2019

# Mathematical models and modular composition rules for synthetic genetic circuits

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS & SCIENCES

Dissertation

# MATHEMATICAL MODELS AND MODULAR COMPOSITION RULES FOR SYNTHETIC GENETIC CIRCUITS

by

**JUNMIN WANG**

B.S., Davidson College, 2012
M.S., The George Washington University, 2014

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2019

## Approved by

First Reader

Calin Belta, PhD
Professor of Mechanical Engineering, Systems Engineering, and Bioinformatics

Second Reader

Samuel A. Isaacson, PhD
Associate Professor of Mathematics

# Acknowledgments

First of all, I would like to express my deep appreciation to my advisor, Dr. Calin Belta for his support, patience, and mentoring over the course of my Ph.D study and research. Without his guidance and continuous support, this dissertation would not be possible. I would also like to thank the rest of my thesis committee: Prof. Samuel Isaacson, Prof. Daniel Segre, Prof. Luis Carvalho, and Prof. Chris Myers, for their guidance and encouragement. They have given me uncountable insightful comments that are invaluable to my work.

Next, I would like to thank National Science Foundation (NSF) for funding my research. This work would not be possible without the financial support from NSF. I would also like to express my gratitude to the Bioinformatics Graduate Program at Boston University and Boston University for miscellaneous types of support, including the travel funding.

Last but not least, I would like to thank all my collaborators at Boston University, at Massachusetts Institute of Technology, at University of Pennsylvania, and every colleague who gave me advice on this work on different occasions. Namely, I would like to give thanks to Dr. Brian Teague, Dr. Jacob Beal, Prof. Ahmad Khalil, and Prof. Eduardo Sontag.

# MATHEMATICAL MODELS AND MODULAR COMPOSITION RULES FOR SYNTHETIC GENETIC CIRCUITS

**JUNMIN WANG**

Boston University, Graduate School of Arts & Sciences, 2019

Major Professor: Calin Belta, PhD
Professor of Mechanical Engineering, Systems Engineering, and Bioinformatics

## ABSTRACT

One major challenge in synthetic biology is how to design genetic circuits with predictable behaviors in various biological contexts. There are two limitations to addressing this challenge in mammalian cells. First, models that can predict circuit behaviors accurately in bacteria cells cannot be directly translated to mammalian cells. Second, upon interconnection, the behavior of a module, the building block of a circuit, may be different from its behavior in a standalone setting. In this thesis, I present a bottom-up modeling framework that can be used to predict circuit behaviors in transiently transfected mammalian cells (TTMC). The first part of the framework is based on a novel bin-dependent ODE model that can describe the behavior of modules in TTMC accurately. The second part of the framework rests upon a method of modular composition that allows model-based design of circuits. The efficacies of the bin-dependent model and the method of modular composition are validated via experimental data. The effects of retroactivity, a loading effect that arises from modular composition, on circuit behaviors are also investigated.

# Contents

1 **Introduction** **1**

1.1 Engineering Biology . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

1.2 Genetic Circuits . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

1.3 Model-Based Circuit Design . . . . . . . . . . . . . . . . . . . . . . . 3

1.4 Circuit Design in Transiently Transfected Mammalian Cells . . . . . . 5

1.5 Robustness and Retroactivity . . . . . . . . . . . . . . . . . . . . . . 8

1.6 Aims of the Dissertation . . . . . . . . . . . . . . . . . . . . . . . . . 10

2 **Bin-Dependent Models** **12**

2.1 Experimental Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

2.2 Protein Concentration vs Plasmid Copy Number . . . . . . . . . . . . 15

2.3 Two-stage Stochastic Gene Expression Models . . . . . . . . . . . . . 18

2.3.1 Model Details . . . . . . . . . . . . . . . . . . . . . . . . . . 26

2.3.2 Exploring Other Plasmid Distributions . . . . . . . . . . . . . 27

2.3.3 Possibility of Experimental Noise as the Cause of Saturation . 31

2.4 Bin-Dependent ODE Model . . . . . . . . . . . . . . . . . . . . . . . 32

2.5 Evaluating Model Performances . . . . . . . . . . . . . . . . . . . . . 33

3 **Modular Composition** **43**

3.1 Components of Circuits . . . . . . . . . . . . . . . . . . . . . . . . . 43

3.1.1 Module . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 43

3.1.2 Reporter . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 45

3.1.3 External Inducer . . . . . . . . . . . . . . . . . . . . . . . . . 45

3.2 Circuits and Models . . . . . . . . . . . . . . . . . . . . . . . . . . 45

3.2.1 Modular Connection . . . . . . . . . . . . . . . . . . . . . . 45

3.2.2 Graph Representation of a Set of Modules . . . . . . . . . . . 46

3.2.3 Composition of Models . . . . . . . . . . . . . . . . . . . . . 46

3.3 Models for Cascades . . . . . . . . . . . . . . . . . . . . . . . . . . 54

3.4 Model Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . 58

**4 Modular Composition, Circuit Behaviors, and Network Topologies 61**

4.1 Adaptation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 61

4.2 Mathematical Models and Modular Composition . . . . . . . . . . . . 62

4.2.1 Transcriptional Regulatory Networks . . . . . . . . . . . . . . 62

4.2.2 Mathematical Models with and without Retroactivity . . . . . 63

4.2.3 Network Enumeration and Simulation . . . . . . . . . . . . . . 65

4.2.4 Models for a type-IV IFFL . . . . . . . . . . . . . . . . . . . . 67

4.3 Adaptive Robustness . . . . . . . . . . . . . . . . . . . . . . . . . . 69

4.3.1 Adaptation . . . . . . . . . . . . . . . . . . . . . . . . . . . . 69

4.3.2 Specification and Adaptive Robustness . . . . . . . . . . . . . 69

4.4 Statistical Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . 71

4.4.1 Investigation of Behavior-Topology Relationships . . . . . . . 71

4.4.2 Networks and Adaptation . . . . . . . . . . . . . . . . . . . . 72

4.4.3 Effects of Retroactivity on Adaptive Robustness . . . . . . . . 72

**5 Conclusions 80**

5.1 Summary of the Thesis . . . . . . . . . . . . . . . . . . . . . . . . . 80

5.2 Impact and Future Directions . . . . . . . . . . . . . . . . . . . . . 81

**List of Journal Abbreviations 83**

**References 86**

**Curriculum Vitae** **95**

# List of Tables

2.1 Parameter values for the two-stage models. $k_1$ and $k_2$ have the units of # of molecules per plasmid per hour. $k_3$ and $k_4$ have the units of # of molecules per mRNA per hour. $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$ have the units of reciprocal hours. In models corresponding to Figure 2·4(a) of the main text and Figure 2·8, $k_1$ and $k_2$ are not constant since transcriptional rates are subject to saturation. In models corresponding to Figure 2·4(b) of the main text and Figure 2·9, $k_3$ and $k_4$ are not constant since translational rates are subject to saturation. NA stands for not applicable. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27

2.2 Optimal parameters and MSE for the traditional Hill-function-based model fit to the complete dataset. All parameter values are rounded to two digits after the decimal point. . . . . . . . . . . . . . . . . . . . 34

2.3 Optimal parameters and MSE over the reduced dataset for the traditional Hill-function-based model fit to the reduced dataset. All parameter values are rounded to two digits after the decimal point. Note that when evaluated over all 21 bins between $10^{5.8}$ MEFL and $10^{7.9}$ MEFL, this model produces MSE of 0.039, 0.062, and 0.038, respectively. . . 35

2.4 Optimal parameters and MSE for the bin-dependent model fit to the complete dataset. All parameter values are rounded to two digits after the decimal point. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 35

2.5 MSE of the models. "Complete" means the traditional Hill-function-based model is fit to the entire dataset, and "reduced" means the Hill-function-based model is fit to data between $10^{5.8}$ MEFL and $10^{7}$ MEFL (TAL14 and TAL21) or data between $10^{5.8}$ MEFL and $10^{7.3}$ MEFL (LmrA). Note, for the reduced model, goodness of fit is only evaluated by comparison to data between $10^{5.8}$ MEFL and $10^{7}$ MEFL(TAL14 and TAL21) or data between $10^{5.8}$ MEFL and $10^{7.3}$ MEFL (LmrA). When error of the reduced model is evaluated over all bins, it is much worse (numbers in parentheses in third column). . . . . . . . . . . . . 39

2.6 Averaged fitting errors of the models within the 12-fold cross-validation. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 40

2.7 Averaged testing errors of the models within the 12-fold cross-validation. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 40

3.1 Values of the rescaled parameters used in the final bin-dependent models for the six cascades. . . . . . . . . . . . . . . . . . . . . . . . 58

4.1 Numbers of NFBL, IFFL, and adaptive networks. Models I, II, III, and IV represent the model without retroactivity, the models with retroactivity assuming $\tilde{\eta} = 0.1$, $\tilde{\eta} = 1$, and $\tilde{\eta} = 10$. . . . . . . . . . . 73

4.2 Results from the perturbation analysis. Models I, II, III, and IV are defined the same as in the caption of Table I. Mean response sensitivity, adaptation errors, and adaptation ratios, which are calculated from 100 randomly selected parameter sets, are shown in the 3rd, 4th, and 5th columns. The adaptive robustness of these networks, which is inferred earlier via the algorithm shown in [Zuliani et al., 2013], is listed in the last columns. . . . . . . . . . . . . . . . . . . . . . . . . . . 78

# List of Figures

2·1 (a) Abstraction of a module encoding a transcriptional regulatory switch and a transfection marker. The induced (input) gene I, activated by an inducer, regulates the expression of O, the regulated (output) gene. Z, the transfection marker, is used to estimate plasmid copy number. (b) Distribution of the transfection marker. The black bins are ignored because they represent untransfected cells (data from [Davidsohn et al., 2015]). (c) Dose-response curves obtained from an experiment (data from [Davidsohn et al., 2015]). Averaged measurements binned by the expression level of Z are shown by color. Cells are separated into bins of width 0.1 on a log scale. Each curve corresponds to a different bin. The 1st bin, represented by the curve at the bottom, contains cells with the lowest plasmid counts. Each dot represents the average concentrations of the induced protein and the regulated protein within a bin at a certain inducer level. Concentrations of the induced and the regulated proteins have units of MEFL. (d) Detailed representation of an inducible switch controlled by doxycycline based on Figure 2(A) of [Davidsohn et al., 2015]. The transcriptional repressor can be TAL14, TAL21, or LmrA. Expressions of the repressors and EYFP are driven by constitutive rtTA and Gal4 proteins, respectively. rtTA and Gal4, which are required for protein activation, are both constitutively expressed and are not considered as limiting factors for the production of the repressors and EYFP. . . 16

2·2 Maximal and minimal expressions of the induced gene I and the regulated gene O for TAL14, TAL21, and LmrA. In the figures, the x-axis corresponds to the concentration of the transfection marker, and the y-axis to the concentration of the input and the output proteins (here concentrations are in units of MEFL). Shown in red is the induced gene I, and in blue the regulated gene O. Each dot is the average protein concentration of cells from one bin. On the top row the circuit is induced at 0nM; on the bottom row, 2000nM. On the top row, least squares regression lines are fit to red dots from $10^7$ to $10^{7.9}$ MEFL (TAL14 and TAL21) or from $10^{7.3}$ to $10^{7.9}$ MEFL (LmrA), and to blue dots from $10^{5.8}$ to $10^7$ MEFL (TAL14 and TAL21) or from $10^{5.8}$ to $10^{7.3}$ MEFL (LmrA). On the bottom row, least squares regression lines are fit to red dots from $10^{5.8}$ to $10^7$ MEFL (TAL14 and TAL21) or from $10^{5.8}$ to $10^{7.3}$ MEFL (LmrA), and to blue dots from $10^{5.8}$ to $10^{7.9}$ MEFL. The dots are calculated from the flow cytometry data of [Davidsohn et al., 2015]. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19

2·3 Simulations of our stochastic model suggest that either increasing translation rates (a) or decreasing transcriptional rates (b) can extend the near-constant induced gene levels at low copy plasmid numbers. X-axis and y-axis stand for number of molecules of the transfection marker and the induced protein in each bin. Best fit horizontal lines are drawn for reference. (a) Comparison of models in which the translational rates decrease in order from 1000 to 1 molecule per mRNA per hour. (b) Comparison of models in which the transcriptional rate of $D_{\text{induced}}$ increases from 0.002 to 1 molecule per plasmid per hour. . 22

2·4 Simulations of our stochastic model suggest that either either saturation of transcriptional kinetics (a) or saturation of translation kinetics (b) can lead to regimes where the induced gene reporter level saturates at high plasmid copy numbers. X-axis and y-axis stand for number of molecules of the transfection marker and the induced protein in each bin. Least squares regression lines are drawn for reference. (a) Comparison of models built under the hypothesis of transcriptional saturation. The half saturation constant $K_{D_{\text{induced}}}$ increases in order from $10^2$ to $10^6$ molecules, and $K_{D_{\text{tm}}}$ is held fixed at $10^4$ molecules. (b) Comparison of models built under the hypothesis of translational saturation. The transcriptional rate of the induced gene decreases in order from 10 to $10^{-3}$ molecule per plasmid per hour, and the transfection marker transcribes at a constant rate of $10^{-1}$ molecule per plasmid per hour. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24

2·5 Simulations of a transcriptional saturation model. X-axis stands for the mid point of each bin, and y-axis number of molecules of the induced protein in each bin. Bin width is chosen to be 0.1, 0.2, and 0.5. Notice, the saturating effect and the general curve are independent of bin size. 25

2·6 Comparison of models in which the translational rates decrease in order from 1000 to 1 molecule per mRNA per hour. . . . . . . . . . . . . . 29

2·7 Comparison of models in which the transcriptional rate of the induced gene increases from 0.002 to 1 molecule per plasmid per hour. . . . . 29

2·8 Comparison of models built under the hypothesis of transcriptional saturation. The half saturation constant $K_{D_{\text{induced}}}$ increases in order from $10^2$ to $10^6$ molecules, and $K_{D_{\text{tm}}}$ is held fixed at $10^4$ molecules. . 30

2·9 Comparison of models built under the hypothesis of translational saturation. The transcriptional rate of the induced gene decreases in order from 10 to $10^{-3}$ molecule per plasmid per hour, and the transfection marker transcribes at a constant rate of $10^{-1}$ molecule per plasmid per hour. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30

2·10 Comparison between complete data and the traditional Hill-function-based TAL14, TAL21, and LmrA models fit to the complete dataset. Plasmid copy number is shown by color. Solid lines are experimental data (data from [Davidsohn et al., 2015]), and dashed lines are model fits. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37

2·11 Comparison between complete data and the bin-dependent TAL14, TAL21, and LmrA models fit to the complete dataset. Plasmid copy number is shown by color. Solid lines are experimental data (data from [Davidsohn et al., 2015]), and dashed lines are model fits. . . . 37

2·12 Comparison between reduced data and the traditional Hill-function-based TAL14, TAL21, and LmrA models fit to the reduced dataset. Plasmid copy number is shown by color. Solid lines are experimental data (data from [Davidsohn et al., 2015]), and dashed lines are model fits. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 38

2·13 Comparison between complete data and the traditional Hill-function-based TAL14, TAL21, and LmrA models fit to the reduced dataset. The models are fit to reduced datasets but are evaluated at all plasmid copy numbers. Plasmid copy number is shown by color. Solid lines are experimental data (data from [Davidsohn et al., 2015]), and dashed lines are model fits. . . . . . . . . . . . . . . . . . . . . . . . . 38

2·14 Testing errors of the best cross-validated models within each bin. . . 40

3·1 Our approach to synthesizing satisfactory circuits. The input and the output of each module are measured by flow cytometries. A model is constructed for each module from data. Models for modules are then assembled into models for circuits. . . . . . . . . . . . . . . . . . . . . . 44

3·2 (a) Graphical representation of a genetic module. The input of a module is the TF, I, while its output is the regulated promoter pO. O is the protein that is expressed by pO. (b) Graphical representation of a regulatory switch. The green dotted box stands for the reporter. The black dotted box stands for the promoter regulated by an external inducer. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 45

3·3 Graphical representation of a two-transcriptional-repressor cascade. $pI_1$ is a promoter that constitutively expresses $I_1$. $I_1$ downregulates $pO_1$, while $O_1/I_2$ downregulates $pO_2$. $pO_1$ and $pO_2$, the promoters that control expression of $O_1$ and $O_2$, have identical sequences of polymerase binding sites, but different sequences of operator binding sites so that $I_1$ and $I_2$ can recognize their targets. The circuit consists of two modules: the $I_1$-$pO_1$ module and the $I_2/O_1$-$pO_2$ module. $I_1$ and $I_2$ can be different combinations of TAL14, TAL21, and LmrA. . . . . . 47

3·4 Graphical representation of a repressilator. $I_1/O_3$ downregulates $pO_1$, $I_2/O_1$ downregulates $pO_2$, and $I_3/O_2$ downregulates $pO_3$. The circuit consists of three modules: the $I_1/O_3$-$pO_1$ module, the $I_2/O_1$-$pO_2$ module and the $I_3/O_2$-$pO_3$ module. . . . . . . . . . . . . . . . . . . . . . 48

3·5 (a) Detailed representations of a cascade controlled by doxycycline based on Figures 2(A) and 3(A) of [Davidsohn et al., 2015]. The transcriptional repressors can be TAL14, TAL21, or LmrA. Expressions of the repressors (TAL14, TAL21, or LmrA) and EYFP are driven by constitutive rtTA and Gal4 proteins, respectively. rtTA and Gal4, which are required for protein activation, are both constitutively expressed and are not considered as limiting factors for the production of the repressors and EYFP. (b) Comparison of the mean-fold errors of the Hill-function-based models [Davidsohn et al., 2015], the bin-dependent models, and EQuiP [Davidsohn et al., 2015] for each cascade. The experimental data the models are validated against are from [Davidsohn et al., 2015]. Numbers on top of the dotted lines represent the average mean-fold errors of six cascades. . . . . . . . . . . . . . . . . . . . . . 52

3·6 Abstract representations of LmrA-TAL14, LmrA-TAL21, TAL14-LmrA, TAL14-TAL21, TAL21-LmrA, and TAL21-TAL14 cascades. . 53

3·7 Comparison between the experimental data and the predictions of the output made by the bin-dependent model. Plasmid copy number is shown by color. Solid curves are experimental data, and dashed curves are model predictions (the experimental data are from [Davidsohn et al., 2015]). From top left to bottom right in the order of left to right and top to bottom are cascades LmrA-TAL14, LmrA-TAL21, TAL14-LmrA, TAL14-TAL21, TAL21-LmrA, and TAL21-TAL14. . . 55

4·1 Response sensitivity, adaptation errors, and adaptation ratios. . . . . 62

4·2 Examples of three-node TRN. Arrows indicate activation, and bars at the ends of edges indicate inhibition. The leftmost network is a negative feedback loop (NFBL), in which the edges traversing B, A, and C accumulate in a negative regulation. The two networks in the center are incoherent feedforward loops (IFFL), in which A directly activates C and indirectly represses C via B. The second from the left is a type-IV IFFL, in which A represses B. The second from the right is a type-I IFFL, in which A activates B. The rightmost network is a "mixture" network composed of a type-IV IFFL and an NFBL between B and C. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 63

4·3 Adaptive robustness of adaptive networks. X-axis represents models without retroactivity, and Y-axis, the counterparts with retroactivity. Blue diamonds represent models assuming $\tilde{\eta} = 0.1$, red pluses, models assuming $\tilde{\eta} = 1$, and black circles, models assuming $\tilde{\eta} = 10$. The black line corresponds to equal adaptive robustness. . . . . . . . . . . . . . 73

4·4 An example of trajectories simulated by non-normalized Type-I IFFL models. Values of the parameters are: $K_{IA} = 0.4$nM, $K_{AB} = 24.62$nM, $K_{AC} = 10$nM, $K_{BC} = 10$nM, $h_{IA} = 1$, $h_{AB} = 2.60$, $h_{AC} = 3.28$, $h_{BC} = 2.77$, $\delta_A = 0.81\text{hr}^{-1}$, $\delta_B = 0.20\text{hr}^{-1}$, $\delta_C = 0.47\text{hr}^{-1}$. DNA concentrations increase and protein production rates per plasmid decrease from the top figure to the bottom figure. For fair comparison, the products of DNA concentrations and protein production rates per plasmid are kept fixed. Top: $\pi_A\eta_A = 144.18\text{nM} \cdot \text{hr}^{-1}$, $\pi_B\eta_B = 31.29\text{nM} \cdot \text{hr}^{-1}$, $\pi_C\eta_C = 80\text{nM} \cdot \text{hr}^{-1}$; 2nd from the top: $\pi_A = 58.61\text{hr}^{-1}$, $\pi_B = 12.72\text{hr}^{-1}$, $\pi_C = 80\text{hr}^{-1}$, $\eta_A = 2.46$nM, $\eta_B = 2.46$nM, $\eta_C = 1$nM; 2nd from the bottom: $\pi_A = 5.86\text{hr}^{-1}$, $\pi_B = 1.27\text{hr}^{-1}$, $\pi_C = 8\text{hr}^{-1}$, $\eta_A = 24.6$nM, $\eta_B = 24.6$nM, $\eta_C = 10$nM; bottom: $\pi_A = 0.59\text{hr}^{-1}$, $\pi_B = 0.13\text{hr}^{-1}$, $\pi_C = 0.8\text{hr}^{-1}$, $\eta_A = 246$nM, $\eta_B = 246$nM, $\eta_C = 100$nM. When normalized, the four models from the top to the bottom represent no retroactivity, $\tilde{\eta} = 0.1$, $\tilde{\eta} = 1$, and $\tilde{\eta} = 10$. All parameters are within reasonable ranges of biological parameters given in [Gyorgy and Del Vecchio, 2014]. . . . . . . . . . . . . . . . . . . . . . . . . . . . 76

# List of Abbreviations

| | |
|---|---|
| API | Application Programming Interfae |
| BLTL | Bounded Linear Temporal Logic |
| Dox | Doxycycline |
| HDL | Hardware Description Language |
| IFFL | Incoherent Feedforward Loop |
| MEFL | Molecules of Equivalent Fluorescein |
| MM | Michaelis-Menten |
| NFBL | Negative Feedback Loop |
| ODE | Ordinary Differential Equation |
| PBLTL | Probabilistic Bounded Linear Temporal Logic |
| PCR | Polymerase Chain Reaction |
| SMC | Statistical Model Checking |
| TF | Transcription Factor |
| TRN | Transcriptional Regulatory Network |
| TTMC | Transiently Transfected Mammalian Cells |

# Chapter 1

# Introduction

## 1.1 Engineering Biology

No matter how technology advances, at the back of our minds the uneasiness about the concept of engineering biology is never completely erased. To some the term "biological engineering" is automatically associated with cutting edge technology, which carries so much hope yet seems so distant. However, in reality bio-engineered products can be found everywhere: from biofuels and bio-materials to agricultural and pharmaceutical products. In total, bio-engineered products already constitute an industry that yields a revenue worth $350 billion per year, accounting for two percent of the U.S. economy [Si and Zhao, 2016].

Engineering biology would not be possible without the ability to manipulate DNA. In 1944, Oswald Avery showed that DNA contains genetic information in bacteria [Avery et al., 1944]. However, the role of DNA as the universal carrier of hereditary information was not generally accepted until James Watson and Francis Crick discovered the three-dimensional double helical structure of DNA in 1953 [Watson and Crick, 1953]. Watson and Crick's discovery paved the way for modern molecular biology, a field that studies the relationship between gene sequences and biological functions. In 1970s, the science community achieved several major breakthroughs in genetic engineering, including the first recombinant DNA molecule [Jackson et al., 1972], the first transgenic organism [Cohen et al., 1973], and the first transgenic mice [Jaenisch and Mintz, 1974]. These events marked the beginning of an era where DNA can be directly manipulated outside of natural breeding.

It is known that many biological activities are mediated not by single genes but by the coordination of multiple interacting genes [Ma and Gao, 2012]. Activities such as multi-stability and oscillations can be observed in networks of specialized gene regulatory elements, including the bacteriophage lambda switch and the Cyanobacteria circadian oscillator [Gardner et al., 2000]. At the turn of the millennium, Gardner et al. and Elowitz et al. created a bistable genetic toggle switch and a biological oscillator by engineering non-specialized genes and repressible promoters in *E. coli* cells [Gardner et al., 2000, Elowitz and Leibler, 2000], leading to the establishment of synthetic biology. The emphasis of synthetic biology is laid upon the the system level of interactions between genes and proteins. Their works showed that similar to components of electronic circuits, transcriptional regulatory elements could be designed and arranged in particular orders to mimic biological functions that previously existed only in nature.

## 1.2 Genetic Circuits

Genetic circuits are at the core of synthetic biology. The term "circuit", which is borrowed from the field of engineering, is emblematic of the interdisciplinary nature of synthetic biology. In essence, a genetic circuit is a collection of interacting genes that are synthesized and encoded on a plasmid DNA(s). Mimicking silicon-based electronic circuits, genetic circuits enable cells to receive chemical or thermal signals as the input, process the signals by carrying out logical functions with biochemical reactions, and generate a biological response(s), such as a change(s) in gene expression levels, as the output [Jusiak et al., 2016]. The relationship between cells and circuits is comparable to the relationship between computers and computer programs. Cells can be regarded as computers that sense and process all sorts of information based on the programs embedded in the circuits. Unlike in a silicon-based circuit, input, output, and intermediate signals in a genetic circuit are entirely biological: information that is transmitted downstream in a genetic circuit is the flow of RNA polymerase on DNA [Brophy and Voigt, 2014]. DNA-binding transcriptional factors (TF) act as switches that recruit or block RNA polymerase to increase or decrease the flow [Brophy and Voigt, 2014].

If carefully designed and successfully delivered, a genetic circuit can enables cells to sense the stimulants in their environment and respond to the stimuli by carrying out desired functions. Unlike a simple knockout or over-expression of a gene, the purpose of constructing a genetic circuit is not merely to turn on or turn off a gene but to let the cells decide when a gene gets turned on or turned off. Building genetic circuits is a common approach to acquiring a minimal network model for mediating a biological function, facilating a deeper understanding of the sophisticated regulatory networks that govern biological activities [Gardner et al., 2000, Elowitz and Leibler, 2000, Basu et al., 2005]. As practical devices, genetic circuits have been showing great promises in biochemical production [Gimpel et al., 2013, Georgianna and Mayfield, 2012], gene therapy [Xie et al., 2011], and environmental protection [Voigt, 2012, Didovyk et al., 2017].

## 1.3 Model-Based Circuit Design

Modern synthetic biology is inseparable from the computational models that guide the construction of synthetic networks. Experimental approaches combined with modeling are an increasingly popular strategy taken by the research community to study systems and synthetic biology. Underlying the popularity is the ever-larger amount of information that otherwise would not be attainable via taking either approach by itself [Endy and Brent, 2001, Ay and Arnosti, 2011, Stark et al., 2003b, Goodwin, 1965, Arkin, 2001]. In systems biology, constructing models from high-throughput experimental data helps identify the components and reveal the relationships among the components in gene networks [Barenco et al., 2006, Stark et al., 2003a, Schlitt and Brazma, 2005, Sontag, 2011]. In synthetic biology, models can be used to simulate temporal behavior of circuits, analyze key features of circuits such as bi-stability [Basu et al., 2005, Gardner et al., 2000] as well as guide circuit construction [Ellis et al., 2009, Del Vecchio, 2007].

