Reverse Engineering and Quantifying Context Effects in Synthetic Gene Networks

Thesis by

Enoch Yeung

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2016

(Defended January 19th, 2016)

Dedicated to my darling Amy and my beloved children, parents, and siblings.

Acknowledgements

I would like to acknowledge my advisor Richard M. Murray for his invaluable guidance and boundless enthusiasm for doing great research. His professional advice, regular feedback, and broad support over all facets of my graduate studies have made it possible to complete this educational journey. Most importantly, his advisement has helped me to become a more conscientious professional and uncompromising scientist and engineer. I am truly grateful for this precious opportunity to study with him.

Next, I would like to express gratitude to my thesis committee. I thank Jim Beck for guiding me to think deeply about local and global identifiability in nonlinear systems, opening my eyes to the world of stochastic system identification, and being patient with me as I tried to achieve the right balance between experiment and theory. Lea, thank you for regularly providing detailed and enthusiastic feedback, for welcoming me in your group meetings, and guiding my scientific journey. John, many thanks for your profound insights, the sudden paradigm shifts that come from our conversations, and your contagious enthusiasm to be a stellar theoretician.

I would also like to thank Ophelia Venturelli for her invaluable inspiration, training, and conversations relating to compositional context, Victoria Hsiao, Jin Park, Anu Thubagere, Adam Rosenthal for wonderful ideas on imaging, and David Younger, Ania Baetica, Vincent Noireaux, Clarmyra Hayes, and Zachary Sun for guidance and assistance with TX-TL experiments,. In addition, I acknowledge Jorge Gonçalves, Jennifer Brophy, Long Cai, Domitilla Del Vecchio, Michael Elowitz, Aaron Dy, Shaobin Guo, Ed Hancock, Jongmin Kim, Eric Klavins, Julius Lucks, Kyle Martin, Andrew Ng, John Palmer, Johann Paulsson, Tom

Prescott, Niles Pierce, Rob Phillips, Erik Winfree, Pan Wei, Vipul Singhal, Eduardo Sontag, Sean Warnick, and Ye Yuan for many insightful conversations. Each of these individuals have played a significant role in shaping the ideas presented in this thesis, though their signature may not be explicit.

Above all, I would like to thank my wonderful wife Amy for her unwavering support and encouragement through my graduate studies. Her faith and confidence in my ability sustained me through countless tough times, including several failed ideas and experiments. Without her, I am certain my research would have been far less fruitful. I also express gratitude to my two little ones, Evelyn and Aaron. Their smiles, giggles, and games added to my motivation even when times were tough. Finally, I am indebted to my parents for the foundation of education that they quietly and consistently laid throughout my life. I cannot begin to enumerate the thousands of teaching moments and educational experiences they have provided I am truly indebted to all my family for this educational experience.

I would also like to acknowledge gracious financial support from multiple organizations. This thesis was supported in part by grants and fellowships from the Charles Lee Powell Foundation, the Kanel Foundation, the National Science Foundation, the Department of Defense, the Air Force Office of Scientific Research, and Defense Advanced Research Projects Agency.

Abstract

In the first part of this thesis, we undertake a quantitative investigation of how compositional context, the spatial arrangement and relative orientation of genes, affects individual gene expression in a genetic network. Taking a synthetic biology approach, we construct a series of simple two-reporter biocircuits, each expressing either an mRNA aptamer or a fluorescent protein, and show that by varying the relative orientation of the two genes we obtain a wide range of gene expression profiles, including context-dependent bimodality. We develop a mathematical model to describe the experimental trends observed based on concepts from DNA supercoiling theory. We validate the model through a series of *in vitro* supercoiling experiments and show that by relaxing positive supercoiling in the plasmids, we can significantly reduce the context effects in gene expression. Most importantly, these insights provide a framework for understanding how compositional context and supercoiling can impose feedback on the intended architecture of a synthetic gene network. As a proof of concept, we engineer a genetic toggle switch exploiting compositional context effects to improve its threshold detection and memory capabilities.

In the second part of this thesis, we examine a series of theoretical and computational tools from dynamical systems theory that assist in engineering novel biochemical reaction networks. We briefly review the concept of dynamical structure functions and network reconstruction as tools for understanding biochemical reaction networks. In particular, we review the concept of resource-loading, show that resource-loading can lead to coupling interactions among biochemical species, and that by estimating a dynamical structure function from experimental data, it is possible to quantify resource loading effects in practice. We

illustrate the importance of knowing these loading effects through several example systems, showing that crosstalk imbalance in feedforward loops can lead to performance limitations. However, since biochemical reaction networks are generally large, in practice, only portions of the global network can be reconstructed at a time. We show, with a combination of theory, simulation, modeling and experiments, it is possible to reconstruct the dynamical structure function of a large-scale biochemical network using a series of network reconstruction experiments. We then demonstrate how the dynamical structure function can be used to analyze context interference and how these perturbations interfere with performance. We illustrate these ideas with several classes of standard biological networks, e.g. autocatalytic systems, cascade systems, and input-coupled systems.

Finally, in the third part of this thesis, we consider models for context interference in stochastic chemical reaction networks. We address the problem of representing a biological system and its environment using a stochastic modeling framework. We first introduce a decomposition of the global chemical reaction system into two systems: a system of interest and its environment. We then present and derive a decomposition of the chemical master equation to achieve a representation describing the dynamics of the system of interest, perturbed by an environmental disturbance. We use this decomposition to model examples of two types of environmental disturbances: the disturbance a system experiences through loading effects from limited resources and the disturbance a system experiences when perturbed by an antibiotic that modifies transcription or translation rates.

Contents

Chapter 1

Introduction

Synthetic gene networks are inherently dependent on their operating context. These context effects can significantly change the way the network dynamics evolve over time. This poses a challenge when engineering novel synthetic gene networks. The synthetic gene networks have to be engineered in such a a way to account to incorporate context interference to their benefit or to insulate against deleterious context effects.

There are many ways in which the context of a synthetic gene network can affect its performance. As we will detail in this thesis, some are well studied and others have been virtually ignored. Context effects can be generally arise in three forms: 1) compositional context, the way in which genetic elements are composed, 2) host context, the way in which the synthetic gene network interacts with the host chassis or organism and 3) environmental context, the way in which environmental parameters such as oxygen levels, redox potentials, diffusion rates, etc. interact with the synthetic gene network.

The first step to understanding these context effects is to quantify the extent of their interference on synthetic gene networks. This thesis addresses this challenge in a variety of ways. First, we focus on compositional context and study how the way in which entire genes are composed affects their intrinsic transcriptional activity. Second, recognizing that context effects are manifold and often confounded with designed network interactions, we use a combination of theory, modeling, and experiments to show that network inference algorithms can be used to quantify context effects. We then invoke classical concepts from control theory

to analyze reconstructed network models. The integration of these two concepts provides fundamental insight into how context effects impact biocircuit performance. Finally, we model context effects at the single cell level, where single cell gene expression is characterized by intrinsic noise and low discrete molecular copy numbers. We develop a new stochastic approach for modeling context interference on synthetic gene networks and illustrate with several relevant examples.

Chapter 2

Quantifying Compositional Context Effects in Synthetic Gene Networks

2.1 Introduction

A fundamental aspect of designing synthetic gene networks is the spatial arrangement and composition of individual genes. With advancements in DNA assembly technology[25, 34, 54, 98], drops in sequencing and prototyping costs [12, 84, 91], and the continual discovery of novel synthetic biological parts [89], synthetic biology is poised to make a leap in the scale and complexity of the networks it builds. So why hasn't it happened yet?

The challenge is that synthetic biological parts can be highly sensitive to context [9], e.g. the physical composition of elements in synthetic genes, conditions of the host chassis, and environmental parameters. Context effects can often be mitigated by engineering principles such as standardization [16, 69] or high-gain feedback [17, 66]. Frequently, it is critical to have an understanding of physical mechanisms underlying context effects before they can be resolved [16, 69]. The key insight is that context effects in synthetic gene networks can rarely be ignored; the study of context effects leads to principle-based design approaches that mitigate their interference.

Complementary to principle-based design approaches are large-library screening approaches [52]. Kosuri et al. showed that it is possible to rapidly screen combinatorial promoter-ribosome-coding sequence libraries for intended gene expression levels and regulatory func-

tion, even if models for individual genetic elements such as promoters and RBSs have limited prediction power [52]. Smanski et al. [87] screened a large combinatorial library for a sixteen gene nitrogen fixation cluster, to explore the effect of genetic permutations in ordering, orientation, and operon occupancy. They discovered there were strong differences in nitrogenase activity, depending on the compositional configuration, but no clear architectural trends emerged from monitoring acetylene reduction. Moreover, the number of compositional variants (more than $\mathcal{O}(10^{19})$) of a sixteen gene cluster made it impossible to exhaustively search and screen for the optimal variant.

These results underscore the complementary role that library screening and principle-based design approaches have in synthetic biology. Library screening approaches can be an extremely effective way to optimize performance in individual parts. However, the number of compositional context variants for larger biological networks quickly mushrooms to scales that are intractable for library-based approaches. If we are to build increasingly larger synthetic biocircuits, including synthetic genomes *designed* from scratch [33], we need deeper physical understanding of how compositional context affects gene expression.

Moreover, while there have been extensive studies on the effects of *intragenic* compositional context on synthetic gene expression (the spatial arrangement of components within a gene, [16, 52, 69]), there have been far fewer studies on the effect of *intergenic* compositional context on synthetic gene expression (the spatial arrangement of entire genes relative to each other). Recently, Chong and coworkers showed that transient accumulation of localized positive supercoiling leads to reduction in gene expression — they showed through *in vitro* transcription experiments that supercoiling could be a physical mechanism behind transcriptional bursting [13]. Their results also suggested that the presence of nearby topological barriers such as DNA-bound proteins or transcriptional activity of neighboring genes can affect local gene expression.

To paraphrase John Donne, the broad implication of these studies is that "no [gene] is an island entire of itself". Clearly, genes with *overlapping* transcripts are subject to

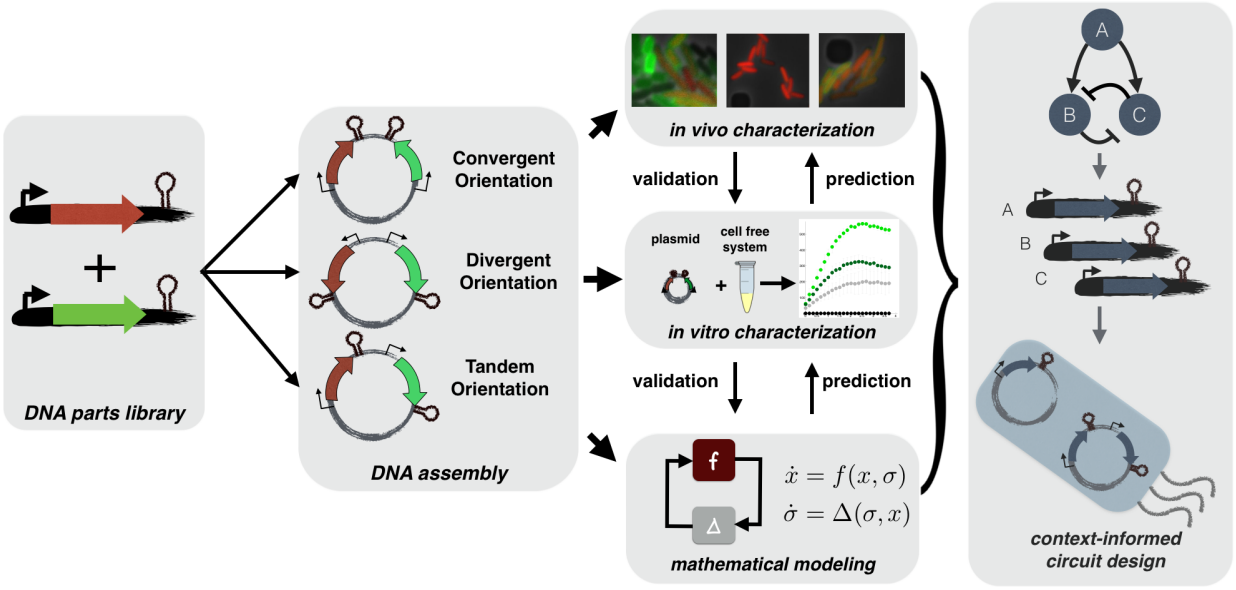

Figure 2.1: **Experimental and theoretical approaches for understanding compositional context effects:** Plasmids are constructed using synthetic biology techniques varying gene orientation, transcript length, coding sequence, replication origin, and antibiotic marker. Each plasmid is characterized thoroughly *in vivo* based on what is appropriate for the fluorescent reporter, e.g. single cell fluorescence microscopy, flow cytometry or plate reader. Plasmids are tested *in vitro* in a cell-free expression system to infer physical hypotheses driving compositional context effects and compared against models capturing the hypotheses inferred from *in vitro* and *in vivo* data. These hypotheses support a conceptual framework for designing synthetic biocircuits that utilize compositional context interactions.

transcriptional interference Rhee et al. [79], Shearwin et al. [83] . However, even in non-overlapping genes, statistical analysis of Korbel et al. [51] on naturally occurring gene pairs suggest there is a strong link between spatial arrangement and co-regulation. Is the same true of synthetic gene networks? If so, how do we use this mode of transcriptional regulation in synthetic gene networks? Even more fundamental, how does intergenic compositional context, i.e. the spatial arrangement of *entire* genes, affect synthetic gene expression?

2.2 Results

2.2.1 Compositional Context Significantly Affects Transcription of Synthetic Genes

To study the effects of compositional context, we constructed a set of plasmids, varying gene orientation, relative orientation, coding sequence identity, and the length of spacing between genes. There are three relative orientations that two genes can assume: 1) convergent orientation, where transcription of both genes proceeds in opposite directions and towards each other, 2) divergent orientation, where transcription of both genes proceeds in opposite directions, away from each other and towards genetic elements on the plasmid backbone, and 3) tandem orientation, where transcription of both genes proceeds in the same direction [57, 83]. We constructed plasmids of each orientation to examine their effect on gene expression *in vivo* and *in vitro*.

Each plasmid incorporated two reporter genes, assembled and inserted in the same locus of a consistent vector backbone. Each gene consisted of an inducible promoter, the Lac or Tet promoter, and a fluorescent reporter. Each plasmid was transformed into MG1655Z1 *E. coli*, which expresses LacI and TetR constitutively from the genome. We chose LacI and TetR since they provide independently inducible systems[10].

We first used mSpinach RNA aptamer and MG RNA aptamer as reporters downstream of the Lac and Tet promoter, respectively. Since mSpinach RNA aptamer is not cytotoxic, it can be used in live-cell imaging to explore how induction response of the Lac promoter varies with compositional context. After equilibrating background levels of fluorescence in mSpinach, we induced the Lac promoter with 1 mM of ispropyl-β-D-1-thiogalactopyroside (IPTG), thus activating expression of mSpinach RNA aptamer. We observed that the induction response of the Lac promoter varied significantly depending on its relative gene orientation, even though the neighboring gene was never activated by aTc (Figure 2.2).

Convergent oriented mSpinach expression produced a ramp-like response to IPTG induc-

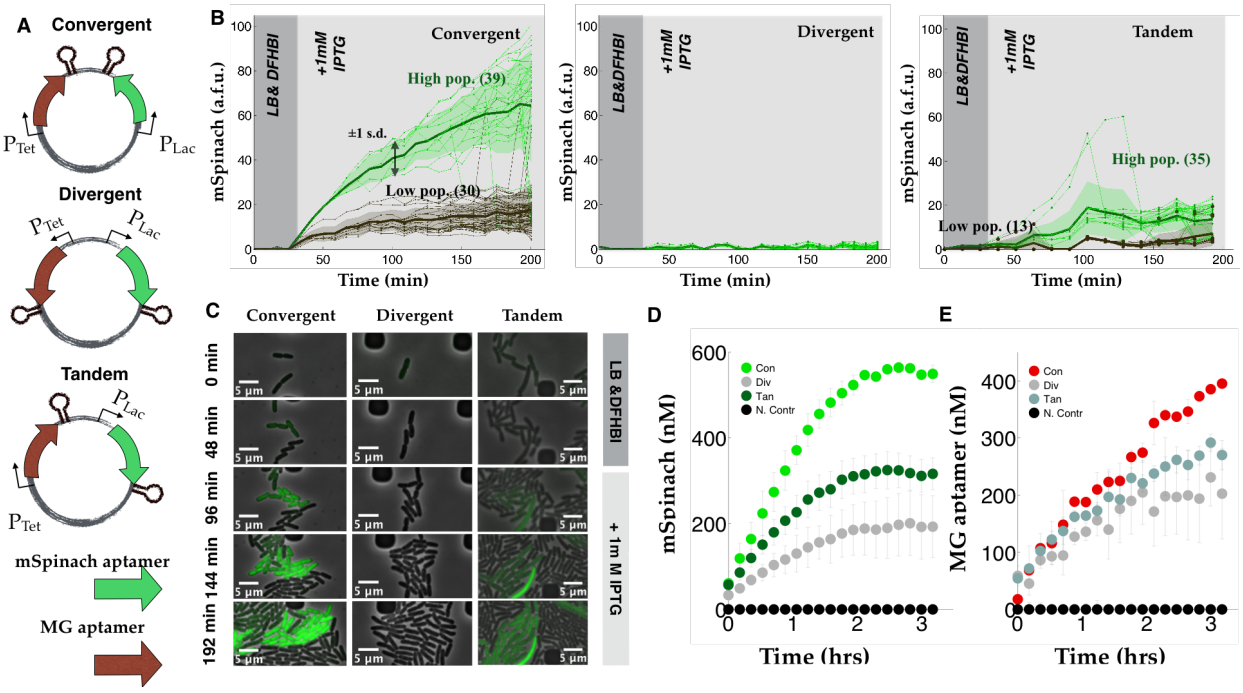

Figure 2.2: **Compositional context alters single cell RNA expression profiles: (A)** Convergent-, divergent-, and tandem-oriented mSpinach and MG aptamer reporters on ColE1 backbone. **(B)** Time-lapse mSpinach expression curves for individual cell traces in response to 1 mM IPTG induction. Solid central lines within a shaded region denote the mean expression across cell lineages within a population, while the shaded area shows one standard deviation from the mean. **(C)** Single cell microscopy images of convergent-, divergent-, and tandem- oriented mSpinach expression. **(D-E)** Convergent-, divergent-, and tandem- oriented mSpinach and MG RNA aptamer expression in an *E. coli* cell free expression system.

tion, rising gradually over the course of three hours to reach a steady-state level of expression coinciding with saturation in the microfluidic chamber (Figure 2.2). Convergent oriented mSpinach also gave a strong bimodal response to IPTG induction, with one group of cells achieving high levels of expression (Figure 2.2B) and another with low expression (Figure 2.2B).

In contrast, divergent oriented mSpinach had a very uniform and weak induction response. Tandem oriented mSpinach had bimodal expression as well, with its brightest population of cells expressing at steady-state levels comparable to the weak population in convergent orientation. The remainder of tandem oriented mSpinach *E. coli* cells showed very weak levels of fluorescence.

Interestingly, cells with tandem oriented mSpinach exhibited pulsatile expression, in contrast to the ramp-like response shown by convergent oriented mSpinach. A few outlier cell traces achieved levels of mSpinach expression comparable to the bright convergent mSpinach population, but only at the peak of their transient pulses. Overall, tandem oriented mSpinach exhibited bursty and weaker gene expression than convergent oriented mSpinach.

Since many intracellular parameters fluctuate stochastically *in vivo* [22], we ran control experiments of each plasmid in a cell-free *E. coli* derived expression system [73]. In this system the effects of single-cell variability are eliminated, e.g. variations in LacI and TetR repressor concentration, polymerase, ribosome, tRNA pools. Also, all deoxynucleotide triphosphates are removed during preparation of cell-extract, thus eliminating any confounding effects of plasmid replication. We prepared separate cell-free reactions for each orientation, assaying mSpinach and MG aptamer expression in a plate reader, using equimolar concentrations for each reaction (Figure 2.2D-E). Because all cell-free reactions were derived from a single batch of well-mixed extract, the variability in LacI repressor concentration was minimal.

Again, we observed that mSpinach was brightest in the convergent orientation, weakest in the divergent orientation and achieved intermediate expression in the tandem orientation. Likewise, MG aptamer expressed strongest in the convergent orientation, weaker in the tandem orientation, and weakest in the divergent orientation. These *in vitro* outcomes were all consistent with the data from *in vivo* single cell experiments. Since the only connection between our *in vitro* tests and and *in vivo* strains is the plasmids themselves, this confirms that compositional context is the reason for differences in gene expression. We hypothesize that compositional context can significantly alter the transcriptional response of a gene to induction.

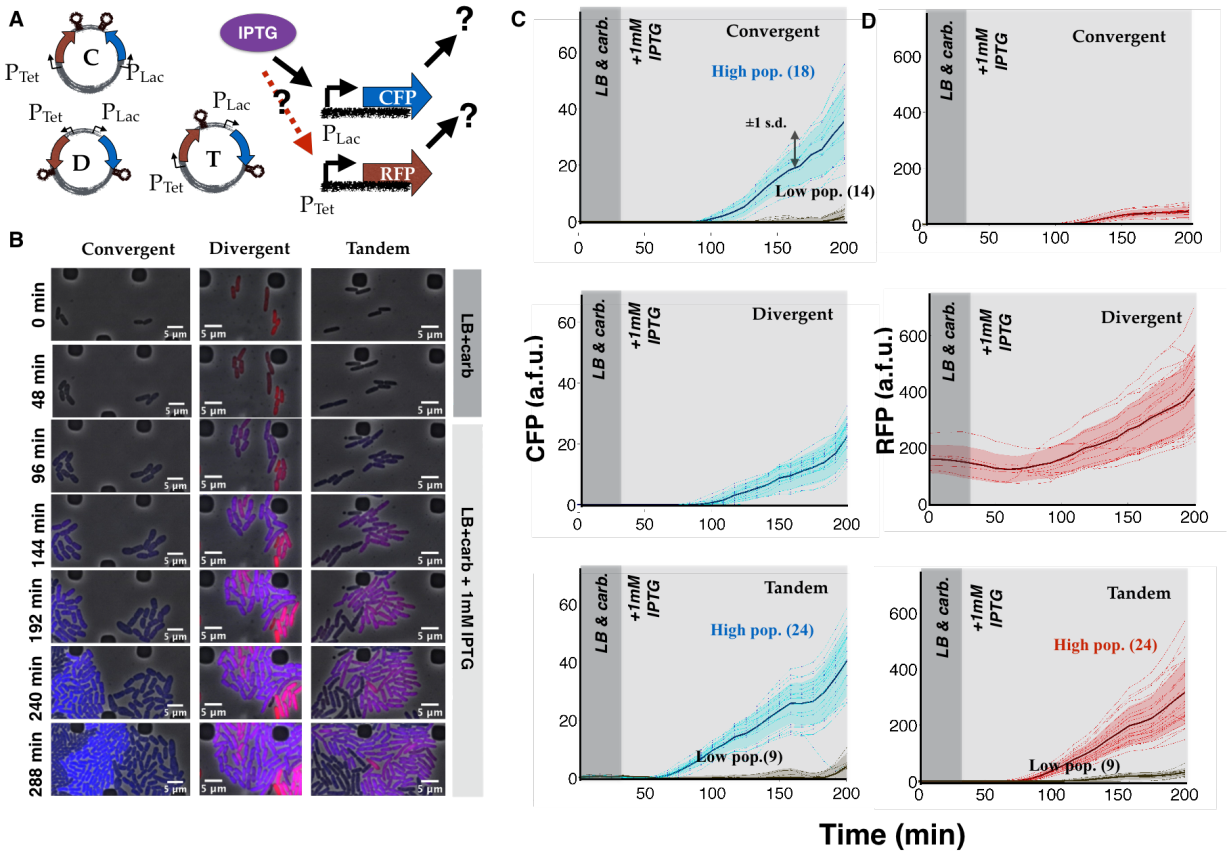

Figure 2.3: **Consistent compositional context effects are observed in translated reporters:** **(A)** Plasmid maps for convergent, divergent, and tandem oriented CFP and RFP on the ColE1 plasmid backbone. **(B)** Single cell microscopy images of convergent-, divergent-, and tandem- oriented CFP and RFP expression superimposed. **(C)** Single cell traces of CFP response to IPTG induction in the convergent, divergent, and tandem orientation. Note that both convergent orientation and tandem orientation exhibit bimodal expression phenotypes. High expressing CFP cells also corresponding to high expressing RFP cells, while low expressing CFP cells (gray traces) correspond to low expressing RFP cells. **(D)** Single cell traces of RFP response to IPTG induction in the convergent, divergent, and tandem orientation. Note that both divergent orientation and tandem orientation respond to IPTG induction with significant RFP expression, while convergent orientation responds slightly with some leaky RFP expression.

