Li, E. Sartorel, J. Thorner, P. Yang, and J. T. Groves, "Membrane-protein binding measured with solution-phase plasmonic nanocube sensors," *Nature Methods*, vol. 9, no. 12, pp. 1189–1191, 2012.

[8] D. Lee, Y. Ding, A. Jayaraman, and J. S. Kwon, "Mathematical modeling and parameter estimation of intracellular signaling pathway: Application to LPS-induced NFκB activation and TNFα production in macrophages," *Processes*, vol. 6, p. 21, 2018.

[9] S. L. Werner, J. D. Kearns, V. Zadorozhnaya, C. Lynch, E. O'Dea, M. P. Boldin, A. Ma, D. Baltimore, and A. Hoffmann, "Encoding NF-κB temporal control in response to TNF: distinct roles for the negative regulators IκBα and A20," *Genes & Development*, vol. 22, pp. 2093–2101, 2008.

[10] J. Hasenauer, S. Waldherr, M. Doszczak, P. Scheurich, N. Radde, and F. Allgöwer, "Analysis of heterogeneous cell populations: A density-based modeling and identification framework," *Journal of Process Control*, vol. 21, pp. 1417–1425, 2011.

[11] S. Lauer, B. Goldstein, R. L. Nolan, and J. P. Nolan, "Analysis of cholera toxin-ganglioside interactions by flow cytometry," *Biochemistry*, vol. 41, no. 6, pp. 1742–1751, 2002.

[12] J. Hasenauer, *Modeling and parameter estimation for heterogeneous cell populations*. PhD thesis, Universität Stuttgart, 2013.

[13] A. Kiparissides, M. Koutinas, C. Kontoravdi, A. Mantalaris, and E. N. Pistikopoulos, "'Closing the loop' in biological systems modeling - from the in silico to the in vitro," *Automatica*, vol. 47, no. 6, pp. 1147–1155, 2011.

[14] F. S. Fritzsch, C. Dusny, O. Frick, and A. Schmid, "Single-cell analysis in biotechnology, systems biology, and biocatalysis," *Annual Review of Chemical and Biomolecular Engineering*, vol. 3, pp. 129–155, 2012.

[15] J. M. Raser and E. K. O'Shea, "Noise in gene expression: Origins, consequences, and control," *Science*, vol. 309, pp. 2010–2013, 2005.

[16] D. Binder, T. Drepper, K. Erich Jaeger, F. Delvigne, W. Wiechert, D. Kohlheyer, and A. Grünberger, "Homogenizing bacterial cell factories: Analysis and engineering of phenotypic heterogeneity," *Metabolic Engineering*, vol. 42, pp. 145–156, 2017.

[17] F. Delvigne, Q. Zune, A. R. Lara, W. Al-Soud, and S. J. Sørensen, "Metabolic variability in bioprocessing: implications of microbial phenotypic heterogeneity," *Trends in Biotechnology*, vol. 32, pp. 608–616, 2014.

[18] D. Imig, N. Pollak, T. Strecker, P. Scheurich, F. Allgöwer, and S. Waldherr, "An individual-based simulation framework for dynamic, heterogeneous cell populations during extrinsic stimulations," *Journal of Coupled Systems and Multiscale Dynamics*, vol. 3, pp. 143–155, 2015.

[19] A. A. Cohen, N. Geva-Zatorsky, E. Eden, M. Frenkel-Morgenstern, I. Issaeva, A. Sigal, R. Milo, C. Cohen-Saidon, Y. Liron, Z. Kam, L. Cohen, T. Danon, N. Perzov, and U. Alon, "Dynamic proteomics of individual cancer cells in response to a drug," *Science*, vol. 322, pp. 1511–1516, 2008.

[20] A. Colman-Lerner, A. Gordon, E. Serra, T. Chin, O. Resnekov, D. Endy, C. G. Pesce, and R. Brent, "Regulated cell-to-cell variation in a cell-fate decision system," *Nature*, vol. 437, p. 699, 2005.

[21] J. T. Mettetal, D. Muzzey, J. M. Pedraza, E. M. Ozbudak, and A. V. Oudenaarden, "Predicting stochastic gene expression dynamics in single cells," *Proceedings of the National Academy of Sciences*, vol. 103, pp. 7304–7309, 2006.

[22] D. T. Gillespie, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," *Journal of Computational Physics*, vol. 22, pp. 403–434, 1976.

[23] C. Kravaris, J. Hahn, and Y. Chu, "Advances and selected recent developments in state and parameter estimation," *Computers & Chemical Engineering*, vol. 51, pp. 111–123, 2013.

[24] O.-T. Chis, J. R. Banga, and E. Balsa-Canto, "Structural identifiability of systems biology models: A critical comparison of cethods," *PLoS ONE*, vol. 6, p. e27755, 2011.

[25] H. L. Pahl, "Activators and target genes of Rel/NF-κB transcription factors," *Oncogene*, vol. 18, pp. 6853–6866, 1999.

[26] H. Kitano, "System biology: a brief overview," *Science*, vol. 295, pp. 1662–1664, 2002.

[27] K. G. Gadkar, R. Gunawan, and F. J. Doyle, "Iterative approach to model identification of biological networks," *BMC Bioinformatics*, vol. 6, p. 155, 2005.

[28] E. Balsa-Canto, A. A. Alonso, and J. R. Banga, "An iterative identification procedure for dynamic modeling of biochemical networks," *BMC Systems Biology*, vol. 4, p. 11, 2010.

[29] M. Rodriguez-Fernandez, M. Rehberg, A. Kremling, and J. R. Banga, "Simultaneous model discrimination and parameter estimation in dynamic models of cellular systems," *BMC Systems Biology*, vol. 7, p. 76, 2013.

[30] D. R. Penas, D. Henriques, P. Gonzalez, R. Doallo, J. Saez-Rodriguez, and J. R. Banga, "A parallel metaheuristic for large mixed-integer dynamic optimization problems, with applications in computational biology," *PLoS ONE*, vol. 12, p. e0182186, 2017.

[31] M. Maurya, S. Bornheimer, V. Venkatasubramanian, and S. Subramaniam, "Mixed-integer nonlinear optimisation approach to coarse-graining biochemical networks," *IET Systems Biology*, vol. 3, pp. 24–39, 2008.

[32] P. J. Verheijen, "Model selection: An overview of practices in chemical engineering," *Computer Aided Chemical Engineering*, vol. 16, pp. 85–104, 2003.

[33] F. Boukouvala, R. Misener, and C. A. Floudas, "Global optimization advances in mixed-integer nonlinear programming, MINLP, and constrained derivative-free optimization, CDFO," *European Journal of Operational Research*, vol. 252, pp. 701–727, 2016.

[34] S. Waldherr, "Estimation method for heterogeneous cell population models in systems biology," *Journal of the Royal Society Interface*, vol. 15, p. 20180530, 2018.

[35] B. Munsky, B. Trinh, and M. Khammash, "Listening to the noise: random fluctuations reveal gene network parameters," *Molecular Systems Biology*, vol. 5, p. 318, 2009.

[36] C. Zechner, J. Ruess, P. Krenn, S. Pelet, M. Peter, J. Lygeros, and H. Koeppl, "Moment-based inference predicts bimodality in transient gene expression," *Proceedings of the National Academy of Sciences*, vol. 109, pp. 8340–8345, 2012.

[37] Z. Cheng, B. Taylor, D. R. Ourthiague, and A. Hoffmann, "Distinct single-cell signaling characteristics are conferred by the MyD88 and TRIF pathways during TLR4 activation," *Science Signaling*, vol. 8, p. ra69, 2015.

[38] J. Liepe, P. Kirk, S. Filippi, T. Toni, C. P. Barnes, and M. P. H. Stumpf, "A framework for parameter estimation and model selection from experimental data in systems biology using approximate bayesian computation," *Nature Protocols*, vol. 9, pp. 439–456, 2014.

[39] P. Nicolas, K.-M. Kim, D. Shibata, and S. Tavaré, "The stem cell population of the human colon crypt: Analysis via methylation patterns," *PLoS Comutational Biology*, vol. 3, p. e28, 2007.

[40] C. Loos and J. Hasenauer, "Mathematical modeling of variability in intracellular signaling," *arXiv*, vol. 1904.08182, 2019.

[41] H. Koeppl, C. Zechner, A. Ganguly, S. Pelet, and M. Peter, "Accounting for extrinsic variability in the estimation of stochastic rate constants," *International Journal of Robust and Nonlinear Control*, vol. 22, pp. 1103–1119, 2012.