Compared to the time when the field of synthetic biology was just established, genetic circuits are expanding rapidly in size and structural complexity. An important enabling technology is the high-throughput DNA synthesis [Kosuri and Church, 2014]. Advancing technologies bring about new opportunities as well as new challenges. Accompanying the rapid characterization of circuit parts is the challenging problem of circuit design. There are many choices to consider when it comes to designing a circuit, whether the choice is about circuit topologies or the DNA sequences for specific circuit parts. The numbers of successfully constructed promoters, coding sequences, terminators, etc. are rising rapidly [Canton et al., 2008]. The total number of circuits that can be built is a combinatorial explosion, so building and testing all possible circuit designs directly via experimental approaches becomes infeasible. The traditional intuition-driven approach is no longer sufficient for selecting a functioning circuit(s) out of all the possibilities. On the other hand, building and simulating predictive models for circuits can often be completed within a reasonable time thanks to today's computational power. In the biology community, there is a growing tendency to use mathematical models to guide circuit design.

Many model-based computational tools that were initially developed for systems biology show great promises in applications for synthetic biology [Marchisio and Stelling, 2009]. Biojade [Goler, 2004] and TABASCO [Kosuri et al., 2007] are among the first tools for circuit design, employing a “drag and drop” user interface, where components of circuits are displayed on a canvas, and users can build *in silico* circuit models by connecting components with wires that represent signal transmission. By connecting the “drag and drop” interface to a simulation environment, Biojade enables the visualization of circuit diagrams and analysis of circuit behaviors [Goler, 2004]. However, the underlying mathematical model of Biojade is sometimes considered too simplistic due to a lack of biological details [Marchisio and Stelling, 2009]. Tinkercell, developed by Chandran et al., comes with an Application Programming Interface (API), allowing users more freedom to choose the appropriate types of mathematical models [Chandran et al., 2009]. To my knowledge, Cello, developed by Nielsen et al., is one of the most comprehensive, well-rounded tools to date for circuit design automation [Nielsen et al., 2016]. Based on the notion of bottom-up circuit assembly, Cello transforms a functional specification of a circuit all the way to DNA sequences via a set of algorithms including parsing the specification, assembling circuit components, and simulating the models. The algorithm underlying Cello for model simulation draws on a library of Boolean logic gates and generates Boolean circuits as directed acyclic graphs from Hardware Description Language (HDL) specifications (Verilog) [Nielsen et al., 2016]. Currently, Cello is oriented towards bacterial cells, operates on Boolean logic, and focuses on simulating the steady state behaviors of circuits. The performance of Cello in eukaryotic cells, especially the ones that do not have steady-state behaviors, is yet to be evaluated [Nielsen et al., 2016].

## 1.4 Circuit Design in Transiently Transfected Mammalian Cells

In synthetic biology, the initial emphasis was put on microbial for synthesis of high-value compounds in biofuel production and environmental applications [Gimpel et al., 2013, Georgianna and Mayfield, 2012]. Many biomedical related proteins, however, can only be correctly folded and hence synthesized in mammalian cells due to their unique glycosylation patterns [Dalton and Barton, 2014, Khan, 2013]. Unlike bacteria cells, mammalian cells developed complex mechanisms to resist the invasion of foreign genetic materials along the course of evolution [Kis et al., 2015]. Mammalian cells are also more compartmentalized, meaning that the interior of the cells is divided into sections, each of which is dedicated to a unique cellular function [Kis et al., 2015]. While compartmentalization allows cells to conduct multiple biological activities simultaneously, it constitutes an additional challenge to the proper expression of synthetic circuits [Kwok, 2010, May et al., 2008]. Over the years, a growing interest in healthcare applications significantly increased the breadth and depth of research on synthetic biology in mammalian cells [Kis et al., 2015].

Transfection is a common procedure for delivering genetic circuits into cells. Many transfection methods have been developed, including biological methods, which use viral vectors to achieve sustained expression of circuits [Roesler et al., 2002, Hacein-Bey-Abina et al., 2002, Pfeifer and Verma, 2001], and chemical methods, which enable nucleic acids to cross cell membranes via the formation of positively charged nucleic acid/chemical complexes [Washbourne and McAllister, 2002, Schenborn and Goiffon, 2000, Holmen et al., 1995]. Depending on the method of transfection, the delivered genetic materials can either exist transiently in cells (transient transfection) or get passed down to later generations (stable transfection) [Kim and Eberwine, 2010]. Compared to stable transfection, transient transfection offers faster expressions of transfected genes, with higher expression levels. It has lower cytotoxicity and induces no mutagenesis [Vink et al., 2014, Kis et al., 2015, Kim and Eberwine, 2010]. It has also been shown to be an effective technique for speeding up the screening of novel synthetic designs [Schaumberg et al., 2016]. Those properties have motivated the investigation of transient transfection in mammalian synthetic biology.

Despite the wide use of transient transfection, a model that attends to the context specificity of transiently transfected mammalian cells (TTMC) was still missing. In the field, there are many more models developed for bacteria cells than for mammalian cells [Mathur et al., 2017]. This can pose a problem to synthetic biologists working with TTMC because existing circuit models for bacteria cells or even stably transfected cells are not sufficient for describing the gene expression mechanism in TTMC. In addition, modeling frameworks based on Boolean models [Wang et al., 2012] are more developed than frameworks based on other model types. Transcriptional regulation often displays on and off switch-like behaviors [Ay and Arnosti, 2011]. By dividing cell states into two states, i.e., zeros and ones, Boolean models are commonly used to capture the steady state behaviors of the circuits, as is shown in [Nielsen et al., 2016]. Boolean models are easy to analyze analytically and implement computationally, but they may overlook the important details in temporal behaviors of circuits and compromise the accuracy of the results. For example, plasmids that are introduced into cells via transient transfection are only expressed temporarily and do not become integrated into the host's genome. Plasmids get partitioned into daughter cells upon cell division, so plasmid counts as well as protein production rates decrease over time, resulting in a system without steady states. On the other hand, ordinary differential equation (ODE) models are often used to represent multicomponent, temporally evolving dynamics systems [Ay and Arnosti, 2011]. Genetic circuits can be represented by a set of differential equations in which kinetic rates are defined as net results of molecular synthesis, degradation, and interactions between molecule species [Ay and Arnosti, 2011]. For circuit design automation, there is a need to develop more descriptive models for a larger variety of cell types. In this dissertation, the focus is on developing an ODE model for circuit behaviors in TTMC.

Once a model is constructed, a method of composition needs to be developed to facilitate predictions of circuit behavior based on modules. Under the assumption of modularity, circuits are decoupled into smaller components also known as modules, each of which can be tested and characterized individually. Based on the characterization of modules, *in silico* models for circuits are constructed, simulated, and validated against wet lab results. Model simulations and experimental data can then be repeatedly compared for iteratively improving the circuit performance [Marchisio and Stelling, 2009]. Though seemingly a trivial problem, integration of individual modules for making circuit predictions is challenging because experimental data of individual modules are generally subject to batch effects [Johnson et al., 2007].

## 1.5 Robustness and Retroactivity

In theory, an accurate modeling framework for circuit behaviors, if applied to high-quality experimental data, should suffice the conditions necessary for an efficient circuit design process. However, factors including variation in experimental conditions and noise of gene expression could cause circuits to deviate substantially from their expected behaviors. Unlike electronic circuits, it is impossible to achieve precise control of chemical kinetic rates in synthetic genetic circuits due to complex cellular environments. Design strategies that improve the chance of success for circuit assembly are an active research area in synthetic biology. It is well known that variation in circuit structures contributes to diverse biological functions inside the cells [Alon, 2007]. Physicists have long speculated that there might be a limited number of network topologies that can execute any particular biological function robustly. In other words, even though chemical kinetic rates vary significantly across a cell population, particular networks are more likely to execute a biological function successfully than most other networks. Investigating the robustness of networks is especially important for synthetic biologists, as understanding the relationship between network topologies and biological functions provides invaluable instructions for how to engineer genetic circuits with a target function robustly. There is an extensive literature about particular topologies like incoherent feedforward loops (IFFL) and functions they enable, but there is a demand to develop a systematic approach to investigating the relationship between network topologies and any biological function(s).

A critical assumption underlying circuit assembly is modularity, that is, behavior of a circuit can be predicted based on its components. For improving circuit performance, in the past much attention was given to the effects of the parameters that are inherent in the modules, including protein production rates and decay rates [Shi et al., 2017, Ma et al., 2009]. However, it should be pointed out that behaviors of circuits are determined not only by behaviors of modules but also by the loading effects that arise from modular interconnections, known as retroactivity [Ventura et al., 2010, Del Vecchio et al., 2008]. Retroactivity refers to the phenomenon where transmitting a biological signal from the upstream system to the downstream system alters the behavior of the upstream system [Gyorgy and Del Vecchio, 2014, Jayanthi et al., 2013, Mou and Del Vecchio, 2015]. Based on the theoretical foundations, Jayanthi et al. proved via experiments the existence of retroactivity in genetic circuits and the feasibility of controlling retroactivity via plasmid copy numbers. According to [Gyorgy and Del Vecchio, 2014] and [Jayanthi et al., 2013], raising plasmid copy numbers and lowering protein production rates per plasmid by the same fold can increase the retroactivity of a system without affecting its steady states. Up till now retroactivity has been shown to impact ultra-sensitivity [Ventura et al., 2010], input-output characteristics [Brewster et al., 2014], response times, etc. [Jayanthi et al., 2013, Jiang et al., 2011]. No previous work has yet been done on the effects of retroactivity on robustness of a network to achieve a particular biological function. Understanding such effects will shed light on methods to control retroactivity and design more robust synthetic circuits.

## 1.6 Aims of the Dissertation

The demand for a more efficient and accurate circuit design process obligates the development of computational tools and the exploration of design strategies for genetic circuits. In this dissertation, I present a series of studies that aim to improve the accuracy of model predictions for circuit behaviors and the robustness of circuits for achieving desired behaviors.

In Chapter 2, I present a novel bin-dependent model that accounts for specific cellular mechanisms of TTMC. Transient transfection of cells can be highly stochastic, resulting in large variations in plasmid counts across a population. Binning cells by plasmid copy number is a common practice for analyzing transient transfection data. In many kinetic models of transfected cells, protein production rates are assumed proportional to plasmid copy number. The validity of this assumption in TTMC is unclear, and models based on this assumption appear unable to reproduce experimental flow cytometry data robustly. We hypothesize that protein saturation at high plasmid copy number is a reason previous models break down and validate our hypothesis by comparing experimental data and a stochastic chemical kinetics model. The model demonstrates that there are multiple distinct physical mechanisms that can cause saturation. Based on these observations, we develop a novel minimal bin-dependent ODE model that assumes different parameters for protein production in cells with low versus high numbers of plasmids.

In Chapter 3, I expand the bin-dependent ODE model for a transcriptional regulatory switch into an ODE modeling framework that can be applied to a wide variety of circuits. I provide a precise definition of genetic modules, from which we then develop a method of modular composition that addresses the cross-batch variation among different flow cytometry datasets, allowing model-based design of circuits in TTMC.

In Chapter 4, I present an investigation of how retroactivity impacts circuit behaviors. Specifically, I focus on adaptation, which refers to a system's ability to respond transiently to an input signal and subsequently recover to the initial states. Adaptive robustness, the ability of a circuit to achieve adaptation, is subject to retroactivity, the loading effects that stem from modular interconnections. Studying the effects of retroactivity on adaptive robustness facilitates the employment of retroactivity to improve circuit performance. To achieve this goal, I provide a definition of adaptive robustness, present a framework for quantifying adaptive robustness via statistical model checking (SMC), and apply this framework to investigate the effects of retroactivity on adaptive robustness.

Collectively, this dissertation seeks to highlight new computational models and offer novel insights into the discipline of circuit design. The intention of the work detailed in this dissertation is to complement existing automation platforms for circuit design, diversifying the application of mathematical models in synthetic biology in terms of model categories (e.g., ODE), cellular contexts (e.g., mammalian cells), and application domains (e.g., robustness and circuit topologies). Driven by a deeper understanding of cellular biology, development of mathemtical models will greatly reduce the cost of circuit construction *in vitro*, making synthetic biology a more promising technology.

# Chapter 2

# Bin-Dependent Models

Transient transfection is a widely adopted technique for delivering foreign genetic materials into eukaryotic cells. The transfected genetic materials utilize the cells' innate transcriptional and translational machinery to get expressed, conferring on cells novel biological functions. In this chapter, I detail a novel bin-dependent model that is capable of describing experimental data in TTMC accurately without incorporating sophisticated mechanistic details. In essence, the bin-dependent model is an ODE model that describes the time evolution of gene expression levels but differs from a regular ODE model in bacteria or stably integrated cells by capturing specific cellular mechanisms in TTMC. The bin-dependent model includes copy number as a predictor and is compatible with the method of binning that is widely used for analyzing TTMC.

## 2.1 Experimental Data

Via transient transfection, we deliver genetic circuits into mammalian cells, creating synthetic transcriptional regulatory systems that enable cells to respond to the external stimuli. The first step in constructing models for such systems is to examine experimental data. In this thesis we focus on the bottom-up approach to building circuits via the assembly of individual modules, where a module is defined as a transcriptional regulatory switch. As an example of the types of modules we will use, consider a module comprising a fluorescent-reporter system involving three fluorescent genes: the induced (input) gene, the regulated (output) gene, and the transfection marker (Figure 2·1(a)). The expression levels of the fluorescent genes are measured via flow cytometries, with the fluorescence intensities used as approximations for the concentrations of the fluorescent proteins. The induced gene is regulated by a constitutive activator protein, and an external inducer whose concentration is under control. The product of the induced gene serves as a TF for the regulated gene, controlling the latter's expression of a fluorescent reporter. The induced gene's product is not fluorescent, but is measured by co-expressing a fluorescent reporter gene of a different color from the same promoter [Kærn et al., 2003]. The expression of the induced gene can be modulated by changing the amount of the inducer. Expression of the induced gene and the regulated gene at various inducer levels constitutes a dose-response curve (Figure 2·1(c)). In TTMC, expression levels largely depend on the numbers of plasmids transfected in individual cells [Glover et al., 2010, Davidsohn et al., 2015], which cannot be controlled and are highly variable across a population. Therefore, it is necessary to estimate the plasmid copy numbers so that the effect of variation in copy numbers on gene expression can be captured. This is often achieved by co-transfecting another constitutively expressed fluorescent protein, which serves as the transfection marker (Figure 2·1(a)). The induced gene, the regulated gene, and the transfection marker can be encoded on either one plasmid or separate plasmids. The former ensures that there is a one-to-one correspondence among the genes. In comparison, the latter is often preferred as separate plasmids can be absorbed by cells more readily due to smaller sizes, interference among the transcriptional units is minimized, and the concentrations of individual proteins can be adjusted more easily [Chen and Xia, 2011, Assur et al., 2012]. In what follows, we assume the transfection marker has been encoded on a separate plasmid for all models and experiments. We also assume the induced gene acts as an inhibitor of the regulated gene.

Fluorescence readings from flow cytometers can be converted to standard units of Molecules of Equivalent Fluorescein (MEFL) via TASBE Control [Davidsohn et al., 2015, Beal, 2015, mef, 2001]. Standardized data are segmented into bins by plasmid counts so that subpopulations of cells with similar plasmid counts can be studied in groups (Figure 2·1(b)) [Davidsohn et al., 2015, Davidsohn, 2013, Siciliano et al., 2018]. Since flow cytometry measurements are typically log-normal distributed or a mixture of two log-normal distributions [Beal, 2017, Hattis and Burmaster, 2006], binning is performed on a log scale to ensure that each bin contains relatively equal numbers of cells. The width of bins is selected depending on the resolution at which analysis is to be conducted.

We use data from [Davidsohn et al., 2015] as an example to demonstrate the construction of a module and the implementation of binning. Davidsohn et al. constructed the circuits using the rtTA and GAL4/UAS system: the input (TAL14, TAL21, or LmrA repressor) is activated by a constitutive rtTA protein and doxycycline, and expression of the output (EYFP), which is inhibited by the input, is driven by a constitutive Gal4 protein [Davidsohn et al., 2015]. A detailed representation of the circuit structure can be found in Figure 2·1(d). rtTA and Gal4, which are indispensable for protein activation, are both constitutively expressed and are not considered as limiting factors for the production of the input and the output. Omitting rtTA and Gal4 leads to an abstraction of the circuit that is depicted in Figure 2·1(a). The strength of repression is modulated by inducing the switch at twelve dosages of doxycycline (Dox), and is indicated by the reporter gene EBFP2. Another fluorescent gene, mKate, is a constitutively expressed gene that serves as a transfection marker. Concentrations of all fluorescent proteins are measured for every single cell by a flow cytometer 72 hours post transfection [Davidsohn et al., 2015]. These data are then standardized into MEFL units and segmented by concentrations of the mKate protein into bins of width 0.1 on a log scale (Figure 2·1(b)). Because bi-modality observed in the concentrations of the mKate protein is believed to be caused by whether individual cells get transfected, only cells with concentrations of mKate centering around the larger mode, ranging from $10^{5.8}$ to $10^{7.9}$ (unit: MEFL), are used for modeling as in [Davidsohn et al., 2015] (Figure 2·1(b)). For data that lie in this range, geometric means of concentrations of the EBFP2 protein and the EYFP protein are calculated within each bin. In this thesis, we focus on the average temporal behavior within each bin, with the goal of developing ODE models that can be directly parametrized from binned flow cytometry data.

## 2.2 Protein Concentration vs Plasmid Copy Number

Hill functions are commonly used to model transcriptional regulation in ODE models (Figure 2·1(a)). Mathematically, a Hill function is defined as:

$$H(I) = \begin{cases} (1-\gamma)\cdot \dfrac{1}{1+\left(\frac{I}{d}\right)^h} + \gamma, & \text{if I is an inhibitor} \\ (1-\gamma)\cdot \dfrac{\left(\frac{I}{d}\right)^h}{1+\left(\frac{I}{d}\right)^h} + \gamma, & \text{if I is an activator,} \end{cases} \tag{2.1}$$

where $I$ is the concentration of the inhibitor/activator. $H(I)$ accounts for the fraction of the promoter that is active. $\gamma$ is the minimum fraction of the promoter that is active: if I is an inhibitor, $\gamma$ is the fraction active given infinite abundance of I; if I is an activator, $\gamma$ is the fraction active in absence of I. $h$ is the Hill coefficient, and $d$ is the dissociation constant.

Davidsohn et al. developed a traditional Hill-function-based model to describe the time evolution of the induced and the regulated proteins in TTMC (Figure 2·1(a))

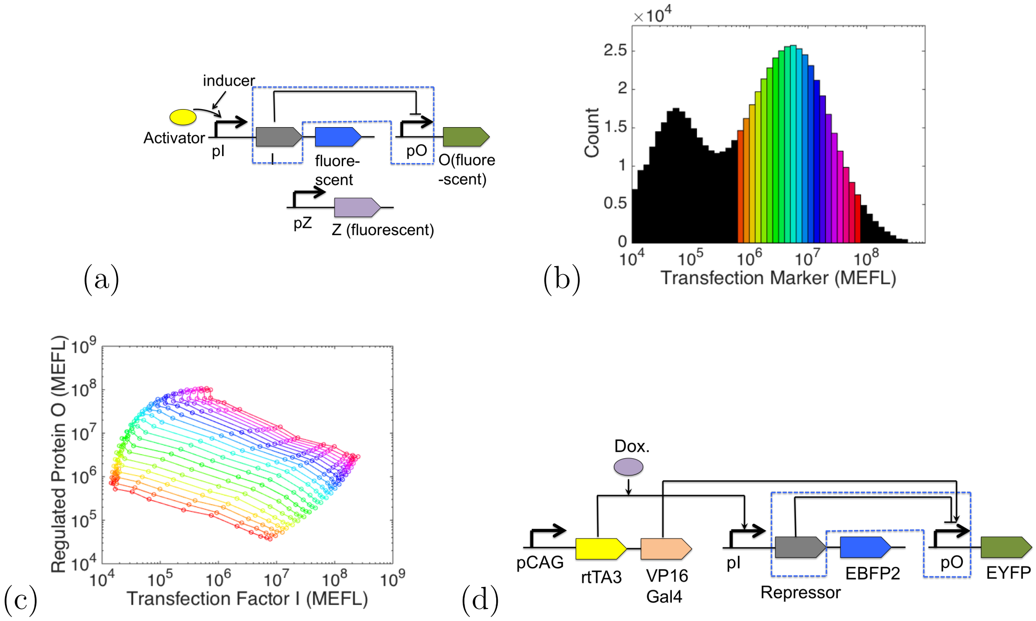

**Figure 2·1:** (a) Abstraction of a module encoding a transcriptional regulatory switch and a transfection marker. The induced (input) gene I, activated by an inducer, regulates the expression of O, the regulated (output) gene. Z, the transfection marker, is used to estimate plasmid copy number. (b) Distribution of the transfection marker. The black bins are ignored because they represent untransfected cells (data from [Davidsohn et al., 2015]). (c) Dose-response curves obtained from an experiment (data from [Davidsohn et al., 2015]). Averaged measurements binned by the expression level of Z are shown by color. Cells are separated into bins of width 0.1 on a log scale. Each curve corresponds to a different bin. The 1st bin, represented by the curve at the bottom, contains cells with the lowest plasmid counts. Each dot represents the average concentrations of the induced protein and the regulated protein within a bin at a certain inducer level. Concentrations of the induced and the regulated proteins have units of MEFL. (d) Detailed representation of an inducible switch controlled by doxycycline based on Figure 2(A) of [Davidsohn et al., 2015]. The transcriptional repressor can be TAL14, TAL21, or LmrA. Expressions of the repressors and EYFP are driven by constitutive rtTA and Gal4 proteins, respectively. rtTA and Gal4, which are required for protein activation, are both constitutively expressed and are not considered as limiting factors for the production of the repressors and EYFP.

[Davidsohn et al., 2015, Alon, 2007]:

$$\begin{aligned} \frac{dI_i}{dt} &= \alpha_i \cdot \phi(t) - \lambda_I \cdot I_i \\ \frac{dO_i}{dt} &= \beta \cdot \phi(t) \cdot \left(\frac{P_i}{P_1}\right)^f \cdot H(I_i) - \lambda_O \cdot O_i \\ \phi(t) &= \left(\frac{1}{2}\right)^{\lfloor \frac{t}{T} \rfloor} \\ H(I_i) &= (1-\gamma) \cdot \frac{1}{1+\left(\frac{I_i}{d}\right)^h} + \gamma. \end{aligned} \tag{2.2}$$

In Equation (2.2), $i$ represents the $i$-th plasmid count bin. $I_i$ and $O_i$ are the average concentrations of the induced and the regulated proteins in the $i$-th bin. $\alpha_i$ is the production rate of the induced protein in the $i$-th bin. $\alpha_i$ is assumed time-invariant because I is induced by a constant concentration of inducer. $\phi(t)$ captures that the population-average plasmid counts decrease due to cell division over time. $T$ is length of the cell cycle; $\lambda_I$ and $\lambda_O$ are dilution/degradation rates of I and O. $\beta$ is the maximal average production rate of the regulated protein for cells in the 1st bin, i.e, cells that have minimal plasmid counts $P_1$. $P_i$ is the mid-point of the $i$-th plasmid count bin. $f$ maps the ratios of the concentrations of transfection markers to the ratios of plasmid counts [Davidsohn et al., 2015]. For the sake of convenience, Equation (2.2) is referred to as the Hill-function-based model.

A key assumption of their model is that the log of the maximal production rate of the regulated protein is a linear function of the log of the transfection marker. This assumption is supported by findings of several other studies in different biological contexts [Glover et al., 2010, Cohen et al., 2009]. However, this assumption is only partially supported by the experimental data in [Davidsohn et al., 2015], shown here in Figure 2·2. When the induced gene is minimally induced (0 nM of inducer), i.e., the regulated protein expressed without repressor, the log of the regulated protein's concentration grows proportionally to the log of the transfection marker between

$10^{5.8}$ and $10^7$ MEFL for TAL14 and TAL21 or between $10^{5.8}$ and $10^{7.3}$ MEFL for LmrA. When the induced gene is fully induced (2000 nM of inducer), the log of the induced protein's concentration also grows linearly in the log of the transfection marker between $10^{5.8}$ and $10^7$ MEFL for TAL14 and TAL21 or between $10^{5.8}$ and $10^{7.3}$ MEFL for LmrA. Figure 2·2 also suggests that when either the induced gene or the regulated gene is maximally expressed, the concentrations of both the induced and the regulated proteins saturate starting from $10^{7.1}$ MEFL for TAL14 and TAL21 or $10^{7.4}$ MEFL for LmrA.

Furthermore, Figure 2·2 and the data in [Davidsohn et al., 2015] suggest that when the induced gene is induced at 0nM, the log of the induced protein's concentration is near-constant for low plasmid copy numbers [Davidsohn et al., 2015]. When the induced gene is fully induced, i.e., the regulated protein fully repressed, the log of the regulated protein's concentration grows linearly across all bins.

## 2.3 Two-stage Stochastic Gene Expression Models

We now develop a detailed stochastic model of the plasmid system, similar to the one Davidsohn et al. constructed experimentally [Davidsohn et al., 2015]. This model will enable us to explore possible mechanisms contributing to the observed saturation of protein concentrations at high plasmid copy number, as well as the near constant protein concentrations at low plasmid copy number. We do not attempt to fit this model to the single-time flow cytometry data directly as it is too complex to fit accurately without the incorporation of additional experimental measurements. Instead, *our purpose here is to use the stochastic model to gain a qualitative understanding of which biological hypotheses, and what ranges of physical gene expression parameters, may contribute to the observed saturation effect.* Our ultimate goal is to develop a simple model that qualitatively describes our limited set of data, avoiding further

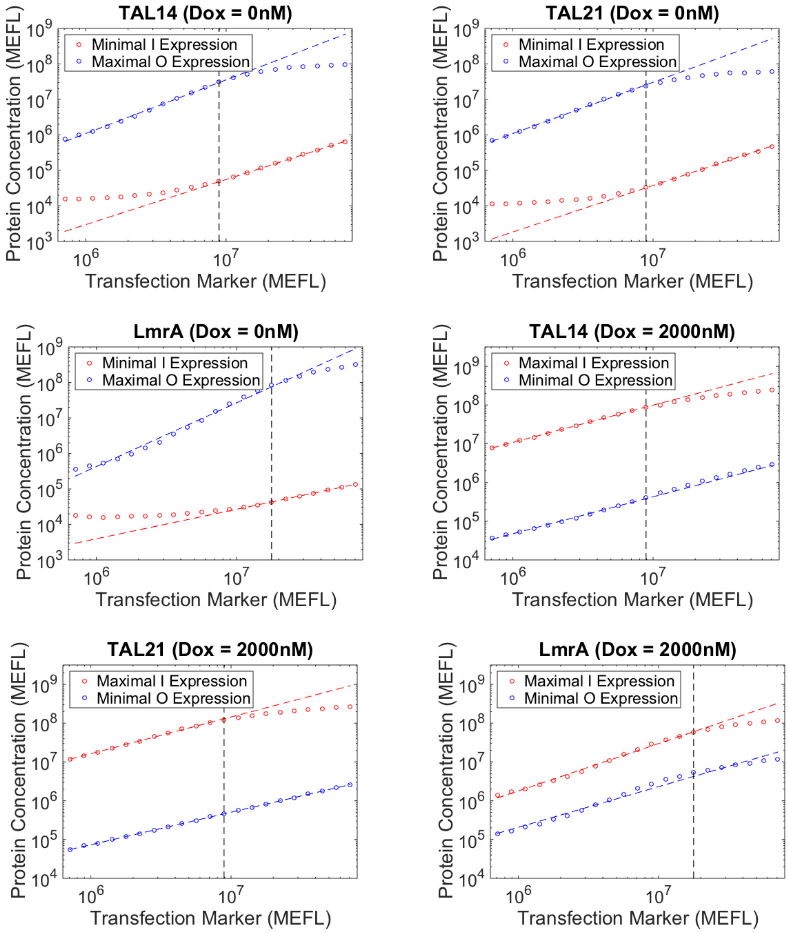

**Figure 2·2:** Maximal and minimal expressions of the induced gene I and the regulated gene O for TAL14, TAL21, and LmrA. In the figures, the x-axis corresponds to the concentration of the transfection marker, and the y-axis to the concentration of the input and the output proteins (here concentrations are in units of MEFL). Shown in red is the induced gene I, and in blue the regulated gene O. Each dot is the average protein concentration of cells from one bin. On the top row the circuit is induced at 0nM; on the bottom row, 2000nM. On the top row, least squares regression lines are fit to red dots from $10^7$ to $10^{7.9}$ MEFL (TAL14 and TAL21) or from $10^{7.3}$ to $10^{7.9}$ MEFL (LmrA), and to blue dots from $10^{5.8}$ to $10^7$ MEFL (TAL14 and TAL21) or from $10^{5.8}$ to $10^{7.3}$ MEFL (LmrA). On the bottom row, least squares regression lines are fit to red dots from $10^{5.8}$ to $10^7$ MEFL (TAL14 and TAL21) or from $10^{5.8}$ to $10^{7.3}$ MEFL (LmrA), and to blue dots from $10^{5.8}$ to $10^{7.9}$ MEFL. The dots are calculated from the flow cytometry data of [Davidsohn et al., 2015].

time-intensive experimental assays. Therefore, in the next subchapter, we develop a more simplified ODE model that can be parametrized from just the limited flow cytometry data, building from the qualitative understanding of the two-plasmid system our stochastic model provides.