2.2.2 Compositional Context Effects Are Pervasive in Translational Reporters

Context interference is only relevant to synthetic gene network design to the extent they alter expression of critical processes, e.g. expression levels of proteins that regulate other

components in the network. To explore if these compositional context effects propagated to translational expression, we replaced the transcript of MG aptamer with the coding sequence for red fluorescent protein (Bba E1010 [106]). We also interchanged the spacer between mSpinach and RFP, to see if our results were dependent on the sequence of the spacer. We then ran an identical experiment, as in Figure 2.2, to see how induction of mSpinach affected and correlated with RFP expression in single cells.

As expected, relative gene orientation had the same effect on mSpinach transcription as in Figure 2.2. Even with RFP in place of MG aptamer, mSpinach expression was highest in the convergent orientation and weakest in the divergent orientation.

We also observed that both convergent and tandem oriented mSpinach expressed with a bimodal phenotype (see Figure 2.9B-C). These results confirmed that the identity of the neighboring gene and spacer sequence content was not the source of these gene expression differences.

Interestingly, RFP expression was extremely leaky in the divergent orientation. In contrast, convergent oriented mSpinach and RFP showed strong XOR logic — any cells that expressed small amounts of RFP did not respond to IPTG induction with mSpinach expression, while cells that did not express any RFP showed strong mSpinach expression. This data suggests compositional context can be exploited to shape co-expression of neighboring genes.

To further show that compositional context effects extend to translated reporters, we replaced mSpinach with cyan fluorescent protein (CFP) [97]. We deliberately used a weak RBS from [69], for CFP and a strong RBS [55] to ensure that any ribosome competition effects would be unidirectional (RFP loading on CFP and not vice-versa) [39]. Thus, if both genes are induced, any differences in RFP expression would elucidate compositional context effects and not competition for translational resources.

We ran a single induction experiment, analogous to experiments run for Figure 2.2 and Figure 2.9. Induction of CFP with IPTG showed that mean CFP expression was again

strongest in the convergent, (slightly) weaker in the tandem orientation, and weakest in the divergent orientation (Figure 2.3, Figures 2.10C-D (IPTG-only condition).

As a control for plasmid backbone, we cloned and induced CFP as a single gene on the exact same plasmid locus (either in sense or anti-sense orientation relative to the plasmid vector). In both control plasmids, 100 bp flanking upstream and downstream sequences were preserved as in the experimental plasmids, to eliminate any promoter sensitivity to upstream sequence perturbation. We noticed a dramatic 5-fold increase in signal over the weakest expressing orientation (compare Figures 2.10B,F-G and 2.10C). In contrast, comparing sense and anti-sense expression of CFP showed only a small (at most 10% difference in expression). This confirmed that the observed compositional context effects could not be attributed to genetic elements within the plasmid backbone. We also tested the effect of changing the plasmid origin (from ColE1 to p15A) and resistance marker (AmpR to CmR), see Figure 2.11B-C. While the quantitative differences in expression changed by varying plasmid vector (most likely reflecting a change in the copy number of the plasmid), the trends were qualitatively identical. This confirmed that plasmid backbone composition was not the primary source of the observed context effects.

Once again, to control for single-cell variability *in vivo*, we tested RFP and CFP expression of each context variant in a cell free expression system [84]. CFP and RFP expressed strongest in convergent orientation, weaker in tandem orientation, and weakest in divergent orientation (Figure 2.5B). These results were consistent with results of our prior *in vitro* tests with mSpinach-MG aptamer plasmids.

Taken in whole, these findings lead us to conclude that the increase in convergent and tandem CFP expression over divergent oriented CFP was unrelated to resource loading effects, plasmid copy number variability or processes related to plasmid replication. These compositional context trends were also consistently observed across multiple coding sequences, transcript lengths, including transcriptional and translational reporters. Therefore, we conclude the compositional context is the primary source of the observed differences in gene

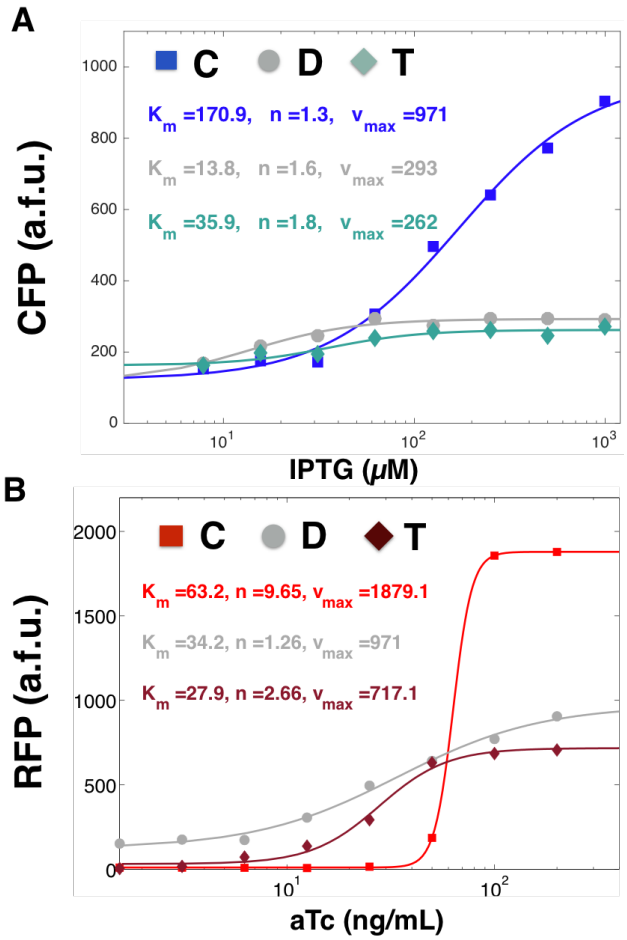

Figure 2.4: **Ultrasensitivity, basal expression, and amplitude of gene induction responses are significantly affected by compositional context: (A)** Induction curves for convergent-, divergent-, and tandem- oriented CFP fitted to a Hill function at aTc = 200 ng/mL and varying concentrations of IPTG. **(B)** Induction curves for convergent-, divergent-, and tandem-oriented RFP fitted a Hill function at IPTG = 1000 nM and varying concentrations of aTc.

expression.

2.2.3 Induction response of genes is affected significantly by compositional context

To see how compositional context altered the induction response over a range of inducer concentrations, we titrated both IPTG and aTc and quantified RFP and CFP expression in bulk culture plate reader experiments (Figure 2.4 and Figure 2.10E). As predicted by our

choice of RBSs (using a strong RBS for RFP and a weak RBS for CFP), increases in RFP expression consistently resulted in decreased CFP expression independent of orientation. As expected, increasing CFP expression did not decrease RFP expression. What was most notable was how gene orientation affected the induction response of RFP expression to varying amounts of aTc inducer.

In the convergent orientation, we saw that the transfer curve of RFP expression exhibited strong ultra-sensitivity, increasing by 120-fold in response to only an 8-fold change in aTc. At 100-200 ng/mL of aTc, RFP expression plateaued in an on-state of expression and below 25 ng/mL, RFP expression plateaued in an off-state of expression. Thus, diluting aTc with an 8x dilution factor had the effect of completely switching RFP from an on to an off state.

In contrast, the tandem orientation required a 64-fold change in aTc concentration to achieve a comparable (100x) fold-change in RFP expression. At 100-200 ng/mL, we also saw RFP expression plateaued in an on-state of expression (for all concentrations of IPTG tested). However, RFP reached an off-state of expression only when aTc was diluted down to 3 ng/mL or lower. Thus, to achieve the same dynamic range as convergent RFP required an 8x increase in dilution factor.

Divergent oriented RFP exhibited the smallest dynamic range. Varying aTc concentration 200-fold produced at most a 2.7 fold change in RFP expression. Even without aTc, RFP expressed at much higher levels than background. We also saw leaky expression at the single cell level, both in the divergent oriented RFP and CFP MG1655Z1 strain (Figure 2.3) and divergent oriented RFP and mSpinach strain (Figure 2.10). Since both strains used different spacing sequences of lengths ranging from 150-350 bp, we concluded these leaky effects were a function of RFP gene orientation and not spacer identity nor proximity to the Lac promoter.

We also fit the induction response of each fluorescent protein while maximally inducing the other gene (Figure 2.4B-C). Our fits characterized induction response in terms of four parameters, leaky expression l, effective cooperativity n, maximum expression V_{max}, and

half-max induction K_m. We noticed that convergent oriented RFP showed significantly increased cooperativity coefficient, nearly four-fold more than tandem orientation and and eight-fold more than divergent orientation. Also, convergent orientation consistently fitted with the highest K_m value in both RFP and CFP induction curves, suggesting that orienting genes convergently raises the induction threshold.

Our experimental data show that compositional context changes gene expression, induction, and repression. Overall, compositional context can dramatically alter canonical properties of synthetic gene expression and thus should not be overlooked when designing synthetic gene networks.

2.2.4 A Dynamic Model Incorporating Supercoiling States Recapitulates Observed Compositional Context Effects

Building on the work of [13, 57, 63] we investigated whether supercoiling can explain the compositional context effects seen in our data. We constructed an ordinary differential equation (ODE) model describing transcription and translation of both genes. To describe the interplay between gene expression and accumulation of supercoiling for each gene, we introduced separate states to keep track of promoter supercoiling and coding sequence supercoiling. This model structure allowed us to study how supercoiling buildup affected both the processes of transcription initiation and elongation [19]).

The kinetic rates of transcriptional initiation and transcriptional elongation are significantly affected by supercoiling buildup [19]. Negative supercoiling relaxes or melts the DNA double helix, facilitating transcription initiation and elongation benefits. However, excessive negative supercoiling can lead to the formation of R-loops, structural complexes that involve DNA binding to nascently produced RNA still attached to RNA polymerase. These R-loops complexes have been shown to cause transcriptional stalling [19].

Conversely, positive supercoiling of DNA introduces torsional stress since positive supercoils naturally oppose the left-handed twist of DNA. Such stress leads to localized regions of

tightly wound DNA that is less likely to be transcribed; positive supercoils downstream of a transcription bubble can also impose torsional resistance against further unwinding of the DNA, thereby stalling transcription. When a gene expresses and produces positive supercoiling downstream of the transcription bubble, the accumulation of positive supercoiling is especially exacerbated by the presence of a topological barrier, e.g. the binding of a DNA binding protein such as a transcription factor, or even the presence of another active gene in negatively supercoiled state. Buildup in positive supercoiling reduces the initial rate of gene transcription [13]. Thus, excessive supercoiling in the DNA double helix in either direction can decrease transcription rates.

In our model we account for the above considerations by encoding a dependency of transcription rate parameters on local supercoiling density. We model the buildup in supercoiling density as a consequence of the presence of DNA binding proteins or torsional stress from transcription of nearby genes. We build on the analysis of Meyer and Beslon [63] and consider transcription initiation rates to be *dynamically* dependent on supercoiling density. We model them as Hill functions of the absolute deviation of the promoter supercoiling state from a natural supercoiling state [79]. In other words, as DNA becomes too twisted in either the positive or negative direction, transcription initiation rates and elongation rates decrease. Similarly, we suppose that the elongation rate of the gene of interest can be modeled as a Hill function of the supercoiling state over the transcript region. Thus, by modeling the dependency of transcription rates on supercoiling, we can model context-specific coupling between neighboring genes (Figure 2.5).

After incorporating these supercoiling hypotheses, our model was able to recapitulate compositional context trends observed in our experimental data (Figure 2.5B). Our simulations showed that convergent oriented mSpinach (and CFP) is able to achieve higher levels of expression than its divergent and tandem counterparts, due to differences in supercoiling levels. These differences arise in our supercoiling model from topological barriers imposed by transcription bubbles and DNA binding proteins. Since mSpinach expresses in

the anti-sense direction, its expression introduces negative-supercoils upstream according to the twin-domain model by Liu & Wang [57]. In moderate amounts, negative supercoiling facilitates the unwinding of DNA and thus enhances the amount of transcriptional initiation and elongation occurring over the Lac promoter and mSpinach transcript. In the divergent and tandem orientation, mSpinach is expressed in the sense direction, which results in positive supercoiling build-up downstream of the promoter (Figure 2.5A). The outcome is that divergent and tandem mSpinach expression is reduced (compared to convergent), since the buildup of positive supercoiling in the presence of adjacently positioned DNA binding proteins or an active transcription bubble inhibits initiation and elongation. This effect is more severe in the divergent orientation, since excessive positive and negative supercoiling generated by initiation of the Tet and Lac promoter can interfere with each other's initiation (Figure 2.5C).

In exploring the parameter space of our model, we also found that gyrase (an enzyme that relaxes positive supercoiling) and topoisomerase (an enzyme that relaxes negative supercoiling) activity are not sufficiently high to counteract the coils introduced by rapid repeated transcription events on DNA with multiple genes. These findings were consistent with the analysis of Chong et al. [13], Liu & Wang [57], and Meyer et al. [63], which argued that buildup of transcription-induced supercoiling far outpaces the activity of supercoiling maintenance enzymes in *E. coli*. This explains why we are able to see compositional context effects both *in vivo* and *in vitro* where gyrase and topoisomerase enzymes are presumably present and active. These results also suggested that extended pre-incubation of plasmids with gyrase would allow us to infer the effect of relaxing positive supercoiling on gene expression in each orientation.

2.2.5 Relaxing positive supercoiling in plasmids significantly reduces compositional context effects

To test the effect of incubating context-variant plasmids with gyrase, we purified plasmids expressing convergent, divergent, and tandem oriented RFP and CFP from uninduced MG1655Z1 *E. coli*. We divided each plasmid sample into two aliquots — one aliquot was used as a control for the absence of gyrase treatment and the second aliquot was incubated with gyrase (NEB) at 37° C overnight. Once again, we tested the expression levels of each plasmid in the cell-free TX-TL system [84].

In the absence of gyrase, convergent orientation expressed higher than divergent and tandem orientation in both RFP and CFP channels. After gyrase treatment, tandem oriented CFP and RFP expressed brighter than their convergent and divergent counterparts. Treating with gyrase changed the relative ordering of expression levels, as opposed to unilaterally shifting all orientations simultaneously. This suggested that supercoiling as an intrinsic driver of context interference, rather than an extrinsic global factor. Also, the disparity in protein expression between the two orientations farthest apart in expression, convergent and divergent, shrunk from 300 nM to 100 nM (66% for CFP) and from 500 nM to 180 nM (64% for RFP). Since gyrase serves only to relax plasmids of positive supercoiling, this confirmed that supercoiling is the mechanism underlying compositional context effects.

We anticipated that treatment with gyrase would release positively supercoiled domains in the downstream region of tandem oriented CFP and RFP and release positive supercoiling buildup from divergently (leaky) expressed RFP (Figure 2.5A) and thereby reduce torsional stress in the promoters of divergently oriented CFP and RFP. Our experimental results confirmed these hypotheses, with divergent orientated CFP and RFP increasing by more than 2 fold and tandem orientated CFP and RFP increasing by 1.4 fold.

Interestingly, gyrase treatment of convergent oriented CFP and RFP appeared to reduce signal slightly, by approximately 10%. This may be because convergent oriented CFP and RFP exhibited little or no leak when uninduced (in contrast to divergent and tandem orien-

tation); thus the purified plasmid for convergent orientation did not have as much positive supercoiling for gyrase to mitigate. Treatments with gyrase may actually have introduced too much negative supercoiling, leading to the small drop in expression observed.

These experimental outcomes are consistent with our model of supercoiling and its impact on compositional context. Gyrase relaxes positively supercoiled domains downstream of convergent and tandem oriented RFP, while in the divergent orientation, gyrase relaxes any positive supercoiling buildup near the promoter region. Once these positive supercoils are removed, the genes are able to express at much higher levels than prior to treatment.

Our data shows that compositional context can have a strong effect on the dynamics of supercoiling within plasmids. Nearby *transcriptionally active* genes or protein-bound genes act as topological barriers to stop migration of supercoils or dispersion of localized torsional stress. Protein-bound genes in particular, act to trap supercoiling in neighboring transcriptionally active genes; this may explain why in our IPTG induction experiments, the mere presence of a repressed RFP and MG gene (respectively) could have such a significant effect on CFP and mSpinach expression. In this way, gene orientation and placement can introduce a fundamentally different form of feedback coupling between neighboring genes. When used appropriately, these feedback effects can be beneficial or detrimental to the intended architecture of the biocircuit, as we illustrate in the next section.

2.2.6 Compositional context improves memory and threshold detection in toggle switch

Synthetic gene networks, for the most part, have been designed primarily to avoid one type of compositional context effect: terminator leakage. Terminator leakage can cause positive correlation between a downstream gene with an upstream genes. While this is a noteworthy consideration in designing synthetic gene networks, we can actually utilize compositional context to improve or reinforce the feedback architecture of synthetic gene networks.

The toggle switch provides an excellent case study of how an informed understanding

of compositional context can improve design. Being one of the first synthetic biocircuits ever made, it was constructed in divergent orientation to avoid terminator leakage effects between two mutually repressing genes, LacI and TetR [31, 50]. From the perspective of protein regulation, two proteins, LacI and TetR, enforce mutual repression by transcriptional repression.

However, we can also build the toggle switch in convergent or tandem orientation. The convergent toggle switch is most appealing, based on several experimental insights: 1) the competing dynamics of positive and negative supercoils between the two genes encodes an additional layer of mutual negative feedback (Figure 2.7A), 2) the coexpression profiles of RFP and CFP in the convergent orientation (Figure 2.10)E and mSpinach and RFP in the convergent orientation (Figure 2.9) was strongly anti-correlated. All of these properties of compositional context have the potential to enhance or strengthen the existing mutual negative feedback in the toggle switch.

Since our previous controls of sense and anti-sense encoded single genes showed that changing orientation of a single gene on a backbone did not affect expression more than 15 %, we thus constructed a two plasmid version of the toggle switch, with LacI and TetR expressed on separate plasmids. This "context-free" version of the toggle acted as a reference for how a toggle switch should function independent of genetic context.

In both versions of the toggle switch, each gene cassette in the toggle switch was bicistronic, with LacI reported by translation of RFP and TetR reported by GFP. We used stronger ribosome binding sites to express LacI and TetR and weak BCDs to express the downstream reporters.

This was done to minimize any ribosomal loading effects from reporter translation and again, to show that even a toggle switch built *de novo* from existing synthetic biological parts with different CDSs, promoters, and RBSs could utilize compositional context (Figure 2.7). We also tested the original Gardner-Collins toggle switch, comparing performance in the original orientation to a convergent variant, see Figure 2.11 and discussion in the

Information.

We first tested the ability of the toggle switch to act as a threshold detector. In theory, the phase portrait of a toggle switch consists of two locally asymptotically stable equilibrium points and a separatrix which drives state trajectories into the basin of attraction of one of the equilibrium points [31]. As a proxy for varying the amount of actively repressing TetR and LacI, we simultaneously varied the concentration of inducers aTc and IPTG, thereby allowing us to attenuate the activity of LacI and TetR repression independently. Most notably, when the toggle switch was configured in the convergent orientation, it exhibited much sharper XOR logic and separation between high GFP-low RFP states and high RFP-low GFP states compared to its two-plasmid (and divergent) counterpart (Figure 2.7 and Figure 2.11A-B).

The two-plasmid toggle exhibits weaker thresholding in two specific parameter regimes — when IPTG and aTc are both present in high concentrations and when IPTG and aTc are both present in low concentrations. When both inducers are present in high concentrations, the majority of Lac and Tet promoters are unrepressed because most repressor proteins are sequestered by inducers, leading to weak feedback. The weak feedback makes it difficult to differentiate which inducer is higher, since all promoters are essentially expressing constitutively (Figure 2.7F-H). When both inducers are present in low concentrations, both promoters are strongly repressed making it difficult for one promoter to gain a dominant foothold over the other sufficient to produce fluorescent signal. Thus, in the low inducer concentration regime, even if one inducer is higher in concentration than the other, neither gene is strong enough to repress the other to the point of producing detectable fluorescence (Figure 2.7F-H).

On the other hand, the convergent toggle shows clear separation between high GFP-low RFP states and low GFP-high RFP states in both of these parameter regimes. This improved performance can be explained by examining the effects of supercoiling and compositional context (Figure 2.7A-B). Suppose, for illustration, that LacI-RFP is slightly more

induced than TetR-GFP. The positive supercoiling from TetR-GFP expression propagates downstream to meet the negative supercoils generated from transcription of LacI-mRFP. As more and more LacI-mRFP expresses, it forces the positive supercoils back into the TetR-GFP coding sequence. When this happens, TetR-GFP is no longer able to express and its transcript region is thus available to LacI-RFP as a downstream region for dissipating its internal torsional stress. Thus, by propagating supercoils into its neighboring gene, LacI-RFP exerts a form of negative feedback independent of transcription factor-mediated repression.

This explains why the convergent toggle is able function in regimes where IPTG and aTc are simultaneously high or low. When IPTG and aTc are both present in high concentrations, the attenuation of transcription factor repression is compensated by the presence of supercoiling mediated repression. Thus, even though LacI and TetR are not as effective in repressing their respective promoters, the extra layer of feedback allows the convergent toggle to decide on a dominant state (LacI-RFP). Similarly, in the low parameter regime, even though both repressors are strong, the additional feedback from supercoiling favors one gene or the other (an enhancement of the winner-takes-all or XOR logic) and evidently improves the ability of the toggle switch to again allow LacI-RFP to dominate over TetR-GFP. Thus, there is a multi-layer feedback effect introduced by supercoiling in the convergent orientation, conformal with the intended feedback architecture of the toggle switch. In this way, we see that compositional context can be a powerful tool for encoding feedback in synthetic gene networks.

2.3 Discussion

2.3.1 The Link Between Compositional Context Effects and Growth Phase: Temporal Aspects of Compositional Context

Our experimental data, as well the outcomes of several gyrase treatment experiments, support a model of how supercoiling dynamics affect transcription. Depending on its com-

positional context, the supercoiling state of a gene can be affected by the propagation of supercoiling from nearby coding regions. In this way, supercoiling couples the activity of two neighboring genes. The strength of that coupling and its impact on the temporal dynamics of gene expression depends on the orientation of the genes and what part of the gene is exposed to torsional stress from the neighboring gene. On the whole, these features of our model are able to recapitulate the *in vitro* and *in vivo* trends observed at steady-state, but do not account for aspects of how gyrase and topoisomerase levels are regulated during different growth states.

An interesting facet of these context effects are the temporal dynamics of supercoiled genes, topoisomerase concentrations, and their dependence on cell culture growth phase. Specifically, as *E. coli* cells transition from exponential to stationary phase, plasmid DNA exhibited significantly less negative supercoiled DNA. Balke & Gralla [3] showed that up to ten negative supercoils could be lost in the pBR322 plasmid in stationary phase cells grown in LB. Thus, gyrase activity (which maintains negative supercoiling) is attenuated as cells approach the end of their exponential growth phase. These findings are corroborated by our data; we also saw that compositional context differences become increasingly dramatic just as cells complete their exponential growth phase (Figure 2.10C-D).

In this work, we have not made a point to model the temporal dynamics of gyrase and topoisomerase as a function of cellular growth phase since doing so would require rigorous characterizations of gyrase and topoisomerase concentrations through the entire growth cycle. Another interesting extension would be to examine how gyrase dynamics and the compositional context of genes in core metabolic systems affect or modulate the dynamics of metabolism.

2.3.2 Supercoiling Dynamics Dominate Genetic Context Effects

Our analysis considered supercoiling as the physical basis for generating expression differences. In past work, the primary context effects considered in designing synthetic biocircuits

are the effects of terminator leakage and transcriptional interference from overlapping promoter and RBS elements. We claim that supercoiling is the dominant source of compositional context effects observed in our data, justified by the following observations.