[42] J. Hasenauer, S. Waldherr, M. Doszczak, N. Radde, P. Scheurich, and F. Allgöwer, "Identification of models of heterogeneous cell populations from population snapshot data," *BMC Bioinformatics*, vol. 12, p. 125, 2011.

[43] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf, "Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems," *Journal of The Royal Society Interface*, vol. 6, pp. 187–202, 2008.

[44] G. Lillacci and M. Khammash, "The signal within the noise: efficient inference of stochastic gene regulation models using fluorescence histograms and stochastic simulations," *Bioinformatics*, vol. 29, pp. 2311–2319, 2013.

[45] J. Yao, A. Pilko, and R. Wollman, "Distinct cellular states determine calcium signaling response," *Molecular Systems Biology*, vol. 12, p. 894, 2016.

[46] O. Lenive1, P. D. W. Kirk, and M. P. H. Stumpf, "Inferring extrinsic noise from single-cell gene expression data using approximate bayesian computation," *BMC Systems Biology*, vol. 10, p. 81, 2016.

[47] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer, "Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood," *Bioinformatics*, vol. 25, pp. 1923–1929, 2009.

[48] S. Zeng, S. Waldherr, C. Ebenbauer, and F. Allgöwer, "Ensemble observability of linear systems," *IEEE Transactions on Automatic Control*, vol. 61, pp. 1452–1465, 2016.

[49] S. Waldherr, S. Zeng, and F. Allgöwer, "Identifiability of population models via a measure theoretical approach," in *Proceedings of 19th World Congress of the International Federation of Automatic Control*, pp. 1717–1722, 2014.

[50] S. Zeng, S. Waldherr, and F. Allgöwer, "An inverse problem of tomographic type in population dynamics," in *53rd IEEE Conference on Decision and Control*, pp. 1643–1648, 2014.

[51] M. Nayhouse, J. S.-I. Kwon, P. D. Christofides, and G. Orkoulas, "Crystal shape modeling and control in protein crystal growth," *Chemical Engineering Science*, vol. 87, pp. 216–223, 2013.

[52] S. C. Ke, L. J. DeLucas, and J. G. Harrison, "Crystal simulation of portein crystal growth using aggregates as the growth unit," *Journal of Physics D: Applied Physics*, vol. 31, pp. 1064–1070, 1998.

[53] J. S.-I. Kwon, M. Nayhouse, P. D. Christofides, and G. Orkoulas, "Modeling and control of protein crystal shape and size in batch crystallization," *AIChE Journal*, vol. 59, pp. 2317–2327, 2013.

[54] J. S.-I. Kwon, M. Nayhouse, P. D. Christofides, and G. Orkoulas, "Modeling and control of crystal shape in continuous protein crystallization," *Chemical Engineering Science*, vol. 107, pp. 47–57, 2014.

[55] J. S.-I. Kwon, M. Nayhouse, G. Orkoulas, and P. D. Christofides, "Crystal shape and size control using a plug flow crystallization configuration," *Chemical Engineering Science*, vol. 119, pp. 30–39, 2014.

[56] Y. Lou and P. D. Christofides, "Estimation and control of surface roughness in thin film growth using kinetic Monte-Carlo models," *Chemical Engineering Science*, vol. 58, pp. 3115–3129, 2003.

[57] A. Varshney and A. Armaou, "Multiscale optimization using hybrid PDE/kMC process systems with application to thin film growth," *Chemical Engineering Science*, vol. 60, pp. 6780–6794, 2005.

[58] A. Varshney and A. Armaou, "Identification of macroscopic variables for low-order modeling of thin-film growth," *Industrial & Engineering Chemistry Research*, vol. 45, pp. 8290–8298, 2006.

[59] M. Crose, J. S.-I. Kwon, M. Nayhouse, D. Ni, and P. D. Christofides, "Multiscale modeling and operation of PECVD of thin film solar cells," *Chemical Engineering Science*, vol. 136, pp. 50–61, 2015.

[60] M. Crose, J. S.-I. Kwon, A. Tran, and P. D. Christofides, "Multiscale modeling and run-to-run control of PECVD of thin film solar cells," *Renewable Energy*, vol. 100, pp. 129–140, 2017.

[61] D. G. Vlachos, "Multiscale integration hybrid algorithms for homogeneous-heterogeneous reactors," *AIChE Journal*, vol. 43, pp. 3031–3041, 1997.

[62] A. Varshney and A. Armaou, "Reduced order modeling and dynamic optimization of multiscale PDE/kMC process systems," *Computers & Chemical Engineering*, vol. 32, pp. 2136–2143, 2008.

[63] M. Stamatakis, Y. Chen, and D. G. Vlachos, "First-principles-based kinetic Monte Carlo simulation of the structure sensitivity of the water-gas shift reaction on platinum surfaces," *The Journal of Physical Chemistry C*, vol. 115, pp. 24750–24762, 2011.

[64] S. Piccinin and M. Stamatakis, "CO oxidation on Pd(111): A first-principles-based kinetic Monte Carlo study," *ACS Catalysis*, vol. 4, pp. 2143–2152, 2014.

[65] J. Li, E. Croiset, and L. A. Ricardez-Sandoval, "Carbon nanotube growth: First-principles-based kinetic Monte Carlo model," *Journal of Catalysis*, vol. 326, pp. 15–25, 2015.

[66] D. Chaffart, S. Rasoulian, and L. A. Ricardez-Sandoval, "Distributional uncertainty analysis and robust optimization in spatially heterogeneous multiscale process systems," *AIChE Journal*, vol. 62, pp. 2374–2390, 2016.

[67] D. Chaffart and L. A. Ricardez-Sandoval, "Robust dynamic optimization in heterogeneous multiscale catalytic flow reactors using polynomial chaos expansion," *Journal of Process Control*, vol. 60, pp. 128–140, 2017.

[68] M. Fallahi-Sichani and J. J. Linderman, "Lipid raft-mediated regulation of G-protein coupled receptor signaling by ligands which influence receptor dimerization: A computational study," *PLoS ONE*, vol. 4, p. e6604, 2009.

[69] K. Mayawala, D. G. Vlachos, and J. S. Edwards, "Spatial modeling of dimerization reaction dynamics in the plasma membrane: Monte Carlo vs. continuum differential equations," *Biophysical Chemistry*, vol. 121, pp. 194–208, 2006.

[70] M. N. Costa, K. Radhakrishnan, B. S. Wilson, D. G. Vlachos, and J. S. Edwards, "Coupled stochastic spatial and non-spatial simulations of ErbB1 signaling pathways demonstrate the importance of spatial organization in signal transduction," *PLoS ONE*, vol. 4, p. e6316, 2009.

[71] J. Chen, S. C. Almo, and Y. Wu, "General principles of binding between cell surface receptors and multi-specific ligands: A computational study," *PLoS Computational Biology*, vol. 13, p. e1005805, 2017.

[72] J. J. Hughey, T. K. Lee, and M. W. Covert, "Computational modeling of mammalian signaling networks," *WIREs Systems Biology and Medicine*, vol. 2, no. 2, pp. 194–209, 2010.

[73] L. N. Handly, J. Yao, and R. Wollman, "Signal transduction at the single-cell level: Approaches to study the dynamic nature of signaling networks," *Journal of Molecular Biology*, vol. 428, no. 19, pp. 3669–3682, 2016.

[74] S. Gaudet and K. Miller-Jensen, "Redefining signaling pathways with an expanding single-cell toolbox," *Trends in Biotechnology*, vol. 34, no. 6, pp. 458–469, 2016.

[75] R. A. Williams, J. Timmis, and E. E. Qwarnstrom, "Computational models of the NF-κB signalling pathway," *Computation*, vol. 2, no. 4, pp. 131–158, 2014.

[76] C. Kravaris and J. H. Seinfeld, "Identification of parameters in distributed parameter systems by regularization," *SIAM Journal on Control and Optimization*, vol. 23, no. 2, pp. 217–241, 1985.

[77] Q. Xue, Y. Lu, M. R. Eisele, E. S. Sulistijo, N. Khan, R. Fan, and K. Miller-Jensen, "Analysis of single-cell cytokine secretion reveals a role for paracrine signaling in coordinating macrophage responses to TLR4 stimulation," *Science Signaling*, vol. 8, no. 381, p. ra59, 2015.