In our stochastic model, cells are co-transfected by a mixture of induced gene plasmids and transfection marker plasmids. We focus on the dynamics of the transfection marker and the induced gene, which are integrated on separate plasmids. The total initial number of plasmids transfected in a given cell is assumed to follow a log-normal distribution [Davidsohn et al., 2015, Beal, 2017]. This assumption is because the shape of the protein distribution is known to reflect the shape of the underlying plasmid distribution [Tal and Paulsson, 2012], and the protein distribution is often observed to be approximately log-normal [Beal, 2017, Hattis and Burmaster, 2006]. The conditional distribution of the number of each of the two types of plasmids, given the total number of plasmids, is assumed to be binomial [Davidsohn et al., 2015]. This is because the plasmids we consider are assumed to be well-mixed, of relatively small and similar sizes, and hence indistinguishable for purposes of co-transfection [Davidsohn et al., 2015]. In the remainder, we choose values for kinetic parameters such that they span the parameter distributions calculated from transcriptomics and proteomics data given in [Schwanhausser et al., 2011]. We select parametric values for the initial plasmid distributions based on the polymerase chain reaction (PCR) findings of [Tachibana et al., 2002, Cohen et al., 2009, B. James and Giorgio, 2000]. The biochemical reactions in our model are shown below:

$$\begin{aligned}
&\mathrm{D}_{\mathrm{tm}} \overset{K_1}{\rightarrow} \mathrm{D}_{\mathrm{tm}} + \mathrm{M}_{\mathrm{tm}} &\qquad& \mathrm{D}_{\mathrm{induced}} \overset{K_2}{\rightarrow} \mathrm{D}_{\mathrm{induced}} + \mathrm{M}_{\mathrm{induced}}\\
&\mathrm{M}_{\mathrm{tm}} \overset{K_3}{\rightarrow} \mathrm{M}_{\mathrm{tm}} + \mathrm{P}_{\mathrm{tm}} && \mathrm{M}_{\mathrm{induced}} \overset{K_4}{\rightarrow} \mathrm{M}_{\mathrm{induced}} + \mathrm{P}_{\mathrm{induced}}\\
&\mathrm{M}_{\mathrm{tm}} \overset{\Lambda_1}{\rightarrow} \emptyset && \mathrm{M}_{\mathrm{induced}} \overset{\Lambda_2}{\rightarrow} \emptyset
\end{aligned}$$

$$\mathrm{P}_{\mathrm{tm}} \xrightarrow{\Lambda_3} \emptyset \qquad\qquad \mathrm{P}_{\mathrm{induced}} \xrightarrow{\Lambda_4} \emptyset,$$

where D, M, and P stand for plasmid, mRNA, and protein. Subscript "tm" stands for the transfection marker, and "induced" for the induced gene that is co-transfected. $\Lambda_i$ ($i = 1-4$) are first order degradation rate constants. Depending on the hypothesis underlying each model, $K_i$ ($i = 1-4$) are defined either as normal first-order rate constants, where $K_1 = k_1 \cdot D_{\mathrm{tm}}$, and $K_2$, $K_3$, and $K_4$ are defined similarly, or as Michaelis-Menten (MM) equations, where a saturated $K_1$ is defined as $K_{1,\mathrm{max}} \cdot \frac{D_{\mathrm{tm}}}{D_{\mathrm{tm}}+K_{D_{\mathrm{tm}}}}$, and saturated $K_2$, $K_3$, and $K_4$ are defined similarly. $K_{1,\mathrm{max}}$ represents the maximal value of $K_1$, and $K_{D_{\mathrm{tm}}}$ the half saturation constant. Formulas of $K_1$, $K_2$, $K_3$, and $K_4$ in each model can be found in Chapter 2.3.1.

Using StochKit and GillesPy, for each fixed set of parameters we simulate this model using the Gillespie method 400,000 times [Thattai and van Oudenaarden, 2004, Gillespie, 1977, Abel et al., 2016, Sanft et al., 2011]. This is comparable to the number of experimental samples generated in [Davidsohn et al., 2015]. Length of the simulation is 50 hours. Cell division takes place every 20 hours, and plasmids are binomially partitioned in daughter cells upon cell division. The initial cell cycle position for a cell is sampled randomly from the uniform distribution unif(0,20). Additional details of the simulation, including the parameter values, can be found in Chapter 2.3.1. After simulation, we divide the simulated data based on the transfection marker into bins of width 0.2, which is comparable to values that are typically chosen in flow cytometry experiments [Davidsohn et al., 2015, Davidsohn, 2013, Siciliano et al., 2018]. We then calculate the geometric mean of the induced protein's concentrations for each bin.

To examine the mechanisms that contribute to the near-constant induced reporter concentrations at low plasmid copy number, and the saturating induced reporter concentrations at high plasmid copy number, we systematically vary individual or pairs

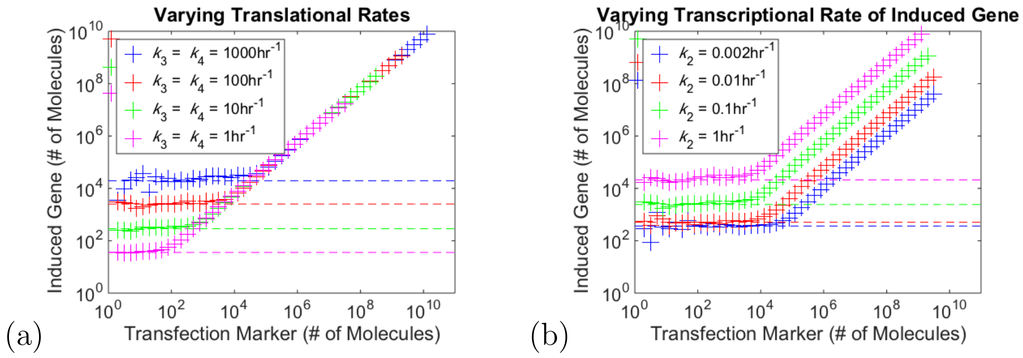

**Figure 2·3:** Simulations of our stochastic model suggest that either increasing translation rates (a) or decreasing transcriptional rates (b) can extend the near-constant induced gene levels at low copy plasmid numbers. X-axis and y-axis stand for number of molecules of the transfection marker and the induced protein in each bin. Best fit horizontal lines are drawn for reference. (a) Comparison of models in which the translational rates decrease in order from 1000 to 1 molecule per mRNA per hour. (b) Comparison of models in which the transcriptional rate of $D_{\text{induced}}$ increases from 0.002 to 1 molecule per plasmid per hour.

of parameters while holding the remaining parameters constant. We begin by examining possible mechanisms that lead to near-constant induced reporter concentrations at low plasmid numbers, creating two cohorts of models. In each cohort we assume the $K_i$ are normal first-order rate expressions, i.e., $K_1 = k_1 D_{\text{tm}}$ with $K_2$, $K_3$, and $K_4$ defined similarly. The first cohort varies only the translational rate constants $k_3$ and $k_4$, while the second cohort varies only the induced gene's transcriptional rate, $k_2$. Simulations of the stochastic model demonstrate that either increasing translation rates, or decreasing transcription rates, can lead to the observed constant induced reporter levels at low plasmid copy numbers (Figure 2·3).

We next investigate mechanisms that may cause protein concentrations to saturate at high plasmid copy numbers. Though the physical mechanism has not been proven, several experimental studies conclude that some steps of the transcription process may saturate in cells expressing large amounts of mRNA [Takahashi et al., 2011,Hama et al., 2006]. It has also been suggested that the cationic liposomes used in transfection inhibit the process of transcription [Tachibana et al., 2001]. Hence, it is possible that a high concentration of liposomes (associated with high plasmid copy numbers) is also a mechanism that induces saturation in transcription rates. Motivated by these possible mechanisms, we modify our stochastic model to incorporate saturation of transcriptional kinetics. We now take the transcription rates, $K_1$ and $K_2$, to be given by saturating MM approximations with MM constants, $K_{D_{\text{tm}}}$ and $K_{D_{\text{induced}}}$ (see Chapter 2.3.1). Here smaller $K_D$ values correspond to saturation beginning at lower plasmid copy numbers. By systematically varying both $K_D$ values (see Chapter 2.3.1) we observe that transcriptional saturation may induce protein saturation when $K_{D_{\text{induced}}} \ll K_{D_{\text{tm}}}$ (see Figure 2·4(a)). That is, protein levels as a function of the amount of plasmid may saturate if the transcriptional rate of the induced reporter saturates at a lower level of plasmid than that at which the transcriptional rate of the transfection marker saturates.

Finally, we now investigate whether translational saturation can also induce saturation in protein levels at high plasmid copy numbers. Tachibana et al. presented experimental evidence which suggests that protein synthesis saturates when a large amount of mRNA is present [Tachibana et al., 2002]. Motivated by this study, we now consider a version of our stochastic model where the transcriptional rates $K_1 = k_1 D_{\text{tm}}$ and $K_2 = k_2 D_{\text{induced}}$ are non-saturating first order reactions as in our first model, but the translation rates $K_3$ and $K_4$ are saturating MM approximations. Since the induced gene and the transfection marker are homologous fluorescent genes, we use the same maximal translation rates and same MM constants in $K_3$ and $K_4$ (see Chapter 2.3.1). This final version of our model suggests that under the hypothesis of translational saturation, protein reporter saturation can be observed if $k_2 \gg k_1$, i.e. if the induced gene transcribes faster than the transfection marker's gene (see Figure

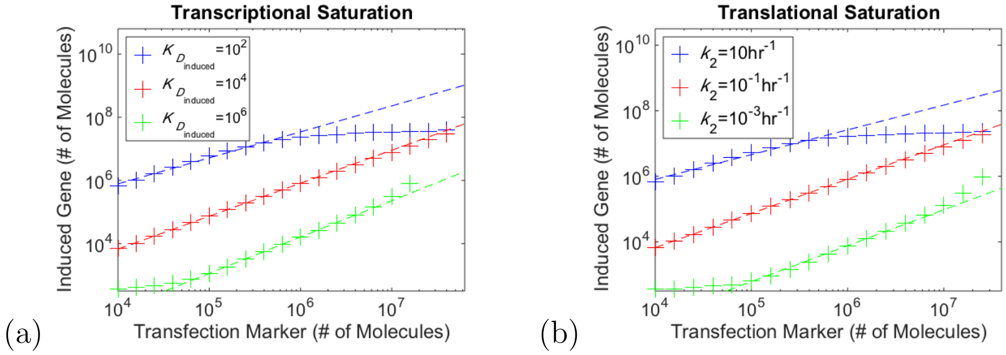

**Figure 2·4:** Simulations of our stochastic model suggest that either either saturation of transcriptional kinetics (a) or saturation of translation kinetics (b) can lead to regimes where the induced gene reporter level saturates at high plasmid copy numbers. X-axis and y-axis stand for number of molecules of the transfection marker and the induced protein in each bin. Least squares regression lines are drawn for reference. (a) Comparison of models built under the hypothesis of transcriptional saturation. The half saturation constant $K_{D_{\text{induced}}}$ increases in order from $10^2$ to $10^6$ molecules, and $K_{D_{\text{tm}}}$ is held fixed at $10^4$ molecules. (b) Comparison of models built under the hypothesis of translational saturation. The transcriptional rate of the induced gene decreases in order from 10 to $10^{-3}$ molecule per plasmid per hour, and the transfection marker transcribes at a constant rate of $10^{-1}$ molecule per plasmid per hour.

2·4(b)).

In summary, we have demonstrated two different physical mechanisms that may induce a near-constant level of the induced gene reporter at low plasmid copy numbers (high translation rates or low transcription rates). We have also demonstrated two different physical mechanisms that may induce a saturating level of induced gene reporter for high plasmid copy numbers (having the induced gene transcription kinetics saturate at lower plasmid levels than needed for saturation of the transfection marker gene transcription kinetics, or having translational saturation with the induced gene transcribing faster than the transfection marker's gene). Note that the results we have derived do not depend on the precise choice of bin width (see Figure 2·5). In Chapter 2.3.2 we show that these results persist when considering an alternative model for

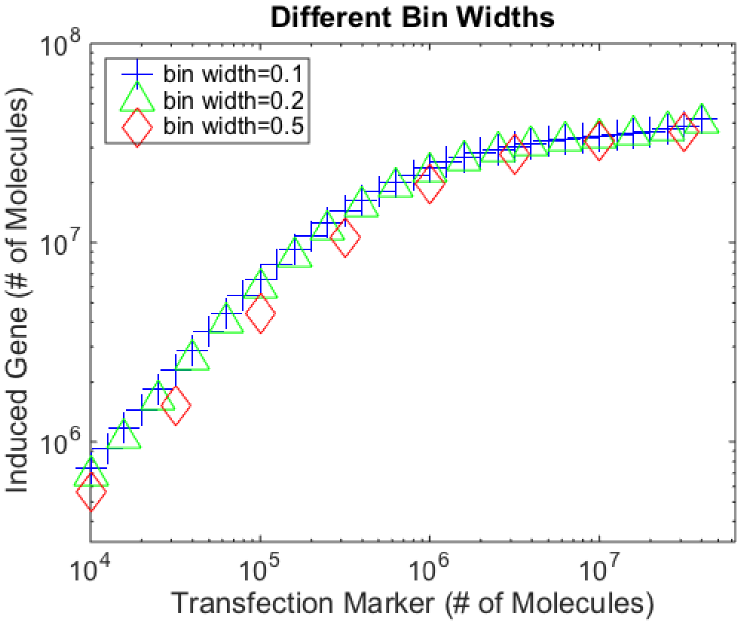

**Figure 2·5:** Simulations of a transcriptional saturation model. X-axis stands for the mid point of each bin, and y-axis number of molecules of the induced protein in each bin. Bin width is chosen to be 0.1, 0.2, and 0.5. Notice, the saturating effect and the general curve are independent of bin size.

the initial plasmid distributions within cells. In Chapter 2.3.3 we explain why the observed saturation region within the flow cytometry data is unlikely to be due to experimental noise.

Our analysis poses a challenge to the characterization of circuit behavior in TTMC. The stochastic models demonstrate there are multiple (physical) mechanisms that can explain the observed saturation (constant levels) of the induced gene reporter at high (low) plasmid copy numbers. Due to the complexity of these models it seems unlikely one could fit them, or even select which is most appropriate, from just single-time-point flow cytometry data.

### 2.3.1 Model Details

Details regarding the two-stage stochastic gene expression models can be found in this subchapter.

The induced gene and the transfection marker are encoded on separate plasmids. Gene expression is modeled as a two-stage process consisting of transcription and translation. Length of the simulation is 50 hours. Cell division takes place every 20 hours, and plasmids are binomially partitioned in daughter cells upon cell division. The initial cell cycle position for a cell is sampled randomly from the uniform distribution unif(0,20). The reaction rates can be expressed as follows:

$$\begin{aligned}
K_1 &= k_1 \cdot D_{\text{tm}}, & K_2 &= k_2 \cdot D_{\text{induced}}, \\
K_3 &= k_3 \cdot D_{\text{tm}}, & K_4 &= k_4 \cdot D_{\text{induced}}, \\
\Lambda_1 &= \lambda_1 \cdot M_{\text{tm}}, & \Lambda_2 &= \lambda_2 \cdot M_{\text{induced}}, \\
\Lambda_3 &= \lambda_3 \cdot P_{\text{tm}}, & \Lambda_4 &= \lambda_4 \cdot P_{\text{induced}},
\end{aligned}$$

where $\lambda_j$ $(j = 1 - 4)$ and $k_j$ $(j = 1 - 4)$ are intrinsic rates. Under the hypothesis of transcriptional saturation,

$$\begin{aligned}
K_1 &= 1000 \cdot \frac{D_{\text{tm}}}{D_{\text{tm}} + K_{D_{\text{tm}}}}, \\
K_2 &= 1000 \cdot \frac{D_{\text{induced}}}{D_{\text{induced}} + K_{D_{\text{induced}}}},
\end{aligned}$$

where $K_{D_{\text{tm}}} = 10^4$, and $K_{D_{\text{induced}}} = 10^2, 10^4,$ or $10^6$. Under the hypothesis of translational saturation,

$$\begin{aligned}
K_3 &= 1000000 \cdot \frac{M_{\text{tm}}}{M_{\text{tm}} + 10000}, \\
K_4 &= 1000000 \cdot \frac{M_{\text{induced}}}{M_{\text{induced}} + 10000}.
\end{aligned}$$

| Parameter Values | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Figure # | $k_1$ | $k_2$ | $k_3$ | $k_4$ | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ |
| Figure 2·3(a) and Figure 2·6 | 0.1 | 0.1 | 1000 | 1000 | 0.01 | 0.01 | 0.01 | 0.01 |
| | 0.1 | 0.1 | 100 | 100 | 0.01 | 0.01 | 0.01 | 0.01 |
| | 0.1 | 0.1 | 10 | 10 | 0.01 | 0.01 | 0.01 | 0.01 |
| | 0.1 | 0.1 | 1 | 1 | 0.01 | 0.01 | 0.01 | 0.01 |
| Figure 2·3(b) and Figure 2·7 | 0.1 | 0.002 | 100 | 100 | 0.01 | 0.01 | 0.01 | 0.01 |
| | 0.1 | 0.01 | 100 | 100 | 0.01 | 0.01 | 0.01 | 0.01 |
| | 0.1 | 0.1 | 100 | 100 | 0.01 | 0.01 | 0.01 | 0.01 |
| | 0.1 | 1 | 100 | 100 | 0.01 | 0.01 | 0.01 | 0.01 |
| Figure 2·4(a) and Figure 2·8 | NA | NA | 100 | 100 | 0.01 | 0.01 | 0.01 | 0.01 |
| | NA | NA | 100 | 100 | 0.01 | 0.01 | 0.01 | 0.01 |
| | NA | NA | 100 | 100 | 0.01 | 0.01 | 0.01 | 0.01 |
| Figure 2·4(b) and Figure 2·9 | 0.1 | 10 | NA | NA | 0.01 | 0.01 | 0.01 | 0.01 |
| | 0.1 | 0.1 | NA | NA | 0.01 | 0.01 | 0.01 | 0.01 |
| | 0.1 | 0.001 | NA | NA | 0.01 | 0.01 | 0.01 | 0.01 |

**Table 2.1:** Parameter values for the two-stage models. $k_1$ and $k_2$ have the units of # of molecules per plasmid per hour. $k_3$ and $k_4$ have the units of # of molecules per mRNA per hour. $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$ have the units of reciprocal hours. In models corresponding to Figure 2·4(a) of the main text and Figure 2·8, $k_1$ and $k_2$ are not constant since transcriptional rates are subject to saturation. In models corresponding to Figure 2·4(b) of the main text and Figure 2·9, $k_3$ and $k_4$ are not constant since translational rates are subject to saturation. NA stands for not applicable.

Values of the parameters in each model are shown in Table 2.1.

For the models detailedly described in the main text, the initial total number of plasmids in a given cell is assumed to follow a log-normal distribution: $N[\log(100), \log(10)]$ [Davidsohn et al., 2015]. The initial copy numbers of each species of plasmid, $D_{\text{tm}}$ and $D_{\text{induced}}$ given the total number of plasmids $P$ are assumed to follow binomial distributions: $B(P, 0.5)$ [Davidsohn et al., 2015].

### 2.3.2 Exploring Other Plasmid Distributions

In [Davidsohn et al., 2015], co-transfected plasmids were pre-mixed before forming complexes with lipofectamine, and according to [Schwake et al., 2010], numbers of co-transfected plasmids in individual cells should be highly correlated. In co-transfection experiments, the correlation between co-transfected plasmids can be adjusted by changing the co-transfection protocol [Schwake et al., 2010]. Besides the models described in the main text, we construct, simulate, and analyze additional cohorts of detailed two-stage models, assuming that numbers of co-transfected plasmids follow a bivariate log-normal distribution, and correlations between co-transfected plasmids can be varied. The initial plasmid copy numbers in a cell, $D_{\text{tm}}$ and $D_{\text{induced}}$, are integer roundups of two continuous variables sampled from a bivariate lognormal distribution,

$$N\left[\begin{pmatrix} \log(100) \\ \log(100) \end{pmatrix}, \begin{pmatrix} [\log(10)]^2 & \rho \cdot [\log(10)]^2 \\ \rho \cdot [\log(10)]^2 & [\log(10)]^2 \end{pmatrix}\right]$$
$$.$$

$\rho$ represents the correlation between $D_{\text{induced}}$ and $D_{\text{tm}}$, and is set to values of 0.25, 0.5, and 0.75 to represent low, medium, and high correlation in different models. Length of the simulation, assumptions about cell division and asynchronicity, and definitions of the reaction rates are kept the same. Values of the rest of the parameters in each model can be found in Table 2.1. Results of the simulation can be found in Figures 2·6, 2·7, 2·8, and 2·9. Irrespective of the underlying plasmid distributions, we reach the same conclusions on biological hypotheses and parameter regions that can explain our experimental observations qualitatively. Another interesting point worth noticing is that as is shown by Figures 2·8 and 2·9, the saturation behavior is only observed when $\rho$ is set to 0.75, indicating the possible role of co-transfection efficiency as a contributing factor.

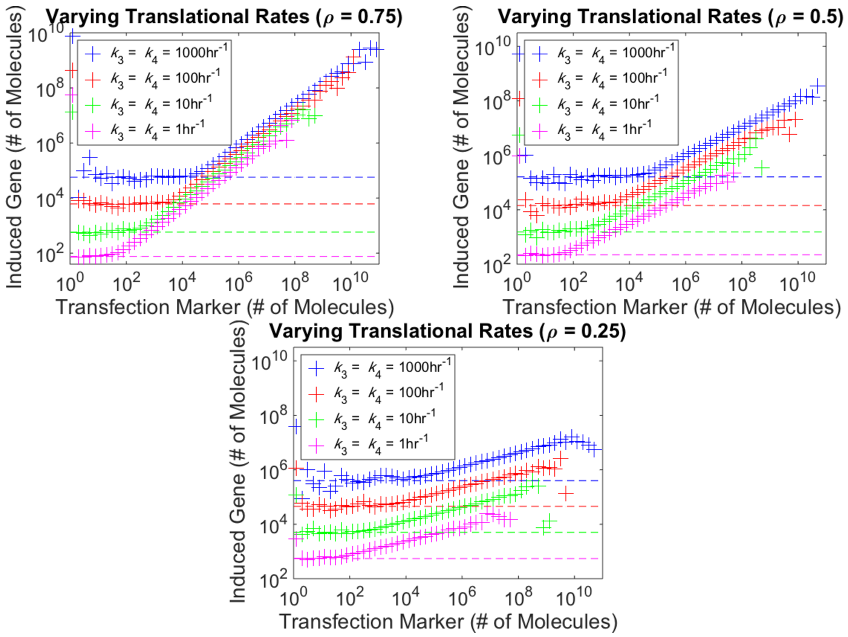

**Figure 2·6:** Comparison of models in which the translational rates decrease in order from 1000 to 1 molecule per mRNA per hour.

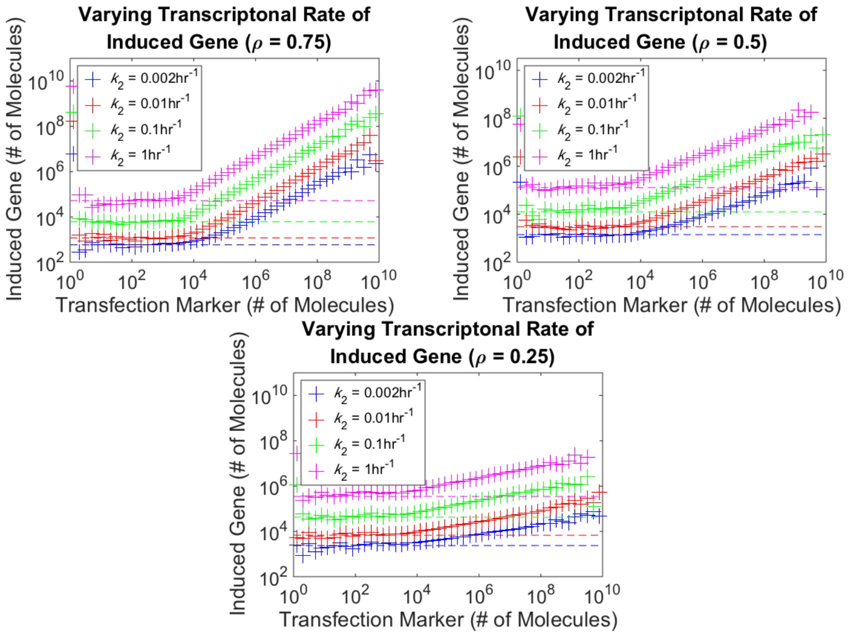

**Figure 2·7:** Comparison of models in which the transcriptional rate of the induced gene increases from 0.002 to 1 molecule per plasmid per hour.

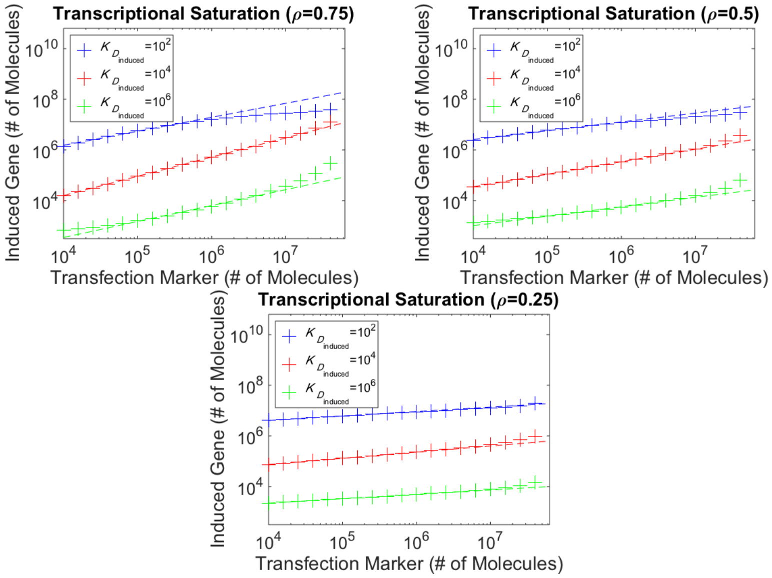

**Figure 2·8:** Comparison of models built under the hypothesis of transcriptional saturation. The half saturation constant $K_{D_{\text{induced}}}$ increases in order from $10^2$ to $10^6$ molecules, and $K_{D_{\text{tm}}}$ is held fixed at $10^4$ molecules.