First, we see consistent differences in gene expression, even when only one gene is induced and the other remains repressed. If terminator leakage and transcriptional interference were the source of compositional context effects, we would not expect to see any effects in the case of single gene induction.

However, there is more than a 2-fold difference in expression between divergent expressed CFP and its single reporter counterpart (sense or anti-sense, compare Figure 2.10C with Figure 2.3F-G). The physical presence of a neighboring gene has an effect, even if it is not transcriptionally active. Thus, transcriptional interference via terminator leakage does not explain the data.

Second, if transcriptional interference were the primary driver for context effects, we would expect convergent oriented genes to achieve far weaker levels of expression than divergent or tandem orientation. In theory, transcribing polymerases that managed to leak through two terminators (Larson et al. [53] characterized the termination efficiency of our terminators at 98%) would collide in the convergent orientation, leading to an increase in abortive transcription events or transcriptional stalling.

Admittedly, we see from our *in vivo* characterizations that early-log phase CFP and RFP expression is weaker in the convergent orientation than the divergent and tandem orientation. The fly in the ointment for this argument is that *both* CFP and RFP expression are *higher* (see Figure 2.4 and Figure 2.10C-E) in the doubly induced case than in the singly induced case in early log phase, which contradicts the predictions of transcriptional interference theory.

Transcriptional interference also does not account for the sudden rise in convergent oriented expression relative to divergent orientation as cells approach the end of their exponential growth phase. Supercoiling theory, on the other hand, predicts that as gyrase activity wanes, the promoter regions of divergently oriented genes become more positively super-

coiled, which inhibits their activity. This positive supercoiling originates from the RFP promoter as it transcribes in the anti-sense direction, thus asymmetrically inhibiting CFP expression and favoring RFP expression (see Figure 2.10E).

Thirdly, consider the differences in expression *in vitro* of convergent, divergent, and tandem transcribed RFP and CFP (Figure 2.6). Our experimental characterizations *in vitro* control for variations for plasmid copy number (as a function of orientation), since plasmid replication does not occur in the transcription-translation (TX-TL) cell-free system [73]. Nonetheless, we see that levels of CFP (and RFP) expression in the convergent and divergent orientation differ by nearly 300 nM (and 500 nM) when purified directly from cells in their natural supercoiled state, whereas treating with gyrase to eliminate positive supercoiling decreases the difference by nearly 70% in both genes. Relaxing positive supercoiling in divergently oriented RFP and CFP with gyrase also allows expression levels comparable to tandem orientation prior to gyrase treatment, while treating tandem oriented RFP and CFP enables expression levels higher than both post-treatment and pre-treatment convergent oriented CFP and RFP. Taken in whole, these observations confirm that the dominant physical process driving the effects of compositional context is supercoiling.

2.3.3 The Role of Compositional Context in Synthetic Biocircuit Design

Our findings show that compositional context significantly alters gene expression in synthetic gene networks. When appropriately harnessed, compositional context can be used to strengthen or enhance existing feedback loops in the intended biocircuit design. These findings validate prior analysis underscoring the value of accounting for compositional context effects in synthetic biocircuit design [9].

Broadly speaking, there are many levels of abstraction and ways to define compositional context. Cox et al. investigated how different regulatory elements in existing promoters could be assembled in distal, core, and proximal sites to define a library of new combina-

torial promoters [15]. Similarly, Mutalik et al. showed that the compositional context of a ribosome binding site, specifically sequences downstream of the ribosome binding site could have a significant impact on the effective binding strength of the ribosome [69]. Using a bicistronic design approach, they showed they were able to better insulate against downstream sequence variability to produce predictable parts. These are examples of the importance of understanding and insulating against *intragenic* compositional context.

The results of our experimental studies emphasize the importance of understanding *intergenic compositional context* effects, i.e. composition of entire genes. We have seen that compositional context effects can cause variations of 3-4 fold of the same gene (promoter, RBS, coding sequence, etc.) simply by rearranging its orientation and the orientation of other neighboring genes. The significance of these outcomes raise an important issue. As *intragenic* context, e.g. choice of BCD, promoter design, polycistronic design, are optimized to produce a functional gene cassette with model-predicted gene expression levels [52, 70], how do we ensure these predictions are not confounded by intergenic context as genes are composed?

One solution is to separate genes that need to have precise regulated expression levels on to different plasmids. However, the drawback of this approach is that separating genes on different plasmids introduces imbalances in gene copy number, which in turn can lead to additional design-build-test cycles to rebalance circuit dynamics. Also, it is often the case that there are too many genes in a biocircuit to isolate individually on separate plasmids. In such settings, the findings of this work are important to consider, as they can be used to inform how to optimally compose adjacent genes.

The effects of adding spacing sequences between genes are complex. Specifically, we varied the amount of spacing between mSpinach and MG aptamer in convergent, divergent, and tandem orientation by adding increments of 100 bps between genes and found that spacing did not have a monotonic effect on decreasing the fold-change across orientations (see Figure 2.8A-C). Most unusual was the sudden drop in signal observed in the divergent and tandem

orientation, but not in the convergent orientation with 450 bp of spacing between the two genes. It is possible that since the persistence length of DNA is 150 base pairs, 450 base pairs of spacing facilitates formation of plectonomes (with DNA loops consisting of three 150 bp domains) in the spacing region, which induce torsional stress and inhibit formation or movement of the transcription bubble.

In general, genes responded well to induction when induced one by one, though their raw expression levels varied depending on compositional context. This may explain why some circuits in the past have been successfully engineered, with little consideration given to the effects of supercoiling and compositional context. For example, the original toggle switch (oriented divergently) was designed to respond and latch to the presence of a single inducer [31] — it did this well and latched to LacI or TetR dominant states. In contrast, the threshold detection abilities of the toggle were not explored.

Likewise, the fold-change in 'off' vs 'on' states of three input and four input AND gates developed by Moon and colleagues [67] was strongest when comparing singly induced expression levels against the corresponding fully induced state. Interestingly, the four layer and three logic gates in these biocircuits were compositionally composed so that no two genes involved in any constituent layer of logic were placed adjacent to each other. Pairs of genes involved in logic gates were always separated by an auxiliary backbone gene or placed on separate plasmids. Overall, the success in this work suggest that genes can be insulated by inserting short 'junk' transcriptional units in between each other. Engineering approaches for attenuating compositional context effects are a subject of future research.

2.4 Experimental Procedures

2.4.1 Plasmid Construction, Assembly, and Strain Curation

Plasmids were designed and constructed using either the Gibson isothermal DNA assembly technique [34] or Golden Gate DNA assembly approach [25] using BsaI type II restriction

enzyme. All plasmids were cloned into JM109 *E. coli* (Zymo Research T3005) or NEB Turbo *E. coli* (NEB C2984H) strains and sequence verified. Sequence verified plasmids were transformed into MG1655Z1 and MG1655ΔLacI (also lacking TetR) strains of *E. coli*. All plasmids with ColE1 replication origin were transformed and cloned at 29 C to maintain low copy number of the ColE1 replication origin. Sequence verified colonies were grown in LB and the appropriate antibiotic and stored as glycerol stocks (17 % glycerol) at -80° C.

2.4.2 Single Cell Fluorescence Microscopy

Based on the principles elucidated by Han et al.[40], we ran all our experiments at 29° C when imaging mSpinach. Cells were revived from glycerol stock overnight at 29° C in LB, diluted to an OD of 0.1 and recovered for 2 hours in log-phase. Cells were then diluted to an density of approximately 5×10^6 cells/mL of LB and loaded into a CellASIC plate. Separate solutions for flowing LB with 200 μM DFHBI and LB with 200 μM DFHBI and 1 mM IPTG were prepared and loaded into reagent wells in the CellASIC ONIX B04A plate for imaging.

Fluorescence and bright field images from time-lapse microscopy were cropped using ImageJ and analyzed in MATLAB with Schnitzcell [104]. For characterizing coexpression of mSpinach and MG RNA aptamer, we used single cell agar pad microscopy, with all cells grown shaking at 29° C in a 96 well plate from overnight recovery until they reached log-phase (∼4 hours). Induction occurred by transferring 10 μL of cultures into another 96 well plate into 350 μL of LB with 1mM IPTG and 200 ng/mL aTc.

2.4.3 Plate Reader Experiments

For plate reader experiments, all cultures were revived from glycerol stock at 37° C in LB and the appropriate antibiotic, followed by redilution to OD 0.05-0.1, recovered at log-phase for 2 hours at 37° C, and then pipetted into a 96 square well glass bottom plate (Brooks Life Sciences MGB096-1-2-LG-L) with the appropriate media, antibiotic and inducer. All

measurements were taken on Biotek Synergy H1 plate readers, using the internal monochromomator with excitation (and emission) wavelengths for mSpinach, MG aptamer, CFP, and RFP at 469 nm (and 501 nm) at gain 100, 625 (and 655 nm) at gain 150, CFP at 430 nm (and 470 nm) at gain 61 and 100, RFP at 580 nm (and 610 nm) at gain 61 and 100. For RNA aptamer imaging, all *in vitro* and *in vivo* experiments were performed at 29° C with 200 μM DFHBI (for mSpinach) and 50 μM of Malachite Green dye.

2.5 Experimental Procedures and Data Analysis

2.5.1 Plasmid Assembly

Initial efforts to characterize orientation effects involved cloning plasmids with no spacing DNA between genes. We used Gibson assembly to build these plasmids and naturally found that in the convergent and divergent orientation, the primers used to amplify overhangs had strong secondary structure, which reduced cloning efficiency. Thus, we inserted a minimum of 150 base pairs of randomly generated DNA. DNA sequences were randomly generated in MATLAB, using a custom script and the function `randi()` and subsequently screened for secondary structure in Geneious, a gene designer software. All spacer sequences between genes were determined to have no hairpins or any predicted secondary structure at 37° C before use in cloning workflows.

To construct mSpinach and MG RNA aptamer reporter plasmids, we ordered 500 bp Integrated DNA Technologies gBlocks containing the mSpinach and MG RNA aptamer coding sequences in convergent, divergent, and tandem orientation. Backbones and DNA inserts were amplified and prepared at equimolar concentrations in an isothermal Gibson Assembly, incubated for one hour at 50 C, following the methods of Gibson et. al [34]. Gibson products were subsequently transformed into JM109 Zymogen *E. coli* using a quick-transform protocol, plated at 29° C on LB agar plates with 100 μg/mL of carbencillin. Colonies were screened using standard colony PCR techniques, sequence verified using Operon Sequenc-

ing's overnight sequencing service (both Standard and Power Read products). All strains were sequence verified both in JM109 and experimental strains of MG1655Z1 and MG1655Δ LacI.

To construct mSpinach and RFP plasmids, we used a similar approach as described above, except that we used an RFP coding sequence derived from BglBrick plasmid (pBbE5k-RFP), amplified as a linear double stranded DNA molecule compatible with Gibson assembly. We used an analogous approach to construct CFP and RFP reporter plasmids on the ColE1 backbone. To switch backbones (p15A with chloremphenicol resistance marker) and construct CFP and RFP sense and anti-sense plasmids, we used Golden Gate assembly with BsaI-HF enzyme from New England Biolabs (NEB R3535L). All Golden Gate parts were constructed using an internal protocol with standardized four base pair overhangs. Colonies were screened and sequence verified following the same techniques used for plasmids built by Gibson Assembly. Finally, all plasmids developed for this thesis were sequence verified both from Qiagen purified plasmid and as glycerol stock (using Operon's DNA prep service). Sequence verified strains were stored in 17 % glycerol stocks at -80 C with LB and either 100 μg/mL of carbencillin or 34 μg/mL of chloremphenicol.

Imaging RNA aptamers: quantitating mSpinach expression using single cell time-lapse fluorescence microscopy, agar pad microscopy and plate readers

In our preliminary tests, we quickly found that mSpinach RNA aptamer is not particularly bright, compared to GFP, RFP, and other standard fluorescent proteins. Moreover, its brightness depends on the operating temperature of the experiment [40], since the steady state folding configuration of the mSpinach RNA aptamer depends on temperature. We found that mSpinach signal at 200 μM DFHBI (Lucerna Technologies) was nearly undetectable at 37° C. To minimize photobleaching of mSpinach, we developed a custom Python script to interface with MicroManager [20], employing the fast shutter of the XFO-citep 120 PC (8 ms resolution) to time exposure of the mSpinach expressing cells to light. To maintain

an operating temperature of 29° C we used a custom-built microscopy incubation chamber with a World Precision Instruments Heater controller.

Once the microfluidic plate (EMD Millipore Cell ASIC ONIX B04A) was thermally equilibrated, cells were loaded into the imaging chamber and trapped using a loading protocol provided by EMD Millipore to a density of about 3 cells per field of view. Fluorescence microscopy imaging was performed on an Olympus IX81 inverted fluorescence microscope using a Chroma wtGFP filter cube (450/50 BP excitation filter, 480 LP dichroic beamsplitter, and 510/50 BP emission filter), with an XFO-citep 120 PC light source at 100 % intensity and a Hamamatsu ORCA-03G camera. Following the recommendations of [40], we limited imaging frequency and exposure to every 10 minutes and for 200 ms, respectively. All experiments were conducted with an untransformed control strain of MG1655Z1 *E. coli* in a parallel microfluidic chamber, to quantify background cell fluorescence in DFHBI. For Figures 2.2B, 2.3C-D, and Figure 2.9B we segmented and tracked single cell traces of mSpinach (or RFP and CFP) fluorescence using Schnitzcell [104] and subtracted background fluorescence from each experimental strain. For each point in time, background fluorescence was defined as the maximum of background chamber fluorescence (quantified using ImageJ as mean fluorescence in a nearby non-occupied area of the microfluidic chamber housing the experimental strain) and background cell fluorescence of the control strain for each frame. The majority of background fluorescence was defined by the background fluorescence in cells, with infrequent fluctuations in background fluorescence due to slight perturbations to the autofocus plane.

We also found that fixing cells with paraformaledehyde lead to inconsistent RNA aptamer fluorescence, with unusually high levels of fluorescence in the negative control (especially in the MG aptamer channel). While fixing cells traditionally allows fixation of protein dynamics, this is not true for imaging mSpinach and MG aptamer. It is possible that fixation alters the permeability of the membrane and enables excessive buildup of MG oxalate dye, which at high concentrations non-specifically binds to RNA molecules in the fixed cell. For this reason, our experimental technique involved imaging of live cells on agar pads, moving

as quickly as possible from agar pad to agar pad, and well after the dynamics of induction had reached steady-state.

In contrast, imaging mSpinach in the cell-free expression system developed by Shin and Noireaux [84] required relatively little effort. Fluorescence quantification was performed on a Synergy H1 Biotek plate reader with 469 nm wavelength excitation, 501 nm wavelength emission. Since TX-TL reactions are typically run at 29 C, this further facilitated formation of the mSpinach RNA aptamer in the 32-2 configuration, see the Online Material for Paige et. al. [76]. We did notice that imaging more frequently than 15 minutes had an effect on the dynamics of mSpinach (presumably due to photobleaching), hence we ran all experiments with 15 minute imaging frequencies.

It is important to note that production of mSpinach in 10 μL bulk volume *in vitro* reactions allows for approximately 10^9 more copies of mSpinach than produced in a cell, we speculate this greatly increases the detectability of mSpinach signal over *in vivo* assays. We found imaging mSpinach in dense cultures (OD ≈ 1) also produced significant signal above background. Thus, the primary challenges of working with mSpinach is its relatively weak signal per single cell. We anticipate that using the latest version of mSpinach (mSpinach2) or dBroccoli in future tests will greatly improve signal [28].

2.5.2 Analysis of Plate Reader Data

To generate the data plotted in Figure 2.4, Figures 2.8B-C, 2.10 and 2.11, we extracted data from the Biotek H1 Synergy plate readers using the Gen5 software package, exported to MATLAB matrices for optical density (OD) and fluorescence intensity in either mSpinach (469 excitation, 501 emission, gain 100), GFP (485 excitation, 525 emission, gain 61) CFP (430 excitation, 470 emission, gain 61) and RFP (580 excitation, 610 emission, gain 61 or 150) channels with inverted (bottom-up) fluorescence acquisition. Each sample was background subtracted, normalized by OD, plotted either as a single time point $t = 9.2$ hours corresponding to the tail-end of exponential growth phase or as complete time traces from

$t = 0$ to $t = 11.7$ hours ($t = 700$ minutes). Each strain was grown in duplicate in MatriPlate (Brooks Life Science Systems MGB096-1-2-LG-L) 96 well square well glass bottom plates at 500 μ L volumes.

Similarly, for toggle switch data analysis we followed the approach outlined above. We note that given our choice of ribosome binding sites for GFP and RFP (BCD1 and BCD9 respectively), expression of GFP and RFP was weaker to avoid ribosomal loading effects. Thus, we did not see significant signal until 8 hours after initial induction. Signal increased monotonically throughout the experiment, varying depending on the balance of IPTG and aTc induction. Data plotted in Figures 2.7 were background subtracted and normalized by OD.

To estimate RNA aptamer and protein expression in the TX-TL system, we used data from prior calibration experiments, titrating purified fluorescent protein or RNA aptamer and quantitating expression in the Biotek. Each Biotek was calibrated independently; the results of the calibration were used to back out fluorescent protein from raw AFUs, after background subtraction.

2.5.3 Flow Cytometry Experiments and Data Analysis

Flow cytometer experiments were performed using a BD Biosciences Flow Assisted Cell Sorter (FACS) Aria II Flow Cytometer to quantify GFP and RFP fluorescence. GFP fluorescence was detected using a 488 nm laser and 530/30 nm internal bandpass filter while RFP fluorescence was detected using 561 nm laser and a 610/20 nm internal bandpass filter. Each plasmid strain (featuring convergent or divergent orientation of the modified pIKE107 Gardner Collins toggle switch) in either MG1655 *E. coli* or MG1655 Δ LacI *E. coli* was plated on cells from clonal glycerol stocks, grown at 37° C on selective media agar plates overnight. Three colonies were picked from each plate to seed replicate cultures for the experiment. All cell cultures were grown in LB media with carbencillin at 100 μg/mL at 37° C. Cultures were induced with either 50 ng/mL of aTc or 1000 μM of IPTG for 5 hours (defined as the

latching period from $t = -5$ to $t = 0$ hours). After latching, cells were diluted with a dilution factor of 1000x, in approximate 8 hour intervals, in selective LB media from $t = 0$ to $t = 48$ hours in the experiment. At $t = 0, 24$, and 48 hours, cell cultures were rediluted and grown for two hours to reach exponential growth phase and rediluted 1:10 in 1x phosphate buffer saline solution. As a negative control, we quantified GFP, RFP, and CFP fluorescence of an untransformed strain of MG1655 *E. coli* as well as cell-free 1x PBS stock to determine forward and side-scatter gating parameters for background particulate matter.

All flow cytometry data was processed using the FlowJo Software. Cells were gated using an ellipsoidal gate of forward and side-scatter values. We utilized live-gating during data acquisition to obtain approximately 20,000 events. All distributions were plotted as modal percentage vs. GFP intensity (in arbitrary fluorescent units). Modal percentage for a given GFP intensity is defined as the ratio of cell count for the given GFP intensity bin normalized by the cell count for the modal GFP intensity bin, multiplied by 100. This method of plotting eliminates the variability of total counts in sub-populations after gating, while still portraying important features of the distribution such as mode, modal variance and modal kurtosis.

2.5.4 Note on Linear DNA Experiments

Since supercoiling buildup is only possible in certain scenarios, e.g. in the presence of chromosomal binding proteins, topologically constrained plasmids or linear DNA tethered to a scaffold [13], we used linear DNA to explore how orientation affects gene transcription in the absence of topological barriers. On linear DNA, divergently oriented mSpinach and MG aptamer have relatively little supercoiling buildup since the linear ends of the DNA enable free rotation of the DNA about the helical axis as transcription occurs.

On tandem oriented DNA, since gene transcription occurs in the same direction for both genes, there is less torsional stress between the two genes and rotation of the downstream gene (mSpinach) enables relaxation of any torsional buildup for itself. However, MG aptamer

expression can be adversely affected if mSpinach is actively transcribed, since the viscous drag of the open complex on the mSpinach sequence may inhibit free rotation of DNA. Thus, tandem oriented MG aptamer can theoretically experience buildup of positive supercoiling downstream of its gene sequence (and upstream of the Lac promoter), depending on the occupancy of open complexes on the mSpinach gene sequence.

Expression of convergent oriented genes on linear DNA can result in buildup of local torsional stress in between the two genes, with positive supercoiling downstream of the sense cassette and negative supercoiling downstream the anti-sense cassette (see Figure 2.5). In theory, the negative supercoiling could facilitate expression of the anti-sense gene, while the positive supercoiling would interfere with expression of sense gene. To test this, we amplified linear DNA fragments of mSpinach and MG aptamer in convergent, divergent, and tandem orientation and gel purified each sample. We also amplified single gene linear DNA controls containing either mSpinach or MG aptamer. After an additional PCR purification step (to wash out any salt content from gel purification), we expressed convergent, divergent, and tandem mSpinach and MG aptamer from equimolar concentrations of linear DNA (see Figure 2.8D-F).

Remarkably, we observed that convergent oriented mSpinach expressed significantly higher than divergent, tandem oriented, or the single gene control linear DNA. The expression of both divergent and tandem oriented mSpinach was comparable to the mSpinach control and to each other, suggesting that in the absence of topological barriers, differences in expression between tandem and divergently oriented mSpinach were significantly attenuated.

In contrast, convergent oriented MG aptamer expressed at levels comparable to the control while divergent oriented MG aptamer was expressed at slightly higher concentrations. Most interesting was the complete shutoff of MG aptamer expression in the tandem orientation. This outcome came as a surprise, since we could think of no other hypothesis involving transcriptional interference that could explain the loss of MG aptamer expression. These results, combined with the strong responses of plasmids to gyrase treatment, further

validated supercoiling as the source of compositional context effects.

2.5.5 Note: Fitting Hill Functions for Different Gene Orientations

We modify the standard Hill Equation to include a term for promoter leakiness that is independent from the dynamic range due to inducer concentration. The equation for expression due to a promoter with some leakiness and Hill function-type response to an inducer chemical is given as:

$$f([I]) = l + \alpha \frac{[I]^n}{K_m^n + [I]^n},$$

where [I] is inducer concentration, l is leaky expression, α is the amplitude of expression due to inducer, n is the apparent cooperativity of the response to inducer, and K_m is the concentration at which induction is half maximal. Thus, the maximum total expression upon full induction is given by:

$$V_{max} = l + \alpha$$

All four parameters were fit using RFP/CFP expression data shown in Figure 2.5. Both RFP/CFP induction functions were fit for the case in which the other gene is fully induced using the Matlab function `nlinfit`. RFP was fit to the data that varies aTc (1.56 ng/mL to 200 ng/mL) while keeping IPTG at 1000 nM (left column of Figure 4). Similarly, CFP was fit to the data that varies IPTG (7.85 nM to 1000 nM) while keeping aTc at 200 ng/mL (top row of Figure 4). Fits along with experimental data points were plotted for all three orientations for both RFP and CFP (Figure S2A-B).

2.6 Note Comparing Toggle Switch Performance of the Original (Divergent) Gardner-Collins Toggle and Convergent Gardner-Collins Toggle

While our experimental results with *de novo* toggle switches showed compositional context can reinforce the toggle's feedback architecture (Figure 2.7), we also wanted to compare performance of the convergent toggle with the canonical toggle switch developed by Gardner et al. [31]. To this end, we modified the original pIKE107 toggle switch, which was assembled with divergent oriented LacI and TetR-GFP genes, to include an RFP coding sequence downstream of LacI. All RBSs and promoter sequences were left as originally cloned. Next, we used restriction digest DNA assembly to convert the original toggle switch into a convergent toggle switch. All cloning was done using the pIKE107 plasmid backbone. Plasmids were grown up in cloning strains, sequence verified, and transformed into both MG1655ΔLacI and MG1655 *E. coli*.

Our first experimental test was to confirm that the improved thresholding properties in the convergent toggle were preserved, independent of plasmid backbone identity. We immediately discovered in preliminary experiments that the RBS for TetR-GFP was significantly stronger than the RBS for LacI-RFP, so to draw a fair comparison with our results in Figure 2.7, we attenuated the concentrations of IPTG by an order of magnitude. We found that convergent oriented Gardner-Collins toggle exhibited strong XOR logic in the high IPTG-high aTC regime and low IPTG-low aTc regime (Figure 2.11A), consistent with our results in Figure 2.7. In contrast, the divergent Gardner-Collins toggle did not exhibit strong XOR logic in the high IPTG-high aTc regime and the distinction between high GFP-low RFP and high RFP-low GFP states was generally not as clear (see Figure2.11B).