[78] S. Tay, J. J. Hughey, T. K. Lee, T. Lipniacki, S. R. Quake, and M. W. Covert, "Single-cell NF-κB dynamics reveal digital activation and analogue information processing," *Nature*, vol. 466, pp. 267–271, 2010.

[79] M. S. Hayden and S. Ghosh, "NF-κB, the first quarter-century: remarkable progress and outstanding questions," *Genes & Development*, vol. 26, no. 6, pp. 203–234, 2012.

[80] Y. Chu and J. Hahn, "Parameter set selection via clustering of parameters into pairwise indistinguishable groups of parameters," *Industrial & Engineering Chemistry Research*, vol. 48, no. 13, pp. 6000–6009, 2009.

[81] K. Z. Yao, B. M. Shaw, B. Kou, and K. B. McAuley, "Modeling ethylene/butene copolymerization with multi-site catalysts: Parameter estimability and experimental design," *Polymer Reaction Engineering*, vol. 3, no. 3, pp. 563–588, 2003.

[82] C. Prussin, "Cytokine flow cytometry: Understanding cytokine biology at the single-cell level," *Journal of Clinical Immunology*, vol. 17, no. 3, p. 195, 1997.

[83] K. R. Schulz, E. A. Danna, P. O. Krutzik, and G. P. Nolan, "Single-cell phospho-protein analysis by flow cytometry," *Current Protocols in Immunology*, pp. 8–17, 2012.

[84] A. Hoffmann, A. Levchenko, M. L. Scott, and D. Baltimore, "The IκB-NF-κB signaling module: temporal control and selective gene activation," *Science*, vol. 298, pp. 1241–1245, 2002.

[85] A. Krikos, C. D. Laherty, and V. M. Dixit, "Transcriptional activation of the tumor necrosis factor α-inducible zinc finger protein, A20, is mediated by κB elements," *Journal of Biological Chemistry*, vol. 267, 1992.

[86] E. G. Lee, D. L. Boone, S. Chai, S. L. Libby, M. Chien, J. P. Lodolce, and A. Ma, "Failure to regulate TNF-induced NF-κB and cell death responses in A20-deficient mice," *Science*, vol. 289, no. 5488, pp. 2350–2354, 2000.

[87] D. L. Boone, E. E. Turer, E. G. Lee, R.-C. Ahmad, M. T. Wheeler, C. Tsui, P. Hurley, M. Chien, S. Chai, O. Hitotsumatsu, E. McNally, C. Pickart, and A. Ma, "The ubiquitin-modifying enzyme A20 is required for termination of Toll-like receptor responses," *Nature Immunology*, vol. 5, pp. 1052–1060, 2004.

[88] P. Chardin and F. McCormick, "Brefeldin A: The advantage of being uncompetitive," *Cell*, vol. 97, pp. 153–155, 1999.

[89] T. H. Ward, R. S. Polishchuk, S. Caplan, K. Hirschberg, and J. Lippincott-Schwartz, "Maintenance of Golgi structure and function depends on the integrity of ER export," *Journal of Cell Biology*, vol. 155, pp. 557–570, 2001.

[90] E. Latz, A. Visintin, E. Lien, K. A. Fitzgerald, B. G. Monks, E. A. Kurt-Jones, D. T. Golenbock, and T. Espevik, "Lipopolysaccharide rapidly traffics to and from the Golgi apparatus with the Toll-like receptor 4-MD-2-CD14 complex in a process that is distinct from the initiation of signal transduction," *Journal of Biological Chemistry*, vol. 49, pp. 47834–47843, 2002.

[91] A. Liaunardy-Jopeace, C. E. Bryant, and N. J. Gay, "The COPII adaptor protein TMED7 is required to initiate and mediate the anterograde trafficking of Toll-like receptor 4 to the plasma membrane," *Science Signaling*, vol. 7, no. 336, p. ra70, 2014.

[92] D. Wang, J. Lou, C. Ouyang, W. Chen, Y. Liu, X. Liu, X. Cao, J. Wanga, and L. Lu, "Ras-related protein Rab10 facilitates TLR4 signaling by promoting replenishment of TLR4 onto the plasma membrane," *Proceedings of the National Academy of Sciences*, vol. 107, no. 31, pp. 13806–13811, 2010.

[93] S. J. Jones, E. C. Ledgerwood, J. B. Prins, J. Galbraith, D. R. Johnson, J. S. Pober, and J. R. Bradley, "TNF recruits TRADD to the plasma membrane but not the trans-golgi network, the principal subcellular location of TNF-R1," *Journal of Immunology*, vol. 162, pp. 1042–1048, 1999.

[94] N. Neznanov, A. Kondratova, K. M. Chumakov, B. Angres, B. Zhumabayeva, V. I. Agol, and A. V. Gudkov, "Poliovirus protein 3a inhibits tumor necrosis factor (TNF)-induced apoptosis by eliminating the TNF receptor from the cell surface," *Journal of Virology*, vol. 75, no. 21, pp. 10409–10420, 2001.

[95] J. Xaus, M. Comalada, A. F. Valledor, J. Lloberas, F. López-Soriano, J. M. Argilés, C. Bogdan, and A. Celada, "LPS induces apoptosis in macrophages mostly through the autocrine production of TNF-α," *Blood*, vol. 95, pp. 3823–3831, 2000.

[96] M. W. Covert, T. H. Leung, J. E. Gaston, and D. Baltimore, "Achieving stability of lipopolysaccharide-induced NF-κB activation," *Science*, vol. 309, no. 5742, pp. 1854–1857, 2005.

[97] E. Lombardo, A. Alvarez-Barrientos, B. Maroto, L. Boscà, and U. G. Knaus, "TLR4-mediated survival of macrophages is MyD88 dependent and requires TNF-α autocrine signalling," *Journal of Immunology*, vol. 178, no. 6, pp. 3731–3739, 2007.

[98] I. Zanoni, R. Ostuni, L. R. Marek, S. Barresi, R. Barbalat, G. M. Barton, F. Granucci, and J. C. Kagan, "CD14 controls the LPS-induced endocytosis of Toll-like receptor 4," *Cell*, vol. 147, no. 4, pp. 868–880, 2011.

[99] Y. Tan, I. Zanoni, T. W. Cullen, A. L. Goodman, and J. C. Kagan, "Mechanisms of Toll-like receptor 4 endocytosis reveal a common immune-evasion strategy used by pathogenic and

commensal bacteria," *Immunity*, vol. 43, pp. 909–922, 2015.

[100] R. Rajaiah, D. J. Perkins, D. D. C. Ireland, and S. N. Vogel, "CD14 dependence of TLR4 endocytosis and TRIF signaling displays ligand specificity and is dissociable in endotoxin tolerance," *Proceedings of the National Academy of Sciences*, vol. 112, no. 27, pp. 8391–8396, 2013.

[101] J. C. Kagan, T. Su, T. Horng, A. Chow, S. Akira, and R. Medzhitov, "TRAM couples endocytosis of Toll-like receptor 4 to the induction of interferon-β," *Nature Immunology*, vol. 9, pp. 361–368, 2008.

[102] Y. Chu, A. Jayaraman, and J. Hahn, "Parameter sensitivity analysis of IL-6 signalling pathways," *IET Systems Biology*, vol. 1, no. 6, pp. 342–352, 2007.

[103] S.-C. Lin, Y.-C. Lo, and H. Wu, "Helical assembly in the MyD88-IRAK4-IRAK2 complex in TLR/IL-1R signalling," *Nature*, vol. 465, pp. 885–890, 2010.

[104] J. Bagnall, C. Boddington, J. Boyd, R. Brignall, W. Rowe, N. A. Jones, L. Schmidt, D. G. Spiller, M. R. White, and P. Paszek, "Quantitative dynamic imaging of immune cell signalling using lentiviral gene transfer," *Integrative Biology*, vol. 7, pp. 713–725, 2015.

[105] C. Moya, Z. Huang, P. Cheng, A. Jayaraman, and J. Hahn, "Investigation of IL-6 and IL-10 signalling via mathematical modelling," *IET Systems Biology*, vol. 5, no. 1, pp. 15–26, 2011.

[106] M.-H. Sung, N. Li, Q. Lao, R. A. Gottschalk, G. L. Hager, and I. D. C. Fraser, "Switching of the relative dominance between feedback mechanisms in lipopolysaccharide-induced NF-κB signaling," *Science Signaling*, vol. 7, no. 308, p. ra6, 2014.