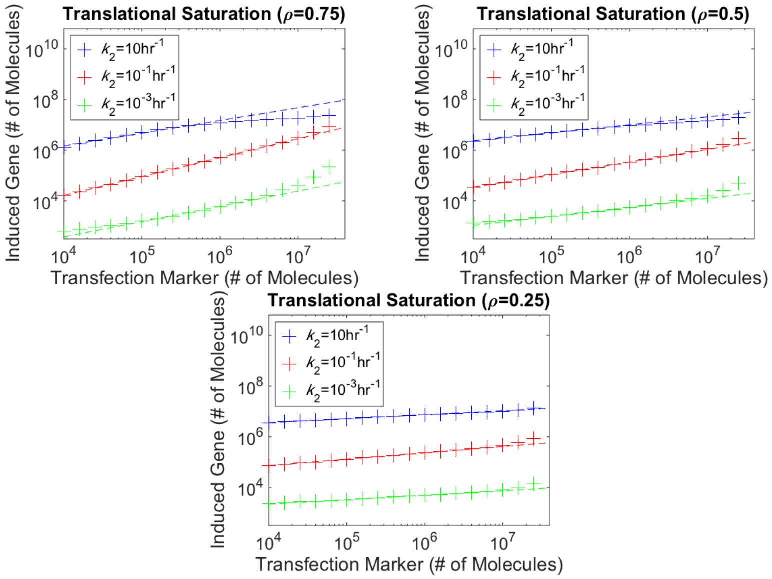

**Figure 2·9:** Comparison of models built under the hypothesis of translational saturation. The transcriptional rate of the induced gene decreases in order from 10 to $10^{-3}$ molecule per plasmid per hour, and the transfection marker transcribes at a constant rate of $10^{-1}$ molecule per plasmid per hour.

### 2.3.3 Possibility of Experimental Noise as the Cause of Saturation

We note that the special regions at low and high plasmid numbers (Figure 2) could be speculated to arise from the limited detection range of the flow cytometer. Data in [Davidsohn et al., 2015] suggest that the upper detection limit is at least $10^{9.2}$ MEFL (Supplementary Figure 24(a) of [Davidsohn et al., 2015]). The possibility of a detection limit can then be ruled out at high plasmid numbers for two reasons. First, the induced and the regulated proteins saturate near $10^8$ and $10^7$ MEFL, respectively (Figure 2·2). Near $10^8$ and $10^7$ MEFL, the geometric standard deviations of (MEFL) concentrations of the induced protein and the regulated protein are between 2 and 2.5. Protein concentrations within each bin are approximately lognormal distributed [Beal, 2017], which means 95% of the cells are within two geometric standard deviations from the geometric means, which is less than $10^{9.2}$ MEFL. In other words, there are fewer than 2.5% of the cells whose fluorescence intensity exceeds $10^{9.2}$ MEFL. Hence, the upper limit of the detection range at $10^{9.2}$ does not have substantial effects on the reported values of our data. Second, saturations due to instrument range often cause protein histograms to have an abrupt cut-off shape, i.e., measurements exceeding the upper detection limit would all gather near a single value (see Supplementary Figure 16(b), Supplementary Figure17(b), and Supplementary Figure18(b) of [Davidsohn et al., 2015]). At low plasmid numbers, autofluorescence is a major obstacle limiting the detection sensitivity [Brahme, 2014]. Despite autofluorescence corrections, data towards the lower end may be susceptible to experimental noise. Our stochastic models provide an alternative approach to studying these systems with low numbers of molecules. The simulations suggest the possibility of near-constant average protein levels in minimally transfected cells when flow cytometry measurement noise is removed.

## 2.4 Bin-Dependent ODE Model

Though mechanistic details cannot be disentangled from single-time flow cytometry measurements, characterization of building blocks such as regulatory switches remains a critical problem to be addressed. This is needed to enable the development of models that can predict the dynamics of circuits/pathways with more components, and, which exhibit more complicated behaviors. To further this goal, we now develop a simple, phenomenological ODE model that can accurately describe single-time transient transfection flow cytometry data. While development of a more physically detailed model would be ideal, as shown in the last subchapter it would require additional experimental data to be uniquely determined.

To account for the observed saturation in protein concentration, we propose replacing the traditional Hill-function-based model (Equation (2.2)) with a bin-dependent model. The bin-dependent model divides flow cytometry data into two subsets based on plasmid copy number, i.e., one with and one without saturation.

$$
\begin{aligned}
\frac{dI_i}{dt} &= \alpha_i \cdot \phi(t) - \lambda \cdot I_i \\
\frac{dO_i}{dt} &= \begin{cases}
\beta \cdot \phi(t) \cdot \left(\frac{P_i}{P_1}\right)^f \cdot \left(\frac{1-\gamma}{1+\left(\frac{I_i}{d}\right)^h} + \gamma\right) - \lambda \cdot O_i, & \text{if } P_i < P_{i'} \\
\beta \cdot \phi(t) \cdot \left(\frac{P_{i'}}{P_1}\right)^f \cdot \left(\frac{P_i}{P_{i'}}\right)^g \cdot \frac{1-\gamma}{1+\left(\frac{I_i}{d}\right)^h} & \\
\quad + \beta \cdot \phi(t) \cdot \left(\frac{P_i}{P_1}\right)^f \cdot \gamma - \lambda \cdot O_i, & \text{if } P_i \geq P_{i'}
\end{cases} \\
\phi(t) &= \left(\frac{1}{2}\right)^{\lfloor \frac{t}{T} \rfloor},
\end{aligned}
\tag{2.3}
$$

In Equation (2.3), $i'$ is the bin that separates high plasmid copy number from the rest. The separating bin is chosen to be the bin at which average concentrations of the co-transfected protein switch from linear growth to plateauing. For high plasmid copy number, we assume the log of the plasmid copy number can be approximated as a linear function of the log of the transfection marker, but with a flatter slope (Figure 2·2). $f$ and $g$ capture the relationship between the concentrations of the transfection marker and the maximal production rates of the output protein for low and high copy numbers, respectively. The rest of the notations follow the Hill-funtion-based model (Equation (2.2)). We do not explicitly characterize the functional form of how $\alpha_i$ depends on the plasmid level as we simply fit a different value of $\alpha_i$ for each bin. Note that the bin-dependent model only requires one additional parameter than a standard Hill-function-based model.

## 2.5 Evaluating Model Performances

We fit the traditional Hill-function-based model (see Equation (2.2)) and the bin-dependent model (see Equation (2.3)) to the TAL14, TAL21, and LmrA datasets from [Davidsohn et al., 2015] for validation (TAL14, TAL21, and LmrA are names of the repressors in the regulatory switches). Model fitting is implemented via minimizing the mean-squared errors (MSE) between the log of observed and predicted concentrations of the regulated proteins. We log-transform the concentrations to reduce the absolute errors that are often associated with measurements of large protein concentrations on a linear scale [Braun et al., 2005].

Assume the regulatory switch is induced at $m$ dosages, and cells are segmented into $n$ bins by their plasmid copy numbers. Let $O_{iu}$ denote the averaged measurements of concentrations of the regulated protein in the $i$-th bin at dose level $u$ at the final time point $t^*$, and $\hat{O}_{iu}$ the counterpart numerically simulated by the model. We fit $\log\left(\hat{O}_{iu}\right)$ to $\log\left(O_{iu}\right)$ by iteratively searching for the set of parameters that minimize

| Optimized fits | | | | | |
|---|---|---|---|---|---|
| Model | $\beta$ (Unit: MEFL/hr) | $f$ | $d$ (Unit: MEFL) | $h$ | Error |
| TAL14 | $5.52 \times 10^4$ | 1.47 | $1.04 \times 10^5$ | 0.73 | 0.013 |
| TAL21 | $6.96 \times 10^4$ | 1.28 | $2.13 \times 10^5$ | 0.68 | 0.015 |
| LmrA | $1.51 \times 10^4$ | 1.72 | $2.34 \times 10^6$ | 0.92 | 0.020 |
| | $\gamma$ | | | | |
| TAL14 | $1.50 \times 10^{-3}$ | | | | |
| TAL21 | $1.91 \times 10^{-5}$ | | | | |
| LmrA | $5.85 \times 10^{-4}$ | | | | |

**Table 2.2:** Optimal parameters and MSE for the traditional Hill-function-based model fit to the complete dataset. All parameter values are rounded to two digits after the decimal point.

the MSE [Carpenter, 1960]:

$$\frac{\sum_{u=1}^{m} \sum_{i=1}^{n} \left[ \log\left(O_{iu}\right) - \log\left(\hat{O}_{iu}\right) \right]^2}{mn - \# \text{ of params}}$$

via the `GlobalSearch` solver in Matlab. `GlobalSearch` uses a scatter-search mechanism to generate start points, initiates a local solver from these start points, and reevaluates the start points during the minimization process. We implement `GlobalSearch` using the local solver `fmincon`, and for `fmincon`, we use the '`sqp`' algorithm. To fit the traditional Hill-function-based models, we set boundaries of $\log_{10}(d)$, $\log_{10}(\beta)$, $\log_{10}(f)$, $\log_{10}(h)$, and $\log_{10}(\gamma)$ to be $[2, 8]$, $[-2, 6]$, $[-1, 1]$, $[-4, 4]$, and $[-5, 0]$, respectively. To fit the bin-dependent models, we keep the above settings and set the boundary of $\log_{10}(g)$ to be $[-3, 1]$. The rest of the search algorithm parameters are set to their default values.

For the bin-dependent model, the bin that separates flow cytometry data into subsets of fast and slow protein production is chosen to be $10^{7.1}$ MEFL for TAL14 and TAL21, and $10^{7.4}$ MEFL for LmrA since in the dataset, saturation in protein production is observed to the right of $10^7$ MEFL and $10^{7.3}$ MEFL, respectively (Figure 2·2). For comparison, a traditional Hill-function-based model is fit to the complete

| Optimized fits | | | | | |
|---|---|---|---|---|---|
| Model | $\beta$ (Unit: MEFL/hr) | $f$ | $d$ (Unit: MEFL) | $h$ | Error |
| TAL14 | $4.92 \times 10^4$ | 1.67 | $6.87 \times 10^4$ | 0.71 | 0.004 |
| TAL21 | $5.20 \times 10^4$ | 1.56 | $2.02 \times 10^5$ | 0.68 | 0.004 |
| LmrA | $1.86 \times 10^4$ | 1.84 | $4.01 \times 10^5$ | 0.55 | 0.008 |
| | $\gamma$ | | | | |
| TAL14 | $1.60 \times 10^{-3}$ | | | | |
| TAL21 | $1.08 \times 10^{-3}$ | | | | |
| LmrA | $5.81 \times 10^{-3}$ | | | | |

**Table 2.3:** Optimal parameters and MSE over the reduced dataset for the traditional Hill-function-based model fit to the reduced dataset. All parameter values are rounded to two digits after the decimal point. Note that when evaluated over all 21 bins between $10^{5.8}$ MEFL and $10^{7.9}$ MEFL, this model produces MSE of 0.039, 0.062, and 0.038, respectively.

| Optimized fits | | | | | |
|---|---|---|---|---|---|
| Model | $\beta$ (Unit: MEFL/hr) | $f$ | $d$ (Unit: MEFL) | $h$ | Error |
| TAL14 | $4.87 \times 10^4$ | 1.74 | $5.39 \times 10^4$ | 0.68 | 0.004 |
| TAL21 | $4.68 \times 10^4$ | 1.58 | $2.90 \times 10^5$ | 0.72 | 0.005 |
| LmrA | $1.66 \times 10^4$ | 1.91 | $3.73 \times 10^5$ | 0.59 | 0.009 |
| | $\gamma$ | $g$ | | | |
| TAL14 | $2.83 \times 10^{-4}$ | 1.10 | | | |
| TAL21 | $1.10 \times 10^{-3}$ | 0.83 | | | |
| LmrA | $2.36 \times 10^{-5}$ | 1.09 | | | |

**Table 2.4:** Optimal parameters and MSE for the bin-dependent model fit to the complete dataset. All parameter values are rounded to two digits after the decimal point.

dataset (all 21 bins between $10^{5.8}$ MEFL and $10^{7.9}$ MEFL), and to a reduced dataset (12 bins between $10^{5.8}$ MEFL and $10^{7.0}$ MEFL for TAL14 and TAL21; 15 bins between $10^{5.8}$ MEFL and $10^{7.3}$ MEFL for LmrA). The parameters in fit models are shown in Tables 2.2 - 2.4, and the fit model values versus the experimental values of the fluorescent reporters are shown in Figures 2·10 and 2·11. When fit to just the reduced dataset, the traditional kinetic model produces much smaller errors than when fit to all bins (Table 2.5). However, the model fit to the reduced dataset only works well on the reduced dataset; evaluated at high plasmid copy numbers, this model deviates substantially from observations (Figures 2·12 and 2·13). Evaluated over all 21 bins, the model fit to the reduced dataset produces MSE of 0.039, 0.062, and 0.038, respectively. In comparison, the bin-dependent model has only one more parameter but fits the data well for all plasmid copy numbers (Table 2.5).

We further compare the Hill-function-based model and the bin-dependent model via cross-validation. We conduct a 12-fold cross-validation by randomly dividing the flow cytometry data into 12 subsets of the same size, fitting the models separately on each combination of 11 subsets, and then testing the models on the single subsets that were left out [Geisser, 1993]. The fitting errors and the testing errors are then averaged over the 12 combinations of subsets. The fitting errors are defined as [Carpenter, 1960]:

$$\frac{\sum_{u=1}^{m}\sum_{i=1}^{n}\left[\log\left(O_{iu}\right)-\log\left(\hat{O}_{iu}\right)\right]^{2}}{mn-\#\text{ of params}}. \tag{2.4}$$

The testing errors are defined as [Carpenter, 1960]:

$$\frac{\sum_{u=1}^{m}\sum_{i=1}^{n}\left[\log\left(O_{iu}\right)-\log\left(\hat{O}_{iu}\right)\right]^{2}}{mn}. \tag{2.5}$$

Our results suggest that both the fitting errors and the testing errors of the bin-dependent models are 1.5 - 2 times better than those of the Hill-function-based models (Tables 2.6 and 2.7). The bin-dependent model shows a less significant improvement for LmrA than for TAL14 and TAL21. A possible explanation is that for LmrA, the saturation effect is observed in six bins to the right of $10^{7.3}$ MEFL rather than in nine bins to the right of $10^{7}$ MEFL. For each repressor, we choose the model that produces the least testing error among 12 cross-validated models to be the best model. We evaluate the best models for each plasmid copy number. The results indicate that the

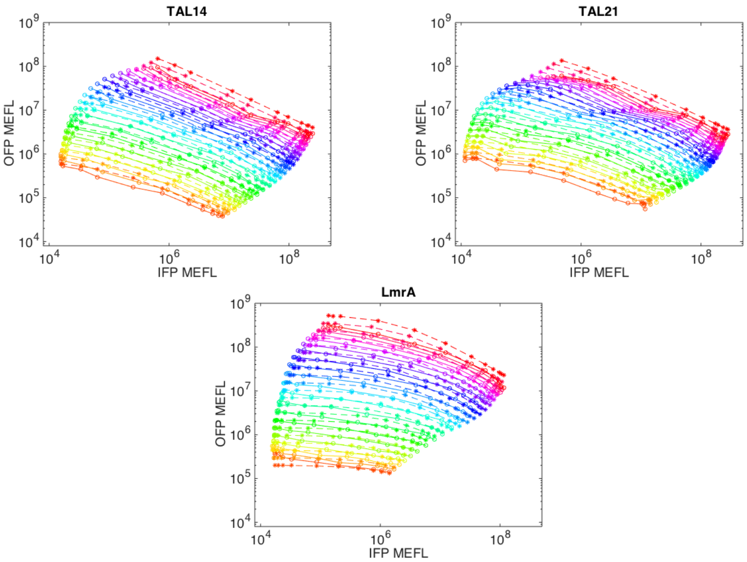

**Figure 2·10:** Comparison between complete data and the traditional Hill-function-based TAL14, TAL21, and LmrA models fit to the complete dataset. Plasmid copy number is shown by color. Solid lines are experimental data (data from [Davidsohn et al., 2015]), and dashed lines are model fits.

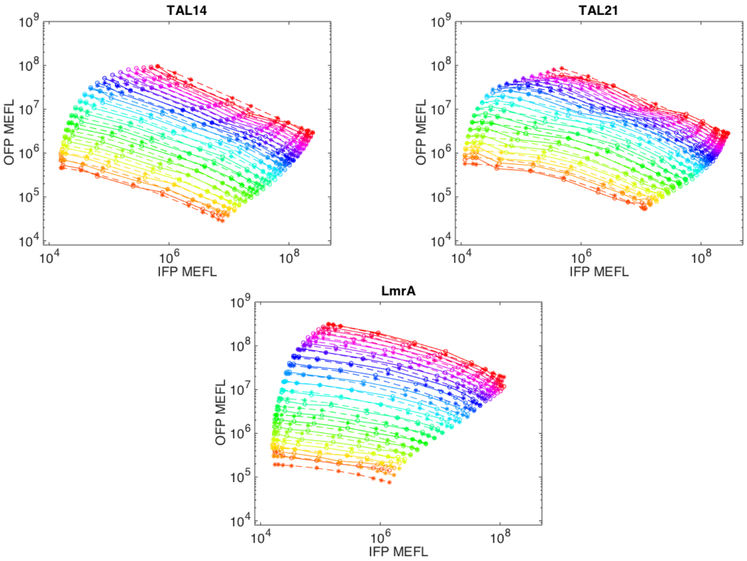

**Figure 2·11:** Comparison between complete data and the bin-dependent TAL14, TAL21, and LmrA models fit to the complete dataset. Plasmid copy number is shown by color. Solid lines are experimental data (data from [Davidsohn et al., 2015]), and dashed lines are model fits.

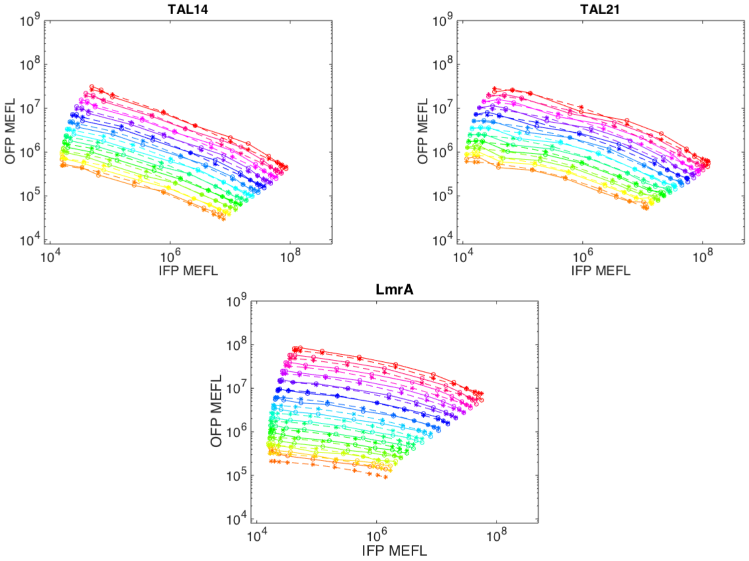

**Figure 2·12:** Comparison between reduced data and the traditional Hill-function-based TAL14, TAL21, and LmrA models fit to the reduced dataset. Plasmid copy number is shown by color. Solid lines are experimental data (data from [Davidsohn et al., 2015]), and dashed lines are model fits.

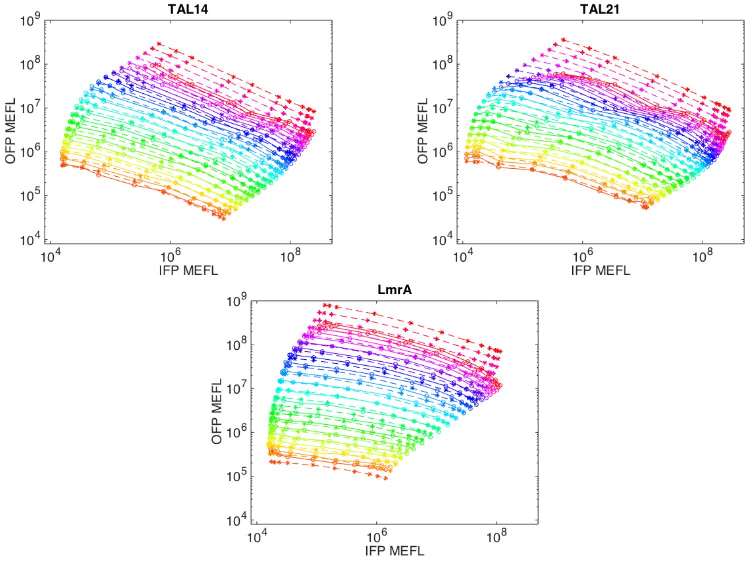

**Figure 2·13:** Comparison between complete data and the traditional Hill-function-based TAL14, TAL21, and LmrA models fit to the reduced dataset. The models are fit to reduced datasets but are evaluated at all plasmid copy numbers. Plasmid copy number is shown by color. Solid lines are experimental data (data from [Davidsohn et al., 2015]), and dashed lines are model fits.

**Table 2.5:** MSE of the models. “Complete” means the traditional Hill-function-based model is fit to the entire dataset, and “reduced” means the Hill-function-based model is fit to data between $10^{5.8}$ MEFL and $10^{7}$ MEFL (TAL14 and TAL21) or data between $10^{5.8}$ MEFL and $10^{7.3}$ MEFL (LmrA). Note, for the reduced model, goodness of fit is only evaluated by comparison to data between $10^{5.8}$ MEFL and $10^{7}$ MEFL(TAL14 and TAL21) or data between $10^{5.8}$ MEFL and $10^{7.3}$ MEFL (LmrA). When error of the reduced model is evaluated over all bins, it is much worse (numbers in parentheses in third column).

| Goodness of fit | | | |
|---|---|---|---|
| Repressor | Hill-function-based (complete) | Hill-function-based (reduced) | bin-dependent |
| TAL14 | 0.013 | 0.004 (0.039) | 0.004 |
| TAL21 | 0.015 | 0.004 (0.062) | 0.005 |
| LmrA | 0.020 | 0.008 (0.038) | 0.009 |

bin-dependent models produce not only lower but also more consistent errors across all bins (Figure 2·14). The errors of the Hill-function-based models get large near $10^{7}$ MEFL and $10^{7.8}$ MEFL for all repressors. This signals that there are patterns in the data that are not explained by the Hill-function-based models [Martin et al., 2017]. The bin-dependent model produces larger errors for LmrA than for TALER repressors because there are slight indications of a near-constant region at low plasmid numbers for LmrA (Figure 2·2). In summary, we find that the bin-dependent model consistently provides significantly better fits to the experimental data than the Hill-function based model.

Note, for high-plasmid-count subsets, our bin-dependent model assumes the log of the maximal protein production rate is approximated as a linear function of the log of the transfection marker. Although the relationship is arguably better fit by other functions, our assumption leads to a model with a good fit across the entire dataset, while only requiring one additional parameter.

The bin-dependent model presented here provides a new solution to characterizing fundamental synthetic constructs quantitatively in TTMC. A stochastic two-stage

**Table 2.6:** Averaged fitting errors of the models within the 12-fold cross-validation.

| Fitting Errors | | |
|---|---|---|
| Repressor | Hill-function-based | bin-dependent |
| TAL14 | 0.013 | 0.006 |
| TAL21 | 0.017 | 0.009 |
| LmrA | 0.018 | 0.013 |

**Table 2.7:** Averaged testing errors of the models within the 12-fold cross-validation.

| Testing Errors | | |
|---|---|---|
| Repressor | Hill-function-based | bin-dependent |
| TAL14 | 0.014 | 0.007 |
| TAL21 | 0.017 | 0.008 |
| LmrA | 0.019 | 0.013 |

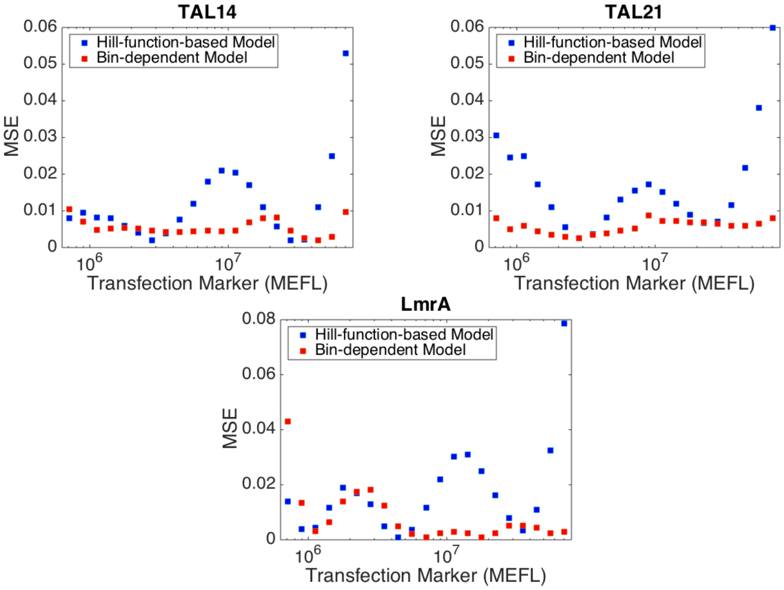

**Figure 2·14:** Testing errors of the best cross-validated models within each bin.

model, if fit to the data, should explicitly account for both the transcriptional rate and the translational rate of the transfection marker, the TF, and the regulated gene. For reproducing the saturation behavior quantitatively, the saturation kinetics may need to be replaced by Hill equations, basal production rates incorporated, and the mean and the variance of the plasmid distribution either fit or experimentally measured. In addition, as we showed above, two stochastic models built under different hypotheses, differing in parameter values and transcription/translation production rate functions, can recapitulate almost identical observations. Moreover, the Gillespie algorithm is a discrete simulation algorithm, which assumes species populations are given by numbers of molecules. In order to use the Gillespie algorithm properly for model fitting, one must also know a functional relationship to convert and quantitatively compare average numbers of molecules within cells to flow-cytometry measurements (in units of MEFL). In comparison, the value of the bin-dependent model lies in its ability to describe the saturation effects in flow cytometry data accurately without addressing the specific mechanistic details. The bin-dependent ODE model only contains six parameters, but a detailed two-stage stochastic model for fitting would contain ten to eighteen parameters, needs model selection studies to determine functional forms of production terms, and requires assumptions on how to convert MEFL units to numbers of molecules. There is also the issue that stochastic models are substantially more computationally expensive; it is not clear how computationally feasible it would be to simultaneously fit all the parameters in such models.

We can apply the bin-dependent model presented here to similar flow cytometry datasets to construct a characterized library of regulatory switches. The quantitative parameters of regulatory switches can then be used for constructing in silico models for the behaviors of more complicated circuits, such as feedback circuits. Accurate characterization of regulatory switches is a major first step towards improving the predictions of circuit behaviors in TTMC.

# Chapter 3

# Modular Composition

This chapter centers around a novel method of composition that enables forward design of complex circuits in TTMC (Figure 3·1). Under the assumption of modularity, we assume that the behavior of circuits can be predicted based on the behaviors of circuit components, also known as modules [syn, 2014, Gyorgy and Del Vecchio, 2014, Del Vecchio and Sontag, 2007, Del Vecchio et al., 2016, Sivakumar and Hespanha, 2013]. The accuracy of predictions is constrained by the cross-batch variation among different modules. Compatible with the binning in TTMC, our method of composition improves the accuracy of predictions by reducing the cross-batch variation. For validation, we apply our method to cascades consisting of two regulatory switches. Predictions of the mathematical models compare well with the experimental data. Our findings suggest reducing batch effects and selecting a proper model both contribute to improving model predictions.