Our second experimental test was a stability test of the memory properties of the toggle. We found that both in the MG1655ΔLacI *E. coli* and MG1655 *E. coli* strain there were significant differences between the convergent and divergent toggle. In particular, we observed

cells transformed with the divergent toggle tended to drift from its high-GFP state into a lower GFP state over time, while the convergent toggle tended to maintain a high-GFP state throughout the course of the entire 53 hour experiment (48 hours post-latching and 5 hours of latching.)

These results can be explained, once again, using our model of supercoiling and its role in strengthening negative feedback in the toggle switch. In the divergent orientation, the supercoiling propagated by proximal promoters generally results in decreased expression of the repressors. This results in weaker repression, since both promoters are affected by the presence of supercoilng (reference the results in Figures 2.2, 2.3, 2.5). This leads to overall reduction in reporter signal, but also leaky repression. As time transpires post-induction (growing in media without inducer), this allows the population of cells to drift from the high-GFP state.

In the convergent orientation, once TetR-GFP is expressed in the high state, it continuously dominates by propagating supercoils into the LacI-RFP transcript region. These supercoils may impose a higher activation threshold for the LacI-RFP state, thus keeping it effectively off throughout the course of the experiment. Remarkably, we observed that even in the presence of constitutively expressed genomic LacI repressor in MG1655 *E. coli*, the convergent toggle did not drift significantly from its initial state. This result can be interpreted as enhanced disturbance rejection capabilities of the convergent toggle; it requires a significant amount of LacI to flip the toggle switch to a high LacI-RFP state. Small amounts of LacI are not sufficient to overcome the combined repression barrier imposed by supercoiling and TetR repression. In this way, the convergent orientation of the toggle reinforces the feedback architecture of the toggle switch, resulting in improved memory, disturbance rejection, and better thresholding performance.

Deriving Supercoiling Dynamics in a ODE Model of mSpinach and MG Aptamer Expression

Here we explore a detailed model for describing the interplay of supercoiling and gene expression. The motivation to do this arises from 1) experimental results which strongly suggest that supercoiling and not transcriptional interference is the primary cause of differences observed in mSpinach, CFP, RFP, and MG aptamer expression across different gene orientations and 2) the need for a mathematical modeling framework that describes how the temporal dynamics of gene expression vary as a function of supercoiling state and neighboring gene activity.

We consider three structural phenomena that arise in supercoiled DNA: positively supercoiled DNA, negatively supercoiled DNA, and R-loop formation [19] of the RNAP-DNA elongation complex in negatively supercoiled DNA. We begin with the basic premises of the twin-domain supercoiling models [57], namely that when a gene is transcribed, negative supercoiling is introduced upstream of the open complex and positive supercoiling is introduced downstream of the open complex. We introduce several concepts from the supercoiling literature [19, 51, 57, 74, 78].

Definition 1. *We define the constant* $h_0 = 10.5$ *to be the number of DNA base pairs involved in a single turn of a B-form DNA molecule in its natural state.*

Definition 2. *We define the linking number* α_{LN} *of a region of DNA to be the number of supercoiling turns in that region.*

Definition 3. *We define the supercoiling density* σ_X *of a region of DNA* X *of* N *base pairs length as* $\sigma = \alpha_{LN}/N$.

Thus, we will assume that the plasmid DNA in our experiments is in its natural B-form configuration. Of course, by simply defining $h_0 = 11$ or $h_0 = 12$, it is possible extend our results to consider DNA in its A and Z form respectively.

It is important to note the notions of positive and negative supercoiling correspond to the notions of left-handed twist and right-handed twist, respectively, and are well defined as long as the direction along which gene expression occurs is specified and fixed. For example, a gene expressing in the sense direction (as considered in the model by Wang and Liu [57] and the recent analysis of Chong and Xie [13]) creates *right-handed twist* or negative supercoiling, conformal with the natural twist or direction of turn in a DNA double helix, downstream of the transcription bubble and *left-handed twist*, or positive supercoiling, upstream of the transcription bubble. Thus, the convention that negative-supercoiling builds upstream of gene expression and positive supercoiling downstream, is sensible only when considering 'sense' transcription.

When a gene expresses in the anti-sense direction, then using the reference frame defined by sense transcription and the right-handed twist of DNA, we note that unless we rotate the axis of the reference frame 180 degrees, the buildup of supercoiling downstream of anti-sense transcription is still *right-handed* (i.e. negative) and the buildup of supercoiling upstream of anti-sense transcription is still *left-handed* (i.e. positive) (see Figure 2.5 for a visual example).

A simple way to prove this is to construct a physical model of a supercoiled double-helix. Take two ropes, twisted into a double helix with right-handed twist. Note that defining the twist of the double helix as right-handed inherently imposes directionality in your rope (e.g. left to right or bottom to top, your thumb pointing in the direction of right or top). Tie both ends to a topological barrier, e.g. by connecting them to form a loop (like a plasmid) or fused to two separate posts, so that the twist internal to the double helix cannot dissipate past these barriers. Simulate a transcription bubble by pulling the two ropes apart and notice that preceding the bubble (opposing the direction that your thumb pointed) you will see the generation of additional right-handed twist and succeeding the bubble you generate left-handed twist (conformal with the direction of your thumb). Notice that the bubble could have been formed by unwinding the double helix left to right (sense transcription) or right to left (anti-sense transcription). However, it does not matter what direction we unwound the

DNA to form the bubble; the end result is the same — negative supercoiling or right-handed twist preceding the bubble and positive supercoiling (or left-handed twist) succeeding the bubble. Thus, the original twist of the DNA, not the direction of bubble propagation, defines what type of supercoiling builds up preceding and succeeding a transcription bubble.

It is important to clarify that we are not declaring the default supercoiling state of DNA *in vivo* as generally negatively supercoiled. Rather, we are referencing the classical convention that states that the double helix inherently has right-handed curl or twist [19]. Moreover, we make no assertions about the exact amount of additional negative or positive supercoiling introduced surrounding a sense transcription bubble or an anti-sense transcription bubble. Various aspects of the nature of supercoiling build-up and propagation have yet to be characterized fully via experiments, such as the rate of propagation of supercoils, the spatial distribution of supercoils succeeding or preceding a transcription bubble, and how DNA promoter and transcript sequence pertain to the rate at which supercoils are introduced. While our model is thus constructed with the capacity for quantitative prediction, until it is supported by robust estimates of physiological parameters, it is meant provide a mechanistic hypothesis for explaining the effects observed in our *in vivo* and *in vitro* experiments as opposed to exact predictions.

When two genes are present, e.g. in the convergent orientation, the intergenic region between the two genes becomes exposed to both left-handed (positive) and right-handed (negative) twist. It is important to note that left-handed (positive) and right-handed twist (negative) do not simply cancel out — the arbitrary nomenclature of positive or negative twist does not confer the same algebraic consequences of adding positive and negative numbers. Rather, when the a right-handed DNA double helix experiences torsional stress from simultaneously introducing both left-handed and right-handed twist from two opposing point sources (e.g. transcription bubbles in convergent orientation), the two twists define opposing forces that meet each other at some kink point between the two point sources. The outcome is not annihilation of positive and negative supercoiling but rather the transition of the kink

along the longitudinal axis of DNA until an *equilibrium* is achieved, i.e. the forces driving left-handed twist through the kink are equally balanced by forces driving right-handed twist through the kink. At equilibrium the *net* force is zero but this does not implicate in any way that the presence of right-handed or left-handed coils have been annihilated. With each transcriptional event or binding of a gyrase or topoisomerase to modulate the surrounding DNA's supercoiling state, the equilibrium position of the kink is correspondingly adjusted.

With these observations in order, we now consider the scenario when two non-overlapping genes are adjacent to each other in varying orientations. For the purposes of our model, three regions of DNA for each gene will be of interest, the promoter of a transcriptional unit, the coding sequence of a transcriptional unit, and the intergenic spacing region between adjacent genes in our constructs. Supercoiling has been experimentally demonstrated to affect both the processes of transcription initiation and transcription elongation. Thus, we make a point to distinguish and keep track of the supercoiling states of both the promoter and coding transcript. For simplicity of exposition, we do not explicitly model the supercoiling density of the intergenic spacing region, however, our models will implicitly assume that the spacing region is able to absorb supercoils propagated from upstream or downstream transcription events up to the kink (if present).

For notation, when modeling the RNA aptamer plasmids, we will use TL_X where $X = G$ or S to denote the length of the MG and mSpinach RNA aptamer transcript respectively, EC_X to denote the elongation complex formed while transcribing gene X, R to denote RNA polymerase, PL_X to denote the length of the p_{Lac} and p_{Tet} promoters, and N_S the length of the intergenic spacing region of noncoding DNA between genes. Similarly, we will use the subscript $X = RF$ or CF to indicate the parameter of interest pertains to the coding sequence for RFP or CFP respectively.

Convergent Orientation Model

In the convergent orientation, promoters face each other and as both genes express, positive supercoiling propagates from the sense transcription bubble into the intergenic spacing region and negative supercoiling propagates from the anti-sense transcription bubble (see Figure 2.4) to form a kink in between the two genes. Where standard transcription translation models of gene expression assume constant rates of transcription initiation and transcription elongation, we now make explicit the dependencies of these rates on supercoiling state. The chemical reaction network for this orientation is given as:

$$\emptyset \xrightarrow{\rho_L} LacI \qquad \emptyset \xrightarrow{\rho_T} TetR$$

$$LacI \xrightarrow{\delta_p} \emptyset \qquad TetR \xrightarrow{\delta_p} \emptyset$$

$$LacI + IPTG \underset{k_{ua,L}}{\overset{k_{a,L}}{\rightleftharpoons}} aLacI \qquad TetR + aTc \underset{k_{ua,T}}{\overset{k_{a,T}}{\rightleftharpoons}} aTetR$$

$$p_{\text{Lac}} + LacI \underset{k_{u,L}}{\overset{k_{seq,L}}{\rightleftharpoons}} pLacC \qquad p_{\text{Lac}} + TetR \underset{k_{u,T}}{\overset{k_{seq,T}}{\rightleftharpoons}} pTetC$$

$$pLacC + IPTG \xrightarrow{k_{a,L}} pLac + aLacI \qquad pTetC + aTc \xrightarrow{k_{a,T}} pTet + aTetR$$

$$R + p_{\text{Lac}} \underset{k_r}{\overset{k_f(\sigma_{p,S})}{\rightleftharpoons}} CC_S \qquad R + p_{\text{Tet}} \underset{k_r}{\overset{k_f(\sigma_{p,G})}{\rightleftharpoons}} CC_G$$

$$CC_S \xrightarrow{k_{open}} EC_S \qquad CC_G \xrightarrow{k_{open}} EC_G$$

$$EC_S \xrightarrow{k_{cat}(\sigma_{t,S})} mS + R + p_{\text{Lac}} \qquad EC_G \xrightarrow{k_{cat}(\sigma_{t,G})} MG + R + p_{\text{Tet}}$$

$$mS \xrightarrow{\delta_m} \emptyset \qquad MG \xrightarrow{\delta_m} \emptyset$$

We note here, that for simplicity, we model LacI and TetR as naturally occurring in their tetrameric and dimeric forms. The complex aLacI and aTetR denote the inducer-bound forms of LacI and TetR that are unable to bind to their target promoter. When LacI and TetR bind their respective promoter, we denote them with pLacC and pTetC to indicate the promoter is sequestered from transcriptional processes.

We now derive an expression for $\sigma_{p,S}(t)$ by first considering the effects of transcription on the supercoiling density of the transcript. Consider the collection of plasmids present in the cell or volume of cell-free extract. Consider a small time interval $[t, t+\epsilon]$ for small $\epsilon > 0$. Suppose that $\Delta_{LN,t}$ turns are introduced with the production of each mSpinach transcript and that $x\Delta_{LN}, t$ number of turns are introduced into the transcript as x mSpinach molecules are produced. Simultaneously, we suppose that if y additional open complexes have been formed, then correspondingly $y\Delta_{LN,t}$ turns have been removed from the transcript region (in order to facilitate open complex formation). Also, although the promoter region is short, each time a transcription initiation event occurs, the promoter region is unwound and propagates supercoiling. However, many transcription initiation events stall or reversibly dissociate. We suppose such events do not introduce significant amounts of supercoiling — rather, only when an elongation complex is formed do we suppose that the promoter region has been unwound and introduced supercoils in the proximal regions. Thus, we suppose that if there are y new elongation complexes, then there are $y\Delta_{LN,p}$ turns. Moreover, once the elongation complex departs, it is not necessarily true that the promoter will resume its normal B-form DNA state. However, we suppose that the reaction event of a new initiation complex finally forming is indicative of a supercoiling state being removed. In this way, turns are gained and lost by the incoming and outgoing of holoenzyme complexes on the promoter and transcript regions. We assume that any transcriptional pausing, abortive initiation, and aborted elongation events are effectively modeled by their respective transcriptional parameters. The dynamics of $\sigma_{t,S}$ can then be expressed as:

$$\begin{aligned}
\sigma_{t,S}(t+\epsilon) &= \sigma_{t,S}(t) + \frac{\Delta_{LN,t}}{n_{f,S}}(x-y) + \frac{\Delta_{LN,p}}{n_{f,S}}(y-z), \\
\sigma_{t,S}(t+\epsilon) &= \sigma_{t,S}(t) + \frac{\Delta_{LN,t}}{n_{f,S}}\left((mS^c(t+\epsilon) - mS^c(t)) - (EC_S^c(t+\epsilon) - EC_S^c(t))\right) \\
&\quad + \frac{\Delta_{LN,p}}{n_{f,S}}(EC_S^c(t+\epsilon) - EC_S^c(t) - (CC_S^c(t+\epsilon) - CC_S^c(t))\,),
\end{aligned}$$

where $\sigma_{t,S}(t)$ denotes the supercoiling density at time t, $mS^c(t)$ denotes the integer molec-

ular *count* of total mSpinach molecules produced by time t, Δ_{LN} denotes the change in the linking number of the mSpinach coding region per mSpinach transcript expressed, and $n_{f,S}$ is the combined length of free mSpinach transcript and spacer that is able to absorb the residual twist introduced by transcription. The amount of free spacer DNA between negatively supercoiled and positively supercoiled DNA available to absorb additional supercoiling depends on the dynamic equilibrium between negative twist and positive twist from mSpinach and MG aptamer transcription, respectively. We suppose that the length

$$n_{f,S} \equiv \max\left(PL_S + TL_S + \frac{N_S p_{Tet}}{2p_{Tet} + pTetC} + (N_S/2 + TL_G)\,\frac{pTetC}{(p_{Tet} + pTetC)} - \Delta_{kink}, 0\right)$$

where

$$\Delta_{kink,0} \equiv (\sigma_{t,S} + \sigma_{t,G})h_0)$$

and the third and fourth terms on the right-hand side constitute a weighted average of the length of DNA spacing available for either 1) the scenario where transcription is active in the adjacent gene or 2) the transcription factor TetR is bound to the p_{Tet} promoter. Similarly, we write

$$n_{f,G} \equiv \max\left(PL_G + TL_G + \frac{N_S p_{Lac}}{2(p_{Lac} + pLacC)} + (NS/2 + TL_S)\frac{pLacC}{(pLac + pLacC)} + \Delta_{kink}, 0\right) \tag{2.1}$$

. When $\sigma_{t,S} = \sigma_{t,G}$ note that the point of transition between negative and positive supercoiling is exactly centered. When $\sigma_{t,S} < \sigma_{t,G}$, i.e. the force from negative twist exceeds the force of positive twist, the kink is forced in the direction of the MG aptamer coding transcript and $n_{f,S} > TL_S + N_S/2$. Conversely, if mSpinach transcription does not produce additional negative supercoils to counteract the positive twist from MG aptamer expression, then $n_{f,S} < TL_S + N_S/2$.

The above equation states that the supercoiling density at time $t + \epsilon$ is the supercoiling density at time t with an additive perturbation term, corresponding to the change in super-

coiling density from transcription of $x = mS^c(t+\epsilon) - mS^c(t)$ transcripts. Normalizing by the reaction volume Ω on both sides, dividing by ϵ, and taking $\epsilon \to 0$, we obtain an expression in terms of the derivative of mSpinach *concentration*:

$$\frac{d(\sigma_{t,S})}{dt} = \left(\frac{d\,(mS)}{dt} + \delta_m mS - \frac{d\,(EC_S)}{dt}\right)\frac{\Delta_{LN,t}}{n_{f,S}} + \left(\frac{d\,(EC_S)}{dt} - \frac{d\,(CC_S)}{dt}\right)\frac{\Delta_{LN,p}}{n_{f,S}}.$$

Notice that the quantity $d(mS)/dt + \delta_m mS$ represents the rate at which total mSpinach RNA aptamer is produced in the system, since it is the state dynamics of mSpinach without mRNA degradation. However, the supercoiling state of DNA is continuously regulated by gyrase, an enzyme that relieves positive supercoiling, and topoisomerase, an enzyme that relieves negative supercoiling. We estimate that gyrase relieves positive supercoiling of the transcript region at roughly $\gamma = 0.5$ turns per second per plasmid, while topoisomerase relieves negative supercoiling of the transcript region at roughly $\tau = 0.25$ turns per second per plasmid [57]. Both enzymes act to maintain the natural physiological (negative) supercoiling density of σ_0. We suppose that in the absence of any transcriptional activity, the balance of these rates tends towards gyrase activity and a steady state of σ_0. For simplicity we suppose that gyrase and topoisomerase binding does not interfere with the transcriptional binding dynamics of polymerase. We incorporate these maintenance dynamics as follows:

$$\frac{d(\sigma_{t,S})}{dt} = \left(\frac{d\,(mS)}{dt} + \delta_m mS - \frac{d\,(EC_S)}{dt}\right)\frac{\Delta_{LN,t}}{n_{f,S}} + \left(\frac{d\,(EC_S)}{dt} - \frac{d\,(CC_S)}{dt}\right)\frac{\Delta_{LN,p}}{n_{f,S}} + m(\sigma_{t,S})$$

where

$$m(\sigma) \equiv T_0\tau\frac{[\sigma-\sigma_0]^-/k_{M,\tau}}{\sigma_0 + (\sigma-\sigma_0)^2/k_{M,\tau}} - G_0\gamma\frac{[\sigma-\sigma_0]^+/k_{M,g}}{\sigma_0 + (\sigma-\sigma_0)^2/k_{M,g}}$$

where ν is the total length of DNA, $x \equiv [x]^- + [x]^+$ denotes an additive decomposition of x into its strictly negative and nonnegative parts, and T_0 and G_0 are the topoisomerase and gyrase concentrations present *in vivo* or *in vitro* cell-free extract.

Next, to obtain an expression for $\Delta_{LN} < 0$, i.e. the number of negative supercoiling

turns introduced by expression of one mSpinach transcript, we argue as follows. As the open complex proceeds along the anti-sense DNA template of mSpinach, it unwinds and displaces the supercoiling of a 17 base pair region [57], corresponding to the DNA footprint of a transcription bubble (i.e. DNA-RNAP open complex). The transcription bubble requires an uncoiled region of DNA to transcribe. Thus, an additional $17/h_o$ turns are introduced into the upstream and downstream regions. We suppose that half of these turns are introduced as negative supercoiling and the other half as positive. Thus, in the wake of the transcription bubble passing through the entire transcript, there are

$$-\frac{17}{h_o}\frac{TL_S}{17}\frac{1}{2} = -\frac{TL_S}{(2h_o)}$$

negative supercoiling turns introduced into intergenic spacer downstream. When transcription termination occurs, the bubble is no longer held open by the open complex and the negative supercoils travel back into the unwound DNA of the mSpinach transcript and spacer, while the positive supercoils dissipate upstream of the promoter. Similarly, as the promoter expresses it also introduces negative supercoils downstream into the transcript region. The expression for $\sigma_{t,S}(t)$ then simplifies to

$$\dot{\sigma}_{t,S} = -\left(\dot{mS} - \delta_m mS - \dot{EC}_S\right)\frac{TL_S}{2h_0 n_{f,S}} - \left(\dot{EC}_S - \dot{CC}_S\right)\frac{PL_S}{2h_0 n_{f,S}} + m(\sigma_{t,S}).$$

Here we use $\dot{\theta}$ notation to denote the derivative of θ. Following similar arguments, we can write the dynamics of $\sigma_{p,S}(t)$ as

$$\dot{\sigma}_{p,S} = -(\dot{EC}_S - \dot{CC}_S)\frac{PL_S}{2h_0 n_{f,S}} + m(\sigma_{p,S}).$$

Similarly, the supercoiling density dynamics for the MG RNA aptamer gene are given as:

$$\dot{\sigma}_{t,G} = \left(\dot{EC}_G - \dot{CC}_G\right)\frac{PL_G}{2h_0 n_{f,G}} + \left(\dot{MG} - \delta_m MG - \dot{EC}_G\right)\frac{TL_G}{2h_0 n_{f,G}} + m(\sigma_{t,G}),$$
$$\dot{\sigma}_{p,G} = -(\dot{EC}_G - \dot{CC}_G)\frac{PL_G}{2h_0 n_{f,G}} + m(\sigma_{p,G}).$$

Notice the change in sign in the MG RNA aptamer dynamics. In this way, the directionality of sense transcription, relative to the right-handed twist of DNA, is encoded. If MG RNA aptamer was expressed in the anti-sense direction (which is the case in our divergently orientated construct), then the supercoiling introduced would be negative.

An important question is how transcription initiation rate $k_f(\cdot)$ and elongation rate $k_{cat}(\cdot)$ depends on supercoiling density. In [63] it was argued that the reaction rate of transcription initiation could be modeled with a Hill function type curve, based on experimental data characterizing the pelA and pelE promoters [75]. Although these results are specific to the bacterium *Dickeya dadantii*, it has been generally postulated that supercoiling density acts as a form of global gene regulation [19, 78] both in prokaryotic and eukaryotic organisms. A study of the ilvY and ilvC promoters [79] in *E. coli* suggest that promoter activity is optimal around around a certain value of σ^* and that activity tapers as σ diverges towards positive or negative infinity. Balke and Gralla [3] argued that *global* supercoiling state forms the basis of a feedback loop for a system of genes in an organism, in response to environmental cues regarding metabolite and resource availability.