[107] H. Tsukamoto, K. Fukudome, S. Takao, N. Tsuneyoshi, and M. Kimoto, "Lipopolysaccharide-binding protein-mediated Toll-like receptor 4 dimerization enables rapid signal transduction against lipopolysaccharide stimulation on membrane-associated CD14-expressing cells," *International Immunology*, vol. 22, no. 4, pp. 271–280, 2010.

[108] J. Sakai, E. Cammarota, J. A. Wright, P. Cicuta, R. A. Gottschalk, N. Li, I. D. C. Fraser, and C. E. Bryant, "Lipopolysaccharide-induced NF-κB nuclear translocation is primarily

dependent on MyD88, but TNFα expression requires TRIF and MyD88," *Scientific Reports*, vol. 7, p. 1428, 2017.

[109] R.-G. Shao, T. Shimizu, and Y. Pommier, "Brefeldin A is a potent inducer of apoptosis in human cancer cells independently of p53," *Experimental Cell Research*, vol. 227, pp. 190–196, 1996.

[110] J. L. Moon, S. Y. Kim, S. W. Shin, and J.-W. Park, "Regulation of brefeldin a-induced ER stress and apoptosis by mitochondrial NADP+-dependent isocitrate dehydrogenase," *Biochemical and Biophysical Research Communications*, vol. 417, pp. 760–764, 2012.

[111] N. Cláudio, A. Dalet, E. Gatti, and P. Pierre, "Mapping the crossroads of immune activation and cellular stress response pathways," *EMBO Journal*, vol. 32, pp. 1214–1224, 2013.

[112] H. Mellor, S. R. Kimball, and L. S. Jefferson, "Brefeldin A inhibits protein synthesis through the phosphorylation of the α-subunit of eukaryotic initiation factor-2," *FEBS Letters*, vol. 350, pp. 143–146, 1994.

[113] A. B. Tam, E. L. Mercado, A. Hoffmann, and M. Niwa, "ER stress activates NF-κB by integrating functions of basal IKK activity, IRE1 and PERK," *PLS ONE*, vol. 7, no. 10, p. e45078, 2012.

[114] P. E. Fishman and P. K. Curran, "Brefeldin A inhibits protein synthesis in cultured cells," *FEBS Letters*, vol. 314, no. 4, pp. 371–374, 1992.

[115] Y. Ando, T. Oku, and T. Tsuji, "Platelet supernatant suppresses LPS-induced nitric oxide production from macrophages accompanied by inhibition of NF-κB signaling and increased arginase-1 expression," *PLoS ONE*, 2016.

[116] J. Selimkhanov, B. Taylor, J. Yao, A. Pilko, J. Albeck, A. Hoffmann, L. Tsimring, and R. Wollman, "Accurate information transmission through dynamic biochemical signaling networks," *Science*, vol. 346, no. 6215, pp. 1370–1373, 2014.

[117] R. A. Kellogg, C. Tian, T. Lipniacki, S. R. Quake, and S. Tay, "Digital signaling decouples activation probability and population heterogeneity," *eLife*, vol. 4, p. e08931, 2015.

[118] A. S. M. Noman, N. Koide, F. Hassan, I. I.-E-Khuda, J. Dagvadorj, G. Tumurkhuu, S. Islam, Y. Naiki, T. Yoshida, and T. Yokochi, "Thalidomide inhibits lipopolysaccharide-induced tumor necrosis factor-a production via down-regulation of MyD88 expression," *Innate Immunity*, vol. 15, no. 1, pp. 33–41, 2009.

[119] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *Journal of Physical Chemistry*, vol. 82, no. 25, pp. 2340–2361, 1977.

[120] E. L. Haseltine and J. B. Rawlings, "Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics," *Journal of Chemical Physics*, vol. 117, pp. 6959–6969, 2002.

[121] Y. Cao and L. R. Petzold, "The numerical stability of leaping methods for stochastic simulation of chemically reacting systems," *Journal of Chemical Physics*, vol. 121, pp. 12169–12178, 2004.

[122] J. S.-I. Kwon, M. Nayhouse, P. D. Christofides, and G. Orkoulas, "Modeling and control of protein crystal shape and size in batch crystallization," *AIChE Journal*, vol. 59, pp. 2317–2327, 2013.

[123] T. Lipniacki, P. Paszek, A. R. Brasier, B. A. Luxon, and M. Kimmel, "Stochastic regulation in early immune response," *Biophysical Journal*, vol. 90, pp. 725–742, 2006.

[124] D. E. Nelson, A. E. C. Ihekwaba, M. Elliott, J. R. Johnson, C. A. Gibney, B. E. Foreman, G. Nelson, V. See, C. A. Horton, D. G. Spiller, S. W. Edwards, H. P. McDowell, J. F. Unitt, E. Sullivan, R. Grimley, N. Benson, D. Broomhead, D. B. Kell, and M. R. H. White, "Oscillations in TNFα signaling control the dynamics of gene expression," *Science*, vol. 306, pp. 704–708, 2004.

[125] D. J. Wilkinson, "Stochastic modelling for quantitative description of heterogeneous biological systems," *Nature Reviews Genetics*, vol. 10, no. 2, p. 122, 2009.

[126] A. Gupta and J. B. Rawlings, "Comparison of parameter estimation methods in stochastic chemical kinetic models: Examples in systems biology," *AIChE Journal*, vol. 60, pp. 1253–1268, 2014.

[127] M. Ashyraliyev, Y. Fomekong-Nanfack, J. A. Kaandorp, and J. G. Blom, "Systems biology: parameter estimation for biochemical models," *The FEBS journal*, vol. 276, pp. 886–902, 2009.

[128] L. Lamoreaux, M. Roederer, and R. Koup, "Intracellular cytokine optimization and standard operating procedure," *Nature Protocols*, vol. 1, no. 3, pp. 1507–1516, 2006.

[129] T. K. Lee, E. M. Denny, J. C. Sanghvi, J. E. Gaston, N. D. Maynard, J. J. Hughey, and M. W. Covert, "A noisy paracrine signal determines the cellular NF-κB response to lipopolysaccharide," *Science Signaling*, vol. 2, no. 93, p. ra65, 2009.

[130] D. Xiu and G. E. Karniadakis, "The Wiener-Askey polynomial chaos for stochastic differential equations," *SIAM Journal on Scientific Computing*, vol. 24, pp. 619–644, 2006.

[131] Y. Misumi, K. M. Yuko Misumi, A. Takatsuki, G. Tamura, and Y. Ikehara, "Novel blockade by brefeldin A of intracellular transport of secretory proteins in cultured rat hepatocytes," *Journal of Biological Chemistry*, vol. 261, pp. 1139–11403, 1986.

[132] C. Bueno, J. Almeida, M. Alguero, M. Sánchez, J. Vaquero, F. Laso, J. S. Miguel, L. Escribano, and A. Orf, "Flow cytometric analysis of cytokine production by normal human peripheral blood dendritic cells and monocytes: Comparative analysis of different stimuli, secretion-blocking agents and incubation periods," *Cytometry Part A*, vol. 46, pp. 33–40, 2001.

[133] R. A. Gottschalk, A. J. Martins, B. R. Angermann, B. Dutta, C. E. Ng, S. Uderhardt, J. S. Tsang, I. D. C. Fraser, M. Meier-Schellersheim, and R. N. Germain, "Distinct NF-κB and MAPK activation thresholds uncouple steady-state microbe sensing from anti-pathogen inflammatory responses," *Cell Systems*, vol. 2, pp. 378–390, 2016.

[134] L. de Haan and T. R. Hirst, "Cholera toxin: A paradigm for multi-functional engagement of cellular mechanisms (review)," *Molecular Membrane Biology*, vol. 21, no. 2, pp. 77–92, 2004.

[135] WHO, "Cholera vaccines: Who position paper," *Weekly Epidemiological Record*, vol. 85, pp. 117–128, 2010.

[136] R.-G. Zhang, D. L. Scott, M. L. Westbrook, S. Nance, B. D. Spangler, G. G. Shipley, and E. M. Westbrook, "The three-dimensional crystal structure of cholera toxin," *Journal of Molecular Biology*, vol. 251, no. 4, pp. 563–573, 1995.

[137] R.-G. Zhang, M. L. Westbrook, E. M. Westbrook, D. L. Scott, Z. Otwinowsk, P. R. Maulik, R. A. Reed, and G. G. Shipley, "The 2.4 A crystal structure of cholera toxin B subunit pentamer: choleragenoid," *Journal of Molecular Biology*, vol. 251, no. 4, pp. 550–562, 1995.