## 3.1 Components of Circuits

### 3.1.1 Module

As is mentioned in Chapter 2.1, in this thesis a transcriptional regulatory module is defined as a switch gene and the promoter it regulates (Figure 3·2a). The input of the module is the switch gene, and the output, the regulated promoter. The strength of the regulated promoter is often indicated by the expression level of the downstream gene. The promoter can be regulated either positively or negatively, depending on

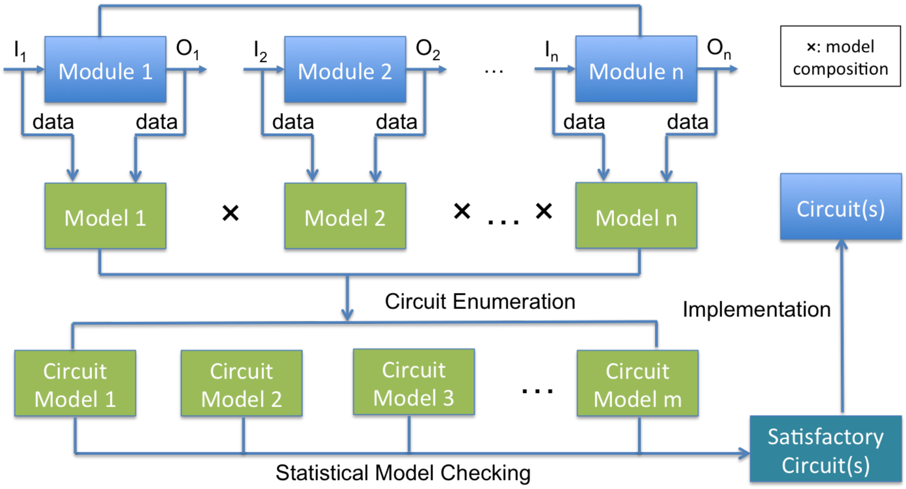

**Figure 3·1:** Our approach to synthesizing satisfactory circuits. The input and the output of each module are measured by flow cytometries. A model is constructed for each module from data. Models for modules are then assembled into models for circuits.

whether the regulator is an activator or an inhibitor. Mathematically, a module M is expressed as: $\mathrm{M} = \{\mathrm{I}, \mathrm{pO}\}$, where I and pO stand for the TF and the promoter, respectively. We assume I is an inhibitor, but similar results can be derived if I is an activator.

The definition we choose is widely used in the community [Ellis et al., 2009, Davidsohn et al., 2015] and has a distinct advantage. Another definition of a module in the community is a transcriptional unit, i.e., the coding sequence for a gene along with the sequences necessary for its transcription [Pierce, 2005]. In comparison, the definition we choose captures the interaction between a TF and a promoter. It maps a module to a transcriptional regulatory model, whose parameters can be directly inferred from experimental data. Based on this definition, models for modules contain all the information needed to quantify signal propagation in a circuit.

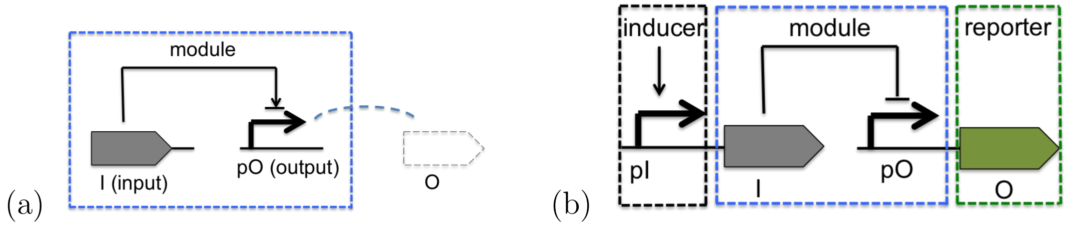

**Figure 3·2:** (a) Graphical representation of a genetic module. The input of a module is the TF, I, while its output is the regulated promoter pO. O is the protein that is expressed by pO. (b) Graphical representation of a regulatory switch. The green dotted box stands for the reporter. The black dotted box stands for the promoter regulated by an external inducer.

### 3.1.2 Reporter

In a circuit, some proteins do not carry regulatory functions. One such example are proteins used as markers for the states of the cells, e.g., fluorescent proteins, antibodies, etc. We refer to these proteins as reporters (Figure 3·2b).

### 3.1.3 External Inducer

Besides modules and reporters, a circuit often contains promoters regulated by external inducers (Figure 3·2b). The connection of TF to these promoters makes it possible to control circuit behaviors via external inducers.

## 3.2 Circuits and Models

### 3.2.1 Modular Connection

Within a set of modules, two are connected if the promoter of one module expresses the TF of the other. Mathematically, the connection between modules M and $\mathrm{M}^*$ can be represented by a tuple $(\mathrm{M}, \mathrm{M}^*)$, where $\mathrm{M} = \{\mathrm{I}, \mathrm{pO}\}$, $\mathrm{M}^* = \{\mathrm{I}^*, \mathrm{pO}^*\}$, and pO expresses $\mathrm{I}^*$.

Similarly, the connection between a module M and a reporter R can be represented by a tuple $(\mathrm{M}, \mathrm{R})$, where $\mathrm{M} = \{\mathrm{I}, \mathrm{pO}\}$, and pO expresses R. The connection between an external-inducer-regulated promoter E and a module M can be represented by a tuple $(\mathrm{E},\mathrm{M})$, where $\mathrm{M}=\{\mathrm{I},\mathrm{pO}\}$, and E expresses I. The connection between E and R can be represented by $(\mathrm{E},\mathrm{R})$, where E expresses R.

### 3.2.2 Graph Representation of a Set of Modules

A set of modules and reporters can be represented by a graph $G$, where modules and reporters are nodes, and connections are edges. Mathematically, $G$ is given by:

$$\begin{aligned} G &= (M, E) \\ M &= \{m_i|\ i = 1,2,...,n\} \cup \{r_k|k = 1,2,...,n'\} \\ m_i &= \{\mathrm{I}_i, \mathrm{pO}_i\} \\ &\forall i,j,k,\ m_i \cap m_j = \emptyset,\ m_i \cap r_k = \emptyset \\ E &= \{(m_i, m_j)|\ \mathrm{pO}_i \text{ expresses } \mathrm{I}_j\} \\ &\cup \{(m_i, r_k)|\ \mathrm{pO}_i \text{ expresses } \mathrm{r}_k\} \end{aligned}$$

### 3.2.3 Composition of Models

Based on the models for modules, we can develop models for general circuit topologies in which each promoter is either constitutively expressed or regulated by one and only one unique TF. We name the circuit to be built the target circuit. Assume the target circuit consists of $m$ modules and $n$ external-inducer-regulated promoters. Let $\{\mathrm{pO}_k\}_{k=1}^{m}$ denote the set of regulated promoters in the target circuit. Because each promoter is regulated by one unique TF, we know for all $k = 1,2,...,m$, there exists a unique gene, also known as the input of the module $\mathrm{I}_k$ such that $\mathrm{I}_k$ regulates $\mathrm{pO}_k$. Similarly, because the strength of the promoter is indicated by the expression level of the downstream gene, we know for all $k = 1,2,...,m$, there exists a unique downstream gene $\mathrm{O}_k$ such that expression of $\mathrm{O}_k$ initiates at $\mathrm{pO}_k$. It is worth mentioning

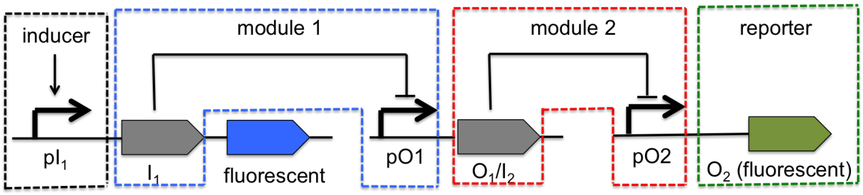

**Figure 3·3:** Graphical representation of a two-transcriptional-repressor cascade. pI$_1$ is a promoter that constitutively expresses I$_1$. I$_1$ downregulates pO$_1$, while O$_1$/I$_2$ downregulates pO$_2$. pO$_1$ and pO$_2$, the promoters that control expression of O$_1$ and O$_2$, have identical sequences of polymerase binding sites, but different sequences of operator binding sites so that I$_1$ and I$_2$ can recognize their targets. The circuit consists of two modules: the I$_1$-pO$_1$ module and the I$_2$/O$_1$-pO$_2$ module. I$_1$ and I$_2$ can be different combinations of TAL14, TAL21, and LmrA.

that through the composition of modules, some TF may be regulated by others, i.e., $\{\mathrm{I}_k\}_{k=1}^m \cap \{\mathrm{O}_k\}_{k=1}^m \neq \emptyset$.

As an example, we derive the following model for a two-transcriptional-repressor cascade. The structure of a two-transcriptional-repressor cascade is illustrated in Figure 3·3.

$$\frac{dI_{1i}}{dt} = \alpha_i \cdot \phi(t) - \lambda \cdot I_{1i}$$

$$\frac{dO_{1i}}{dt} = \begin{cases} \beta_1 \cdot \phi(t) \cdot \left(\frac{P_i}{P_1}\right)^{f_1} \cdot \left(\frac{1-\gamma_1}{1+\left(\frac{I_{1i}}{d_1}\right)^{h_1}} + \gamma_1\right) & \\ -\lambda_1 \cdot O_{1i}, & \text{if } P_i < P_{1i'} \\ \beta_1 \cdot \phi(t) \cdot \left(\frac{P_{i'}}{P_1}\right)^{f_1} \cdot \left(\frac{P_i}{P_{i'}}\right)^{g_1} \cdot \frac{1-\gamma_1}{1+\left(\frac{I_{1i}}{d_1}\right)^{h_1}} & \\ +\gamma_1 \cdot \phi(t) \cdot \left(\frac{P_i}{P_1}\right)^{f_1} - \lambda_1 \cdot O_{1i}, & \text{if } P_i \geq P_{1i'} \end{cases} \tag{3.1}$$

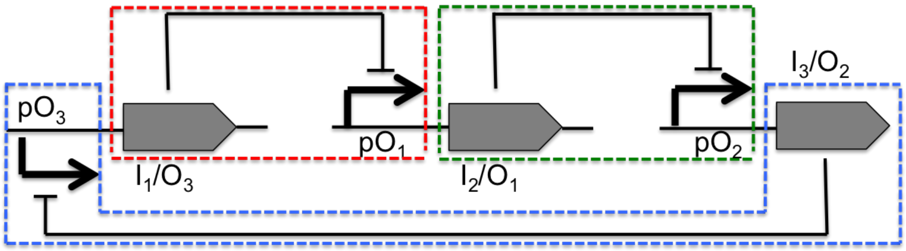

**Figure 3·4:** Graphical representation of a repressilator. I$_1$/O$_3$ downregulates pO$_1$, I$_2$/O$_1$ downregulates pO$_2$, and I$_3$/O$_2$ downregulates pO$_3$. The circuit consists of three modules: the I$_1$/O$_3$-pO$_1$ module, the I$_2$/O$_1$-pO$_2$ module and the I$_3$/O$_2$-pO$_3$ module.

$$\frac{dO_{2i}}{dt} = \begin{cases} \beta_2 \cdot \phi(t) \cdot \left(\frac{P_i}{P_1}\right)^{f_2} \cdot \left(\frac{1-\gamma_2}{1+\left(\frac{O_{1i}}{d_2}\right)^{h_2}} + \gamma_2\right) & \text{if } P_i < P_{2i'} \\ -\lambda_2 \cdot O_{2i}, & \\ \beta_2 \cdot \phi(t) \cdot \left(\frac{P_{i'}}{P_1}\right)^{f_2} \cdot \left(\frac{P_i}{P_{i'}}\right)^{g_2} \cdot \frac{1-\gamma_2}{1+\left(\frac{O_{1i}}{d_2}\right)^{h_2}} & \text{if } P_i \geq P_{2i'} \\ +\gamma_2 \cdot \phi(t) \cdot \left(\frac{P_i}{P_1}\right)^{f_2} - \lambda_2 \cdot O_{2i}, & . \end{cases}$$

where $I_{1i}$, $O_{1i}$ and $O_{2i}$ are average concentrations of the respective parts in the $i$-th bin. $\beta_2$, $f_2$, $d_2$, $h_2$, $\gamma_2$, $\lambda_2$, $g_2$, and $P_{2i'}$ are counterparts of $\beta_1$, $f_1$, $d_1$, $h_1$, $\gamma_1$, $\lambda_1$, $g_1$, and $P_{1i'}$ for the I$_2$/O$_1$-pO$_2$ module.

Via the same approach, a model can be derived for a repressilator, in which a negative feedback loop results in temporal oscillations in the expression levels of the genes. The structure of a repressilator is illustrated in Figure 3·4.

$$\begin{aligned}
\frac{dO_{1i}}{dt} &= \begin{cases} \beta_1 \cdot \phi(t) \cdot \left(\dfrac{P_i}{P_1}\right)^{f_1} \cdot \left(\dfrac{1-\gamma_1}{1+\left(\frac{O_{3i}}{d_1}\right)^{h_1}} + \gamma_1\right) & \\ - \lambda_1 \cdot O_{1i}, & \text{if } P_i < P_{1i'} \\ \beta_1 \cdot \phi(t) \cdot \left(\dfrac{P_{i'_1}}{P_1}\right)^{f_1} \cdot \left(\dfrac{P_i}{P_{i'_1}}\right)^{g_1} \cdot \dfrac{1-\gamma_1}{1+\left(\frac{O_{3i}}{d_1}\right)^{h_1}} & \\ + \beta_1 \cdot \phi(t) \cdot \left(\dfrac{P_i}{P_1}\right)^{f_1} \cdot \gamma_1 - \lambda_1 \cdot O_{1i}, & \text{if } P_i \geq P_{1i'} \end{cases} \\
\frac{dO_{2i}}{dt} &= \begin{cases} \beta_2 \cdot \phi(t) \cdot \left(\dfrac{P_i}{P_1}\right)^{f_2} \cdot \left(\dfrac{1-\gamma_2}{1+\left(\frac{O_{1i}}{d_2}\right)^{h_2}} + \gamma_2\right) & \\ - \lambda_2 \cdot O_{2i}, & \text{if } P_i < P_{2i'} \\ \beta_2 \cdot \phi(t) \cdot \left(\dfrac{P_{i'_2}}{P_1}\right)^{f_2} \cdot \left(\dfrac{P_i}{P_{i'_2}}\right)^{g_2} \cdot \dfrac{1-\gamma_2}{1+\left(\frac{O_{1i}}{d_2}\right)^{h_2}} & \\ + \beta_2 \cdot \phi(t) \cdot \left(\dfrac{P_i}{P_1}\right)^{f_2} \cdot \gamma_2 - \lambda_2 \cdot O_{2i}, & \text{if } P_i \geq P_{2i'} \end{cases} \\
\frac{dO_{3i}}{dt} &= \begin{cases} \beta_3 \cdot \phi(t) \cdot \left(\dfrac{P_i}{P_1}\right)^{f_3} \cdot \left(\dfrac{1-\gamma_3}{1+\left(\frac{O_{2i}}{d_3}\right)^{h_3}} + \gamma_3\right) & \\ - \lambda_3 \cdot O_{3i}, & \text{if } P_i < P_{3i'} \\ \beta_3 \cdot \phi(t) \cdot \left(\dfrac{P_{i'_3}}{P_1}\right)^{f_3} \cdot \left(\dfrac{P_i}{P_{i'_3}}\right)^{g_3} \cdot \dfrac{1-\gamma_3}{1+\left(\frac{O_{2i}}{d_3}\right)^{h_3}} & \\ + \beta_3 \cdot \phi(t) \cdot \left(\dfrac{P_i}{P_1}\right)^{f_3} \cdot \gamma_3 - \lambda_3 \cdot O_{3i}, & \text{if } P_i \geq P_{3i'} \end{cases}
\end{aligned} \tag{3.2}$$

where $O_{1i}$, $O_{2i}$, and $O_{3i}$ are average concentrations of the respective parts in the $i$-th bin. $\beta_2$, $f_2$, $d_2$, $h_2$, $\gamma_2$, $\lambda_2$, $g_2$, and $P_{2i'}$ are counterparts of $\beta_1$, $f_1$, $d_1$, $h_1$, $\gamma_1$, $\lambda_1$, $g_1$, and $P_{1i'}$ for the $I_2/O_1$-p$O_2$ module, and $\beta_3$, $f_3$, $c_3$, $d_3$, $h_3$, $\gamma_3$, $\lambda_3$, $g_3$, and $P_{3i'}$, counterparts for the $I_3/O_2$-p$O_3$ module.

The model for a target circuit is a collection of the models for all modules and external-inducer-regulated promoters. Like most biological data, flow cytometry measurements are subject to noise. This noise may originate from imperfect experimental conditions as well as data calibration [Davidsohn et al., 2015]. In order to make accurate quantitative predictions of circuit behaviors, we need to reduce batch effects by bringing different batches to the same scale, based on the approach taken in [Davidsohn et al., 2015]. The scaling factors among batches can be calculated by comparing the means and the tightness of the data of different batches (details can found in [Davidsohn et al., 2015]). Once the scaling factors are calculated, we use these scaling factors to rescale the parameters of the bin-dependent models since rescaling in our context is first-order linear compensation [Davidsohn et al., 2015], i.e. there is no difference between rescaling the parameters and fitting the parameters to rescaled data. Details of rescaling can be found in Chapter 3.3.

Using the above method, we develop models for the six two-repressor cascades shown in [Davidsohn et al., 2015]: LmrA-TAL14, LmrA-TAL21, TAL14-LmrA, TAL14-TAL21, TAL21-LmrA, and TAL21-TAL14. The bin-dependent cascade models are constructed, and their agreement with experimental measurements are compared with that of the Hill-function-based and the EQuiP models developed in [Davidsohn et al., 2015]. The equations and parameters for the bin-dependent models can be found in Chapter 3.3. The bin-dependent circuit models are developed by composing together the individual module models that were *individually fit* in the previous section. We *do not re-fit* the equations for each model to data for the complete two-module cascades. In this way we can assess how well models fit to individual modules can predict circuit behavior when composed together. To offer a comparable study to [Davidsohn et al., 2015], we use the parameters Davidsohn et al. fit for the Hill-function-based models.

In each of the cascades the downstream reporter EYFP is down-regulated by a repressor, which itself is inhibited by another repressor. The exact representation of the cascade structure can be found in Figure 5(A) of [Davidsohn et al., 2015] or Figure 3·5(a), with Figure 3·6 providing abstractions that highlight the key parts of the circuits. In this figure, $O_2$ corresponds to the EYFP reporter. We compare simulations of the cascade models to experimental data by measuring the differences between simulated and observed concentrations of EYFP 72 hours post transfection (experimental data from [Davidsohn et al., 2015]). Full details of the experimental protocol can be found in [Davidsohn et al., 2015].

The agreement between experimental measurements and model predictions for the six cascades is illustrated in Figure 3·7. For all six cascades, the bin-dependent model is able to capture the positive association between the input and the output (Figure 3·7). It also captures the buffer-like behavior of the cascades, i.e., the dynamic range of the output is narrower compared to that of the input due to low cooperativity of the regulatory modules (Figure 3·7). [Ferrell and Ha, 2014]

To further investigate how well our composed circuit models fit the experimental data, we examined the average mean fold error, defined as the average over all six cascades of the mean-fold errors over all induction levels of each individual cascade. The mean-fold error is defined as $e^{\frac{\sum_{u=1}^{M}\sum_{i=1}^{N}\left|\log\left(\frac{O'_{ui}}{\hat{O}'_{ui}}\right)\right|}{MN}}$, where $\hat{O}'_{ui}$ and $O'_{ui}$ denote the predicted and the observed concentrations of EYFP at hour 72, $M$ the number of inducer levels, and $N$ the number of bins. The rescaled bin-dependent model is found to outperform the Hill-function-based model presented in [Davidsohn et al., 2015], with an average mean-fold error of 1.6 fold for the former vs 3.0 fold for the latter. Moreover, for five out of six cascades the bin-dependent model also produces smaller mean-fold errors than the Hill-Function model [Davidsohn et al., 2015] (Figure 3·5(b)). The accuracy of the bin-dependent model varies relative to EQuiP, achieving a smaller

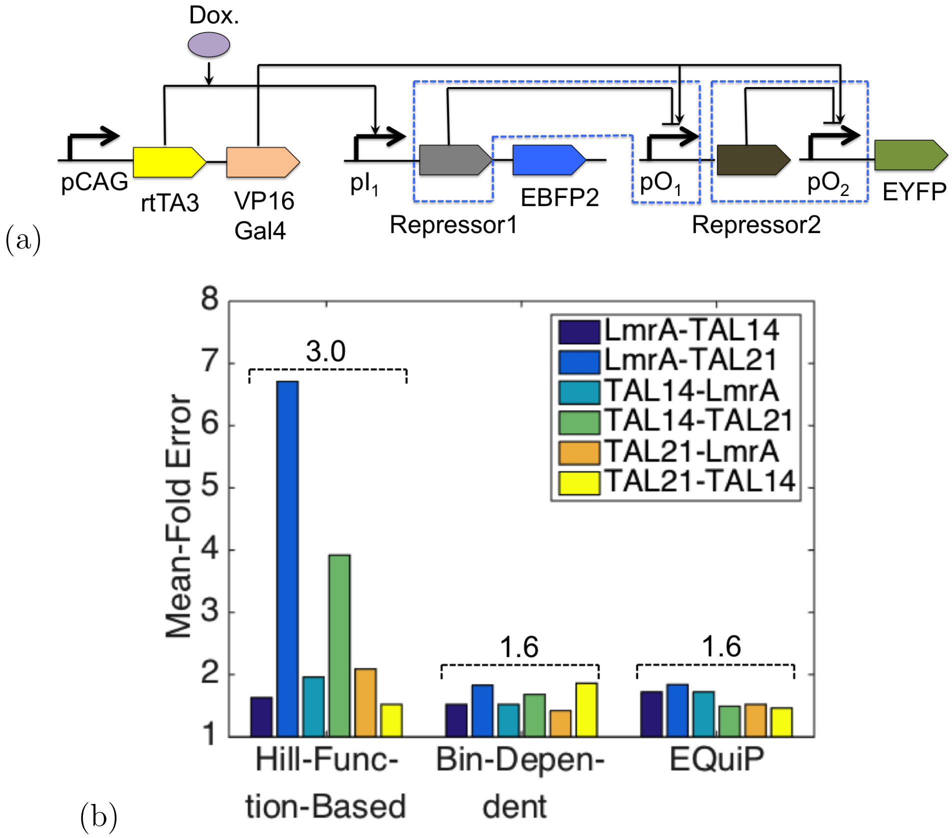

**Figure 3·5:** (a) Detailed representations of a cascade controlled by doxycycline based on Figures 2(A) and 3(A) of [Davidsohn et al., 2015]. The transcriptional repressors can be TAL14, TAL21, or LmrA. Expressions of the repressors (TAL14, TAL21, or LmrA) and EYFP are driven by constitutive rtTA and Gal4 proteins, respectively. rtTA and Gal4, which are required for protein activation, are both constitutively expressed and are not considered as limiting factors for the production of the repressors and EYFP. (b) Comparison of the mean-fold errors of the Hill-function-based models [Davidsohn et al., 2015], the bin-dependent models, and EQuiP [Davidsohn et al., 2015] for each cascade. The experimental data the models are validated against are from [Davidsohn et al., 2015]. Numbers on top of the dotted lines represent the average mean-fold errors of six cascades.

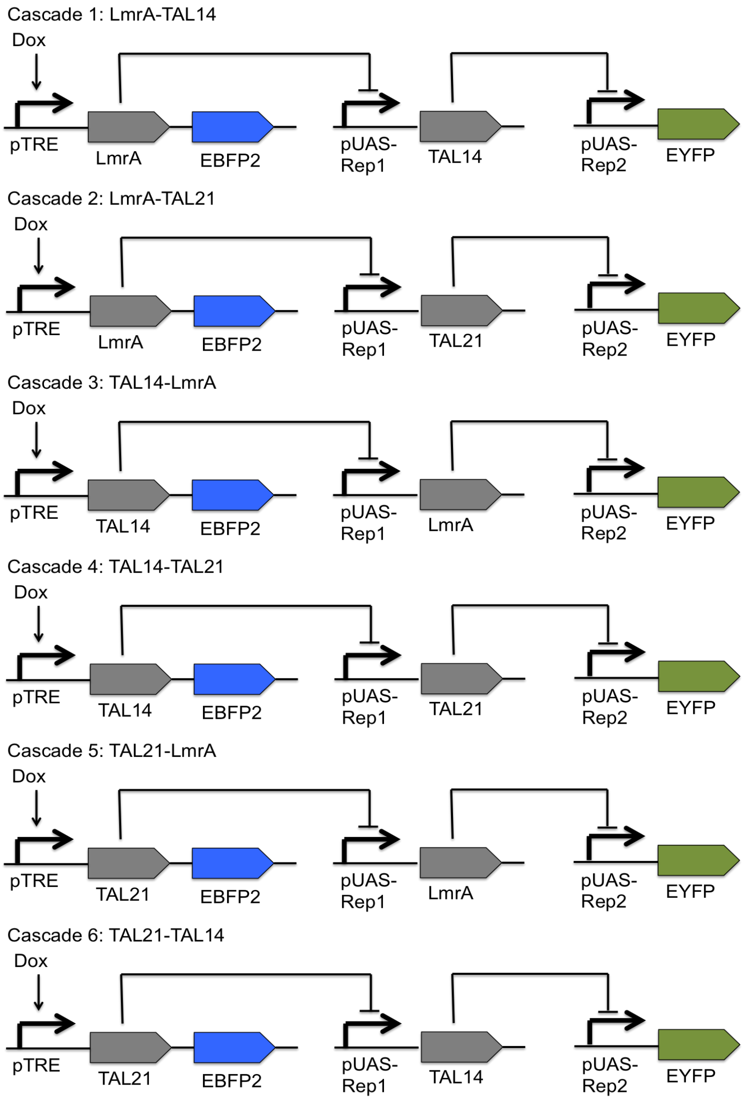

**Figure 3·6:** Abstract representations of LmrA-TAL14, LmrA-TAL21, TAL14-LmrA, TAL14-TAL21, TAL21-LmrA, and TAL21-TAL14 cascades.

mean-fold error for some cascades and larger error others (see Figure 3·5(b)). The average over all six cascades is the same as EQuiP (1.6).