Broadly speaking, it is difficult to draw general conclusions regarding the relation of supercoiling state and promoter activity — all experimental measurements in the studies described above were of the *global* supercoiling density. In these studies, the common approach was to treat a purified plasmid with topoisomerase to introduce additional twist. Whether this twist was introduced uniformly across the plasmid or non-uniformly is unclear. However, we can suppose that when a topoisomerase was used to treat plasmid, it introduced a monotone amount of twist (gyrase introduced *only* negative coils and Topo I

introduced *only* positive coils). We thus proceed supposing that incubation and treatment with a topoisomerase had a monotonic effect on promoter supercoiling state and that the overall *qualitative* trends observed regarding the ilvY, ilvC, and pelE, and pelA promoters can be used to inform the *qualitative* or phenomenological model of how *local* supercoiling density and promoter activity are related. Drawing from physical intuition, we argue that a promoter cannot initiate transcription if it is excessively wound with positive or negative twist. We suppose that transcription initiation is thus optimal at a particular value of local supercoiling density σ^*. Moreover, we suppose that for a given promoter of length PL_X, $X = S, G, RF$, or CF the optimal *local* supercoiling density optimum roughly is related to the optimal *global* supercoiling density σ_0 via the following approximation:

$$\sigma^* \approx \sigma_0 P_L / PL_X,$$

where P_L is the length of the plasmid. We model the rate of transcription initiation as a second-order symmetric Hill function, with an optimum centered around σ^*.

$$k_{f,X}(t) = \frac{\zeta}{1 + (\sigma_{p,X}(t) - \sigma^*)^2 / k_{M,\sigma}}, \tag{2.2}$$

where $X = G$ or S for MG and mSpinach transcription respectively and ζ is the optimal putative forward reaction rate of transcription initiation assuming the supercoiling state $\sigma_{p,X}$ is optimal for transcription initiation. Similarly, we suppose in the case of transcriptional elongation that the optimum $\sigma^* = \sigma_0 P_L / TL_X$ and the elongation rate is defined by the functions

$$k_{cat,X}(t) = \frac{\beta}{1 + (\sigma_{t,X}(t) - \sigma^*)^2 / k_{M,\sigma}}, \tag{2.3}$$

where $X = G$ or S for MG and mSpinach respectively and β is the putative transcription elongation rate when the supercoiling state $\sigma_{t,X}$ is optimal for transcription. Finally, we note the following conservation laws hold since DNA and RNAP concentration are constant

in our *in vitro system*

$$
\begin{aligned}
R^{tot} &= R + EC_S + EC_G + CC_S + CC_G, \\
p^{tot}_{\text{Lac}} &= p_{\text{Lac}} + CC_S + EC_S + pLacC, \\
p^{tot}_{\text{Tet}} &= p_{\text{Tet}} + CC_G + EC_G + pTetC. \\
LacI^{tot} &= LacI + aLacI + pLacC \\
TetR^{tot} &= TetR + aTetR + pTetC \\
IPTG^{tot} &= IPTG + aLacI \\
aTc^{tot} &= aTc + aTetR
\end{aligned}
$$

Using these laws, we can write a reduced order dynamical system model for the *convergent* biocircuit:

$$
\begin{aligned}
\dot{mS} &= k_{cat,S}(\sigma_{t,S})EC_S - \delta_m mS, \\
\dot{MG} &= k_{cat,G}(\sigma_{t,G})EC_G - \delta_m MG, \\
\dot{EC}_S &= k_{open}CC_S - k_{cat}(\sigma_{t,S})EC_S, \\
\dot{EC}_G &= k_{open}CC_G - k_{cat}(\sigma_{t,G})EC_G, \\
\dot{CC}_S &= k_f(\sigma_{p,S})(R^{tot} - EC_S - EC_G - CC_S - CC_G)(p^{tot}_{\text{Lac}} - CC_S - EC_S - pLacC) \\
&\quad - (k_r + k_{open})CC_S \\
\dot{CC}_G &= k_f(\sigma_{p,G})(R^{tot} - EC_S - EC_G - CC_S - CC_G)(p^{tot}_{\text{Tet}} - CC_G - EC_G - pTetC) \\
&\quad - (k_r + k_{open})CC_G
\end{aligned}
\tag{2.4}
$$

$$\begin{aligned}
\dot{LacI} &= \rho_l + k_{ua,L}(IPTG^{tot} - IPTG) + k_{u,L}(LacI^{tot} - LacI - IPTG^{tot} + IPTG) \\
&\quad - k_{aL} LacI\, IPTG - k_{seq,L} p_{\mathrm{Lac}}\, LacI - \delta_p LacI \\
\dot{TetR} &= \rho_t + k_{ua,T}(aTc^{tot} - aTc) + k_{u,T}(TetR^{tot} - TetR - aTc^{tot} + aTc) \\
&\quad - k_{aL} TetR\, aTc - k_{seq,T} p_{\mathrm{Tet}}\, TetR - \delta_p TetR \\
\dot{IPTG} &= -k_{a,L}(LacI\ + pLacC) IPTG + k_{ua,L}(LacI^{tot} - LacI - pLacC) \\
\dot{aTc} &= -k_{a,T}(TetR\ + pTetC) aTc + k_{ua,T}(TetR^{tot} - TetR - pTetC) \\
\dot{\sigma}_{t,S} &= -\left(\dot{mS} - \delta_m mS - \dot{EC}_S\right) \frac{TL_S}{2h_0 n_{f,S}} - \left(\dot{EC}_S - \dot{CC}_S\right) \frac{PL_S}{2h_0 n_{f,S}} + m(\sigma_{t,S}) \\
\dot{\sigma}_{t,G} &= \left(\dot{EC}_G - \dot{CC}_G\right) \frac{PL_G}{2h_0 n_{f,G}} + \left(\dot{MG} - \delta_m MG - \dot{EC}_G\right) \frac{TL_G}{2h_0 n_{f,G}} + m(\sigma_{t,G}), \\
\dot{\sigma}_{p,S} &= -(\dot{EC}_S - \dot{CC}_S) \frac{PL_S}{2h_0 n_{f,S}} + m(\sigma_{p,S}) \\
\dot{\sigma}_{p,G} &= -(\dot{EC}_G - \dot{CC}_G) \frac{PL_G}{2h_0 n_{f,G}} + m(\sigma_{p,G})
\end{aligned}$$

In simulating the supercoiling dynamics we noticed that the magnitude of our *local* supercoiling states settle around steady-state values much higher than the traditionally accepted range of global supercoiling density. In practice, experiments have determined that DNA is negatively supercoiled with a *global* supercoiling density of -0.065 and can drop to as low as -0.1. This parameter does not reflect the *local* supercoiling density of the regions of interest in our model, namely the supercoiling density of the transcript and the promoter.

For example, a small region of DNA can maintain a positively coiled plectonome while the rest of the DNA is relatively relaxed. The global supercoiling density will reflect the overall twist, as opposed to the high density in either of the two regions. Wang and Liu estimated that expression of a single transcript can introduce supercoils into DNA at a rate of 4 supercoils per second per transcript. Assuming gyrase introduces $\gamma = 1$ negative supercoils per second and $\tau = .5$ positive supercoils per second on a given plasmid, if half the supercoils introduced propagate upstream and the over half downstream, over the course of just five minutes [57] the region downstream (such as the 150 bp spacer sequence in our plasmids) of the transcript could achieve a *local* supercoiling density of $\sigma = 2.0$. A

measurement of the *global* supercoiling state of the plasmid, say 3.5 kbp in length, would yield a *global* estimate of only $\sigma = 0.08$! Therefore, it is important to note the distinction between *local* and *global* supercoiling density; the *local* supercoiling density of a region of DNA can reach much higher magnitudes despite a relatively low (and conventionally acceptable) *global* supercoiling density.

Divergent Orientation Model

In our divergently oriented plasmid, the Tet promoter and Lac promoter express in opposing directions, but the transcription bubbles diverge or move away from each other. Thus, the only torsional stress introduced comes from backward propagation of coils from unwinding the regions of DNA encoding promoters into the intergenic spacing region between the two genes. The Tet promoter back propagates positive supercoils into the intergenic spacing region while the Lac promoter back propagates negative supercoils. The position of dynamic equilibrium between the positively supercoiled region upstream of the Tet promoter and the negatively supercoiled region upstream of the Lac promoter is determined by the balance of forces arising from positive and negative twist (diametrically opposing each other) in the promoter supercoiling states $\sigma_{p,S}$ and $\sigma_{p,G}$. If $\sigma_{p,S}$ is much larger than $\sigma_{p,G}$ then the equilibrium shifts in favor of the Lac promoter and the positive coils are pushed closer to the actual Tet promoter (or vice-versa). We model the amount of spacer available to the promoters as $n_{f,S}$ and $n_{f,G}$ where

$$\begin{aligned} n_{f,S} &= \max\{PL_S + N_S/2 - \Delta_K, 0\} \equiv \max\{PL_S + N_S/2 - (\sigma_{p,S} + \sigma_{p,G})h_0, 0\} \\ n_{f,G} &= \max\{PL_G + N_S/2 + \Delta_K, 0\} \equiv \max\{PL_G + N_S/2 + (\sigma_{p,S} + \sigma_{p,G})h_0, 0\} \end{aligned} \tag{2.5}$$

again noting that $n_{f,S} + n_{f,G} = PL_S + PL_G + NS$ base pairs defines the total length of DNA in which localized supercoiling buildup can propagate. We suppose that all other supercoils arising from transcription elongation are dissipated within regions downstream of the promoters.

The supercoiling dynamics of the divergently oriented construct for mSpinach and MG RNA aptamer are thus given as

$$\begin{aligned}
\dot{\sigma}_{t,S} &= \left(\dot{EC}_S - \dot{CC}_S\right)\frac{PL_S}{2h_0TL_S} + m(\sigma_{t,S}),\\
\dot{\sigma}_{t,G} &= \left(\dot{EC}_G - \dot{CC}_G\right)\frac{PL_G}{TL_G} + m(\sigma_{t,G}),\\
\dot{\sigma}_{p,S} &= -\left(\dot{EC}_S - \dot{CC}_S\right)\frac{PL_S}{2h_0n_{f,S}} + m(\sigma_{p,S}),\\
\dot{\sigma}_{p,G} &= \left(\dot{EC}_G - \dot{CC}_G\right)\frac{PL_G}{2h_0n_{f,G}} + m(\sigma_{p,G}).
\end{aligned}$$

while the rest of the system dynamics are as presented in the convergent model. Any differences in expression are thus a function of the supercoiling dynamics above, the initial conditions of these four states, and their effects on k_f, k_{cat} and the topoisomer maintenance function $m(\cdot)$.

Tandem Orientation Model

In the tandem orientation, negative supercoiling backpropagates from the p_{Lac} promoter into the intergenic spacing region between the MG aptamer coding sequence and the mSpinach promoter. The torsional stress introduced by downstream propagation of positive supercoils from MG aptamer elongation and upstream propagation of negative supercoils from mSpinach transcription initiation again defines a dynamic equilibrium that is determined by the balance of $\sigma_{t,G}$ and $\sigma_{p,S}$. When the Lac promoter for mSpinach is much more active relative to the transcriptional activity of the MG aptamer coding sequence, $\sigma_{p,S}$ can dominate $\sigma_{t,G}$ such that any residual positive supercoils from MG aptamer transcription are pushed back into the coding sequence for MG aptamer. Excessive negative supercoiling from the mSpinach promoter, likewise, can make their way into the MG aptamer coding sequence. This is especially likely if the transcript region of MG aptamer is short (since it generates less positive supercoils to counteract the twisting force of negative supercoiling from mSpinach expression). The presence of excessive negative supercoiling in a transcript region can result

in the formation of a R-loop complex, a hybrid of the RNAP-DNA open complex and the nascent mRNA chain with upstream DNA [19]; this complex stalls the elongation process indefinitely and impedes subsequent transcription events. These effects are accounted for in the function $k_{cat}(\sigma)$, which tapers off towards 0 if $\sigma \to -\infty$. The sensitivity of k_{cat} to -σ is determined by the parameter $k_{M,\sigma}$.

Alternatively, if MG aptamer expression is high or leaky, it can likewise propagate positive supercoils downstream into the spacer region, which subsequently shutoff promoter activity of mSpinach. The decrease in promoter activity in mSpinach only further enables MG aptamer expression, which leads to MG aptamer dominant expression. This is particularly relevant if the coding sequence for MG aptamer is long, or replaced with a long coding sequence for a protein, e.g. RFP. In such a scenario, the expression of RFP can repress future expression of mSpinach.

It is thus important to consider the dynamic equilibrium of $\sigma_{t,G}$ and $\sigma_{p,S}$ and how the balance of these forces impact the positioning of positive and negative supercoils in the transcript region of MG aptamer and the Lac promoter. We suppose that the length of DNA available for positive supercoiling buildup (from MG aptamer expression) is given as:

$$n_{f,G} = \max\{TL_G + N_S/2 + \Delta_K, 0\}$$

and the length of DNA available for negative supercoiling buildup (from mSpinach expression) is given as:

$$n_{f,S} = \max\{PL_S + p_{Tet}/(pTetC + p_{Tet})N_S/2 + pTetC/(pTetC + p_{Tet})(TL_G + N_S/2) - \Delta_K, 0\}$$

where $\Delta_K \equiv (\sigma_{p,S} + \sigma_{t,G})h_0$. The supercoiling dynamics are thus given by

$$\begin{aligned}
\dot{\sigma}_{t,S} &= \left(\dot{EC}_S - \dot{CC}_S\right) \frac{PL_S}{2h_0 n_{f,S}} + m(\sigma_{t,S}), \\
\dot{\sigma}_{t,G} &= \left(\dot{m}_G + \delta_m m_G - \dot{EC}_G\right) \frac{TL_G}{2h_0 n_{f,G}} + \left(\dot{EC}_G - \dot{CC}_G\right) \frac{PL_G}{2h_0 TL_G} + m(\sigma_{t,G}), \\
\dot{\sigma}_{p,S} &= -\left(\dot{EC}_S - \dot{CC}_S\right) \frac{PL_S}{2h_0 n_{f,S}} + m(\sigma_{p,S}), \\
\dot{\sigma}_{p,G} &= \left(\dot{EC}_G - \dot{CC}_G\right) \frac{PL_G}{2h_0 n_{f,G}} + m(\sigma_{p,G}).
\end{aligned}$$

RFP CFP reporter models

Modeling the expression of RFP and CFP instead of mSpinach and RFP does not change any of the preceding arguments for deriving dynamics of supercoiling states. The only difference at the transcriptional level is that we use different length parameters, TL_R and TL_C (see Table 2.1) to define the length of transcript regions and denote the transcriptional products of each transcription elongation reaction as m_C (for the CFP mRNA transcript) and m_R (for the RFP mRNA transcript).

The primary source of genetic context effects in our model is supercoiling at the DNA level. Therefore, in this work we do not consider the effects of secondary structure in mRNA or superhelicity of mRNA-DNA hybrids. Thus, we deliberately model translation reactions simplistically, with the following chemical reactions:

$$
\begin{aligned}
m_{CFP} &\xrightarrow{k_{B16}k_{TL}R^{tot}} P_{C,u} \\
m_{RFP} &\xrightarrow{k_{Bgl}k_{TL}R^{tot}} P_{R,u} \\
P_{C,u} &\xrightarrow{k_{f,C}} P_{C,f} \\
P_{R,u} &\xrightarrow{k_{f1,R}} P_{R,f} \\
P_{R,f} &\xrightarrow{k_{f2,R}} P_R \\
CFP &\xrightarrow{\delta_p} \emptyset \\
RFP &\xrightarrow{\delta_p} \emptyset
\end{aligned}
$$

We suppose these translation reactions are the same for convergent, divergent, and tandem oriented genes. Notice the inclusion of maturation reactions for CFP and RFP. We suppose that CFP matures through a one-step process while RFP matures through a two-step process [106]. Here we do not necessarily assume that RFP is dimeric, since the variant of dsRed1 that we used in our experiments is actually monomeric. However, we suppose that there is an intermediate stage between unfolded RFP and the final folded RFP. We found that including this intermediate stage recapitulated the significant delay observed in RFP expression in the cell-free expression system, that was not seen in CFP.

Moreover, the cell-free expression system is typically run in a bulk reaction setting, as a closed biochemical reaction system with a finite and limited amount of ATP, NTPs, and energy molecules to carry out transcription and translation. It has been observed empirically and shown through experiments that as ADP levels build up relative to ATP, enzymatic reactions become increasingly unfavorable (otherwise fluorescent reporters not subject to degradation would express in unbounded and increasing concentrations). Throughout our experiments, we observed these effects of resource depletion, beginning at $t_0 \approx 2$ hours onwards. To be consistent with the modeling approaches of Tuza and Singhal, [86, 94], we suppose that the translation rate $k_{TL}(t)$ decays with time as a first order process, beginning

at time t_0, and with decay parameter $\alpha_d = \log(2)/(480)$

$$k_{TL}(t) \equiv k_{TL}^0 e^{(-\alpha_d(t-t_0)\mathbf{1}_{t>t_0})}.$$

where k_{TL}^0 is the nominal translation rate assuming an open system with limitless ATP and energy.

The additional reaction dynamics in the state-space model are thus specified as follows:

$$\begin{aligned}
\dot{P}_{C,u} &= k_{B16} k_{f,C} k_{TL} R^{tot} m_{CFP} - k_{f,C} P_{C,u} - \delta_p P_C \\
\dot{P}_{C,f} &= k_{f,C} P_{C,u} \\
\dot{P}_{R,u} &= k_{Bgl} k_{f,R} k_{TL} R^{tot} m_{RFP} - k_{f,R} P_{R,u} - \delta_p P_R \\
\dot{P}_{R,f} &= k_{f1,R} P_{R,u} - k_{f2,R} P_{R,f} \\
\dot{P}_R &= k_{f,2R} P_{R,f}
\end{aligned}$$

The outcomes of our simulations are plotted in Figure 2.5 using parameters from Table 2.1. We see that RFP and CFP expression varies depending on orientation, initial condition of supercoiling states, and that the model is able to recapitulate the trends observed in the data.

It is important to remark that while our model is able to describe the effects observe, it is the gyrase experiments that definitively confirm the validity of supercoiling as a working hypothesis for the physical mechanism driving compositional context effects. Our model serves to validate supercoiling as a hypothesis for compositional context, but not necessarily to prove it.

In conclusion, we have constructed three versions of a simple biocircuit to motivate the need to model compositional context in biocircuit assembly. Our initial data suggests that promoter orientation between pairs of promoters has a salient effect on gene expression. We developed a nonlinear model incorporating various phenomena resulting from compositional context and show it captures the patterns seen in experiments. We emphasize that these

Parameter Name	Description	Numerical Value	Source
h_0	Bps per right-handed turn in B-DNA	10.5 bp/turn	Berg et al.
TL_S	mSpinach-tRNA and T500 term. length	203 bp	Larson et al., Paige et al.
TL_G	MG aptamer and T500 terminator transcript length	68 bp	Babendure et al.
N_S	Intergenic spacer length	150 bp	NA
PL_S	Lac promoter length	40 bp	Lutz & Bujard
PL_G	Tet promoter length	44 bp	Lutz & Bujard
p_{Len}	Length of ColE1-mSpinach-MG reporter plasmid	2892 bp	NA
$K_{M,\sigma}$	Michaelis constant for supercoiling Hill functions	50 nM	NA
R^{tot}	Total RNAP concentration	18.931 μM (RNAP)	Bremer & Dennis
$Ribo^{tot}$	Total Ribosome concentration	11.291 μM (ribosome)	Bremer & Dennis
p_{Lac}^{tot}	Total Plasmid Concentration	11 nM	NA
p_{Tet}^{tot}	Total Plasmid Concentration	11 nM	NA
G_0	Gyrase Concentration	12 nM	Maier et al.
T_0	Topoisomerase Concentration	2 nM	Maier et al.
$LacI^{tot}$	LacI concentration	10 nM	Kalisky et al.
$TetR^{tot}$	TetR concentration	0 nM (TX-TL)	Kalisky et al.
$IPTG^{tot}$	IPTG concentration	0 nM (TX-TL)	NA
aTc^{tot}	aTc concentration	0 nM (TX-TL)	NA
$k_{f,max,l}$	Leaky forward transcription initiation rate	7×10^{-2} nM/s	Siegal-Gaskins et al.
k_r	Reverse transcription initation rate	550 s^{-1}	Bintu et al.
$k_{cat,max}$	mSpinach-MG averaged transcription rate	$k_{cat,g}/(105.5)$	NA
$k_{cat,g}$	Per base-pair transcription rate	85 nt/s	Bremer & Dennis
k_{open}	Rate of open complex formation	0.04/s	Buc & McClure
k_l	Fraction of terminator-escaped transcripts	0.02	Larson et al.
ρ_l	Constitutive production rate of LacI	0 nM/s (TX-TL)	NA
ρ_t	Constitutive production rate of TetR	0 nM/s (TX-TL)	NA
k_{TL}^0	Averaged translation rate of RFP/CFP	$21/(TL_C + TL_R)$ /s	Bremer & Dennis
$k_{f,C}$	Folding rate of CFP	$1/(30 \times 60)$/s	NA
$k_{f,R}$	Folding rate of RFP	$1/(110 \times 60)$/s	Zhang et al.
k_{B16}	Relative Ribosomal Affinity	0.75	NA
k_{Bgl}	Relative Ribosomal Affinity	1.25	NA
$\delta_{m,S}$	mSpinach degradation rate	$\log(2)/(30 \times 60)$ /s	NA
δ_{MG}	MG degradation rate	$\log(2)/(60 \times 60)$ /s	NA
τ	Negative coils introduced per sec. per TopoI	0.5 /s	Liu & Wang
γ	Positive coils introduced per sec. per Gyrase	0.5 /s	Liu & Wang
σ_0	Natural superhelical density of DNA	-0.065	Rhee et al.
$k_{M,gyr}$	Hill constant for Gyr. Maintenance Function	200 nM	NA
$k_{a,L}$	IPTG binding rate to free/DNA bound LacI	6×10^3 /s	Kalisky et al.
$k_{ua,L}$	IPTG-LacI disassociation rate	1 /s	Xie et al.
$k_{seq,L}$	Binding rate of LacI to promoter	10 /s	Kalisky et al.
$k_{u,L}$	Fall off rate of LacI to DNA	0.022 /s	Nelson & Sauer
$k_{a,T}$	aTc binding rate to free and DNA bound TetR	$k_{a,L}$	NA
$k_{ua,T}$	aTc-TetR disassociation rate	$k_{ua,L}$	NA
$k_{seq,T}$	Binding rate of TetR to promoter	$k_{seq,L}$	NA
$k_{u,T}$	Fall off rate of TetR to DNA	$k_{u,L}$	NA
δ_p	Degradation rate for untagged proteins	0/s (TX-TL)	Shin & Noireaux

Table 2.1: Parameters used for the deterministic ODE model for convergent, divergent, and tandem oriented reporters

results are wholly the consequences of compositional context. There is no designed interaction in the biocircuit, yet different expression biases arise depending on how genes are arranged. Therefore, with any biocircuit comprised of multiple parts, modeling the effects of compositional context should be a chief consideration during the design and prototyping process.

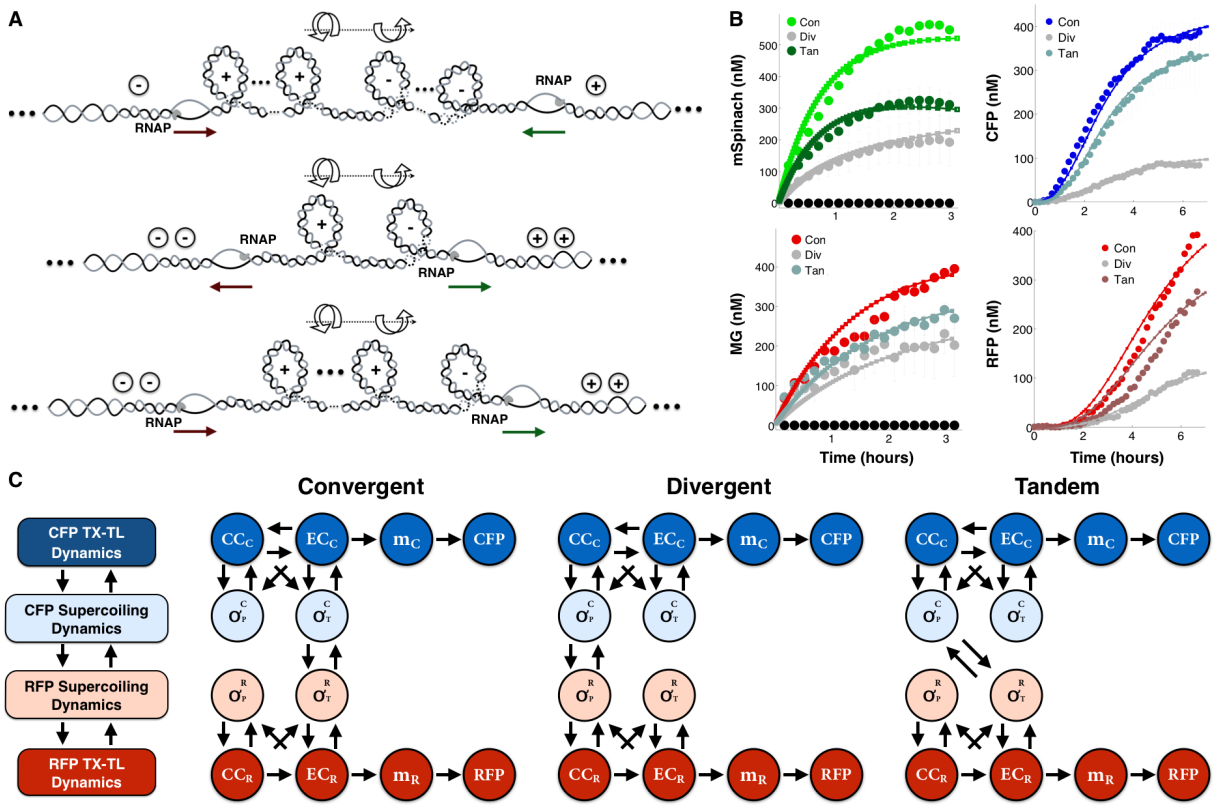

Figure 2.5: **Mathematical models incorporating supercoiling dynamics are able to recapitulate experimental data: (A)** A diagram showing how positive supercoiling builds up downstream of transcription bubbles and negative supercoiling builds up upstream of transcription bubbles. When two genes are adjacently placed, the intermediate region is exposed to opposing forces of torsional stress from positive (left-handed twist) and negative (right-handed twist) supercoiling. These forces do not cancel out each other, but rather oppose each other to achieve a dynamic equilibrium dependent on the transcriptional activity of nearby genes. **(B)** Expression curves of a mathematical model, integrating supercoiling dynamics of promoter and transcript states with gene expression, with supercoiling parameters fit to experimental data from the TX-TL cell free expression system [84]. CC_X and EC_X denote the closed complex and elongation complex states of gene X, respectively. σ_P^X and σ_T^X denote the supercoiling density of the promoter and transcript for gene X, respectively. **(C)** Schematic illustrating the state-dependencies between traditional transcriptional states and supercoiling states in convergent-, divergent-, and tandem- oriented RFP and CFP.