[138] W. B. Turnbull, B. L. Precious, and S. W. Homans, "Dissecting the cholera toxin-ganglioside GM1 interactions by isothermal titration calorimetry," *Journal of the American Chemical Society*, vol. 126, pp. 1047–1054, 2004.

[139] J. Shi, T. Yang, S. Kataoka, Y. Zhang, A. J. Diaz, and P. S. Cremer, "GM1 clustering inhibits cholera toxin binding in supported phospholipid membranes," *Journal of the American Chemical Society*, vol. 129, no. 18, pp. 5954–5961, 2007.

[140] R. Šachl, M. Amaro, G. Aydogan, A. Koukalová, I. I. Mikhalyov, I. A. Boldyrev, J. Humpolíčkovă, and M. Hof, "On multivalent receptor activity of GM1 in cholesterol containing membranes," *Biochimica et Biophysica Acta*, vol. 1853, no. 4, pp. 850–857, 2015.

[141] A. Schön and E. Freire, "Thermodynamics of intersubunit interactions in cholera toxin upon binding to the oligosaccharide portion of its cell surface receptor, ganglioside GM1," *Biochemistry*, vol. 28, no. 12, pp. 5019–5024, 1989.

[142] D. E. Schafer and A. K. Thakur, "Quantitative description of the binding of GM1 oligosaccharide by cholera enterotoxin," *Cell Biophysics*, vol. 4, no. 1, pp. 25–40, 1982.

[143] A. A. Wolf, M. G. Jobling, S. Wimer-Mackin, M. Ferguson-Maltzman, J. L. Madara, R. K. Holmes, and W. I. Lencer, "Ganglioside structure dictates signal transduction by cholera toxin and association with caveolae-like membrane domains in polarized epithelia," *The Journal of Cell Biology*, vol. 141, no. 4, pp. 917–927, 1998.

[144] A. Fujita, J. Cheng, M. Hirakawa, K. Furukawa, S. Kusunoki, and T. Fujimoto, "Gangliosides GM1 and GM3 in the living cell membrane form clusters susceptible to cholesterol depletion and chilling," *Molecular Biology*, vol. 18, no. 6, pp. 2112–2122, 2007.

[145] C. Yuan and L. J. Johnston, "Atomic force microscopy studies of ganglioside GM1 domains in phosphatidylcholine and phosphatidylcholine/cholesterol bilayers," *Biophysical Journal*, vol. 81, no. 2, pp. 1059–1069, 2001.

[146] B. R. Caré and H. A. Soula, "Impact of receptor clustering on ligand binding," *BMC Systems Biology*, vol. 5, no. 1, p. 48, 2011.

[147] C. R. MacKenzie, T. Hirama, K. K. Lee, E. Altman, and N. M. Young, "Quantitative analysis of bacterial toxin affinity and specificity for glycolipid receptors by surface plasmon resonance," *The Journal of Biological Chemistry*, vol. 272, pp. 5533–5538, 1997.

[148] G. M. Kuziemko, M. Stroh, and R. C. Stevens, "Cholera toxin binding affinity and specifictiy for gangliosides determined by surface plasmon resonance," *Biochemistry*, vol. 35, pp. 6375–6384, 1996.

[149] J. H. Seo, C. S. Kim, and H. J. Cha, "Structural evaluation of GM1-related carbonhydrate-cholera toxin interactions through surface plasmon resonance kinetic analysis," *Analyst*, vol. 138, no. 22, p. 6924, 2013.

[150] M. A. Cooper, A. Hansson, S. Löfås, and D. H. Williams, "A vesicle capture sensor chip for kinetic analysis of interactions with membrane-bound receptors," *Analytical Biochemistry*, vol. 277, no. 2, pp. 196–205, 2000.

[151] K. L. Sly and J. C. Conboy, "Determination of multivalent protein-ligand binding kinetics by second-harmonic correlation spectroscopy," *Analytical Chemistry*, vol. 86, pp. 11045–11054, 2014.

[152] M. Yanagisawa, T. Ariga, and R. K. Yu, "Letter to the Glyco-Forum: Cholera toxin B subunit binding does not correlate with GM1 expression: a study using mouse embryonic neural precursor cells," *Glycobiology*, vol. 16, pp. 19G–22G, 2006.

[153] M. E. Breimer, G. C. Hansson, K.-A. Karlsson, G. Larson, and H. Leffler, "Glycosphingolipid composition of epithelial cells isolated along the villus axis of small intestine of a single human individual," *Glycobiology*, vol. 22, no. 12, pp. 1721–1730, 2012.

[154] A. M. Wands, A. Fujita, J. E. McCombs, J. Cervin, B. Dedic, A. C. Rodriguez, N. Nischan, M. R. Bond, M. Mettlen, D. C. Trudgian, A. Lemoff, M. Quiding-Järbrink, B. Gustavsson, C. Steentoft, H. Clausen, H. Mirzaei, S. Teneberg, U. Yrlid, and J. J. Kohler, "Fucosylation and protein glycosylation create functional receptors for cholera toxin," *eLife*, vol. 4, p. e09545, 2015.

[155] P. Krishnan, A. Singla, C.-A. Lee, J. D. Weatherston, N. C. Worstell, and H.-J. Wu, "Hetero-multivalent binding of cholera toxin subunit B with glycolipid mixtures," *Colloids and Surfaces B: Biointerfaces*, vol. 160, pp. 281–288, 2017.

[156] Y. Chu and J. Hahn, "Parameter set selection for estimation of nonlinear dynamic systems," *AIChE Journal*, vol. 53, pp. 2858–2870, 2007.

[157] D. Shoup and A. Szabo, "Role of diffusion in ligand binding to macromolecules and cell-bound receptors," *Biophysical Journal*, vol. 40, no. 1, pp. 33–39, 1982.

[158] B. Goldstein and M. Dembo, "Approximating the effects of diffusion on reversible reactions at the cell surface: ligand-receptor kinetics," *Biophysical Journal*, vol. 68, no. 4, pp. 1222–1230, 1995.

[159] R.-X. Gu, H. I. Ingólfsson, A. H. de Vries, S. J. Marrink, and D. P. Tieleman, "Ganglioside-lipid and ganglioside-protein interactions revealed by coarse-grained and atomistic molec-

ular dynamics simulations," *Journal of Physical Chemistry B*, vol. 121, no. 15, pp. 3262–3275, 2017.

[160] D. S. Patel, S. Park, E. L. Wu, M. S. Yeom, G. Widmalm, J. B. Klauda, and W. Im, "Influence of ganglioside GM1 concentration on lipid clustering and membrane properties and curvature," *Biophysical Journal*, vol. 111, no. 9, pp. 1987–1999, 2016.

[161] H. I. Ingólfsson, M. N. Melo, F. J. van Eerden, C. Arnarez, C. A. Lopez, T. A. Wassenaar, X. Periole, A. H. de Vries, D. P. Tieleman, and S. J. Marrink, "Lipid organization of the plasma membrane," *Journal of the American Chemical Society*, vol. 136, no. 41, pp. 14554–14559, 2014.

[162] D. A. Lauffenburger and J. J. Linderman, *Receptors: Models for Binding, Trafficking, and Signaling*. New York, USA: Oxford University Press, 1993.

[163] K. A. P. McLean and K. B. McAuley, "Mathematical modeling of chemical processes - obtaining the best model predictions and parameter estimates using identifiability and estimability procedures," *Canadian Journal of Chemical Engineering*, vol. 90, pp. 351–366, 2012.

[164] W. Lin, L. T. Biegler, and A. M. Jacobson, "Modeling and optimization of a seeded suspension polymerization process," *Chemical Engineering Science*, vol. 65, pp. 4350–4362, 2010.

[165] B. Lund and B. Foss, "Parameter ranking by orthogonalization-applied to nonlinear mechanistic models," *Automatica*, vol. 44, no. 1, pp. 278–281, 2008.

[166] S. Vajda, H. Rabitz, E. Walter, and Y. Lecourtier, "Qualitative and quantitative identifiability analysis of nonlinear chemical kinetic models," *Chemical Engineering Communications*, vol. 83, no. 1, pp. 191–219, 1989.

[167] V. M. Zavala, C. D. Laird, and L. T. Biegler, "Interior-point decomposition approaches for parallel solution of large-scale nonlinear parameter estimation problems," *Chemical Engineering Science*, vol. 63, no. 19, pp. 4834–4845, 2008.