## 3.3 Models for Cascades

Upon modular connection, parameters that are fit to input-output curves of individual modules need to be corrected for batch effects. As is shown in Supporting Information Section 12 of [Davidsohn et al., 2015], the rescaling factors for the input protein I, the output protein O, and the transfection marker are TAL14: 0.29, 0.93, 0.89; TAL21: 0.20, 1, 1.12; LmrA: 1, 0.41, 1 [Davidsohn et al., 2015]. For example, for output protein O, TAL14 has a scaling factor of 0.93, and TAL21, a factor of 1. This means to compare the output protein between TAL14 and TAL21, data for TAL14 need to be multiplied by 0.93 so that the two are brought to the same scale. The scaling factors are used to rescale the parameters in the bin-dependent models before the models are connected into a chain. $d$ is rescaled with the input, $\beta$ rescaled with the output, and $P_i$ rescaled with the transfection marker. Mathematically speaking, if $c_I$, $c_O$, and $c_P$ are the scaling factors of the input, the output, and the transfection marker, then the rescaled bin-dependent model is formulated as follows:

$$\begin{aligned}
\frac{dI'_i}{dt} &= \alpha'_i \cdot \phi(t) - \lambda \cdot I'_i, \\
\frac{dO'_i}{dt} &= \begin{cases} \beta' \cdot \phi(t) \cdot \left(\dfrac{P'_i}{P'_1}\right)^f \cdot \left(\dfrac{1-\gamma}{1+\left(\frac{I'_i}{d'}\right)^h} + \gamma\right) - \lambda \cdot O'_i, & \text{if } P'_i < P'_{i'} \\ \beta' \cdot \phi(t) \cdot \left(\dfrac{P'_{i'}}{P'_1}\right)^f \cdot \left(\dfrac{P'_i}{P'_{i'}}\right)^g \cdot \dfrac{1-\gamma}{1+\left(\frac{I'_i}{d'}\right)^h} & \\ \qquad & \text{if } P'_i \geq P'_{i'} \\ +\, \beta' \cdot \phi(t) \cdot \left(\dfrac{P'_i}{P'_1}\right)^f \cdot \gamma - \lambda \cdot O'_i, & \end{cases} \qquad (3.3) \\
\phi(t) &= \left(\frac{1}{2}\right)^{\lfloor \frac{t}{T} \rfloor},
\end{aligned}$$

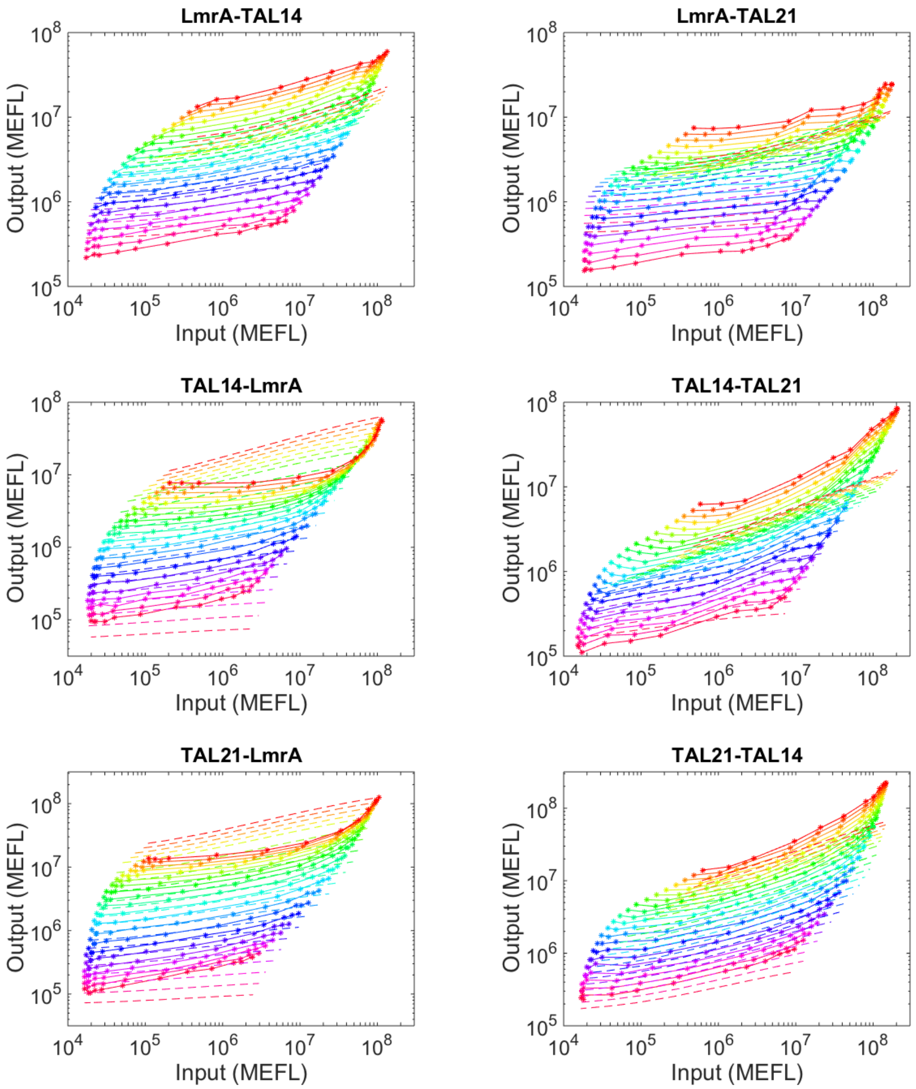

**Figure 3·7:** Comparison between the experimental data and the predictions of the output made by the bin-dependent model. Plasmid copy number is shown by color. Solid curves are experimental data, and dashed curves are model predictions (the experimental data are from [Davidsohn et al., 2015]). From top left to bottom right in the order of left to right and top to bottom are cascades LmrA-TAL14, LmrA-TAL21, TAL14-LmrA, TAL14-TAL21, TAL21-LmrA, and TAL21-TAL14.

where the prime variables represent the variables without batch effects:

$$\begin{aligned} I'_i &= I_i \cdot c_I & \beta' &= \beta \cdot c_O \cdot c_P & P'_i &= P_i \cdot c_P & \alpha'_i &= \alpha_i \cdot c_I \\ O'_i &= O_i \cdot c_O \cdot c_P & d' &= d \cdot c_I & P'_{i'} &= P_{i'} \cdot c_P. \end{aligned}$$

As is shown in Supporting Information Section 12 of [Davidsohn et al., 2015], scaling factors of the transfection marker for the cascades, $\tilde{c}_P$ are $\{1.51, 1.07, 0.68, 0.78, 0.71, 0.79\}$ for TAL14-TAL21, TAL14-LmrA, TAL21-TAL14, TAL21-LmrA, LmrA-TAL14, and LmrA-TAL21, respectively. For all these cascades, $\tilde{c}_I = 1$, and $\tilde{c}_O = 1$. Since the prime variables involve no batch effects, to convert to a cascade, we must divide all the prime variables by the corresponding cascade scaling factors ($\tilde{c}_I$, $\tilde{c}_O$, $\tilde{c}_P$). In addition, to offer a comparable study to [Davidsohn et al., 2015], we follow similar implementation details as are shown in [Davidsohn et al., 2015] by multiplying the dissociation constant of the the second repressor by three (see the fourth to last paragraph of the Supporting Information Section 5 of [Davidsohn et al., 2015]). This is because the plasmids for the second repressor are transfected at one-third the concentration of the first repressor [Davidsohn et al., 2015]. This suggests that production of the second repressor should scale like one-third the activation level of the first repressor during the initial transient, when much of the repressor is produced for the system [Davidsohn et al., 2015]. The final bin-dependent model for cascades is expressed as:

$$\begin{aligned}
\frac{dI''_i}{dt} &= \alpha''_i \cdot \phi(t) - \lambda \cdot I''_i \\
\frac{dO''_{1i}}{dt} &= \begin{cases} \beta''_1 \cdot \phi(t) \cdot \left(\dfrac{P''_{1i}}{P''_{11}}\right)^{f_1} \cdot \left(\dfrac{1-\gamma_1}{1+\left(\frac{I''_i}{d''_1}\right)^{h_1}} + \gamma_1\right) - \lambda \cdot O''_{1i}, & \text{if } P''_{1i} < P''_{1i'} \\ \beta''_1 \cdot \phi(t) \cdot \left(\dfrac{P''_{1i'}}{P''_{11}}\right)^{f_1} \cdot \left(\dfrac{P''_{1i}}{P''_{1i'}}\right)^{g_1} \cdot \dfrac{1-\gamma_1}{1+\left(\frac{I''_i}{d''_1}\right)^{h_1}} & \text{if } P''_{1i} \geq P''_{1i'} \\ \quad + \beta''_1 \cdot \phi(t) \cdot \left(\dfrac{P''_{1i}}{P''_{11}}\right)^{f_1} \cdot \gamma_1 - \lambda \cdot O''_{1i}, & \end{cases} \\
\frac{dO''_{2i}}{dt} &= \begin{cases} \beta''_2 \cdot \phi(t) \cdot \left(\dfrac{P''_{2i}}{P''_{21}}\right)^{f_2} \cdot \left(\dfrac{1-\gamma_2}{1+\left(\frac{O''_{1i}}{3 \cdot d''_2}\right)^{h_2}} + \gamma_2\right) - \lambda \cdot O''_{2i}, & \text{if } P''_{2i} < P''_{2i'} \\ \beta''_2 \cdot \phi(t) \cdot \left(\dfrac{P''_{2i'}}{P''_{21}}\right)^{f_2} \cdot \left(\dfrac{P''_{2i}}{P''_{2i'}}\right)^{g_2} \cdot \dfrac{1-\gamma_2}{1+\left(\frac{O''_{1i}}{3 \cdot d''_2}\right)^{h_2}} & \text{if } P''_{2i} \geq P''_{2i'} \\ \quad + \beta''_2 \cdot \phi(t) \cdot \left(\dfrac{P''_{2i}}{P''_{21}}\right)^{f_2} \cdot \gamma_2 - \lambda \cdot O''_{2i}, & \end{cases} \\
\phi(t) &= \left(\frac{1}{2}\right)^{\lfloor \frac{t}{T} \rfloor},
\end{aligned} \tag{3.4}$$

where

$$I''_i = \frac{I'_i}{\tilde{c}_I} \quad \alpha''_i = \frac{\alpha'_i}{\tilde{c}_I},$$

and for the $k$-th module ($k = 1, 2$) and the $j$-th cascade ($j = 1 - 6$),

$$\begin{aligned}
\beta''_k = \frac{\beta'_k}{\tilde{c}_{Pj} \cdot \tilde{c}_O} \quad P''_{ki} = \frac{P'_{ki}}{\tilde{c}_{Pj}} \quad d''_k = \frac{d'_k}{\tilde{c}_I} \\
O''_{ki} = \frac{O'_{ki}}{\tilde{c}_{Pj} \cdot \tilde{c}_O} \quad P''_{ki'} = \frac{P'_{ki'}}{\tilde{c}_{Pj}}.
\end{aligned}$$

The double prime variables represent variables that account for the batch effects of the cascades. Values of the parameters used in the final bin-dependent models for six cascades are shown in Table 3.1.

| Parameters | | | | |
|---|---|---|---|---|
| Cascade | $\beta_1''$ (MEFL/hr) | $f_1$ | $d_1''$ (MEFL) | $h_1$ |
| LmrA-TAL14 | $9.78 \times 10^3$ | 1.91 | $3.73 \times 10^5$ | 0.59 |
| LmrA-TAL21 | $8.67 \times 10^3$ | 1.91 | $3.73 \times 10^5$ | 0.59 |
| TAL14-LmrA | $4.70 \times 10^4$ | 1.74 | $1.58 \times 10^4$ | 0.68 |
| TAL14-TAL21 | $3.34 \times 10^4$ | 1.74 | $1.58 \times 10^4$ | 0.68 |
| TAL21-LmrA | $5.34 \times 10^4$ | 1.58 | $5.77 \times 10^4$ | 0.72 |
| TAL21-TAL14 | $6.20 \times 10^4$ | 1.58 | $5.77 \times 10^4$ | 0.72 |
| | $\gamma_1$ | $g_1$ | $\beta_2''$ (MEFL/hr) | $f_2$ |
| LmrA-TAL14 | $2.36 \times 10^{-5}$ | 1.09 | $7.24 \times 10^4$ | 1.74 |
| LmrA-TAL21 | $2.36 \times 10^{-5}$ | 1.09 | $5.28 \times 10^4$ | 1.58 |
| TAL14-LmrA | $2.83 \times 10^{-4}$ | 1.10 | $6.35 \times 10^3$ | 1.91 |
| TAL14-TAL2 | $2.83 \times 10^{-4}$ | 1.10 | $2.75 \times 10^4$ | 1.58 |
| TAL21-LmrA | $1.10 \times 10^{-3}$ | 0.83 | $8.76 \times 10^3$ | 1.91 |
| TAL21-TAL14 | $1.10 \times 10^{-3}$ | 0.83 | $7.54 \times 10^4$ | 1.74 |
| | $d_2''$ (MEFL) | $h_2$ | $\gamma_2$ | $g_2$ (MEFL/hr) |
| LmrA-TAL14 | $1.58 \times 10^4$ | 0.68 | $2.83 \times 10^{-4}$ | 1.10 |
| LmrA-TAL21 | $5.77 \times 10^4$ | 0.72 | $1.10 \times 10^{-3}$ | 0.83 |
| TAL14-LmrA | $3.73 \times 10^5$ | 0.59 | $2.36 \times 10^{-5}$ | 1.09 |
| TAL14-TAL21 | $5.77 \times 10^4$ | 0.72 | $1.10 \times 10^{-3}$ | 0.83 |
| TAL21-LmrA | $3.73 \times 10^5$ | 0.59 | $2.36 \times 10^{-5}$ | 1.09 |
| TAL21-TAL14 | $1.58 \times 10^4$ | 0.68 | $2.83 \times 10^{-4}$ | 1.10 |
| | $\lambda$ (hr$^{-1}$) | $P_{1i'}''$ (MEFL) | $P_{2i'}''$ (MEFL) | |
| LmrA-TAL14 | $3.41 \times 10^{-2}$ | $10^{7.55}$ | $10^{7.31}$ | |
| LmrA-TAL21 | $3.41 \times 10^{-2}$ | $10^{7.51}$ | $10^{7.15}$ | |
| TAL14-LmrA | $3.41 \times 10^{-2}$ | $10^{7.12}$ | $10^{7.37}$ | |
| TAL14-TAL21 | $3.41 \times 10^{-2}$ | $10^{6.97}$ | $10^{6.87}$ | |
| TAL21-LmrA | $3.41 \times 10^{-2}$ | $10^{7.16}$ | $10^{7.51}$ | |
| TAL21-TAL14 | $3.41 \times 10^{-2}$ | $10^{7.22}$ | $10^{7.32}$ | |

**Table 3.1:** Values of the rescaled parameters used in the final bin-dependent models for the six cascades.

## 3.4 Model Summary

We have developed a bin-dependent ODE model that describes regulatory mechanisms via the use of standard Hill function type terms, while offering comparable accuracy to the EQuiP model of [Davidsohn et al., 2015]. Parametrized, bin-dependent models of individual modules should be relatively straightforward to integrate as subcomponents within larger existing ODE models of circuits. Moreover, it should also be relatively straightforward to modify a parametrized bin-dependent model to incorporate *additional*, previously-characterized regulatory components (i.e. for studying promoters co-regulated by multiple transcription factors). In this way we expect that bin-dependent models for individual modules should be able to be composed with a variety of existing, well-characterized ODE models that describe components of synthetic and systems biology networks.

Another benefit to the bin-dependent-model-based approach is that it is fairly robust to sampling noise in experimental data. The input-output datasets, which the ODE models are fit to, comprise the geometric means of measured protein concentrations within each bin. These data points may not be well separated, and hence appear noisy, when using sparse flow cytometry datasets. The model fitting step helps overcome this sampling noise by using deterministic ODEs based on widely-used biochemical relationships (such as Hill-functions).

The bin-dependent model presented here establishes a framework for characterizing fundamental synthetic constructs and predicting circuit behaviors quantitatively in TTMC. As we demonstrated with the stochastic model, there are different mechanisms that may contribute to saturation in protein production, a common phenomenon in TTMC. The value of the bin-dependent model lies in both its easy integrability with other ODE models, and in its ability to describe the saturation effect in flow cytometry data accurately without specifying precise mechanistic details for how saturation occurs. The method presented here should be applicable to similar flow cytometry datasets, allowing the possibility to construct a well-characterized library of in silico models for regulatory switches. The quantitative parameters of such regulatory switches could then be used in constructing new predictive models for the behaviors of more complicated circuits and cascades. Our work represents one more step towards building a systematic workflow that can guide circuit design in TTMC.

# Chapter 4

# Modular Composition, Circuit Behaviors, and Network Topologies

The work in the previous two chapters results in an ODE modeling framework for predicting circuit behaviors. Modularity, a critical assumption of the framework, is violated when the behavior of one component depends on other components of the circuit. The resulting phenomenon, known as retroactivity, may impact the behavior of genetic circuits, constituting another potential source of inaccuracy of model predictions. In this chapter, I present an investigation of how retroactivity affects robustness of circuit behaviors. Specifically, I focus on adaptation, a biological function concerning the temporal dynamics of gene expression. We develop a systematic approach for quantifying adaptive robustness via statistical SMC and use this framework to examine the relationship between circuit topologies and adaptations, followed by the effects of retroactivity on adaptive robustness. Note that findings from this study are not confined to circuits in synthetic biology but can be applied to general transcriptional regulatory networks (TRN).

## 4.1 Adaptation

Adaptation consists of a response phase, where the expression level of a gene responds transiently to an external stimulus, and a recovery phase, where the expression level adapts gradually to the initial value (Figure 4·1) [Alon, 2007, Shi et al., 2017]. Examples of adaptation include signal transduction [Behar et al., 2007, Cohen

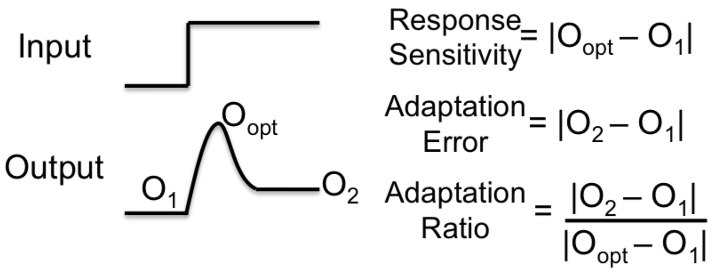
Input
Output
$O_1$
$O_{opt}$
$O_2$
Response Sensitivity = $|O_{opt} - O_1|$
Adaptation Error = $|O_2 - O_1|$
Adaptation Ratio = $\frac{|O_2 - O_1|}{|O_{opt} - O_1|}$

**Figure 4·1:** Response sensitivity, adaptation errors, and adaptation ratios.

et al., 2009, Takeda et al., 2012, Muzzey et al., 2009], bacteria chemotaxis [Alon et al., 1999, Barkai and Leibler, 1997, Macnab and Koshland, 1972], and homeostasis [El-Samad et al., 2002]. It is well known that TRN with certain topologies such as IFFL and negative feedback loops (NFBL) can mediate adaptations robustly.

## 4.2 Mathematical Models and Modular Composition

### 4.2.1 Transcriptional Regulatory Networks

A TRN consisting of N genes can be represented by an N-node graph-like object, where each node is a gene (please refer to Figure 4·2 for examples). There is a directed edge from node i to j if gene i regulates the expression of gene j. In this case, node i is also known as the parent of node j. The time evolution of a TRN is defined as a sequence of concentrations of proteins in the TRN, also known as a trajectory.

In this chapter, we limit ourselves to TRN that contain three nodes. As is shown in several studies, a three-node TRN is a minimum network that facilitates adaptations, and a larger TRN can typically be reduced to a three-node TRN [Ma et al., 2009, Shi et al., 2017]. We denote the three nodes by A, B, and C (Figure 4·2). A is the input node activated by an external inducer. B is the node that transmits the signal from A to C. C is the output node, which is also the node of interest. Examples of three-node TRN are given in Figure 4·2. Focusing on three-node TRN allows us to conduct an

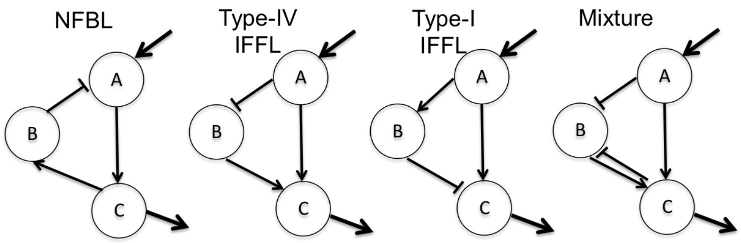

**Figure 4·2:** Examples of three-node TRN. Arrows indicate activation, and bars at the ends of edges indicate inhibition. The leftmost network is a negative feedback loop (NFBL), in which the edges traversing B, A, and C accumulate in a negative regulation. The two networks in the center are incoherent feedforward loops (IFFL), in which A directly activates C and indirectly represses C via B. The second from the left is a type-IV IFFL, in which A represses B. The second from the right is a type-I IFFL, in which A activates B. The rightmost network is a "mixture" network composed of a type-IV IFFL and an NFBL between B and C.

exhaustive search of all possible network topologies.

### 4.2.2 Mathematical Models with and without Retroactivity

In a TRN, the time evolution of any node $i$ that is regulated by other node(s) and/or an external inducer can be described by the following ODE:

$$\dot{x_i} = f_i(x_i, \vec{y_i}), \tag{4.1}$$

where $x_i$ and $\vec{y_i}$ represent the concentrations of node $i$ and the parent(s) of node $i$, respectively. $\vec{y_i}$ includes the concentration of the external inducer if node i is regulated by an external inducer. $f_i$ is expressed as:

$$f_i(x_i, \vec{y_i}) = H_i(\vec{y_i}) - \delta_i x_i, \tag{4.2}$$

where $\delta_i$ denotes the protein degradation rate. $H_i(\vec{y_i})$ is the Hill function that describes the regulated production rate of $x_i$. In this paper, we consider an AND logic for coregulation by multiple TF, i.e., the regulated gene is turned on only when all the activators are abundant and all the repressors are scarce. Under the assumption of the AND logic, $H_i(\vec{y_i})$ can be expressed as [Gyorgy and Del Vecchio, 2014]:

$$H_i(\vec{y_i}) = \eta_i \frac{\sum_{X \subset \{1,2,\ldots,m_i\}} \pi_X \prod_{j \in X} \left(\frac{y_{ij}}{K_{ij}}\right)^{h_{ij}}}{\sum_{X \subset \{1,2,\ldots,m_i\}} \prod_{j \in X} \left(\frac{y_{ij}}{K_{ij}}\right)^{h_{ij}}}, \tag{4.3}$$

where $\eta_i$ stands for the total concentration of the promoter that expresses node $i$, and $m_i$ is the number of parents of node $i$. Similar to [Gyorgy and Del Vecchio, 2014], we assume that no parents of the same node are identical. $X$ corresponds to each complex formed by a different combination of TF; $\pi_X$ denotes the production rate of the corresponding complex per plasmid; $y_{ij}$, $h_{ij}$, and $K_{ij}$ represent the concentration, the Hill coefficient, and the dissociation constant of the $j$-th parent of node $i$. An example of a Hill-function with the AND logic is given in subchapter 4.2.4.

As a TRN is a collection of regulatory interactions among genes, the dynamics of a TRN can be described by:

$$\dot{\vec{x}} = f(\vec{x}), \tag{4.4}$$

where $\vec{x} = [x_1 \ x_2 \ \ldots \ x_N \ u]^T$, $u$ is the concentration of the external inducer, and $f$ is the collection of functions $f_i$ $(i = 1, 2, \ldots, N)$. We assume the inducer does not get produced or degraded, so the concentration of the inducer stays constant, i.e., $\dot{u} = 0$.

With retroactivity considered, the equations for the dynamics of a TRN change from (4.4) to [Gyorgy and Del Vecchio, 2014]:

$$\dot{\vec{x}} = [I + R(\vec{x})]^{-1} f(\vec{x}), \tag{4.5}$$

where $R(\vec{x})$ is known as the retroactivity matrix [Gyorgy and Del Vecchio, 2014].

$R(\vec{x})$ can be calculated via the following equation [Gyorgy and Del Vecchio, 2014]:

$$R(\vec{x}) = \sum_i V_i^T R_i(\vec{y_i}) V_i, \tag{4.6}$$

where $V_i$ is binary, containing as many rows as the length of $\vec{y_i}$ and as many columns as the number of nodes in the network. The element in the $j$-th row and $k$-th column of $V_i$ is 1 if the $j$-th parent of node $i$ is node $k$, 0 otherwise. Under the assumption of the AND logic, $R_i(\vec{y_i})$ is a diagonal matrix, where the $k$-th entry on the diagonal $r_{ik}$ is [Gyorgy and Del Vecchio, 2014]:

$$r_{ik} = \eta_i \frac{h_{ik}^2 y_{ik}^{h_{ik}-1}}{K_{ik}^{h_{ik}}} \left(1 + \left(\frac{y_{ik}}{K_{ik}}\right)^{h_{ik}}\right)^{-2}. \tag{4.7}$$

In Equation (4.7), $\eta_i$ stands for the total DNA concentration of node $i$. $y_{ik}$, $h_{ik}$, and $K_{ik}$ are the protein concentration, the Hill coefficient, and the dissociation coefficient of the $k$-th parent of node $i$. It is easy to show that $V_i^T R_i(\vec{y_i}) V_i$ is always a diagonal matrix. Hence, $R(\vec{x})$ is also diagonal. More details about retroactivity, including its derivation can be found in [Gyorgy and Del Vecchio, 2014].

### 4.2.3 Network Enumeration and Simulation

Similar to [Shi et al., 2017], we first enumerate all possible topologies of three-node TRN. Each node in the network may interact with up to three nodes (two other nodes and itself). One node may activate, inhibit, or simply not regulate another node. There are altogether $3^9 = 19,683$ possible topologies, 3,645 of which have no direct or indirect links between the input A and the output C. With these 3,645 topologies excluded, we consider the remaining $16,038$ topologies in our study [Shi et al., 2017].

To investigate the effects of retroactivity on adaptive robustness, we construct and compare ODE models with and without retroactivity for each topology. To reduce the dimensions of parameter space, we normalize our models via methods shown in [Cao et al., 2016]. An example of normalization is given in subchapter 4.2.4. The normalized protein concentrations of A, B, and C, denoted by $\tilde{x}_A$, $\tilde{x}_B$, and $\tilde{x}_C$, are dimensionless and between values of 0 and 1. For simplicity of analysis, we assume all normalized DNA concentrations have equal values denoted by $\tilde{\eta}$. For each enumerated topology, one ODE model without retroactivity is constructed, together with three models with retroactivity assuming $\tilde{\eta} = 0.1$, $\tilde{\eta} = 1$, and $\tilde{\eta} = 10$, corresponding to systems with low, medium, and high retroactivity. This is because as $\tilde{\eta}$ increases, the diagonal entries of the retroactivity matrices increase, giving rise to higher retroactivity (Equation (4.7)).

Trajectories are generated via the integration of ODE models. The initial states of the trajectories are set to the steady states of the networks prior to the induction. The kinetic parameters are sampled uniformly from the same ranges of values used in [Cao et al., 2016]: $K \sim 0.001 - 1$ (sampled on the log scale), $h \sim 1 - 4$ (sampled on the linear scale), and $\delta \sim 0.01 - 1$ (sampled on the log scale). At $t_0$, the network is induced by such a large concentration of the external inducer I that the expression of A is fully driven. The concentration of the inducer I is much larger than the binding affinity $K_{IA}$. For convenience, we set the concentration of I, $x_I$, equal to 10.