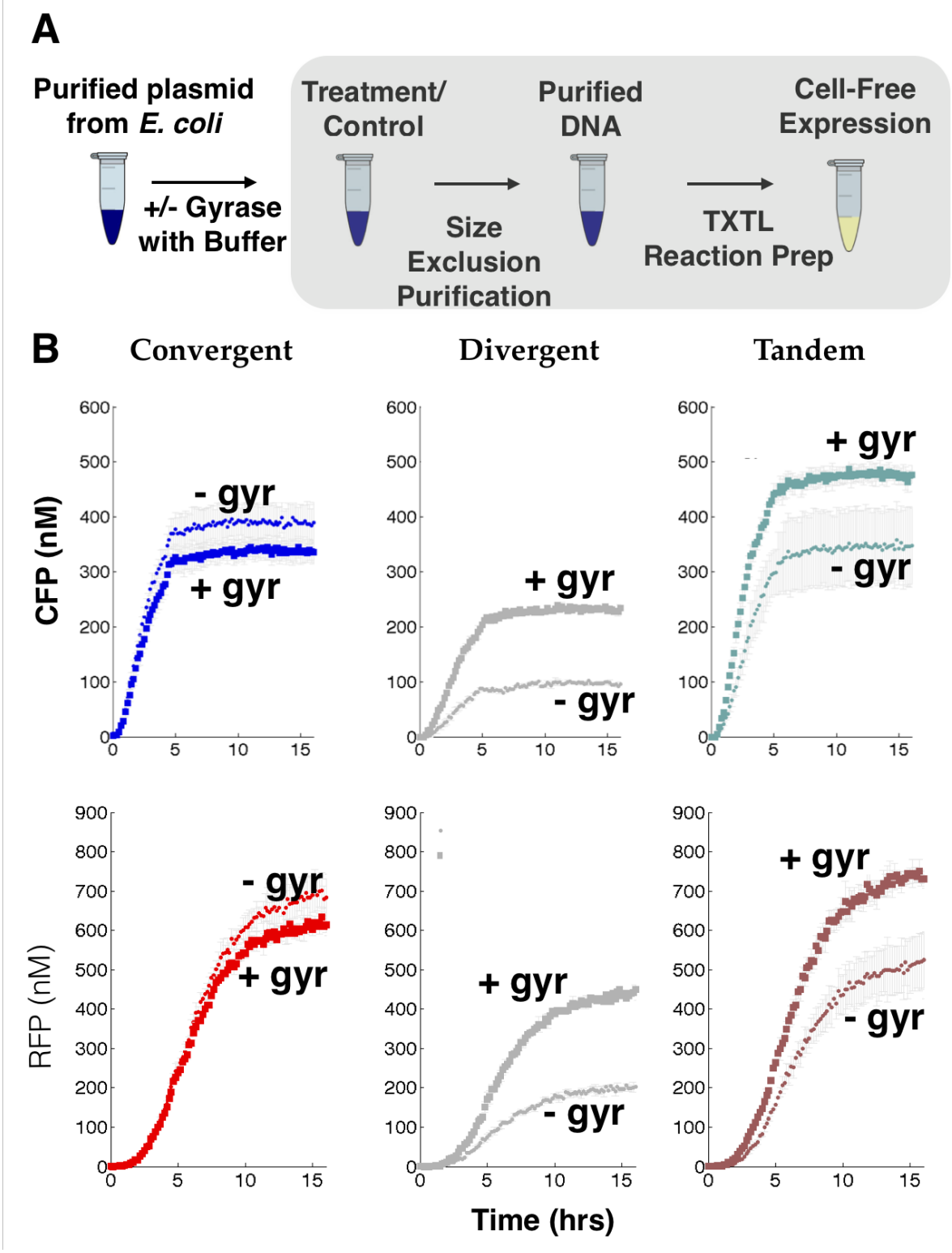

Figure 2.6: **Relaxation of positive supercoiling in plasmids with gyrase enzyme significantly reduces compositional context effects on gene expression: (A)** Workflow for gyrase treatment experiments. **(B)** Expression of CFP and RFP for convergent, divergent, and tandem oriented ColE1 plasmids prior (small dots) and post treatment (large dots) with gyrase.

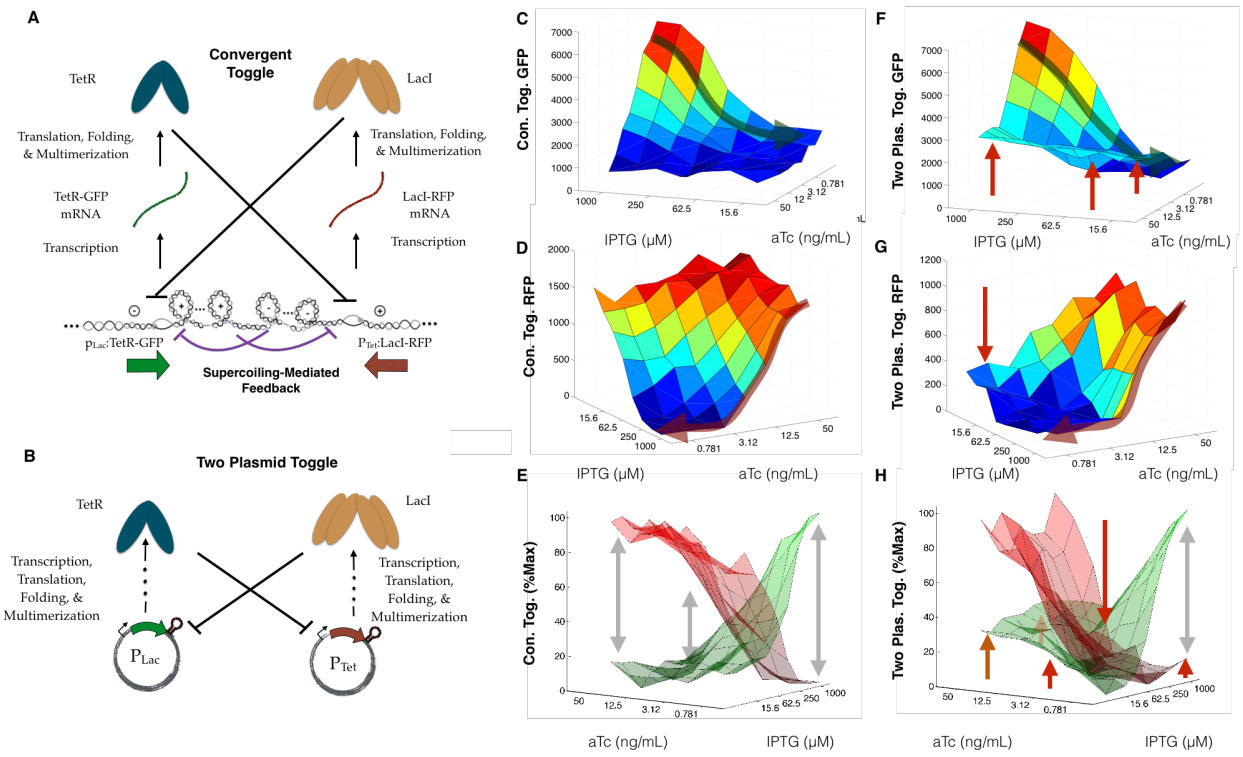

Figure 2.7: **Compositional context can be used to introduce supercoiling-mediated feedback, improving sharpness of threshold in toggle switch: (A)** Diagram of feedback architecture in a convergent toggle switch. **(B)** Diagram of feedback architecture in a convergent toggle switch TetR-GFP (ColE1) and LacI-RFP (p15A).**(C-E)** Experimental data of convergent toggle GFP expression in response to titrating IPTG and aTc concentration. **(F-H)** Experimental data of the two-plasmid toggle GFP expression in response to titrating IPTG and aTc concentration.

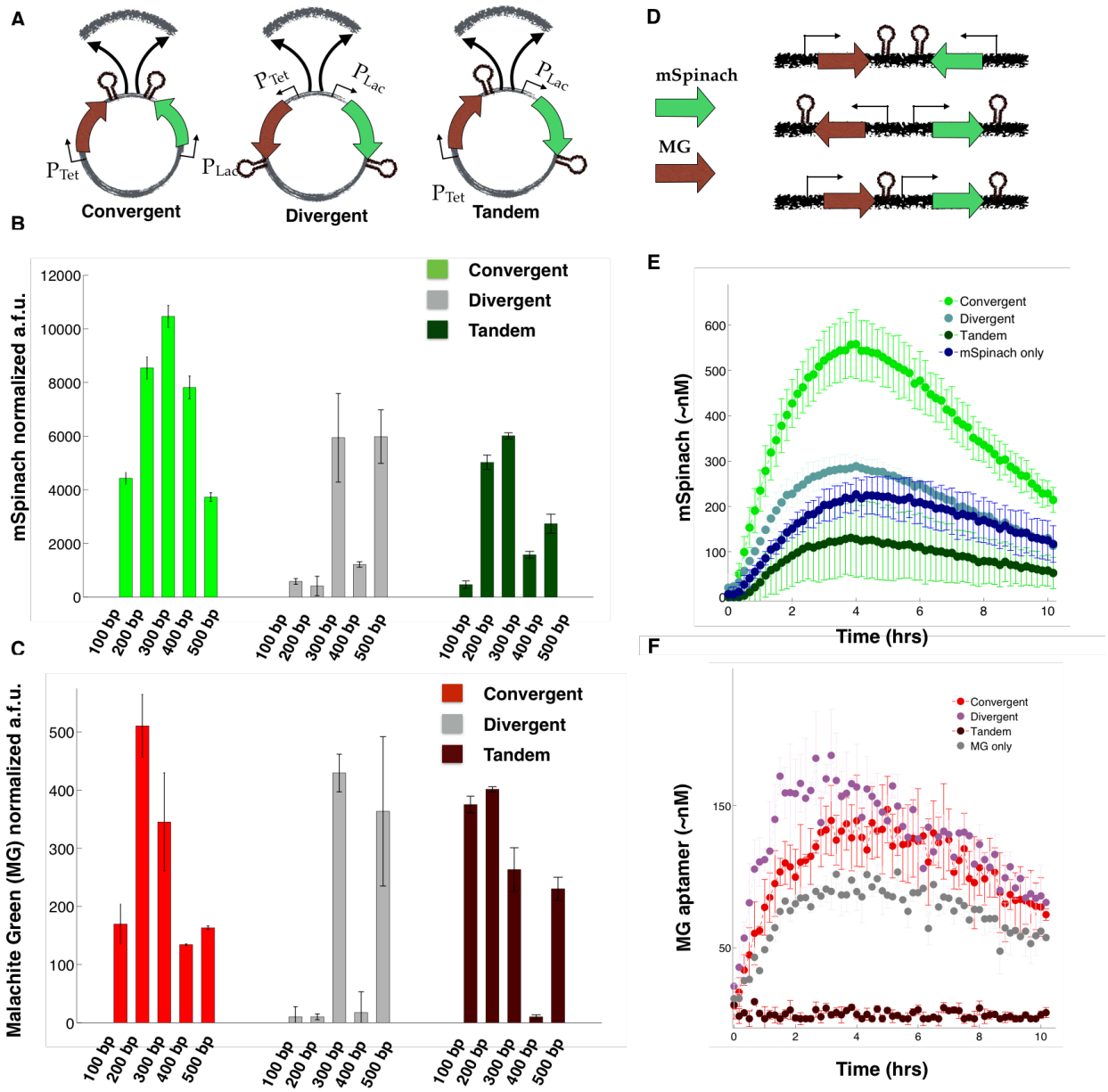

Figure 2.8: Related to Figure 2.2. **(A)** A schematic showing the point of insertion of intergenic spacing sequences of length $n = 100, 200, 300, 400$, and 500 bp. **(B)** Steady-state *in vivo* expression of mSpinach from overnight induction in 1 mM IPTG and 200 ng/mL aTc in convergent, divergent, and tandem orientation, varied as a function of spacer length. **(C)** Steady-state expression of MG RNA aptamer from overnight induction in 1 mM IPTG and 200 ng/mL aTc in convergent, divergent, and tandem orientation, varied as a function of spacer length. **(D)** Diagram of linear DNA fragments with mSpinach and MG RNA aptamers in convergent, divergent, and tandem orientation.**(E)** Cell-free *in vitro* expression of equimolar concentrations of linear mSpinach in convergent, divergent, tandem orientation and as a single gene on a linear DNA. **(F)** Cell-free *in vitro* expression of equimolar concentrations of linear MG RNA aptamer in convergent, divergent, tandem orientation, and as a single gene on linear DNA.

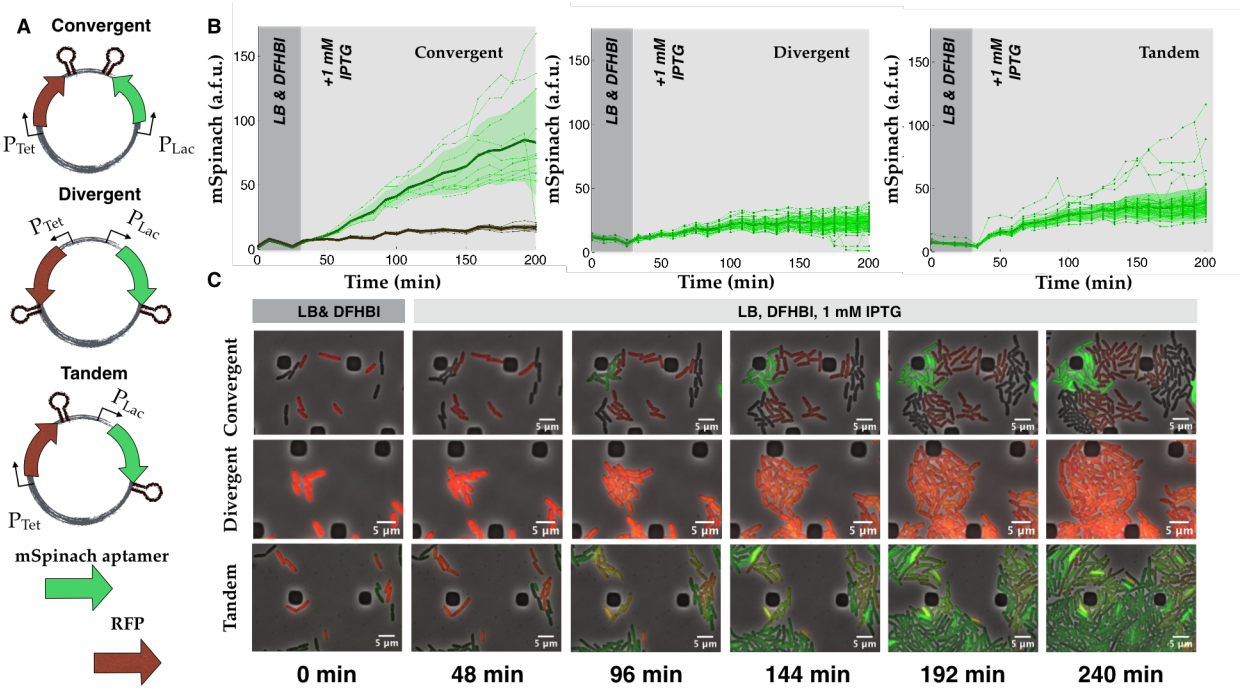

Figure 2.9: Related to Figures 2.2 and 2.3. **(A)** Convergent, divergent, and tandem oriented mSpinach and RFP reporters on ColE1 backbone. **(B)** Time-lapse mSpinach expression curves for individual cell traces in response to 1 mM IPTG induction. Notice that even though mRFP is not induced, its presence significantly affects the magnitude and shape of gene expression. **(C)** Single cell microscopy images of convergent oriented (top) mSpinach and RFP expression, cells responded with a strong bimodal phenotype, (middle) divergent oriented RFP and mSpinach, and (bottom) tandem oriented RFP and mSpinach.

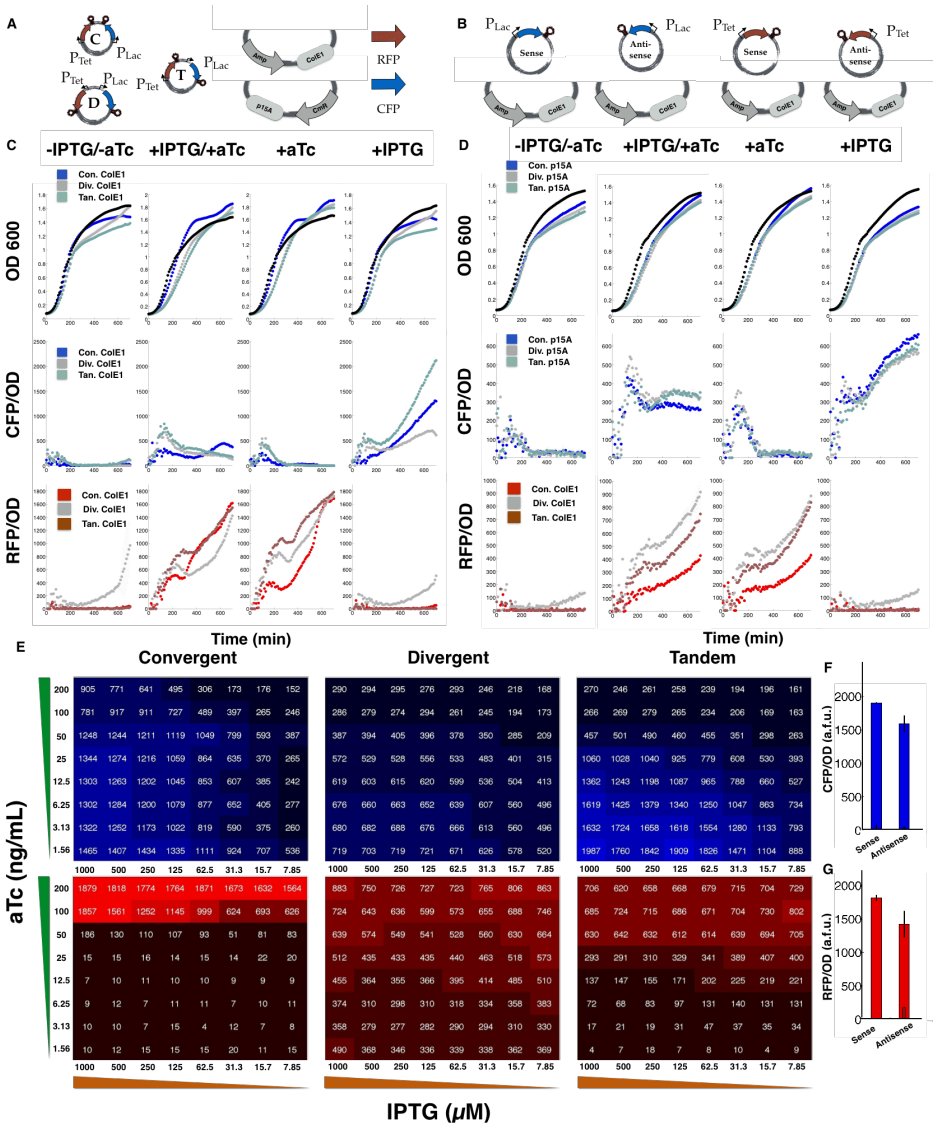

Figure 2.10: Related to Figures 2.3 and 2.4. **(A)** Plasmid layouts for RFP and CFP in convergent, divergent, and tandem orientation, the composition of the plasmid backbone for the ColE1 and p15A backbones used in collecting data for **C-D**. **(B)** A diagram showing the sense and anti-sense CFP and RFP single gene cassette controls, expressed on the ColE1 backbone. **(C-D)** Time lapse *in vivo* plate reader expression of RFP and CFP and growth curves, induced with either 1 mM IPTG, 200 ng/mL aTc, or both, on either ColE1 plasmid or p15A plasmid backbone. **(E)** Quantitative heat-maps of CFP and RFP expression in two variable titration assays of IPTG and aTc for convergent, divergent, and tandem oriented ColE1 plasmids (Figure 2.4). IPTG is titrated left to right with 2x dilutions starting from 1 mM IPTG (far left) while aTc is titrated top to bottom with 2x dilutions starting from 200 ng/mL. **(F-G)** Expression at $t = 550$ minutes for CFP and RFP in sense and anti-sense single gene plasmid controls. Notice that varying orientation with respect to genetic elements on the backbone only produces a small effect, suggesting that the primary source of context interference is from promoters with DNA binding sites (pLac and pTet).

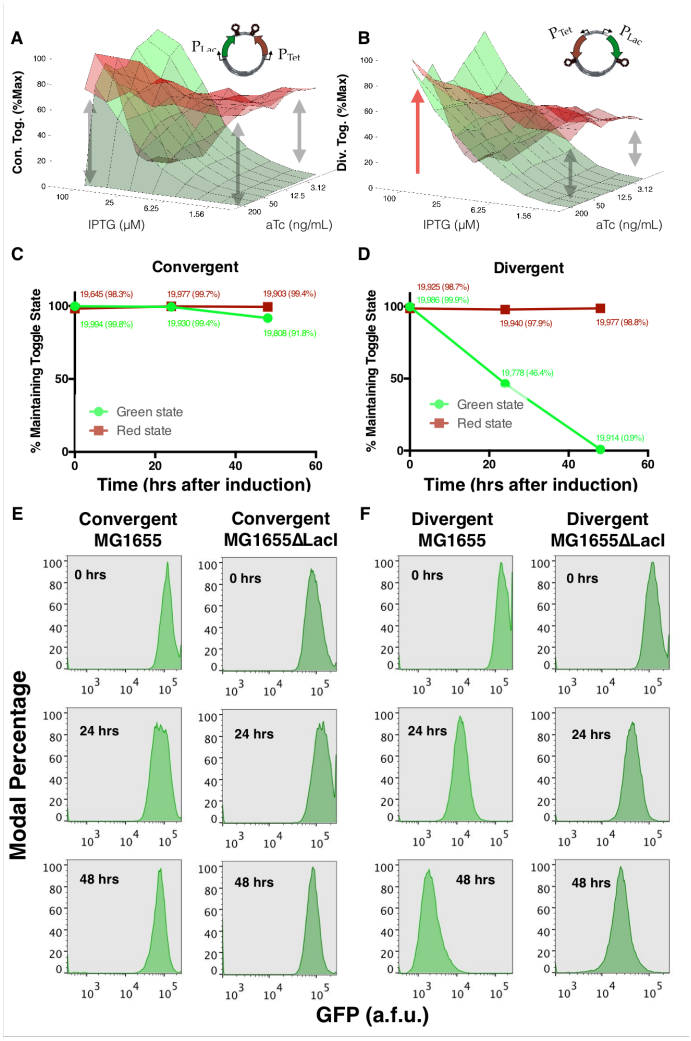

Figure 2.11: Related to Figure 2.7. **(A)** Experimental data from a dual reporter expression assay, titrating both IPTG and aTc concentrations to evaluate threshold behavior of the convergent Gardner-Collins toggle switch in MG1655ΔLacI *E. coli*. **(B)** Experimental data from a dual reporter expression assay, titrating both IPTG and aTc concentrations to evaluate threshold behavior of the divergent Gardner-Collins toggle switch in MG1655ΔLacI *E. coli*. **(C-D)** A stability test of the original Gardner-Collins toggle switch and its convergent counterpart in MG1655 *E. coli*. Cells were latched for 24 hours prior to the start of the experiment ($t = -24$ to $t = 0$) and subsequently rediluted in inducer-free media to assess stability of the toggle. The fraction of cells maintaining the original on-state are plotted against time. **(E)** Distributions showing stability of convergent toggle in the high GFP state in cell populations of MG1655 *E. coli* and MG1655ΔLacI *E. coli* plotted at $t = 0, 24$, and 48 hours. **(F)** Distributions showing stability of divergent toggle in the high GFP state in cell populations of MG1655 *E. coli* and MG1655ΔLacI *E. coli* plotted at $t = 0, 24$, and 48 hours.