[168] W. Dai, L. Bansal, J. Hahn, and D. Word, "Parameter set selection for dynamic systems under uncertainty via dynamic optimization and hierarchical clustering," *AIChE Journal*, vol. 601, no. 1, pp. 181–192, 2013.

[169] M. Mckay, R. J. Beckman, and W. J. Conover, "Comparison of 3 methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, vol. 21, no. 2, pp. 239–245, 1979.

[170] I.-B. Tjoa and L. T. Biegler, "Simultaneous solution and optimization strategies for parameter estimation of differential-algebraic equation systems," *Industrial & Engineering Chemistry Research*, vol. 30, no. 2, pp. 376–385, 1991.

[171] S. Maiti, W. Dai, R. C. Alaniz, J. Hahn, and A. Jayaraman, "Mathematical modeling of pro- and anti-inflammatory signaling in macrophages," *Processes*, vol. 3, no. 1, pp. 1–18, 2015.

[172] S. Kameswaran and L. T. Biegler, "Simultaneous dynamic optimization strategies: recent advances and challenges," *Computers & Chemical Engineering*, vol. 30, no. 10-12, pp. 1560–1575, 2006.

[173] M. G. Jobling, Z. Yang, W. R. Kam, W. I. Lencer, and R. K. Holmes, "A single native ganglioside gm1-binding site is sufficient for cholera toxin to bind to cells and complete the intoxication pathway," *mBio*, vol. 3, no. 6, pp. e00401–12, 2012.

[174] C. A. Day and A. K. Kenworthy, "Functions of cholera toxin B-subunit as a raft cross-linker," *Essays in Biochemistry*, vol. 57, pp. 135–145, 2015.

[175] M. Amaro, R. Šachl, G. Aydogan, I. I. Mikhalyov, R. Vácha, and M. Hof, "Gm1 ganglioside inhibits β-amyloid oligomerization induced by sphingomyelin," *Angewandte Chemie International Edition*, vol. 55, no. 32, pp. 9411–9415, 2016.

[176] E. R. Middleton and E. Rhoades, "Effects of curvature and composition on α-synuclein binding to lipid vesicles," *Biophysical Journal*, vol. 99, no. 7, pp. 2279–2288, 2010.

[177] C. O. Mellet, J.-F. Nierengarten, and J. M. G. Fernández, "Multivalency as an action principle in multimodal lectin recognition and glycosidase inhibition: a paradigm shift driven

by carbon-based glyconanomaterials," *Journal of Materials Chemistry B*, vol. 5, no. 32, pp. 6428–6436, 2017.

[178] T. K. Dam and C. F. Brewer, "Lectins as pattern recognition molecules: the effects of epitope density in innate immunity," *Glycobiology*, vol. 20, no. 3, pp. 270–279, 2009.

[179] B. Engelhardt, H. Fröhlich, and M. Kschischo, "Learning (from) the errors of a systesm biology approach," *Scientific Reports*, vol. 6, p. 20772, 2016.

[180] Z. Lin and M. B. Beck, "On the identification of model structure in hydrological and environmental systems," *Water Resources Research*, vol. 43, p. W02402, 2007.

[181] C. A. Tiemann, J. Vanlier, M. H. Oosterveer, A. K. Groen, P. A. Hilbers, and N. A. van Riel, "Parameter trajectory analysis to identify treatment effects of pharmacological interventions," *PLoS Computational Biology*, vol. 9, p. e1003166, 2013.

[182] E. Balsa-Canto, A. A. Alonso, and J. R. Banga, "Computational procedures for optimal experimental design in biological systems," *IET Systems Biology*, vol. 2, pp. 163–172, 2007.

[183] S. Dobre, T. Bastogne, C. Profeta, M. Barber-Heyob, and A. Richard, "Limits of variance-based sensitivity analysis for non-identifiability testing in high dimensional dynamic models," *Automatica*, vol. 48, pp. 2740–2749, 2012.

[184] G. Lillacci and M. Khammash, "A distribution-matching method for parameter estimation and model selection in computational biology," *International Journal of Robust and Nonlinear Control*, vol. 22, pp. 1065–1081, 2012.

[185] A. Narasingam, P. Siddhamshetty, and J. S. Kwon, "Temporal clustering for order reduction of nonlinear parabolic PDE systems with time-dependent spatial domains: Application to a hydraulic fracturing process," *AIChE Journal*, vol. 63, pp. 3818–3831, 2017.

[186] Y. Jung, H. Park, D. Du, and B. L. Drake, "A decision criterion for the optimal number of clusters in hierarchical clustering," *Journal of Global Optimization*, vol. 25, pp. 91–111, 2003.

[187] M. P. Tan, J. R. Broach, and C. A. Floudas, "A novel clustering approach and prediction of optimal number of clusters: global optimum search with enhanced positioning," *Journal of Global Optimization*, vol. 39, pp. 323–346, 2007.

[188] H. Pohjanpalo, "System identifiability based on the power series expansion of the solution," *Mathematical Biosciences*, vol. 41, pp. 21–33, 1978.

[189] L. Ljung and T. Glad, "On global identifiability for arbitrary model parametrizations," *Automatica*, vol. 30, pp. 265–276, 1994.

[190] S. Vajda, K. R. Godfrey, and H. Rabitz, "Similarity transformation approach to identifiability analysis of nonlinear compartmental models," *Mathematical Biosciences*, vol. 93, pp. 217–246, 1989.

[191] D. Lee, A. Singla, H. Wu, and J. S. Kwon, "An integrated numerical and experimental framework for modeling of CTB and GD1b ganglioside binding kinetics," *AIChE Journal*, vol. 64, pp. 3882–3893, 2018.

[192] M. D. Morris, "Factorial sampling plans for preliminary computational experiments," *Technometrics*, vol. 33, pp. 161–174, 1991.

[193] I. M. Sobol', "Global sensitivity indices for nonlinear mathematical models and their monte-carlo estimates," *Mathematics and Computers in Simulation*, vol. 55, pp. 271–280, 2001.

[194] A. Saltelli, M. Ratto, S. Tarantola, and F. Campolongo, "Sensitivity analysis for chemical models," *Chemical Reviews*, vol. 105, pp. 2811–2827, 2005.

[195] A. Saltelli, D. Gatelli, F. Campolongo, J. Cariboni, M. Ratto, M. Saisana, S. Tarantola, and T. Andres, *Global Sensitivity Analysis: The Primer*. John Wiley & Sons, 2008.

[196] T. Homma and A. Saltelli, "Importance measures in global sensitivity analysis of nonlinear models," *Reliability Engineering & System Safety*, vol. 52, pp. 1–17, 1996.

[197] C. Kontoravdi, S. P. Asprey, E. N. Pistikopoulos, and A. Mantalaris, "Application of global sensitivity analysis to determine goals for design of experiments: An example study on antibody-producing cell cultures," *Biotechnology Progress*, vol. 21, pp. 1128–1135, 2005.

[198] G. Ghosh, V. Y. Wang, D. Huang, and A. Fusco, "NF-κB regulation: lessons from structures," *Immunological Reviews*, vol. 246, pp. 36–58, 2012.

[199] N. Parameswaran and S. Patial, "Tumor necrosis factor-α signaling in macrophages," *Critical Reviews in Eukaryotic Gene Expression*, vol. 20, pp. 87–103, 2010.

[200] T. Kawai and S. Akira, "The role of pattern-recognition receptors in innate immunity: update on Toll-like receptors," *Nature Immunology*, vol. 11, pp. 373–384, 2010.

[201] H. L. Pahl and P. A. Baeuerle, "A novel signal transduction pathway from the endoplasmic reticulum to the nucleus is mediated by transcription factor NF-κB," *EMBO Journal*, vol. 14, pp. 2580–2588, 1995.

[202] J. A. Smith, "Regulation of cytokine production by the unfolded protein response; implications for infection and autoimmunity," *Frontiers in Immunology*, vol. 9, p. 422, 2018.

[203] K. Erguler, M. Pieri, and C. Deltas, "A mathematical model of the unfolded protein stress response reveals the decision mechanism for recovery, adaptation and apoptosis," *BMC Systems Biology*, vol. 7, p. 16, 2013.

[204] D. R. Diedrichs, J. A. Gomez, C. Huang, D. T. Rutkowski, and R. Curtu, "A data-entrained computational model for testing the regulatory logic of the vertebrate unfolded protein response," *Molecular Biology of the Cell*, vol. 29, pp. 1502–1517, 2018.