### 4.2.4 Models for a type-IV IFFL

The topology of a type-IV IFFL is given in Figure 4·2. We can describe the dynamics of a type-IV IFFL without retroactivity via the following model:

$$\begin{aligned}
\frac{dx_A}{dt} &= f_A = \eta_A \frac{\pi_A \left(\frac{x_I}{K_{IA}}\right)^{h_{IA}}}{1+\left(\frac{x_I}{K_{IA}}\right)^{h_{IA}}} - \delta_A x_A \\
\frac{dx_B}{dt} &= f_B = \eta_B \frac{\pi_B}{1+\left(\frac{x_A}{K_{AB}}\right)^{h_{AB}}} - \delta_B x_B \\
\frac{dx_C}{dt} &= f_C = \eta_C \frac{\pi_C \left(\frac{x_A}{K_{AC}}\right)^{h_{AC}} \left(\frac{x_B}{K_{BC}}\right)^{h_{BC}}}{\left(1+\left(\frac{x_A}{K_{AC}}\right)^{h_{AC}}\right)\left(1+\left(\frac{x_B}{K_{BC}}\right)^{h_{BC}}\right)} \\
&\quad - \delta_C x_C.
\end{aligned} \tag{4.8}$$

With retroactivity, the dynamics can be described by:

$$\begin{bmatrix} \frac{dx_A}{dt} \\ \frac{dx_B}{dt} \\ \frac{dx_C}{dt} \end{bmatrix} = \begin{bmatrix} \frac{1}{1+b+a} & 0 & 0 \\ 0 & \frac{1}{1+c} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f_A \\ f_B \\ f_C \end{bmatrix}, \tag{4.9}$$

where

$$\begin{aligned}
a &= \eta_B \frac{h_{AB}^2 {x_A}^{h_{AB}-1}}{K_{AB}^{h_{AB}}} \left(1+\left(\frac{x_A}{K_{AB}}\right)^{h_{AB}}\right)^{-2} \\
b &= \eta_C \frac{h_{AC}^2 {x_A}^{h_{AC}-1}}{K_{AC}^{h_{AC}}} \left(1+\left(\frac{x_A}{K_{AC}}\right)^{h_{AC}}\right)^{-2} \\
c &= \eta_C \frac{h_{BC}^2 {x_B}^{h_{BC}-1}}{K_{BC}^{h_{BC}}} \left(1+\left(\frac{x_B}{K_{BC}}\right)^{h_{BC}}\right)^{-2}.
\end{aligned} \tag{4.10}$$

Following methods in [Cao et al., 2016], we let $\tilde{x}_A = \frac{x_A \delta_A}{\eta_A \pi_A}$, $\tilde{x}_B = \frac{x_B \delta_B}{\eta_B \pi_B}$, $\tilde{x}_C = \frac{x_C \delta_C}{\eta_C \pi_C}$, $\tilde{K}_{AB} = \frac{K_{AB} \delta_A}{\eta_A \pi_A}$, $\tilde{K}_{AC} = \frac{K_{AC} \delta_A}{\eta_A \pi_A}$, and $\tilde{K}_{BC} = \frac{K_{BC} \delta_B}{\eta_B \pi_B}$. The model without retroactivity shown in Equation (4.8) is normalized to:

$$\begin{aligned}
\frac{d\tilde{x}_A}{dt} &= f_{\tilde{A}} = \delta_A \left( \frac{\left(\frac{x_I}{K_{IA}}\right)^{h_{IA}}}{1+\left(\frac{x_I}{K_{IA}}\right)^{h_{IA}}} - \tilde{x}_A \right) \\
\frac{d\tilde{x}_B}{dt} &= f_{\tilde{B}} = \delta_B \left( \frac{1}{1+\left(\frac{\tilde{x}_A}{\tilde{K}_{AB}}\right)^{h_{AB}}} - \tilde{x}_B \right) \\
\frac{d\tilde{x}_C}{dt} &= f_{\tilde{C}} = \delta_C \left( \frac{\left(\frac{\tilde{x}_A}{\tilde{K}_{AC}}\right)^{h_{AC}} \left(\frac{\tilde{x}_B}{\tilde{K}_{BC}}\right)^{h_{BC}}}{\left(1+\left(\frac{\tilde{x}_A}{\tilde{K}_{AC}}\right)^{h_{AC}}\right)\left(1+\left(\frac{\tilde{x}_B}{\tilde{K}_{BC}}\right)^{h_{BC}}\right)} \right. \\
&\qquad \left. -\tilde{x}_C \right).
\end{aligned} \tag{4.11}$$

By letting $\tilde{\eta}_{AB} = \frac{\eta_B}{K_{AB}}$, $\tilde{\eta}_{AC} = \frac{\eta_C}{K_{AC}}$, and $\tilde{\eta}_{BC} = \frac{\eta_C}{K_{BC}}$, we normalize the model with retroactivity shown in Equation (4.9) to:

$$\begin{bmatrix} \frac{d\tilde{x}_A}{dt} \\ \frac{d\tilde{x}_B}{dt} \\ \frac{d\tilde{x}_C}{dt} \end{bmatrix} = \begin{bmatrix} \frac{1}{1+\tilde{b}+\tilde{a}} & 0 & 0 \\ 0 & \frac{1}{1+\tilde{c}} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f_{\tilde{A}} \\ f_{\tilde{B}} \\ f_{\tilde{C}} \end{bmatrix}, \tag{4.12}$$

where

$$\begin{aligned}
\tilde{a} &= \tilde{\eta}_{AB} h_{AB}^2 \left(\frac{\tilde{x}_A}{\tilde{K}_{AB}}\right)^{h_{AB}-1} \left(1+\left(\frac{\tilde{x}_A}{\tilde{K}_{AB}}\right)^{h_{AB}}\right)^{-2} \\
\tilde{b} &= \tilde{\eta}_{AC} h_{AC}^2 \left(\frac{\tilde{x}_A}{\tilde{K}_{AC}}\right)^{h_{AC}-1} \left(1+\left(\frac{\tilde{x}_A}{\tilde{K}_{AC}}\right)^{h_{AC}}\right)^{-2} \\
\tilde{c} &= \tilde{\eta}_{BC} h_{BC}^2 \left(\frac{\tilde{x}_B}{\tilde{K}_{BC}}\right)^{h_{BC}-1} \left(1+\left(\frac{\tilde{x}_B}{\tilde{K}_{BC}}\right)^{h_{BC}}\right)^{-2}.
\end{aligned} \tag{4.13}$$

Based on our assumption of equal normalized DNA concentrations, $\tilde{\eta}_{AB} = \tilde{\eta}_{AC} = \tilde{\eta}_{BC} = \tilde{\eta}$.

## 4.3 Adaptive Robustness

### 4.3.1 Adaptation

Adaptation is a property concerning the time evolution of a network. The quality of adaptation is quantified via the adaptation ratio, defined as the ratio of the adaptation error to the response sensitivity (Figure 4·1). Response sensitivity is defined as the difference between the output response and the initial value. An adaptation error is defined as the difference between the initial value and the steady-state value post the induction. A lower adaptation ratio indicates a higher quality of adaptation.

### 4.3.2 Specification and Adaptive Robustness

We use Bounded Linear Temporal Logic (BLTL) formulas over linear inequalities over concentrations of proteins to specify the functions of a network. A BLTL formula is built on a finite set of predicates over protein concentrations using Boolean operators: $\neg$ (negation), $\vee$ (disjunction), $\wedge$ (conjunction), $\Rightarrow$ (implication) and a temporal operator $\cup^k$ (until) with bound $k$ [Zuliani et al., 2013]. More details about the syntax and the semantics of BLTL can be found in [Zuliani et al., 2013]. An example of a BLTL specification looks like:

$$\Phi_E = \big(x < 20 \cup^5 x = 20\big)\,, \tag{4.14}$$

where $x$ is the concentration of a protein. $\Phi_E$ means that the concentration of protein $x$ should reach 20 within five time units and remain less than 20 at all preceding time units. Satisfaction of $\Phi_E$ by a trajectory $\sigma$ is written as $\sigma \models \Phi_E$.

Assume the trajectory spans a time period of $T$. Let $\vec{x}$ denote the vector of protein concentrations of A, B, and C. The property of adaptation can be formally stated as:

$$\begin{gathered}\Phi = \left(\left(\dot{x}_C \geq 0 \cup^T \left(\dot{x}_C < 0 \cup^T \left(\dot{\vec{x}} = \vec{0}\right)\right)\right) \vee \right.\\ \left.\left(\dot{x}_C \leq 0 \cup^T \left(\dot{x}_C > 0 \cup^T \left(\dot{\vec{x}} = \vec{0}\right)\right)\right)\right) \wedge (r < r^*),\end{gathered} \tag{4.15}$$

where $\dot{x}_C$ represents the rate of change of concentration of C, and $\dot{\vec{x}}$, the vector that contains the rates of changes of all species' concentrations. $r$ is the adaptation ratio, and $r^*$ is the threshold on the adaptation ratio.

Equation (4.15) can be separated into two parts: everything before $r < r^*$ and $r < r^*$. The former requires that the concentration of C first respond to the initial stimuli either by rising or falling and then switch to recovery before the system eventually reaches the steady state. The latter imposes a restriction on the adaptation ratio, excluding trajectories that have weak pulses or do not return to values that are close to the initial states.

We use Probabilistic BLTL (PBLTL) to specify the adaptive robustness of a TRN. The biochemical kinetic rates of the TRN are allowed to vary. A PBLTL formula that describes the ability of a network to achieve adaptations is expressed in the form $P_{\geq\theta}(\Phi)$, where $\Phi$ is the BLTL formula described in Equation (4.15), and $\theta$ is a probability. A network satisfies the PBLTL formula if and only if a trajectory of the TRN satisfies the BLTL formula $\Phi$ in Equation (4.15) with a probability greater than or equal to $\theta$. We call $\theta$ the adaptive robustness of this network. A network is more robust than another if the former satisfies $\Phi$ with a higher probability than the latter. The problem we consider is as follows: given a TRN and a BLTL formula $\Phi$, compute the adaptive robustness of the TRN by calculating the probability $\theta$ with which the TRN satisfies the BLTL formula $\Phi$.

Solving the above problem establishes a standard criterion for comparing different networks and models. It provides us a framework for examining the effects of retroactivity on adaptive robustness of TRN.

## 4.4 Statistical Analysis

### 4.4.1 Investigation of Behavior-Topology Relationships

Zuliani et al. presented an algorithm for estimating the Bayesian interval that contains the true probability of adaptation with an arbitrarily high probability [Zuliani et al., 2013]. The algorithm samples trajectories from a stochastic system iteratively and checks each trajectory against the specification. At each stage, the posterior mean, which is the Bayes estimator for the probability, is updated. The algorithm terminates and returns the probability estimate upon achieving the coverage goal. The estimate is in the form of a Bayesian confidence interval. Otherwise, the algorithm continues by sampling another trajectory.

We use the above algorithm to estimate the probability that a random execution trace of the TRN satisfies the property of adaptation specified by a BLTL formula. Due to practical concerns of simulation times, the BLTL formula we implement differs slightly from Equation (4.15) and is specified as follows:

$$\begin{aligned}\Phi = &\left(\left(\dot{\tilde{x}}_C \geq 0 \cup^{2000} \left(\dot{\tilde{x}}_C < 0 \cup^{2000} \left(||\dot{\vec{\tilde{x}}}||_\infty \leq 10^{-3}\right)\right)\right) \vee \right.\\ &\left.\left(\dot{\tilde{x}}_C \leq 0 \cup^{2000} \left(\dot{\tilde{x}}_C > 0 \cup^{2000} \left(||\dot{\vec{\tilde{x}}}||_\infty \leq 10^{-3}\right)\right)\right)\right) \wedge \\ &\left(r < 0.1\right).\end{aligned} \tag{4.16}$$

In Equation (4.16), $\dot{\tilde{x}}_C$ represents the rate of change of the normalized concentration of C, and $||\dot{\vec{\tilde{x}}}||_\infty$ the infinity norm of the vector containing the rates of changes of all normalized species' concentrations. The maximum simulation duration is set to 2000, and the threshold on the adaptation ratio is set to 0.1 (we experiment with other cut-off values and arrive at the same conclusions). Equation (4.16) excludes networks that spend too much time approaching steady states or have oscillatory behaviors.

For the algorithm, a beta prior with $\alpha = \beta = 1$ is used. For the algorithm parameters, half interval size $\delta$ is set to 0.01, and coverage goal $c$ is set to 0.99. Explanations of the algorithm parameters can be found in [Zuliani et al., 2013]. The probability estimate the algorithm returns, $\hat{\theta}$, is the adaptive robustness of the network.

### 4.4.2 Networks and Adaptation

Our simulations above suggest that the "mixture" network shown in Figure 4·2 has the highest adaptive robustness among all the models without retroactivity, 0.6329. This value means that the unknown probability $\theta$ that the "mixture" network satisfies the property of adaptation lies in $[0.6329 - 0.01, 0.6329 + 0.01]$ with probability $1 - \frac{(1-0.99)\times 0.02}{0.99\times 0.98}$ if retroactivity is not considered.

Here we consider a network model to be adaptive if its adaptive robustness exceeds 0.1. There are 148 adaptive models without retroactivity, 149 with retroactivity assuming $\tilde{\eta} = 0.1$, 150 with retroactivity assuming $\tilde{\eta} = 1$, and 131 with retroactivity assuming $\tilde{\eta} = 10$ (Table 4·3). 106 out of these adaptive models share the same topologies, which suggests that most adaptive networks remain adaptive even when retroactivity is considered. Nevertheless, there is little doubt retroactivity affects the adaptive robustness of the network. The adaptive robustness of these 106 networks with and without retroactivity is compared in Figure 4·3. As is suggested by Figure 4·3, higher retroactivity in general brings about stronger effects on adaptive robustness, as red pluses ($\tilde{\eta} = 0.1$) are on average more distant from the reference line than blue diamonds ($\tilde{\eta} = 1$), and black circles ($\tilde{\eta} = 10$) more distant than red pluses.

### 4.4.3 Effects of Retroactivity on Adaptive Robustness

The data points in Figure 4·3 are scattered on both sides of the reference line, indicating that increasing retroactivity can either enhance or reduce adaptive robustness depending on the circuit topologies. To investigate the mixed effects of retroactivity

**Table 4.1:** Numbers of NFBL, IFFL, and adaptive networks. Models I, II, III, and IV represent the model without retroactivity, the models with retroactivity assuming $\tilde{\eta} = 0.1$, $\tilde{\eta} = 1$, and $\tilde{\eta} = 10$.

| Model | $\tilde{\eta}$ | NFBL | IFFL | Adaptive Networks |
|---|---|---|---|---|
| I | N/A | 92 | 148 | 148 |
| II | 0.1 | 92 | 147 | 149 |
| III | 1 | 94 | 148 | 150 |
| IV | 10 | 84 | 118 | 131 |

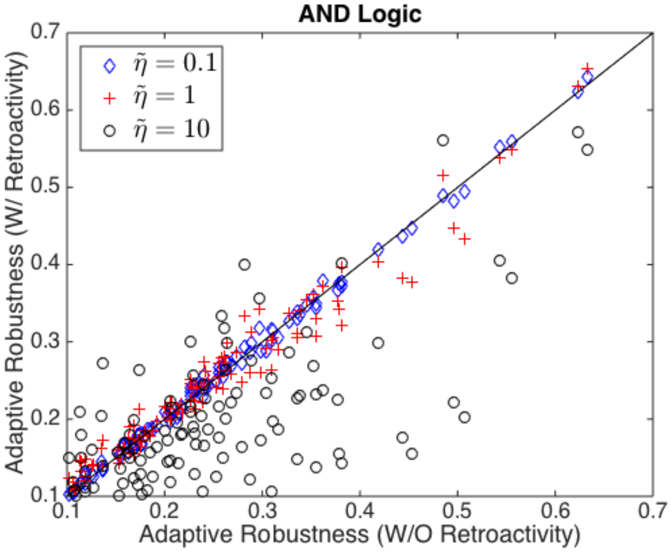

**Figure 4·3:** Adaptive robustness of adaptive networks. X-axis represents models without retroactivity, and Y-axis, the counterparts with retroactivity. Blue diamonds represent models assuming $\tilde{\eta} = 0.1$, red pluses, models assuming $\tilde{\eta} = 1$, and black circles, models assuming $\tilde{\eta} = 10$. The black line corresponds to equal adaptive robustness.

on adaptive robustness, we perform a single parameter perturbation analysis. Specifically, we randomly select 100 parameters that facilitate adaptations in models without retroactivity. Keeping these parameters fixed, we set $\tilde{\eta}$ to 0.1, 1, and 10 and simulate the counterpart models with retroactivity. The perturbation analysis is performed on the type-IV IFFL, the type-I IFFL, and the "mixture" network shown in Figure 4·2. These networks are chosen because they have the overall highest adaptive robustness among all networks, with and without retroactivity. The average response sensitivity and adaptation errors are calculated for each model (Table 4.2). Formulas of the type-IV IFFL models are given in Equations (4.11) and (4.12), and formulas of the type-I IFFL models can be similarly derived. The model for the "mixture" network without retroactivity is:

$$\begin{aligned}
\frac{d\tilde{x}_A}{dt} &= f_{\tilde{A}} = \delta_A \left( \frac{\left(\frac{x_I}{K_{IA}}\right)^{h_{IA}}}{1+\left(\frac{x_I}{K_{IA}}\right)^{h_{IA}}} - \tilde{x}_A \right) \\
\frac{d\tilde{x}_B}{dt} &= f_{\tilde{B}} = \delta_B \left( \frac{1}{\left(1+\left(\frac{\tilde{x}_A}{\tilde{K}_{AB}}\right)^{h_{AB}}\right)\left(1+\left(\frac{\tilde{x}_C}{\tilde{K}_{CB}}\right)^{h_{CB}}\right)} \right. \\
&\quad \left. - \tilde{x}_B \right) \\
\frac{d\tilde{x}_C}{dt} &= f_{\tilde{C}} = \delta_C \left( \frac{\left(\frac{\tilde{x}_A}{\tilde{K}_{AC}}\right)^{h_{AC}}\left(\frac{\tilde{x}_B}{\tilde{K}_{BC}}\right)^{h_{BC}}}{\left(1+\left(\frac{\tilde{x}_A}{\tilde{K}_{AC}}\right)^{h_{AC}}\right)\left(1+\left(\frac{\tilde{x}_B}{\tilde{K}_{BC}}\right)^{h_{BC}}\right)} \right. \\
&\quad \left. - \tilde{x}_C \right).
\end{aligned} \tag{4.17}$$

Its counterpart with retroactivity is:

$$\begin{bmatrix} \frac{d\tilde{x}_A}{dt} \\ \frac{d\tilde{x}_B}{dt} \\ \frac{d\tilde{x}_C}{dt} \end{bmatrix} = \begin{bmatrix} \frac{1}{1+\tilde{b}+\tilde{a}} & 0 & 0 \\ 0 & \frac{1}{1+\tilde{c}} & 0 \\ 0 & 0 & \frac{1}{1+\tilde{d}} \end{bmatrix} \begin{bmatrix} f_{\tilde{A}} \\ f_{\tilde{B}} \\ f_{\tilde{C}} \end{bmatrix}, \tag{4.18}$$

where $f_{\tilde{A}}$, $f_{\tilde{B}}$, and $f_{\tilde{C}}$ are defined as in Equation (4.17), and $\tilde{a}$, $\tilde{b}$, and $\tilde{c}$ are defined as in Equation (4.13). $\tilde{d}$ can be expressed as:

$$\tilde{d} = \tilde{\eta}_{CB} h_{CB}^2 \left( \frac{\tilde{x}_C}{\tilde{K}_{CB}} \right)^{h_{CB}-1} \left( 1 + \left( \frac{\tilde{x}_C}{\tilde{K}_{CB}} \right)^{h_{CB}} \right)^{-2}, \tag{4.19}$$

where $\tilde{\eta}_{CB} = \frac{\eta_B}{K_{CB}}$, and $\tilde{K}_{CB} = \frac{K_{CB}\delta_C}{\eta_C \pi_C}$. Table 4.2 suggests that increasing $\tilde{\eta}$ enhances response sensitivity in IFFL networks. The underlying causes are rooted in the retroactivity matrices. Since IFFL networks differ merely in the types of regulation, the retroactivity matrix given in Equations (4.12) and (4.13) is the same for all IFFL. From Equation (4.12), it is easy to see that increasing $\tilde{\eta}$ decreases $\dot{\tilde{x}}_A$ and $\dot{\tilde{x}}_B$, i.e., changes of protein concentrations of both A and B become slower (Figure 4·4). Inhibition of B by A takes a longer time till B reaches a sufficiently low concentration such that B can no longer activate C. Simultaneously, $\tilde{\eta}$ does not affect $\dot{\tilde{x}}_C$. Consequently, C accumulates a larger response since $\dot{\tilde{x}}_C$ is unaffected by $\tilde{\eta}$, and the growth time of $\tilde{x}_C$ becomes longer. Higher response sensitivity can lead to higher adaptive robustness as a trajectory becomes more likely to satisfy Equation (4.16) due to a lower adaptation ratio.

Simulation of the "mixture" network suggests a different story. Table 4.2 indicates that $\tilde{\eta}$ is negatively associated with mean response sensitivity in the "mixture" network. Due to the NFBL between B and C, increasing $\tilde{\eta}$ decreases the rates of changes of protein concentrations for all nodes. While C has more time to grow, the growth rate of $\tilde{x}_C$ decreases.

Table 4.2 also suggests that increasing retroactivity may increase the overall adaptation errors by raising the risks of destabilization of the initial steady states. In a few simulations assuming $\tilde{\eta} = 10$, $\tilde{x}_C$ skips recovery phases and converges to new steady states. Perfect adaptation requires that the initial steady states be stable [Shi et al., 2017]. Mou et al. find that increasing retroactivity decreases stability radii around

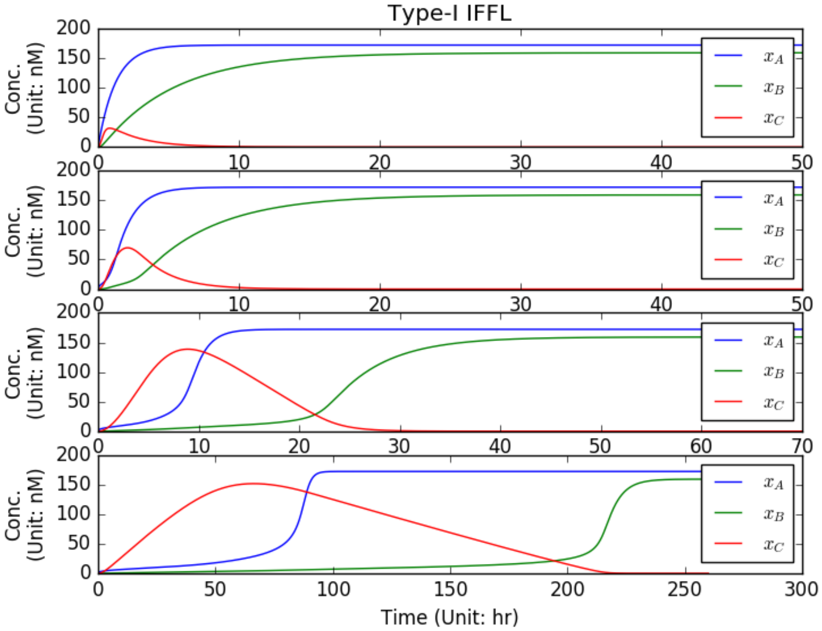

**Figure 4·4:** An example of trajectories simulated by non-normalized Type-I IFFL models. Values of the parameters are: $K_{IA} = 0.4$nM, $K_{AB} = 24.62$nM, $K_{AC} = 10$nM, $K_{BC} = 10$nM, $h_{IA} = 1$, $h_{AB} = 2.60$, $h_{AC} = 3.28$, $h_{BC} = 2.77$, $\delta_A = 0.81\text{hr}^{-1}$, $\delta_B = 0.20\text{hr}^{-1}$, $\delta_C = 0.47\text{hr}^{-1}$. DNA concentrations increase and protein production rates per plasmid decrease from the top figure to the bottom figure. For fair comparison, the products of DNA concentrations and protein production rates per plasmid are kept fixed. Top: $\pi_A\eta_A = 144.18\text{nM}\cdot\text{hr}^{-1}$, $\pi_B\eta_B = 31.29\text{nM}\cdot\text{hr}^{-1}$, $\pi_C\eta_C = 80\text{nM}\cdot\text{hr}^{-1}$; 2nd from the top: $\pi_A = 58.61\text{hr}^{-1}$, $\pi_B = 12.72\text{hr}^{-1}$, $\pi_C = 80\text{hr}^{-1}$, $\eta_A = 2.46$nM, $\eta_B = 2.46$nM, $\eta_C = 1$nM; 2nd from the bottom: $\pi_A = 5.86\text{hr}^{-1}$, $\pi_B = 1.27\text{hr}^{-1}$, $\pi_C = 8\text{hr}^{-1}$, $\eta_A = 24.6$nM, $\eta_B = 24.6$nM, $\eta_C = 10$nM; bottom: $\pi_A = 0.59\text{hr}^{-1}$, $\pi_B = 0.13\text{hr}^{-1}$, $\pi_C = 0.8\text{hr}^{-1}$, $\eta_A = 246$nM, $\eta_B = 246$nM, $\eta_C = 100$nM. When normalized, the four models from the top to the bottom represent no retroactivity, $\tilde{\eta} = 0.1$, $\tilde{\eta} = 1$, and $\tilde{\eta} = 10$. All parameters are within reasonable ranges of biological parameters given in [Gyorgy and Del Vecchio, 2014].

initial steady states [Mou and Del Vecchio, 2015]. Their findings imply that there are certain parameter perturbations under which the Jacobian of a model without retroactivity is stable at a given steady state while the Jacobian of a model with retroactivity is not [Mou and Del Vecchio, 2015]. The results of our simulation are in agreement with [Mou and Del Vecchio, 2015]. Moreover, our results indicate that the effects of retroactivity on adaptation errors also depend on the network topologies. The adaptation errors of the type-I IFFL are hardly affected by $\tilde{\eta}$ while the adaptation errors of the type-IV IFFL and the "mixture" networks are affected by $\tilde{\eta}$ to varying degrees (Table 4.2). Higher adaptation errors can lead to lower adaptive robustness as a trajectory now becomes less likely to satisfy Equation (4.16) due to a higher adaptation ratio.

Our analysis above can be generalized to other networks. It is easy to prove that if C is not a regulator in a network, the bottom right entry of $R(\vec{x})$ in Equation (4.6) is always 0. In these networks, the rate of change of C is not slowed down by retroactivity. Increasing DNA concentrations enhances response sensitivity in these networks, leading to higher adaptive robustness. As is shown in Figure 4·3, there are 16 networks that show consistently decreasing adaptive robustness as $\tilde{\eta}$ increases. Among these 16 networks, the three most frequent motifs are IFFL, NFBL between A and B, and negative self-feedback loops on B. A common feature of these motifs is that C is not a regulator, which confirms our hypothesis.

One direct biological implication of our findings is that changing plasmid copy numbers can enhance adaptive robustness. One approach to increasing retroactivity while keeping the steady-state behavior of the network unchanged is to raise the plasmid copy number and lower the protein production rate per plasmid by the same fold. An example of such an approach is given in Figure 4·4, where the plasmid copy number and the protein production rate per plasmid are allowed to vary, but the

**Table 4.2:** Results from the perturbation analysis. Models I, II, III, and IV are defined the same as in the caption of Table I. Mean response sensitivity, adaptation errors, and adaptation ratios, which are calculated from 100 randomly selected parameter sets, are shown in the 3rd, 4th, and 5th columns. The adaptive robustness of these networks, which is inferred earlier via the algorithm shown in [Zuliani et al., 2013], is listed in the last columns.

| Type-IV IFFL | | | | | |
|---|---|---|---|---|---|
| Model | $\tilde{\eta}$ | Response | Error | Ratio | Robustness |
| I | N/A | 0.5967 | 0.0076 | 0.0165 | 0.6235 |
| II | 0.1 | 0.6013 | 0.0077 | 0.0143 | 0.6241 |
| III | 1 | 0.6250 | 0.0081 | 0.0130 | 0.6310 |
| IV | 10 | 0.6579 | 0.0249 | 0.0351 | 0.5721 |

| Type-I IFFL | | | | | |
|---|---|---|---|---|---|
| Model | $\tilde{\eta}$ | Response | Error | Ratio | Robustness |
| I | N/A | 0.2507 | 0.0052 | 0.0161 | 0.4843 |
| II | 0.1 | 0.2612 | 0.0052 | 0.0154 | 0.4891 |
| III | 1 | 0.3222 | 0.0052 | 0.0123 | 0.5159 |
| IV | 10 | 0.4614 | 0.0053 | 0.0085 | 0.5604 |

| Mixture | | | | | |
|---|---|---|---|---|---|
| Model | $\tilde{\eta}$ | Response | Error | Ratio | Robustness |
| I | N/A | 0.6215 | 0.0055 | 0.0114 | 0.6329 |
| II | 0.1 | 0.6187 | 0.0056 | 0.0114 | 0.6427 |
| III | 1 | 0.6005 | 0.0066 | 0.0140 | 0.6544 |
| IV | 10 | 0.5568 | 0.0878 | 0.1795 | 0.5495 |

total protein production rate is kept fixed. It is clear from Figure 4·4 that A and B maintain the same steady states, whereas C accumulates different levels of response due to different degrees of retroactivity. Experiment-wise, plasmid copy number can be raised via an increase in plasmid dose, and protein production rate per plasmid can be lowered by methods such as adding nucleotides between the promoter and the transcription start site.