Chapter 3

Reverse-Engineering Context Effects with Dynamical Structure Functions

3.1 Introduction

Two key properties that often determine the behavior of a dynamical system are its network structure and parametric realization. The structure of the network generally is determined by how states in the system causally depend on each other; edges in the network are determined by causal dependence while nodes are determined by the states of the system [88]. Network structure alone does not determine dynamical behavior, though, parametric information is also important in determining what dynamical behaviors a system can achieve [32]. Rather, network structure, or topology, often defines or narrows the possible behaviors a system can achieve. Without any structural constraints, a dynamical system can have arbitrary input-output behavior. Once network structure is imposed, the set of realizable input-output trajectories can be reduced [62, 99].

This is particularly evident in biological networks; certain network topologies are referred to as network motifs [62, 95]. In systems and synthetic biology, these network motifs are broadly accepted as enabling useful dynamical behavior. For example, an incoherent feedforward loop can be used for fold-change detection or adaptation, a cyclic network of repressors is associated with either oscillations or multi-stability, and a dual negative feedback network of two nodes is used as memory module or toggle switch. Network structure is thus an im-

portant aspect of designing synthetic biological circuits. By selecting an appropriate network motif and validating its functionality in practice, synthetic biologists are able to guide the phenotype of biological systems to match desired performance specifications.

What network structure to use when interconnecting physical components or entire engineered modules is an important design question. How components are interconnected implicitly defines network structure, which in turn constrains dynamical behavior of the system. Certain network structures can give rise to undesirable dynamic behavior [101]. Choosing the right network structure is thus an important problem in the synthesis of robust engineered dynamical systems.

Similarly, once a dynamical system has been designed and implemented, verifying that the network structure of a dynamical system is operating as designed is an equally important problem. This is especially critical when the engineered system does not behave as expected (a pervasive challenge in current efforts to implement synthetic biocircuits) [9]. The problem of verifying or reverse-engineering a system's network structure from measurement data is called a network reconstruction problem [37]. Network reconstruction problems are a specific class of system identification problems [58], where the model class of interest not only encodes parametric but structural information. In the next section we motivate and formulate the network reconstruction problem for different network representation models and argue that one particular representation is well suited for biochemical reaction networks: the dynamical structure function.

3.2 Motivation: Reconstructing Representations of Network Structure

The network structure of nonlinear dynamical systems is often implicitly defined by the state-space realization. Thus, the process of network reconstruction for the full system becomes a nonlinear parameter estimation or state-space realization problem. Such network

reconstruction problems are non-convex, only locally identifiable at best, under-constrained due to the sampling limits of experimental data, and even ill-posed at times.

A class of dynamical systems where the concept of network structure is well-defined and reconstruction results are readily available are linear time-invariant (LTI) dynamical systems [37]. The most intricate description of network structure of LTI systems refers to the network defined by interactions between every state in the system. Reconstructing the system's network structure is equivalent to finding a unique solution for the state-space realization. However, it is well known that uniquely determining the state-space realization requires full-state feedback, otherwise the problem is ill-posed. [107]. It is thus valuable to find different representations of network structure, consistent with the state-space realization, that encode essential structural information, but that impose less stringent constraints on network reconstruction.

Arguably the simplest yet most broadly employed representation of network structure is the system transfer function. The transfer function describes the closed-loop causal dependencies of system outputs on system inputs. As such, it imposes weak information constraints on the process of network reconstruction. As long as it is possible to perturb the system with each input and measure each output, it is possible to reconstruct the transfer function of the system. Still, the transfer function contains very little structural information; the price of relatively relaxed constraints on the network reconstruction problem is that very little information about the actual network structure of the system, e.g. how states in the system depend on each other and interact, is encoded in the transfer function.

The tradeoffs between cost of network reconstruction and the "informativity" of the structural representation are especially clear in synthetic and systems biology research. In this area, finding or verifying the network of a biological system is an important problem. However, discovering the entire chemical reaction network is typically an ill-posed problem, since additional reactions may be introduced due to host or environmental context, loading effects, or unanticipated retroactivity effects [17]. Even without these effects, the reconstruc-

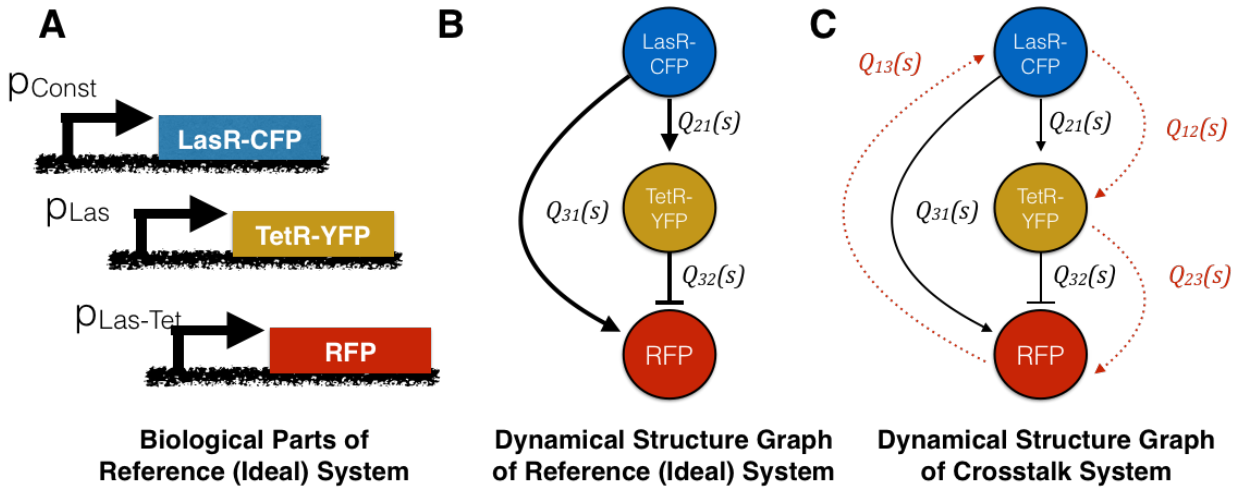

Figure 3.1: **Dynamical structure functions can be used to analyze synthetic gene networks: (A)** The dynamical structure of system (3.7). Nodes represent measured chemical species, with black edges denoting causal dependencies stemming from designed interactions, and red edges denoting causal dependencies arising from crosstalk or loading effects. Notice that the dynamical structure captures network models interactions that are not described by the system transfer function $G(s)$. **(B)** The input-output response of the nonlinear system (3.7). Standard parameters from the literature [65] were used to generate the simulation. As the size of the load Δ_{load} increases, the ability of the IFFL to respond with a pulse decreases. **(C)** The maximum fold-change in the $\mathcal{H}_\infty$ norm of the crosstalk entries in $Q(s)$. The H_∞ norm of $Q_{31}(s)$, plotted as a function of Δ_{load}. As the amplitude of directed crosstalk of x_1 (LasR-CFP) on x_3 (RFP) increases, the pulse height of RFP expression increases since the increased gain of Q_{31} allows RFP to achieve higher expression before TetR repression activates. However, the increase in crosstalk also means that TetR repression is less effective, resulting in steady-state drift as Clp-XP load increases.

tion problem is equivalent to finding a unique realization for the dynamical system, which is ill-posed without measurements of every chemical species in the system. On the other hand, there are many inputs that can be used to perturb the system of interest, e.g. silencing RNA, genetic knock-outs, and small chemical inducers. Using these inputs, it is straightforward to reconstruct the transfer function of the system. However, the transfer function contains virtually no information about how chemical species within the system are interacting.

An intermediate representation of network structure that addresses this trade-off is the dynamical structure function [37]. It is a more detailed description of network structure than the transfer function since it models the causal interactions between measured outputs, in addition to the causal dependencies of outputs on input variables. At the same time, it does not require complete state feedback for reconstruction, since it only models the interactions among output states. In biological systems, this is especially applicable since the output variables of a system are also a subset of the state variables. All unmeasured states are subsumed in the edge-weight functions that describe interactions between measured variables. It is thus possible to experimentally target specific chemical species to measure and verify that the network structure of a biological system is functioning as intended.

We briefly review the theory of dynamical structure functions, as they pertain to biochemical reaction networks. In practice, the state of the dynamical system $x = \begin{bmatrix} y^T & x_h^T \end{bmatrix}^T \in \mathbb{R}^n$, where $y \in \mathbb{R}^p$ are the measured chemical states of the dynamical system, corresponding to components of the biochemical reaction network tagged with fluorescent reporters, and $x_h \in \mathbb{R}^{n-p}$ are the unmeasured chemical states. It is also the cases that there are exogenous inputs $u \in \mathbb{R}^m$ that can be introduced to influence the dynamics of the state x. With the exception of oscillators, many biochemical reaction networks converge to a steady state. Moreover, it is generally the case that the parameters of biochemical reaction networks are time-invariant, so long as macroscopic experimental settings of the system such as temperature, growth media, and dissolved oxygen content remain fixed. Therefore, while the general

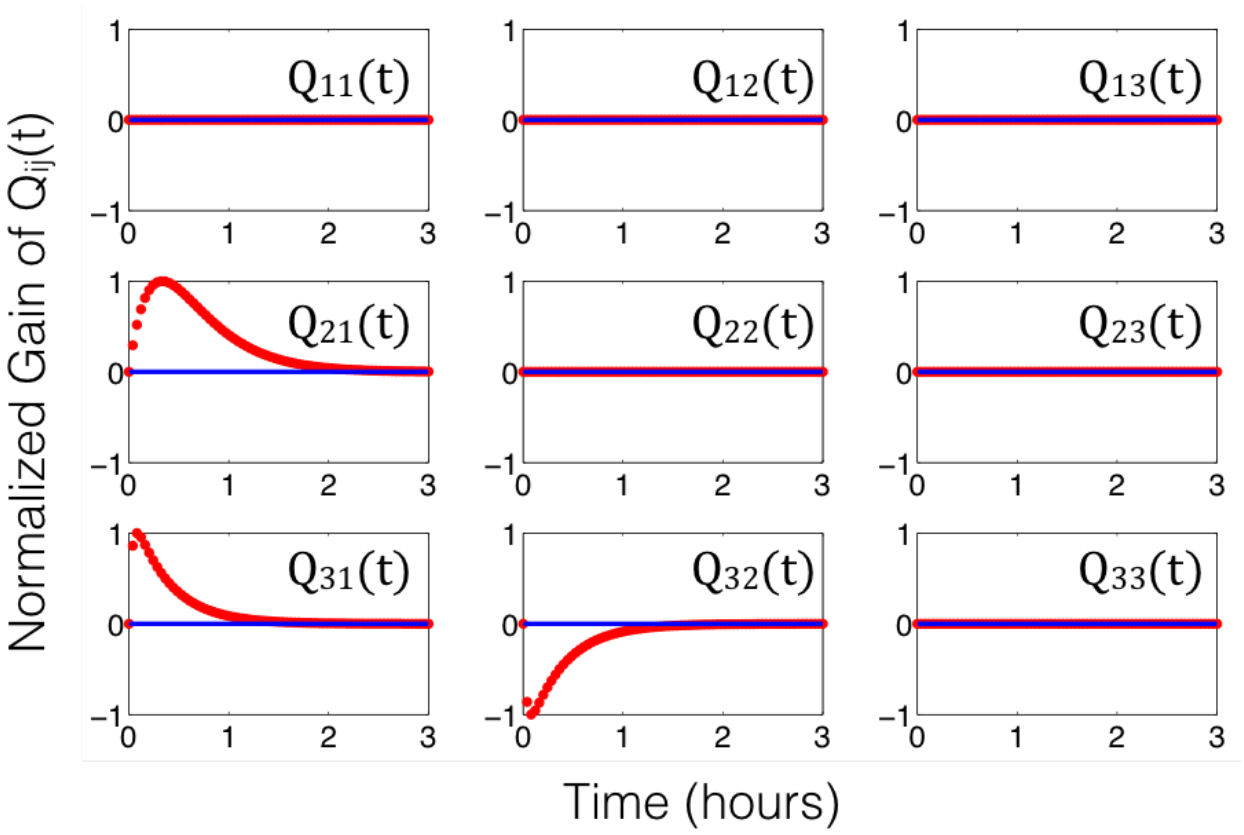

Figure 3.2: **Dynamical structure functions describe how network structure evolves over time (and as a function of frequency):** The time-lapse response of the dynamical structure convolution kernel $Q^a(t) = \mathcal{L}^{-1}(Q^a(s))$ for the incoherent feedforward loop in system (3.6). By examining the functional response of each entry in $Q(t)$ (or $Q(s)$), we see that the network structure of the incoherent feedforward loop in Example 3.2.1 is a time-evolving, or *dynamic*, entity.

model of a biochemical reaction network is of the form

$$\begin{aligned} \dot{y} &= f_y(y, x_h, u), \quad y(0) = y_0 \\ \dot{x}_h &= f_{x_h}(y, x_h, u), \quad x_h(0) = x_{h,0}, \\ y &= \begin{bmatrix} \mathbf{I}_{p\times p} & 0 \end{bmatrix} \begin{bmatrix} y \\ x_h \end{bmatrix} \end{aligned} \tag{3.1}$$

we will suppose that we can linearize the system about an equilibrium point, to write it in

the form:

$$\begin{aligned}\begin{bmatrix}\dot{y}\\ \dot{x}_h\end{bmatrix} &= \begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}\begin{bmatrix}y\\ x_h\end{bmatrix} + \begin{bmatrix}B_1\\ B_2\end{bmatrix} u\\ y &= \begin{bmatrix}\mathbf{I}_{p\times p} & 0\end{bmatrix}\begin{bmatrix}y\\ x_h\end{bmatrix}.\end{aligned} \tag{3.2}$$

We assume the initial condition of the linearized system is $x(0) = 0$, and the entries in $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are calculated as

$$\begin{aligned} A_{11} &\equiv \tfrac{\partial f_y(y,x_h,u)}{\partial y}|_{x=x_e,u=u_e}, & A_{12} &\equiv \tfrac{\partial f_y(y,x_h,u)}{\partial x_h}|_{x=x_e,u=u_e}\\ A_{21} &\equiv \tfrac{\partial f_{x_h}(y,x_h,u)}{\partial y}|_{x=x_e,u=u_e}, & A_{22} &\equiv \tfrac{\partial f_{x_h}(y,x_h,u)}{\partial x_h}|_{x=x_e,u=u_e}\\ B_1 &\equiv \tfrac{\partial f_y(y,x_h,u)}{\partial u}|_{x=x_e,u=u_e}, & B_2 &\equiv \tfrac{\partial f_{x_h}(y,x_h,u)}{\partial u}|_{x=x_e,u=u_e}\end{aligned}$$

Taking Laplace transforms and solving for $X_h(s)$, taking $x(0) = 0$, we obtain

$$sY = W(s)Y(s) + V(s)U(s) \tag{3.3}$$

where

$$\begin{aligned} W(s) &= A_{11} + A_{12}(sI - A_{22})^{-1}A_{21}\\ V(s) &= B_1 + A_{12}(sI - A_{22})^{-1}B_2.\end{aligned} \tag{3.4}$$

Defining $D(s) = \text{diag}\,(W(s))$ and subtracting $D(s)$ from both sides of equation (3.3) and solving for $Y(s)$ we obtain the following equation

$$Y = Q(s)Y(s) + P(s)U(s) \tag{3.5}$$

where $Q(s) = (sI-D)^{-1}(W-D)$ is a $p\times p$ transfer function matrix and $P(s) = (sI-D)^{-1}V$ is a $p\times m$ transfer function matrix. Each entry $Q_{ij}(s)$ is a transfer function that describes the causal dependency of measured state $Y_i(s)$ on measured state $Y_j(s)$. Similarly, the transfer function $P_{ij}(s)$ describes the causal dependency of measured state $Y_i(s)$ on input $U_j(s)$. The matrix pair $(Q(s), P(s))$ is known as the *dynamical structure function*, where $Q(s)$ is referred

to as the network structure and $P(s)$ as the control structure. We illustrate these concepts with an example biochemical reaction network.

3.2.1 The Dynamical Structure Function of an Idealized Incoherent Feedforward Loop

Consider the following synthetic biology design problem: design and implement a novel incoherent feedforward loop. Specifically, we consider implementing a feedforward loop using the synthetic parts pLac-LasR-CFP-LVA, pLas-TetR-YFP-LVA, and pLas-Tet-RFP-LVA and IPTG, $C_3O_6H_{12} - \text{HSL}$, and aTc as inputs. A simple model for this system without any loading effects is given as:

$$\begin{aligned}
\dot{x}_1 &= \rho_1 m_1 - \frac{C_0 x_1/k_{1,d}}{1 + x_1/k_{1,d}} \\
\dot{x}_2 &= \rho_2 m_2 - \frac{C_0 x_2/k_{2,d}}{1 + x_2/k_{2,d}} \\
\dot{x}_3 &= \rho_3 m_3 - \frac{C_0 x_3/k_{3,d}}{1 + x_3/k_{3,d}}
\end{aligned} \tag{3.6}$$

$$\begin{aligned}
\dot{m}_1 &= \frac{\alpha_1 u_1}{k_{M,u1} + u_1} - \delta_m m_1 \\
\dot{m}_2 &= \frac{\alpha_2 (x_1 u_2/k_{M,u2})}{1 + x_1/k_{M,1} + x_1 u_2/k_{M,u2}} - \delta_m m_2 \\
\dot{m}_3 &= \frac{\alpha_3 x_1 u_2}{1 + x_1 u_2/k_{Mu2} + x_2/\left(k_{M,2} + u_3/k_{M,u3}\right)} - \delta_m m_3 \\
y &= \begin{bmatrix} \mathbf{I}_{3\times 3} & \mathbf{0}_{3\times 3} \end{bmatrix} \begin{bmatrix} \vec{x}^T & \vec{m}^T \end{bmatrix}^T
\end{aligned}$$

The dynamical structure function for this system is derived by taking Laplace transforms and eliminating the hidden states m_1, m_2, m_3, see [37] or [1] for a detailed derivation of dynamical structure functions. The network and control structure matrix transfer functions

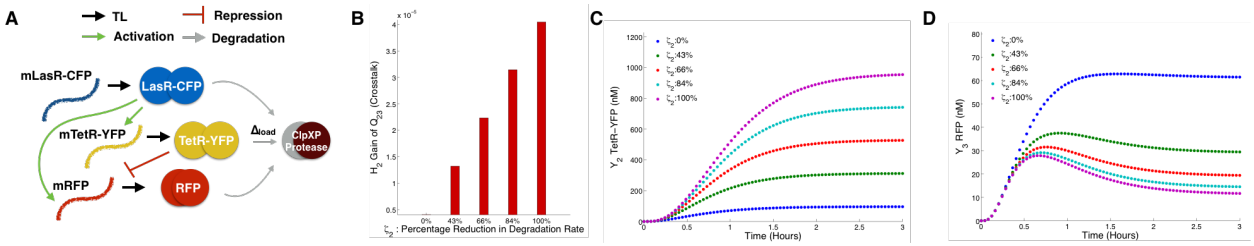

Figure 3.3: **Dynamical structure functions quantify biomolecular crosstalk:(A)** A schematic illustrating the design of this simulation example. The crosstalk and reference model of the incoherent feedforward loop from Examples 3.2.1 and 3.2.2 are simulated accordingly to satisfy internal equivalence, for varying values of $k_{d,2}$. **(B-C)** Time lapse responses of the incoherent feedforward loop: for each value of $k_{d,2}$ the value of ζ_2 at $t = 3$ hours is calculated and used to label curves (as percentage of maximum load). Notice the monotonic relationship between $k_{d,2}$, ζ and the output responses of Y_2 and Y_3 (negatively monotonic). **(D)** The $\mathcal{H}_2$ gain of $Q^c_{23}(s)$ is plotted as a function of ζ. Notice that $Q^c_{23}(s)$ is a pure crosstalk term, since $Q^a_{23}(s) \equiv 0$. As the effective crosstalk in ζ_2 increases, $Q^c_{23}(s)$ mirrors that increase, as shown in Proposition 1.

are written $(Q_a(s), P_a(s))$ where $Q_a(s)$ is written as

$$\begin{pmatrix} 0 & 0 & 0 \\ \frac{0.045}{s^2+1.5\,s+5.7\cdot10^{-4}} & 0 & 0 \\ \frac{1.5\cdot10^{-4}}{s^2+1.7\,s+0.2} & -\frac{1.5\cdot10^{-4}}{s^2+1.7\,s+0.2} & 0 \end{pmatrix}$$

and $P_a(s)$ is

$$\begin{pmatrix} \frac{6.7\cdot10^{-7}}{s^2+1.5\,s+5.7\cdot10^{-4}} & 0 & 0 \\ 0 & \frac{244.0}{s^2+1.5\,s+5.7\cdot10^{-4}} & 0 \\ 0 & \frac{0.78}{s^2+1.7\,s+0.2} & \frac{7.8}{s^2+1.7\,s+0.2} \end{pmatrix}.$$

Notice that $P_a(s)$ is a lower-triangular matrix and satisfies sufficient conditions for network reconstruction [37]. The network, with edge weight functions corresponding to the entries of $Q_a(s)$, is drawn in Figure 3.1A. Notice that if we take $s \in \mathbb{R}_{>0}$, the sign of the entries in $Q_a(s)$ coincides with the form of transcriptional regulation implemented by TetR and LasR, respectively. In [102] it was shown that the sign definite properties of entries in $Q(\mathbb{R}_{>0})$ are useful for reasoning about the monotonicity of interactions between measured outputs and how fundamental limits in system performance relate to network structure.