[205] H. Cho, M. Wu, L. Zhang, R. Thompson, A. Nath, and C. Chan, "Signaling dynamics of palmitate-induced ER stress responses mediated by ATF4 in HepG2 cells," *BMC Systems Biology*, vol. 7, p. 9, 2013.

[206] T. Lipniacki and M. Kimmel, "Deterministic and stochastic models of NF-κB pathway," *Cardiovascular Toxicology*, vol. 7, pp. 215–234, 2007.

[207] R. Cheong, A. Hoffmann, and A. Levchenko, "Understanding NF-κB signaling via mathematical modeling," *Molecular Systems Biology*, vol. 4, p. 192, 2008.

[208] M. von Stosch, N. Carinhas, and R. Oliveira, "Hybrid modeling for systems biology: Theory and practice," in *Large-Scale Networks in Engineering and Life Sciences* (P. B. Benner, R. Findeisen, D. Flockerzi, U. Reichl, and K. Sundmacher, eds.), pp. 367–388, Switzerland: Birkh auser, 2014.

[209] J. Pollard, M.R.Broussard, D.B.Garrison, and K. San, "Process identification using neural networks," *Computers & Chemical Engineering*, vol. 16, pp. 253–270, 1992.

[210] A. Narasingam and J. S. Kwon, "Development of local dynamic mode decomposition with control: Application to model predictive control of hydraulic fracturing," *Computers & Chemical Engineering*, vol. 106, pp. 501–511, 2017.

[211] A. Narasingam, P. Siddhamshetty, and J. S. Kwon, "Temporal clustering for order reduction of nonlinear parabolic PDE systems with time-dependent spatial domains: Application to a hydraulic fracturing process," *AIChE Journal*, vol. 63, pp. 3818–3831, 2017.

[212] M. L. Thompson and M. A. Kramer, "Modeling chemical processes using prior knowledge and neural networks," *AIChE Journal*, vol. 40, pp. 1328–1340, 1994.

[213] D. C. Psichogios and L. H. Ungar, "A hybrid neural network-first principles approach to process modeling," *AIChE Journal*, vol. 38, pp. 1499–1511, 1992.

[214] A. Narasingam and J. S. Kwon, "Data-driven identification of interpretable reduced-order models using sparse regression," *AIChE Journal*, vol. 119, pp. 101–111, 2018.

[215] M. S. F. Bangi and S. Kwon, "Deep hybrid modeling of chemical process: Application to hydraulic fracturing," *Computers & Chemical Engineering*, vol. 134, p. 106696, 2020.

[216] J. Pinto, C. R. de Azevedo, R. Oliveira, and M. von Stosch, "A bootstrap-aggregated hybrid semi-parametric modeling framework for bioprocess development," *Bioprocess and Biosystems Engineering*, vol. 42, pp. 1853–1865, 2019.

[217] J. Schubert, R. Simutis, M. Dors, I. Havlik, and A. Lübbert, "Hybrid modelling of yeast production processes - combination of a priori knowledge on different levels of sophistication," *Chemical Engineering & Technology*, vol. 14, pp. 10–20, 1994.

[218] M. von Stosch, J. Peres, E. F. de Azevedo, and R. Oliveira, "Modeling biochemical networks with intrinsic time delays: A hybrid semi-parameteric approach," *BMC Systems Biology*, vol. 4, p. 131, 2010.

[219] I. M. Chamseddine and K. A. Rejniak, "Hybrid modeling frameworks of tumor development and treatment," *WIREs Systems Biology and Medicine*, vol. 2, p. e1461, 2020.

[220] F. Hamilton, A. L. Lloyd, and K. B. Flores, "Hybrid modeling and prediction of dynamical systems," *PLoS Computational Biology*, vol. 13, p. e1005655, 2017.

[221] J. Lagergren, A. Reeder, F. Hamilton, R. C. Smith, and K. B. Flores, "Forecasting and uncertainty quantification using a hybrid of mechanistic and non-mechanistic models for an age-structured population model," *Bulletin of Mathematical Biology*, vol. 80, pp. 1578–1595, 2018.

[222] B. Engelhardt, M. Kschischo, and H. Fröhlich, "A bayesian approach to estimating hidden variables as well as missing and wrong molecular interactions in ordinary differential equation-based mathematical models," *Journal of the Royal Society of Interface*, vol. 14, p. 20170332, 2017.

[223] M. von Stosch, R. Oliveira, J. Peres, E. F. de Azevedo, and R. Oliveira, "Hybrid semi-parametric modeling in process systems engineering: Past, present and future," *Computers & Chemical Engineering*, vol. 60, pp. 86–101, 2014.

[224] B. Bhadriraju, A. Narasingam, and J. S. Kwon, "Machine learning-based adaptive model identification of systems: Application to a chemical process," *Chemical Engineering Research and Design*, vol. 152, pp. 372–383, 2019.

[225] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 113, pp. 3932–3937, 2016.

[226] D. Bampou, *Polynomial Approximations for Infinite-Dimensional Optimization Problems*. PhD thesis, Imperial College London, 2012.

[227] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems*, vol. 2, pp. 303–314, 1989.

[228] D. M. Himmelblau, "Applications of artificial neural networks in chemical engineering," *Korean Journal of Chemical Engineering*, vol. 17, pp. 373–392, 2000.

[229] D. Lee, A. Jayaraman, and J. S. Kwon, "Identification of cell-to-cell heterogeneity through systems engineering approaches," *AIChE Journal*, vol. 66, p. e16925, 2020.

[230] D. P. Howsmon and J. Hahn, "Regularization techniques to overcome overparameterization of complex biochemical reaction networks," *IEEE life sciences letters*, vol. 2, pp. 31–34, 2016.

[231] A. Yazdani, M. Raissi, and G. E. Karniadakis, "Systems biology informed deep learning for inferring parameters and hidden dynamics," *bioRxiv*, vol. 865063, 2019.

[232] L. Bansal, Y. Chu, C. Laird, and J. Hahn, "Regularization of inverse problems to determine transcription factor profiles from fluorescent reporter systems," *AIChE Journal*, vol. 58, pp. 3751–3762, 2012.

[233] P. Daoutidis and C. Kravaris, "Inversion and zero dynamics in nonlinear multivariable control," *AIChE Journal*, vol. 37, pp. 527–538, 1991.

[234] D. Kahl, P. Wendland, M. Neidhardt, A. Weber, , and M. Kschischo, "Structural invertibility and optimal sensor node placement for error and input reconstruction in dynamic systems," *Physical Review X*, vol. 9, p. 041046, 2019.

[235] P. Daoutidis and C. Kravaris, "Structural evaluation of control configurations for multivariable nonlinear processes," *Chemical Engineering Science*, vol. 47, pp. 1091–1107, 1992.

[236] B. Lee, Y. Kim, D. Shin, and E. S. Yoon, "A study on the evaluation of structural controllability in chemical processes," *Chemical Engineering Science*, vol. 25, pp. 85–95, 2001.

[237] M. Soroush, "Evaluation of achievable control quality in nonlinear processes," *Computers & Chemical Engineering*, vol. 20, pp. 357–364, 1996.

[238] S. Heo and P. Daoutidis, "Control-relevant decomposition of process networks viaoptimization-based hierarchical clustering," *AIChE Journal*, vol. 62, pp. 3177–3188, 2016.

[239] L. A. Herzenberg, J. Tung, W. A. Moore, L. A. Herzenberg, and D. R. Parks, "Interpreting flow cytometry data: a guide for the perplexed," *Nature Immunology*, vol. 7, pp. 681–685, 2006.

[240] P. Weber, J. Hasenauer, F. Allgöwer, and N. Radde, "Parameter estimation and identifiability of biological networks using relative data," in *Proceedings of the 18th IFAC World Congress*, pp. 11648–11653, 2011.

[241] A. M. Pérez-López, M. L. Soria-Gila, E. R. Marsden, A. Lilienkampf, and M. Bradley, "Fluorogenic substrates for in situ monitoring of caspase-3 activity in live cells," *PLoS ONE*, vol. 11, p. e0153209, 2016.

[242] A. Küper, R. Dürr, and S. Waldherr, "Dynamic density estimation in heterogeneous cell populations," *IEEE Control Systems Letters*, vol. 3, pp. 242–247, 2019.

[243] D. Lee, A. Jayaraman, and J. S. Kwon, "Identification of a time-varying intracellular signaling model through data clustering and parameter selection: Application to NFκB signaling pathway induced by LPS in the presence of BFA," *IET Systems Biology*, vol. 13, pp. 169–179, 2019.