# Chapter 5

# Conclusions

## 5.1 Summary of the Thesis

In this thesis, I first presented an ODE modeling framework that predicts circuit behaviors in TTMC. At the core of this framework is a novel bin-dependent model. Compared to a Boolean model, the bin-dependent model not only captures the analog behaviors and wide dynamic ranges of the circuits but also describes the time evolution of the circuit components. Detailed two-stage gene expression models facilitate a relatively thorough qualitative investigation of the mechanism underlying the experimental observations but should not be fit to the data quantitatively, as models based on different hypotheses can explain the same phenomena. In comparison, the bin-dependent model maintains a better balance of accuracy and simplicity for modular characterization. To predict behaviors of circuits, we developed a method of composition that enables model-based design of circuits. The combination of a proper model and batch-effect reduction leads to the improved accuracy of circuit behavior predictions in TTMC. Besides the ODE modeling framework, this thesis also presented a systematic approach based on modular composition, model simulation, and SMC to investigating the relationship between circuit topologies and circuit behaviors. The approach was applied to study how retroactivity, a phenomenon arising from modular composition, impacts the ability of circuits to achieve adaptations.

## 5.2 Impact and Future Directions

Although technology of high-throughput DNA synthesis is advancing rapidly, a design process solely based on manual labor will ultimately be outperformed by a computation-based design workflow. Development of automated workflows for designing genetic circuits is an active research area in the synthetic biology community. Cello represents one of the first closed-loop computational tools that automate the circuit design process by transforming a user-defined functional specification of circuits all the way to the DNA sequences of plasmids that should be synthesized [Nielsen et al., 2016]. Simple as they are, the underlying Boolean models [Nielsen et al., 2016] limit the application of Cello to biological circuits and contexts that exhibit relatively sophisticated temporal dynamics. The ODE modeling framework we developed provides a solution to predictions of circuit behaviors in TTMC. Supported by the validation of experimental data, our framework is likely to improve the speed and the outcome of high-throughput circuit assembly in TTMC, if integrated as a part of the automation workflow. In addition, the bin-dependent model may also be applied to study circuit behaviors in bacteria cells. Relatively recent results in literature suggest that plasmid copy number variation in bacteria cells has long been overlooked [Brynildsrud et al., 2016]. It will be interesting to build circuits similar to the one shown in Figure 2·1(a), which contains an input, an output, and a transfection maker, in bacteria cells and examine the effect of copy number on protein production.

Chapter 2 shows that flow cytometry data are sufficient for fitting the bin-dependent ODE model. For determining the exact mechanism of protein saturation, additional experimental data are required. It is shown in Chapter 2 that two stochastic models built under different hypotheses (transcriptional and translational saturation), differing in parameter values and kinetic rate functions, can recapitulate almost identical observations. To choose between these different cases, one would (minimally) also want to have data on mRNA expression levels or even on DNA levels within individual cells. This would better enable the resolution of the functional forms to use for the various transcription and translation relationships, and (possibly) allow estimation of the many additional parameters of the stochastic models.

The approach presented in Chapter 4 can be further applied to study the effect of retroactivity on other biological functions, such as Turing-pattern formation. It has long been speculated that circuits with activator-inhibitor loops are capable of generating Turing patterns if kinetic parameters including the diffusion rates of morphogens fulfill particular conditions [Turing, 1952]. It will be of great novelty to examine whether retroactivity enlarges or shrinks the parameter space that facilitates pattern formation, as results from such an examination may shed light on recruiting retroactivity as a strategy for designing pattern-forming circuits.

# List of Journal Abbreviations

| | |
|---|---|
| ACS Synth. Biol. | ACS Synthetic Biology |
| Adv. Drug Deliv. Rev. | Advanced Drug Delivery Reviews |
| Adv. Enzyme Regul. | Advances in Enzyme Regulation |
| Adv. Pharm. Bull. | Advanced Pharmaceutical Bulletin |
| Anal. Bioanal. Chem. | Analytical and Bioanalytical Chemistry |
| Annu. Rev. Biomed. Eng. | Annual Review of Biomedical Engineering |
| Annu. Rev. Genomics Hum. Genet. | Annual Review of Genomics and Human Genetics |
| Biochem. Soc. Trans. | Biochemical Society Transactions |
| Bioinformatics | Bioinformatics |
| Biophys. J. | Biophysical Journal |
| Biostatistics | Biostatistics |
| Biotechnol. Bioeng. | Biotechnology and Bioengineering |
| Blood | Blood |
| BMC Bioinformatics | BMC Bioinformatics |
| Brief. Funct. Genomics | Briefings in Functional Genomics |
| Cell | Cell |
| Chem. of Life | Chemistry of Life |
| Curr. Opin. Biotechnol. | Current Opinion in Biotechnology |
| Curr. Opin. Chem. Biol. | Current Opinion in Chemical Biology |
| Curr. Opin. Neurobiol. | Current Opinion in Neurobiology |
| Crit. Rev. Biochem. Mol. | Critical Reviews in Biochemistry and Molecular Biology |
| Eng. Biol. | Engineering Biology |
| FEBS Lett. | FEBS Letters |
| Form. Methods Syst. Des. | Formal Methods in System Design |
| Front. Bioeng. Biotechnol. | Frontiers in Bioengineering and Biotechnology |
| Genetics | Genetics |
| Genome Biol. | Genome Biology |
| IEEE Life. Sci. Lett. | IEEE Life Sciences Letters |
| IEEE Trans. Med. Imag. | IEEE Transactions on Medical |

| | |
|---|---|
| | Imaging |
| In Vitro Cell. Dev. Biol.-Anim. | In Vitro Cellular & Developmental Biology - Animal |
| INFORMS J. Comput. | INFORMS Journal on Computing |
| J. Biol. Eng. | Journal of Biological Engineering |
| J. Cell Biol. | Journal of Cell Biology |
| J. Control. Release | Journal of Controlled Release |
| J. Exp. Med. | Journal of Experimental Medicine |
| J. Gene Med. | Journal of Gene Medicine |
| J. Phys. Chem. | Journal of Physical Chemistry |
| J. R. Soc. Interface | Journal of the Royal Society Interface |
| J. Theor. Biol. | Journal of Theoretical Biology |
| Methods | Methods |
| Methods Mol. Biol. | Methods in Molecular Biology |
| Microb. Genom. | Microbial Genomics |
| Mol. Syst. Biol. | Molecular Systems Biology |
| Mol. Ther. | Molecular Therapy |
| N. Engl. J. Med. | New England Journal of Medicine |
| Nat. Biotechnol. | Nature Biotechnology |
| Nat. Commun. | Nature Communications |
| Nat. Methods | Nature Methods |
| Nature | Nature |
| Pharm. Res. | Pharmaceutical Research |
| Phil. Trans. R. Soc. Lond. B | Philosophical Transactions of the Royal Society of London B: Biological Sciences |
| Phys. Biol. | Physical Biology |
| Plasmid | Plasmid |
| PLOS Comput. Biol. | PLOS Computational Biology |
| PLOS One | PLOS One |
| Proc. American Control Conf. | Proceedings of American Control Conference |
| Proc. IEEE Conf. Decis. Control | Proceedings of IEEE Conference on Decision and Control |
| Proc. IEEE. Int. Conf. Acoust. Speech Signal Process | Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing |
| Proc. Natl. Acad. Sci. U.S.A. | Proceedings of the National Academy of Sciences of the United States of America |
| Protein Sci. | Protein Science |

| | |
|---|---|
| Risk Anal. | Risk Analysis |
| Sci. Rep. | Scientific Reports |
| Sci. Signal | Science Signaling |
| Science | Science |
| Synth. Biol. | Synthetic Biology |
| Synth. Syst. Biotechnol. | Synthetic and Systems Biotechnology |
| Trends Biochem. Sci. | Trends in Biochemical Sciences |
| Trends Biotechnol. | Trends in Biotechnology |

# References

(2001). Measuring molecules of equivalent fluorescein (mefl), pe (mepe), and rpe-cy5 (mepcy) using sphero rainbow calibration particles. Technical report, Technical Report SpheroTechnical Notes: STN-9, Spher- oTech, Libertyville, IL.

(2014). Synthetic biology: back to the basics. *Nat. Methods*, 11(5):463.

Abel, J. H., Drawert, B., Hellander, A., and Petzold, L. R. (2016). Gillespy: a python package for stochastic model building and simulation. *IEEE Life Sci. Lett.*, 2(3):35–38.

Alon, U. (2007). *An Introduction to Systems Biology - Design Principles of Biological Circuits*. Chapman and Hall.

Alon, U., Surette, M. G., Barkai, N., and Leibler, S. (1999). Robustness in bacterial chemotaxis. *Nature*, 397(6715):168–171.

Arkin, A. (2001). Synthetic cell biology. *Curr. Opin. Biotechnol.*, 12:638–644.

Assur, Z., Hendrickson, W. A., and Mancia, F. (2012). Tools for coproducing multiple proteins in mammalian cells. *Methods Mol. Biol.*, 801:173–187.

Avery, O. T., MacLeod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J. Exp. Med.*, 79(2):137–158.

Ay, A. and Arnosti, D. N. (2011). Mathematical modeling of gene expression: a guide for the perplexed biologist. *Crit. Rev. Biochem. Mol.*, 46(2):137–151.

B. James, M. and Giorgio, T. (2000). Nuclear-associated plasmid, but not cell-associated plasmid, is correlated with transgene expression in cultured mammalian cells. *Mol. Ther.*, 1:339–346.

Barenco, M., Tomescu, D., Brewer, D., Callard, R., Stark, J., and Hubank, M. (2006). Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biol.*, 7(3):R25.

Barkai, N. and Leibler, S. (1997). Robustness in simple biochemical networks. *Nature*, 387:913.

Basu, S., Gerchman, Y., Collins, C., Arnold, F., and Weiss, R. (2005). A synthetic multicellular system for programmed pattern formation. *Nature*, 434:1130–1134.

Beal, J. (2015). Bridging the gap: a roadmap to breaking the biological design barrier. *Front. Bioeng. Biotechnol.*, 2(87).

Beal, J. (2017). Biochemical complexity drives log-normal variation in genetic expression. *Eng. Biol.*, 1(1):55–60.

Behar, M., Hao, N., Dohlman, H. G., and Elston, T. C. (2007). Mathematical and computational analysis of adaptation via feedback inhibition in signal transduction pathways. *Biophys. J.*, 93(3):806–821.

Brahme, A., editor (2014). *Comprehensive Biomedical Physics*. Elsevier.

Braun, D., Basu, S., and Weiss, R. (2005). Parameter estimation for two synthetic gene networks: a case study. In *Proc. IEEE. Int. Conf. Acoust. Speech Signal Process.*, volume 5, pages 769–772.

Brewster, R. C., Weinert, F. M., Garcia, H. G., Song, D., Rydenfelt, M., and Phillips, R. (2014). The transcription factor titration effect dictates level of gene expression. *Cell*, 156(6):1312–1323.

Brophy, J. A. N. and Voigt, C. A. (2014). Principles of genetic circuit design. *Nat. Methods*, 11(5):508–520.

Brynildsrud, O., Gulla, S., Feil, E. J., Nørstebø, S. F., and Rhodes, L. D. (2016). Identifying copy number variation of the dominant virulence factors msa and p22 within genomes of the fish pathogen renibacterium salmoninarum. *Microb. Genom.*, 2(4):e000055.

Canton, B., Labno, A., and Endy, D. (2008). Refinement and standardization of synthetic biological parts and devices. *Nat. Biotechnol.*, 26:787–793.

Cao, L.-H., Jing, B.-Y., Yang, D., Zeng, X., Shen, Y., Tu, Y., and Luo, D.-G. (2016). Distinct signaling of drosophila chemoreceptors in olfactory sensory neurons. *Proc. Natl. Acad. Sci. U.S.A*, 113(7):902–911.

Carpenter, R. G. (1960). *Principles and Procedures of Statistics, with Special Reference to the Biological Sciences*, volume 52. McGraw-Hill.

Chandran, D., Bergmann, F. T., and Sauro, H. M. (2009). Tinkercell: modular cad tool for synthetic biology. *J. Biol. Eng.*, 3:19.

Chen, H. and Xia, H. (2011). The research and application of tet-induced regulatory systems. *Chem. of Life*, (2):31.

Cohen, R. N., van der Aa, M., Macaraeg, N., Lee, A., and Jr., F. C. S. (2009). Quantification of plasmid dna copies in the nucleus after lipoplex and polyplex transfection. *J. Control. Release*, 135(2):166–174.

Cohen, S. N., Chang, A. C., Boyer, H. W., and Helling, R. B. (1973). Construction of biologically functional bacterial plasmids in vitro. *Proc. Natl. Acad. Sci. U.S.A*, 70(11):3240–3244.

Dalton, A. and Barton, W. (2014). Over-expression of secreted proteins from mammalian cell lines. *Protein Sci.*, 23(5):517–525.

Davidsohn, N. (2013). *Foundational platform for mammalian synthetic biology.* PhD thesis, Massachusetts Institute of Technology.

Davidsohn, N., Beal, J., Kiani, S., Adler, A., Yaman, F., Li, Y., Xie, Z., and Weiss, R. (2015). Accurate predictions of genetic circuit behavior from part characterization and modular composition. *ACS Synth. Biol.*, 4(6):673–681.

Del Vecchio, D. (2007). Design and analysis of an activator-repressor clock in e. coli. In *Proc. American Control Conf.*

Del Vecchio, D., Ninfa, A. J., and Sontag, E. D. (2008). Modular cell biology: retroactivity and insulation. *Mol. Syst. Biol.*, 4:161.

Del Vecchio, D., Qian, Y., and Dy, A. (2016). Control theory meets synthetic biology. *J. R. Soc. Interface.*

Del Vecchio, D. and Sontag, E. D. (2007). Dynamics and control of synthetic biomolecular networks. In *Proc. American Control Conf.*, New York.

Didovyk, A., Tonooka, T., Tsimring, L., and Hasty, J. (2017). Rapid and scalable preparation of bacterial lysates for cell-free gene expression. *ACS Synth. Biol.*, 6(12):2198–2208.

El-Samad, H., Goff, J. P., and Khammash, M. (2002). Calcium homeostasis and parturient hypocalcemia: an integral feedback perspective. *J. Theor. Biol.*, 214(1):17–29.

Ellis, T., Wang, X., and Collins, J. (2009). Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nat. Biotechnol.*, 27(5).

Elowitz, M. B. and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature*, 403:335.

Endy, D. and Brent, R. (2001). Modelling cellular behaviour. *Nature*, 409:391–395.

Ferrell, J. E. and Ha, S. H. (2014). Ultrasensitivity part iii: cascades, bistable switches, and oscillators. *Trends Biochem. Sci.*, 39(12):612–618.

Gardner, T., Cantor, C., and Collins, J. (2000). Construction of a genetic toggle switch in escherichia coli. *Nature*, 403:339–342.

Geisser, S. (1993). *Predictive Inference: an Introduction*. Chapman and Hall, New York, NY.

Georgianna, D. R. and Mayfield, S. P. (2012). Exploiting diversity and synthetic biology for the production of algal biofuels. *Nature*, 488:329–335.

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

Gimpel, J. A., A., S. E., Georgianna, D. R., and Mayfield, S. P. (2013). Advances in microalgae engineering and synthetic biology applications for biofuel production. *Curr. Opin. Chem. Biol.*, 17(3):489–495.

Glover, D. J., Leyton, D. L., Moseley, G. W., and Jans, D. A. (2010). The efficiency of nuclear plasmid dna delivery is a critical determinant of transgene expression at the single cell level. *J. Gene Med.*, 12(1):77–85.

Goler, J. (2004). Biojade: A design and simulation tool forsynthetic biological systems. Master's thesis, Massachusetts Institute of Technology.

Goodwin, B. (1965). Oscillatory behavior in enzymatic control processes. *Adv. Enzyme Regul.*, 3:425–438.

Gyorgy, A. and Del Vecchio, D. (2014). Modular composition of gene transcription networks. *PLOS Comput. Biol.*

Hacein-Bey-Abina, S., Le Deist, F., Carlier, F., Bouneaud, C., Hue, C., De Villartay, J.-P., Thrasher, A. J., Wulffraat, N., Sorensen, R., Dupuis-Girod, S., Fischer, A., Davies, E. G., Kuis, W., Leiva, L., and Cavazzana-Calvo, M. (2002). Sustained correction of x-linked severe combined immunodeficiency by ex vivo gene therapy. *N. Engl. J. Med.*, 346(16):1185–1193.

Hama, S., Akita, H., Ito, R., Mizuguchi, H., Hayakawa, T., and Harashima, H. (2006). Quantitative comparison of intracellular trafficking and nuclear transcription between adenoviral and lipoplex systems. *Mol. Ther.*, 13(4):786–794.

Hattis, D. and Burmaster, D. E. (2006). Assessment of variability and uncertainty distributions for practical risk analyses. *Risk Anal.*, 14(5):713–730.

Holmen, S. L., Vanbrocklin, M. W., Eversole, R. R., Stapleton, S. R., and Ginsberg, L. C. (1995). Efficient lipid-mediated transfection of dna into primary rat hepatocytes. *In Vitro Cell. Dev. Biol.-Anim.*, 31(5):347–351.

Jackson, D. A., Symons, R. H., and Berg, P. (1972). Biochemical method for inserting new genetic information into dna of simian virus 40: circular sv40 dna molecules containing lambda phage genes and the galactose operon of escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.*, 69(10):2904–2909.

Jaenisch, R. and Mintz, B. (1974). Simian virus 40 dna sequences in dna of healthy adult mice derived from preimplantation blastocysts injected with viral dna. *Proc. Natl. Acad. Sci. U.S.A*, 71(4):1250–1254.

Jayanthi, S., Nilgiriwala, K. S., and Del Vecchio, D. (2013). Retroactivity controls the temporal dynamics of gene transcription. *ACS Synth. Biol.*, 2(8):431–441.

Jiang, P., Ventura, A. C., Sontag, E. D., Merajver, S. D., Ninfa, A. J., and Del Vecchio, D. (2011). Load-induced modulation of signal transduction networks. *Sci. Signal.*, 4(194).

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127.

Jusiak, B., Cleto, S., Perez-Piñera, P., and Lu, T. K. (2016). Engineering synthetic gene circuits in living cells with crispr technology. *Trends Biotechnol.*, 34(7):535–547.

Kærn, M., Blake, W., and Collins, J. (2003). The engineering of gene regulatory networks. *Annu. Rev. Biomed. Eng.*, 5:179–206.

Khan, K. (2013). Gene expression in mammalian cells and its applications. *Adv. Pharm. Bull.*, 3(2):257–263.

Kim, T. and Eberwine, J. (2010). Mammalian cell transfection: the present and the future. *Anal. Bioanal. Chem.*, 397(8):3173–3178.

Kis, Z., Pereira, H. S., Homma, T., Pedrigi, R. M., and Krams, R. (2015). Mammalian synthetic biology: emerging medical applications. *J. R. Soc. Interface*, 12(106).

Kosuri, S. and Church, G. M. (2014). Large-scale de novo dna synthesis: technologies and applications. *Nat. Methods*, 11:499–507.

Kosuri, S., Kelly, J. R., and Endy, D. (2007). Tabasco: A single molecule, base-pair resolved gene expression simulator. *BMC Bioinformatics*, 8:480.

Kwok, R. (2010). Five hard truths for synthetic biology. *Nature*, 463(7279):288–290.

Ma, W., Trusina, A., El-Samad, H., Lim, W. A., and Tang, C. (2009). Defining network topologies that can achieve biochemical adaptation. *Cell*, 138(4):760–773.

Ma, X. and Gao, L. (2012). Biological network analysis: insights into structure and functions. *Brief. Funct. Genomics*, 11(6):434–442.

Macnab, R. M. and Koshland, D. E. (1972). The gradient-sensing mechanism in bacterial chemotaxis. *Proc. Natl. Acad. Sci. U.S.A.*, 69(9):2509–2512.

Marchisio, M. A. and Stelling, J. (2009). Computational design tools for synthetic biology. *Curr. Opin. Biotechnol.*, 20(4):479–485.

Martin, J., Daffos, D., de Adana, R., and Asuero, A. G. (2017). *Fitting Models to Data: Residual Analysis, a Fitting Models to Data: Residual Analysis, a Primer, Uncertainty Quantification and Model Calibration.* InTech.

Mathur, M., Xiang, J. S., and Smolke, C. D. (2017). Mammalian synthetic biology for studying the cell. *J. Cell Biol.*, 216(1):73–82.

May, T., Eccleston, L., Herrmann, S., Hauser, H., Goncalves, J., and Wirth, D. (2008). Bimodal and hysteretic expression in mammalian cells from a synthetic gene circuit. *PlOS One*, 3(6):e2372.

Mou, S. and Del Vecchio, D. (2015). How retroactivity impacts the robustness of genetic networks. In *Proc. IEEE Conf. Decis. Control*, pages 1551–1556.

Muzzey, D., Gómez-Uribe, C. A., Mettetal, J. T., and van Oudenaarden, A. (2009). A systems-level analysis of perfect adaptation in yeast osmoregulation. *Cell*, 138(1):160–171.

Nielsen, A. A. K., Der, B. S., Shin, J., Vaidyanathan, P., Paralanov, V., Strychalski, E. A., Ross, D., Densmore, D., and Voigt, C. A. (2016). Genetic circuit design automation. *Science*, 352(6281).

Pfeifer, A. and Verma, I. M. (2001). Gene therapy: promises and problems. *Annu. Rev. Genomics Hum. Genet.*, 2:177–211.

Pierce, B. (2005). *Genetics: A conceptual approach.* W. H. Freeman and Company.

Roesler, J., Brenner, S., Bukovsky, A. A., Whiting-Theobald, N., Dull, T., Kelly, M., Civin, C. I., and Malech, H. L. (2002). Third-generation, self-inactivating gp91(phox) lentivector corrects the oxidase defect in nod/scid mouse-repopulating peripheral blood-mobilized cd34+ cells from patients with x-linked chronic granulomatous disease. *Blood*, 100(13):4381–4390.

Sanft, K. R., Wu, S., Roh, M., Fu, J., Lim, R. K., and Petzold, L. R. (2011). Stochkit2: software for discrete stochastic simulation of biochemical systems with events. *Bioinformatics*, 27(17):2457–2458.

Schaumberg, K. A., Antunes, M. S., Kassaw, T. K., Xu, W., Zalewski, C. S., Medford, J. I., and Prasad, A. (2016). Quantitative characterization of genetic parts and circuits for plant synthetic biology. *Nat. Methods*, 13(1):94–100.

Schenborn, E. T. and Goiffon, V. (2000). Deae-dextran transfection of mammalian cultured cells. *Methods Mol. Biol.*, 130:147–153.

Schlitt, T. and Brazma, A. (2005). Modelling gene networks at different organisational levels. *FEBS Lett.*, 579(8):1859–1866.

Schwake, G., Youssef, S., Kuhr, J.-T., Gude, S., David, M. P., Mendoza, E., Frey, E., and Radler, J. O. (2010). Predictive modeling of non-viral gene transfer. *Biotechnol. Bioeng.*, 105(4):805–813.

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342.

Shi, W., Ma, W., Xiong, L., Zhang, M., and Tang, C. (2017). Adaptation with transcriptional regulation. *Sci. Rep.*, 7:42648.

Si, T. and Zhao, H. (2016). A brief overview of synthetic biology research programs and roadmap studies in the united states. *Synth. Syst. Biotechnol.*, 1(4):258–264.

Siciliano, V., DiAndreth, B., Monel, B., Beal, J., Huh, J., Clayton, K. L., Wroblewska, L., McKeon, A., Walker, B. D., and Weiss, R. (2018). Engineering modular intracellular protein sensor-actuator devices. *Nat. Commun.*, 9(1):1881.

Sivakumar, H. and Hespanha, J. (2013). Towards modularity in biological networks while avoiding retroactivity. In *Proc. American Control Conf.*

Sontag, E. D. (2011). Modularity, retroactivity, and structural identification. In *Design and Analysis of Biomolecular Circuits*, pages 183–202. Springer-Verlag.

Stark, J., Brewer, D., Barenco, M., Tomescu, D., Callard, R., and Hubank, M. (2003a). Reconstructing gene networks: what are the limits? *Biochem. Soc. Trans.*, 31(6):1519–1525.

Stark, J., Callard, R., and Hubank, M. (2003b). From the top down: towards a predictive biology of signalling networks. *Trends Biotechnol.*, 21:290–293.

Tachibana, R., Harashima, H., Ide, N., Ukitsu, S., Ohta, Y., Suzuki, N., Kikuchi, H., Shinohara, Y., and Kiwada, H. (2002). Quantitative analysis of correlation between number of nuclear plasmids and gene expression activity after transfection with cationic liposomes. *Pharm. Res.*, 19(4):377–381.

Tachibana, R., Harashima, H., Shinohara, Y., and Kiwada, H. (2001). Quantitative studies on the nuclear transport of plasmid dna and gene expression employing nonviral vectors. *Adv. Drug Deliv. Rev.*, 52(3):219–226.

Takahashi, Y., Nishikawa, M., Takiguchi, N., Suehara, T., and Takakura, Y. (2011). Saturation of transgene protein synthesis from mrna in cells producing a large number of transgene mrna. *Biotechnol. Bioeng.*, 108(10):2380–2389.

Takeda, K., Shao, D., Adler, M., Charest, P. G., Loomis, W. F., Levine, H., Groisman, A., Rappel, W.-J., and Firtel, R. A. (2012). Incoherent feedforward control governs adaptation of activated ras in a eukaryotic chemotaxis pathway. *Sci. Signal.*, 5(205).

Tal, S. and Paulsson, J. (2012). Evaluating quantitative methods for measuring plasmid copy numbers in single cells. *Plasmid*, 67(2):167–173.

Thattai, M. and van Oudenaarden, A. (2004). Stochastic gene expression in fluctuating environments. *Genetics*, 167(1):523–530.

Turing, A. (1952). The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 237(641):37–72.

Ventura, A. C., Jiang, P., Van Wassenhove, L., Del Vecchio, D., Merajver, S. D., and Ninfa, A. J. (2010). Signaling properties of a covalent modification cycle are altered by a downstream target. *Proc. Natl. Acad. Sci. U.S.A.*, 107(22):10032–10037.

Vink, T., Oudshoorn-Dickmann, M., Roza, M., Reitsma, J.-J., and de Jong, R. N. (2014). A simple, robust and highly efficient transient expression system for producing antibodies. *Methods*, 65(1):5–10.

Voigt, C. A. (2012). Bacteria collaborate to sense arsenic. *Nature*, 481:33–34.

Wang, R.-S., Saadatpour, A., and Albert, R. (2012). Boolean modeling in systems biology: an overview of methodology and applications. *Phys. Biol.*, 9(5):055001.

Washbourne, P. and McAllister, A. K. (2002). Techniques for gene transfer into neurons. *Curr. Opin. Neurobiol.*, 12(5):566–573.

Watson, J. D. and Crick, F. H. C. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171:737–738.

Xie, Z., Wroblewska, L., Prochazka, L., Weiss, R., and Benenson, Y. (2011). Multi-input rnai-based logic circuit for identification of specific cancer cells. *Science*, 333(6047):1307–1311.

Zuliani, P., Platzer, A., and Clarke, E. M. (2013). Bayesian statistical model checking with application to stateflow/simulink verification. *Form. Methods Syst. Des.*, 43(2):338–367.

# CURRICULUM VITAE

[redacted]

[redacted]