[244] S. Steinmeyer, D. P. Howsmon, R. C. Alaniz, J. Hahn, and A. Jayaraman, "Empirical modeling of T cell activation predicts interplay of host cytokines and bacterial indole," *Biotechnology and Bioengineering*, vol. 114, pp. 2660–2667, 2017.

[245] M. Chaves, T. Eissing, and F. Allgöwer, "Bistable biological systems: A characterization through local compact input-to-state stability," *IEEE Transactions on Automatic Control*, vol. 53, pp. 87–100, 2008.

[246] M. Schliemann, E. Bullinger, S. Borchers, F. Allgöwer, R. Findeisen, and P. Scheurich, "Heterogeneous reduces sensitivity of cell death for TNF-stimuli," *BMC Systems Biology*, vol. 5, p. 204, 2011.

[247] T. Eissing, H. Conzelmann, E. D. Gilles, F. Allgöwer, E. Bullinger, and P. Scheurich, "Bistability analysis of a caspase activation model for receptor-induced apoptosis," *Journal of Biological Chemistry*, vol. 279, pp. 36892–36897, 2004.

[248] F. Boukouvala and C. A. Floudas, "ARGONAUT: Algorithms for global optimization of constrained grey-box computational problems," *Optimization Letters*, vol. 11, pp. 895–913, 2017.

[249] B. Beykal, F. Boukouval, C. A. Floudas, N. Sorek, H. Zalavadia, and E. Gildin, "Global optimization of grey-box computational systems using surrogate functions and application to highly constrained oil-field operations," *Computers & Chemical Engineering*, vol. 114, pp. 99–110, 2018.

[250] J. Eason and S. Cremaschi, "Adaptive sequential sampling for surrogate model generation with artificial neural networks," *Computers & Chemical Engineering*, vol. 68, pp. 220–232, 2014.

[251] A. Milias-Argeitis, M. Rullan, S. K. Aoki, P. Buchmann, and M. Khammash, "Automated optogenetic feedback control for precise and robust regulation of gene expression and cell growth," *Nature Communications*, vol. 7, p. 12546, 2016.

[252] E. M. Zhao, Y. Zhang, J. Mehl, H. Park, M. A. Lalwani, J. E. Toettcher, and J. L. Avalos, "Optogenetic regulation of engineered cellular metabolism for microbial chemical production," *Nature*, vol. 555, p. 683, 2018.

[253] L. H. Reyes, J. M. Gomez, and K. C. Kao, "Improving carotenoids production in yeast via adaptive laboratory evolution," *Metabolic Engineering*, vol. 21, pp. 26–33, 2014.

[254] D. J. Wilkinson, "Stochastic modelling for quantiative description of heterogeneous biological systems," *Nature Reviews Genetics*, vol. 10, pp. 122–133, 2009.

[255] L. Han, L. C. Morales, M. R. Richards, E. N. Kitova, S. Sipione, and J. S. Klassen, "Investigating the influence of membrane composition on protein-glycolipid binding using nanodiscs and proxy ligand electrospray ionization mass spectrometry," *Analytical Chemistry*, vol. 89, pp. 9330–9338, 2017.

[256] H. C. Kang and W. H. Weinberg, "Dynamic Monte Carlo simulations of surface-rate processes," *Accounts of Chemical Research*, vol. 25, pp. 253–259, 1992.

[257] F. C. Collins and G. E. Kimball, "Diffusion-controlled reaction rates," *Journal of Colloid Science*, vol. 4, pp. 425–437, 1949.

[258] R. L. Rowley, *Statistical Mechanics for Thermophysical Property Calculations*. Englewood Cliffs, NJ: Prentice-Hall, 1994.

[259] S. Tristram-Nagle, H. I. Petrache, and J. F. Nagle, "Structure and interactions of fully hydrated dioleoylphosphatidylcholine bilayers," *Biophysical Journal*, vol. 75, pp. 917–925, 1998.

[260] C. E. Miller, J. Majewski, R. Faller, S. Satija, and T. L. Kuhl, "Cholera toxin assault on lipid monolayers containing ganglioside GM1," *Biophysical Journal*, vol. 86, pp. 3700–3708, 2004.

[261] A. Chatterjee and D. G. Vlachos, "An overview of spatial microscopic and accelerated kinetic Monte Carlo methods," *Journal of computer-aided materials design*, vol. 14, pp. 253–308, 2007.

[262] A. R. Burns, D. J. Frankel, and T. Buranda, "Local mobility in lipid domains of supported bilayers characterized by atomic force microscopy and fluorescence correlation spectroscopy," *Biophysical Journal*, vol. 89, pp. 1081–1093, 2005.

[263] M. Stamatakis and D. G. Vlachos, "Equivalence of on-lattice stochastic chemical kinetics with the well-mixed chemical master equation in the limit of fast diffusion," *Computers & Chemical Engineering*, vol. 35, pp. 2602–2610, 2011.

[264] G. Lindblom and G. Orädd, "Lipid lateral diffusion and membrane heterogeneity," *Biochimica et Biophysica Acta*, vol. 1788, pp. 234–244, 2009.

[265] E. A. Merritt, P. Kuhn, S. Sarfaty, J. L. Erbe, R. K. Holmes, and W. G. J. Hol, "Structural studies of receptor binding by cholera toxin mutants," *Protein Science*, vol. 6, pp. 1516–1528, 1997.

[266] L. Moshe, G. Saper, O. Szekely, Y. Linde, C. Gilon, D. Harries, and U. Raviv, "Modulating the structure and interactions of lipid-peptide complexes by varying membrane composition and solution conditions," *Soft Matter*, vol. 9, p. 7117, 2013.

[267] E. M. Danielsen and G. H. Hansen, "Lipid raft organization and function in brush borders of epithelial cells (review)," *Molecular Membrane Biology*, vol. 23, pp. 71–79, 2006.

[268] R. Srivastava and J. B. Rawlings, "Parameter estimation in stochastic chemical kinetic models using derivative free optimization and bootstrapping," *Computers & Chemical Engineering*, vol. 63, pp. 152–158, 2014.

[269] G. Hu, Y. Lou, and P. D. Christofides, "Model parameter estimation and feedback control of surface roughness in a sputtering process," *Chemical Engineering Science*, vol. 63, pp. 1800–1816, 2008.

[270] S. Raimondeau, P. Aghalayam, A. B. Mhadeshwar, and D. G. Vlachos, "Parameter optimization of molecular models: Application to surface kinetics," *Industrial & Engineering Chemistry Research*, vol. 42, pp. 1174–1183, 2003.

[271] N. C. Worstell, A. Singla, P. Saenkham, T. Galbadage, P. Sule, D. Lee, A. Mohr, J. S.-I. Kwon, J. D. Cirillo, and H.-J. Wu, "Hetero-multivalency of Pseudomonas aeruginosa lectin Leca binding to model membranes," *Scientific Reports*, vol. 8, p. 8419, 2018.

[272] G. Adam and M. Delbruck, "Reduction of dimensionality in biological diffusion processes," in *Structural Chemistry and Molecular Biology* (A. Rich and N. Davidson, eds.), pp. 198–215, San Francisco: W.H. Freeman and Comany, 1968.

[273] A. Płóciennikowska, A. Hromada-Judycka, K. Borzęcka, and K. Kwiatkowska, "Co-operation of TLR4 and raft proteins in LPS-induced pro-inflammatory signaling," *Cellular and Molecular Life Sciences*, vol. 72, pp. 557–581, 2015.

[274] N. L. B. Wernick, D. J. F. Chinnapen, J. A. Cho, and W. I. Lencer, "Cholera toxin: An intracellular journey into the cytosol by way of the endoplasmic reticulum," *Toxins*, vol. 2, pp. 310–325, 2010.

[275] D. E. Saslowsky and W. I. Lencer, "Conversion of apical plasma membrane sphingomyelin to ceramide attenuates the intoxication of host cells by cholera toxin," *Cellular Microbiology*, vol. 10, pp. 67–80, 2007.

[276] A. A. Wolf, Y. Fujinaga, and W. I. Lencer, "Uncoupling of the cholera toxin-GM1 ganglioside receptor complex from endocytosis, retrograde Golgi trafficking, and downstream signal transduction by depletion of membrane cholesterol," *Journal of Biological Chemistry*, vol. 277, pp. 16249–16256, 2002.

[277] M. Stamatakis, *Stochasticity and cell population heterogeneity in an artificial lac operon genetic network*. PhD thesis, Rice University, 2009.