Let us now consider the inverse Laplace transform of $\mathcal{L}^{-1}\left(Q^a(s)\right)$, we remark that $Y(t) = \int_0^t Q^a(t)Y(t-\tau)$ follows from the equation

$$\mathcal{L}^{-1}\left(Y(s)\right) = \mathcal{L}^{-1}\left(Q^aY(s) + P^aU(s)\right)$$

whenever $u(t) \equiv 0$ such that $U(s)$ is 0. This argument holds in general for any system of the form (3.2). In particular, the entries $Q^a(t)$ act as convolution kernels, and taken with the integral, define an operator for mapping $y_j(t)$ to $y_i(t)$. Most interestingly, we can see that the network structure of this incoherent feedforward loop is *dynamical*, hence our usage of the term *dynamical structure* function to describe the network structure among the measured chemical species $y(t)$. In this particular case, the time-domain analogue of the dynamical structure (or dynamical structure convolution kernel) is given as

$$Q^a(t) \equiv \begin{pmatrix} 0 & 0 & 0 \\ Q_{21}(t) & 0 & 0 \\ Q_{31}(t) & Q_{32}(t) & 0 \end{pmatrix}$$

where

$$Q_{21}(t) = \left(5.5 \cdot 10^{-3}\right)\, \mathrm{e}^{-\left(8.3 \cdot 10^{-4}\right) t}\, \sinh\left(\left(6.0 \cdot 10^{-5}\right)\, t\right)$$

$$Q_{31}(t) = \left(4.1 \cdot 10^{-7}\right)\, \mathrm{e}^{-\left(4.9 \cdot 10^{-3}\right) t}\, \sinh\left(\left(4.1 \cdot 10^{-3}\right)\, t\right)$$

$$Q_{32}(t) = -\left(9.7 \cdot 10^{-8}\right)\, \mathrm{e}^{-\left(4.9 \cdot 10^{-3}\right) t}\, \sinh\left(\left(4.1 \cdot 10^{-3}\right)\, t\right)$$

3.2.2 The Dynamical Structure Function of an Incoherent Feed Forward Loop with Crosstalk

To truly prototype a novel feedforward loop, it is important to anticipate *in vivo* context effects. In this biocircuit, the components are particularly susceptible to loading effects [17]. In synthetic biological circuits, a protease called Clp-XP targets and degrades LVA-tagged

proteins. This protease can be found in limited supply when there are too many LVA-tagged proteins [27]. Modifying the above model to account for these type of loading effects yields:

$$\begin{aligned}
\dot{x}_1 &= \rho_1 m_1 - \frac{C_0 x_1/k_{1,d}}{1 + x_1/k_{1,d} + x_2/k_{2,d} + x_3/k_{3,d}} \\
\dot{x}_2 &= \rho_2 m_2 - \frac{C_0 x_2/k_{2,d}}{1 + x_1/k_{1,d} + x_2/k_{2,d} + x_3/k_{3,d}} \\
\dot{x}_3 &= \rho_3 m_3 - \frac{C_0 x_3/k_{3,d}}{1 + x_1/k_{1,d} + x_2/k_{2,d} + x_3/k_{3,d}}
\end{aligned} \tag{3.7}$$

$$\begin{aligned}
\dot{m}_1 &= \frac{\alpha_1 u_1}{k_{M,u1} + u_1} - \delta_m m_1 \\
\dot{m}_2 &= \frac{\alpha_2 (x_1 u_2/k_{M,u2})}{1 + x_1/k_{M,1} + x_1 u_2/k_{M,u2}} - \delta_m m_2 \\
\dot{m}_3 &= \frac{\alpha_3 x_1 u_2/k_{M,u2}}{1 + x_1 u_2/k_{Mu2} + x_2/\left(k_{M,2} + u_3/k_{M,u3}\right)} - \delta_m m_3 \\
y &= \begin{bmatrix} \mathbf{I}_{3\times 3} & \mathbf{0}_{3\times 3} \end{bmatrix} \begin{bmatrix} \vec{x}^T & \vec{m}^T \end{bmatrix}^T
\end{aligned} \tag{3.8}$$

Computing the dynamical structure function, we obtain $Q_c(s)$

$$\begin{pmatrix}
0 & \frac{1.6\cdot 10^{-3}}{s+2.1\cdot 10^{-3}} & \frac{0.041}{s+2.1\cdot 10^{-3}} \\
\frac{\left(1.6\cdot 10^{-3}\right) s+0.048}{s^2+1.5\, s+3.3\cdot 10^{-3}} & 0 & \frac{0.041}{s+2.1\cdot 10^{-3}} \\
\frac{\left(3.8\cdot 10^{-4}\right) s+7.4\cdot 10^{-4}}{s^2+1.6\, s+0.13} & \frac{\left(3.8\cdot 10^{-4}\right) s+4.4\cdot 10^{-4}}{s^2+1.6\, s+0.13} & 0
\end{pmatrix}$$

and $P_c(s)$

$$\begin{pmatrix}
\frac{6.7\cdot 10^{-7}}{s^2+1.5\, s+3.2\cdot 10^{-3}} & 0 & 0 \\
0 & \frac{244.0}{s^2+1.5\, s+3.3\cdot 10^{-3}} & 0 \\
0 & \frac{0.78}{s^2+1.6\, s+0.13} & \frac{7.8}{s^2+1.6\, s+0.13}
\end{pmatrix}.$$

Notice that $Q_c(s)$ is no longer lower-triangular, but fully connected. Introducing loading effects creates additional coupling between nodes in the network. If the coupling is significant, the *designed* network interactions of the incoherent feedforward loop are overcome by the *crosstalk* network interactions [102]. Thus, the coupling that is introduced into the biochemical reaction network by loading effects is reflected in the structure of $(Q_c, P_c)(s)$.

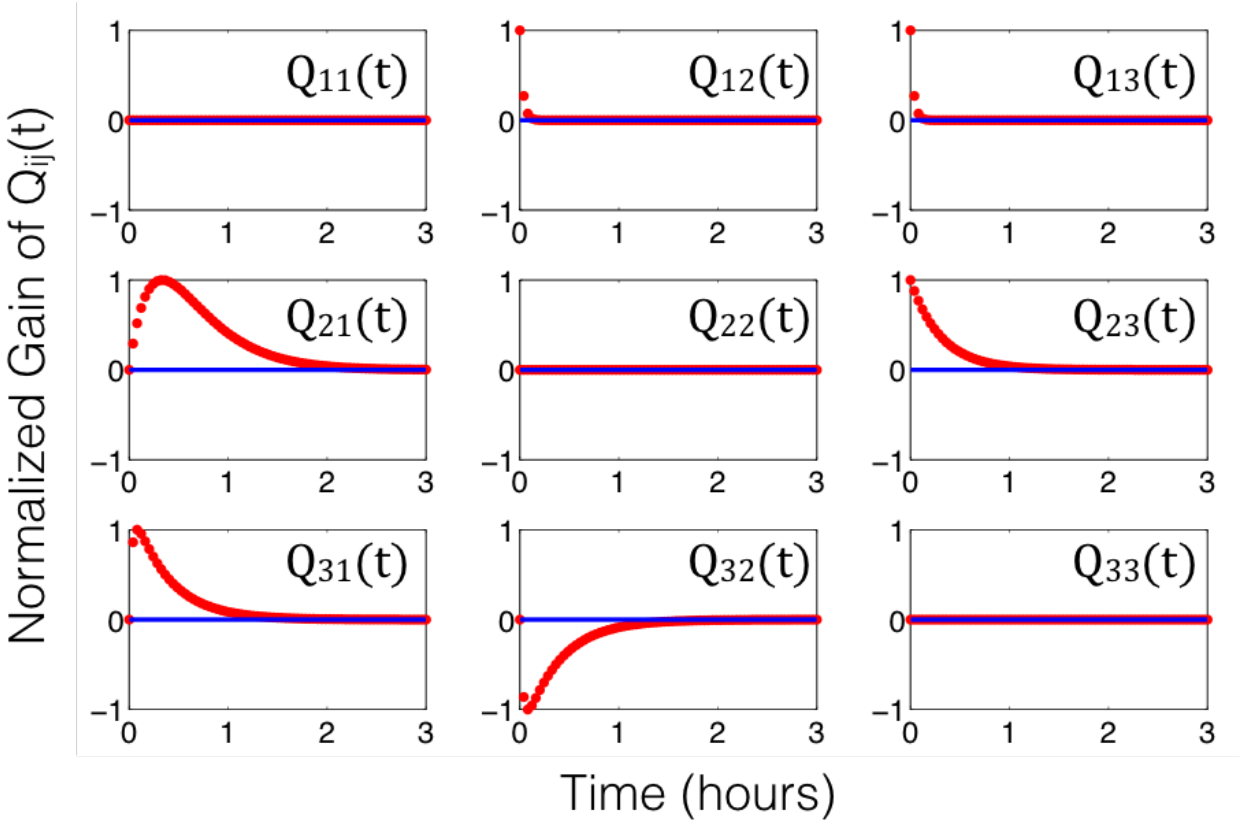

Figure 3.4: **Dynamical structure functions describe how network structure evolves over time (and as a function of frequency):** The time-lapse response of the dynamical structure convolution kernel $Q^a(t) = \mathcal{L}^{-1}(Q^c(s))$ for the incoherent feedforward loop in system (3.6). By examining the functional response of each entry in $Q(t)$ (or $Q(s)$), we see that the network structure of the incoherent feedforward loop in Example 3.2.1 is a time-evolving, or *dynamic*, entity.

In contrast, the transfer function of the crosstalk system only characterizes how system outputs causally depend on inputs. In particular, $G(s)$ is also a full matrix like $Q_c(s)$ of 6th order SISO transfer functions

$$\begin{bmatrix} G_{11}(s) & G_{12}(s) & G_{13}(s) \\ G_{21}(s) & G_{22}(s) & G_{23}(s) \\ G_{31}(s) & G_{32}(s) & G_{33}(s) \end{bmatrix} (s)$$

but all structural information about how loading effects cause interference *among* system states is mixed with the information about how outputs causally depend on inputs in $G(s)$.

An identification algorithm of entries in $G(s)$ will thus be *unable* to quantify the size of crosstalk or interference among system states.

To what extent can the entries of $(Q(s), P(s))$ can be used to quantify the size of crosstalk in a synthetic gene networks? To address this question, we review and expand on the prior results in [102] for quantifying crosstalk in biochemical reaction networks and highlight the relationship of these nonlinear models with $(Q(s), P(s))$.

3.3 Quantifying Crosstalk in Biochemical Reaction Networks

A common way that crosstalk arises in biochemical reaction networks is when species compete for commonly shared enzymes. When this occurs, the sequestration of an enzyme by one competing species makes the enzyme less accessible to other competing species. For example, when two mRNA are competing for a single ribosome, the binding of one mRNA to the ribosome during translation makes it less accessible to other mRNA. At the core of any such crosstalk is a sudden increase in the dependency of one biochemical state on another. Though enzyme loading may be a common source of crosstalk, such interactions can be modeled at a higher level of abstraction, namely how the dynamics of a given state are affected by the movement of nearby states.

Nearly every synthetic gene network implements causal dependencies among states. Often, these "designed" interactions take the form of transcription factor binding, sense-antisense mRNA regulation, and sequestration events. In practice, every physical system exhibits trajectories that are a mixture of the consequences of both interaction types: designed and crosstalk interactions. Throughout the course of this chapter, we will denote the physical system of interest in our models as

$$\begin{aligned} \dot{y} &= f_y^c(y, x_h, u), \quad y(0) = y_0 \\ \dot{x}_h &= f_{x_h}^c(y, x_h, u), \quad x_h(0) = x_{h,0}, \\ y &= \begin{bmatrix} \mathbf{I}_{p\times p} & 0 \end{bmatrix} \begin{bmatrix} y \\ x_h \end{bmatrix} \end{aligned} \tag{3.9}$$

To quantify crosstalk in such systems, we can compare the dynamics of system (3.9) against the dynamics of a reference or alternative system that is free of crosstalk. Such a reference system will still retain all crosstalk-free interaction dynamics and reflects the idealized model often used to design a synthetic gene network, e.g. the feedforward loop model in Example 3.2.1. Moreover, it can represent the desired behavior of the system in a regime where the magnitude of crosstalk effects are supposed to be minimal or engineered in such a way that they are suppressed [17]. We write the reference system as

$$\begin{aligned} \dot{y} &= f_y^a(y, x_h, u), \quad y(0) = y_0 \\ \dot{x}_h &= f_{x_h}^a(y, x_h, u), \quad x_h(0) = x_{h,0}, \\ y &= \begin{bmatrix} \mathbf{I}_{p\times p} & 0 \end{bmatrix} \begin{bmatrix} y \\ x_h \end{bmatrix}. \end{aligned} \tag{3.10}$$

Remark 1. *For the comparison between the alternative and crosstalk system to be fair, it is important that (3.10) satisfy internal equivalence [81]. Specifically, we will suppose that any parameters or dynamics unassociated with crosstalk, e.g. interaction dynamics, catalytic reactions, or anabolic reactions with no loading effects, are held fixed. Thus, as we compare the behavior of both systems, any differences in the hidden state x_h or output y dynamics are purely due to effects of crosstalk.*

With the definition of an alternative system in place, it becomes possible to reason about the size of crosstalk, by comparing the dynamics of both systems. In particular, we can develop a rigorous notion for describing the amount of crosstalk arising from the difference

of trajectories in both systems.

Definition 4 (Crosstalk Trajectory)**.** *For each initial condition* $x(0) = (y(0), x_h(0)) \in \mathbb{R}^n$ *and input trajectory* $u(t)$ *we define the crosstalk trajectory* $\zeta(t)$ *as*

$$\zeta(t) = x^a(t) - x^c(t)$$

The crosstalk trajectory is a time-evolving vector that describes the deviation of the physical system (subject to crosstalk) from the reference system's trajectory. With this notion of crosstalk, we can also make precise the concept of crosstalk between states. We note that in writing the following quantity of interest $\frac{\partial}{\partial x_j}\zeta_i$, it is with a slight abuse of notation, since $\zeta_i(x^a(t), x^c(t))$. Mathematically, we are computing the j^{th} partial derivative of each term in $\zeta_i = f^a(x^a, u) - f^c(x^c, u)$. Thus, to be clear, when we write $\frac{\partial}{\partial x_j}\zeta_i$, it will be implicit that we mean $\frac{\partial}{\partial x_j^a} f_i^a(x^a, u) - \frac{\partial}{\partial x_j^c} f_i^c(x^c, u)$.

Definition 5 (Directed Crosstalk)**.** *We say that a chemical species* x_j *exerts a crosstalk effect on chemical species* x_i*if the* i^{th} *component of the crosstalk trajectory* $\zeta(t)$ *has nonzero partial derivative*

$$\frac{\partial}{\partial x_j}\zeta_i(t) \neq 0$$

for some initial condition of $(x(0), y(0))$ *and input trajectory* $u(t)$*. In general, we will refer to* $\frac{\partial}{\partial x_j}\zeta_i(t)$ *as the crosstalk sensitivity of* x_i *to* x_j*.*

Example 1. *Consider two mRNA species* m_1 *and* m_2 *competing for the same degradation enzyme* D *in a physical system. For simplicity of exposition, suppose their production dynamics do not depend on each other and can be modeled as* $P_1(t)$ *and* $P_2(t)$ *respectively. The crosstalk system is given as*

$$\dot{m}_1 = P_1(t) - \frac{D_0 m_1/k_{M,1}}{1 + m_1/k_{M,1} + m_2/k_{M,2}}$$
$$\dot{m}_2 = P_2(t) - \frac{D_0 m_2/k_{M,2}}{1 + m_1/k_{M,1} + m_2/k_{M,2}}$$

while the reference system is given as

$$\dot{m}_1 = P_1(t) - \frac{D_0 m_1/k_{M,1}}{1 + m_1/k_{M,1}}$$
$$\dot{m}_2 = P_2(t) - \frac{D_0 m_2/k_{M,2}}{1 + m_2/k_{M,2}}.$$

In both systems, we have supposed that time has been rescaled so that the customary parameter k_{cat} for degradation is unity. The crosstalk sensitivity of m_1 and m_2 (with respect to each other) are given as

$$\frac{\partial \zeta_1}{\partial m_2} = \frac{\partial}{\partial m_2} \int_0^t \frac{-D_0\, m_1/k_{M,1}\ m_2/k_{M,2}}{(1 + m_1/k_{M,1})(1 + m_1/k_{M,1} + m_2/k_{M,2})}$$

$$\frac{\partial \zeta_2}{\partial m_1} = \frac{\partial}{\partial m_1} \int_0^t \frac{-D_0 m_1/k_{M,1}\ m_2/k_{M,2}}{(1 + m_2/k_{M,2})(1 + m_1/k_{M,1} + m_2/k_{M,2})}$$

respectively. Clearly the crosstalk between m_1 and m_2 is nonzero.

Remark 2. *In synthetic biocircuit design, two chemical species x_i and x_j are often declared orthogonal when there is no designed interaction between them. Mathematically, in the absence of crosstalk, this corresponds to*

$$\frac{\partial}{\partial x_j} f_i^a(t) \equiv 0$$

for all $x(0)$ and $u(t)$. In such a situation, $\zeta(x_i, x_j) \neq 0$ if and only if

$$\frac{\partial}{\partial x_j} x_i^c(t) = \int_0^t f_i^c(y, x_h, u) d\tau \neq 0.$$

This condition is interesting in experimental settings since a computational estimate of $\frac{\partial}{\partial x_j} \int_0^t f_i^c(t)$ from perturbation experiments coincides with a direct estimate of the sensitivity of the crosstalk $\frac{\partial}{\partial x_j} \zeta_i$. More specifically, when x_i and x_j are measured outputs of the system, we will show in the sequel that quantifying $||Q_{i,j}^c(s)||$ is directly related to an estimate of $\Delta(x_i, x_j)$ near the equilibrium point x_e^c.

Remark 3. *In general, estimating the crosstalk $\Delta(x_i, x_j)$ for the nonlinear systems (3.9) and (3.10) can be challenging if either x_i and x_j are not measured directly. Firstly, if experimental data is available, it will often consist of data for the measured species y in the crosstalk-system, but not the reference system. Second, if only one of the species x_i (or none) is available for measurement, even if perturbation of x_j is possible, a nonlinear observer is required to estimate the trajectory of $x_j(t)$. Unless the parameters of $f_i(x, u)$ are known a priori (which is generally not the case), this then also requires system identification of the parameters of $f_c(x, u)$ and $f_a(x, u)$ which often results in a non-convex optimization problem.*

Thus, our goal is to estimate the observed crosstalk between measured species Y_i and Y_j. This crosstalk estimate will invariably include the dynamics of unmeasured chemical species (such as ATP, RNAP, untagged mRNA and protein species, DNA-protein complexes etc.). From a synthetic biology design standpoint, this is not a disadvantage, since the goal is to design a synthetic gene network with a *system-verified* feedback architecture operating reliably in the context of many unmeasured species. There will always be additional chemical species that are unmeasured. Our goal is to validate that a biocircuit (e.g. an IFFL, repressilator, or completely novel biocircuit) still manifests the intended network structure even in the presence of unmeasured dynamics.

Proposition 1. *Let $\mathcal{L}$ denote the Laplace operator. Suppose the states x^c and x^a of the systems (3.9) are (3.10) are shifted, so that the origin is a locally asymptotically stable equilibrium point and Q^c and Q^a are the respective dynamical structure functions calculated for each linearized system about the origin. Then*

$$\frac{\partial \mathcal{L}\left(\zeta_i\right)}{\partial Y_j} = Q^a_{ij}(s) - Q^c_{ij}(s) + \mathcal{L}\left(O(x^2)\right) \tag{3.11}$$

and in particular, if

$$Q^a_{ij}(s) \equiv 0$$

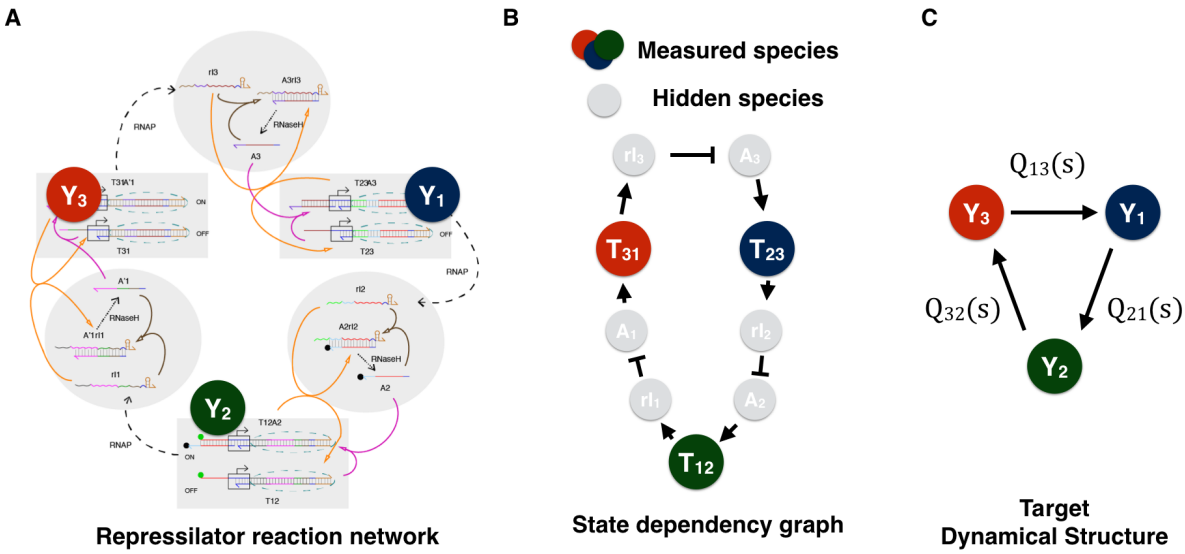

Figure 3.5: **Network representations of a synthetic genelet repressilator: (A)** A reaction network using push-arrow reaction notation of the synthetic genelet repressilator. **(B)** A diagram representing the reaction dynamics in panel **(A)** as state dependencies from a nonlinear ODE model in [49]. **(C)** The dynamical structure of the repressilator (without inputs), with nodes representing measured chemical species and edge weights corresponding to entries in $Q(s)$.

then

$$\frac{\partial \mathcal{L}(\zeta_i)}{\partial Y_j} = Q^c_{ij}(s) + \mathcal{L}\left(O(x^2)\right)$$

and can be estimated from input output data $(Y(s), U(s))$.

Proof. First, notice that the Laplace transform of $\mathcal{L}(\zeta(t)) = \mathcal{L}(x^a - x^c) \triangleq X^a(s) - X^c(s)$, which can be decomposed into its measured and unmeasured states

$$\begin{aligned}
&\begin{bmatrix} Y^a \\ X^a_h \end{bmatrix}(s) - \begin{bmatrix} Y^c \\ X^c_h \end{bmatrix}(s) \\
&= \begin{bmatrix} Q^a Y^a(s) - Q^a Y^c(s) + (P^a - P^c)U(s) + \mathcal{L}\left(O(x^2)\right) \\ X^a_h(s) - X^c_h(s) + \mathcal{L}\left(O(x^2)\right) \end{bmatrix}.
\end{aligned}$$

Examining the i^{th} component equation and taking partials along $Y_j(s)$ yields equation (3.11).

□

Remark 4. *This result is important, since it tells us when estimating $Q^c(s)$ from experimental data will correspond to estimating crosstalk between measured states in $Y(s)$. Since necessary and sufficient conditions for identifying $Q(s)$ and $P(s)$ have been already characterized [37], this immediately yields a proof of identifiability for inferring crosstalk from input-output data.*

More generally, even if parameters for $f^a(x,u)(t)$ are unknown, the structure of $Q^a(s)$ can be analytically calculated (using a symbolic algebra package). For every zero entry in $Q^a(s)$ (coinciding with designed orthogonality between measured states), we can then estimate $Q^c(s)$ directly.

Remark 5. *In practice, estimation of $Q^c(s)$ is also confounded by noise. In our analysis in this thesis, we suppose that a series of filters can be applied to eliminate the noise in the data. This may not be the case for biological systems that have been characterized as inherently stochastic, e.g. single cell gene expression dynamics. In such settings, the estimated dynamical structure $Q^c(s)$ is a mixture of the process noise in the system and the crosstalk. From the standpoint of synthetic biocircuit prototyping, both are undesirable in the ultimate iteration of the biocircuit and thus need to be quantified. In this thesis, we will focus on integrating theory and experimental results for in vitro systems with strong signal-to-noise ratios and only measurement noise. For a theoretical treatment of how to reverse engineer $Q^c(s)$ in the presence of process noise or system perturbation, see [105].*

An advantage of using $Q^c(s)$ to estimate the crosstalk is that we can use the $\mathcal{H}_\infty$ norm of $Q^c_{i,j}(s)$ to calculate the worst-case crosstalk magnitude and $\mathcal{H}_2$ of $Q^c_{i,j}(s)$ to calculate the average crosstalk across all frequencies.

Example 2 (Quantifying Crosstalk with $Q^c(s)$)**.** *Recall the incoherent feedforward loop in Examples 3.2.1 and 3.2.2. In particular, comparing $Q^a(s)$ and $Q^c(s)$ we see that $Q_c(s)$ is a*

full transfer function matrix

$$\begin{pmatrix} 0 & \frac{1.6\cdot10^{-3}}{s+2.1\cdot10^{-3}} & \frac{0.041}{s+2.1\cdot10^{-3}} \\ \frac{\left(1.6\cdot10^{-3}\right)s+0.048}{s^2+1.5\,s+3.3\cdot10^{-3}} & 0 & \frac{0.041}{s+2.1\cdot10^{-3}} \\ \frac{\left(3.8\cdot10^{-4}\right)s+7.4\cdot10^{-4}}{s^2+1.6\,s+0.13} & \frac{\left(3.8\cdot10^{-4}\right)s+4.4\cdot10^{-4}}{s^2+1.6\,s+0.13} & 0 \end{pmatrix}$$

and $Q^a(s)$ is lower-triangular, reflecting the network structure of the intended IFFL. By examining the upper triangular entries in $Q^c(s)$, we can directly examine the effects of degradation crosstalk. In the lower entries of $Q^c(s)$, these crosstalk effects are confounded with the direct interactions modeled in $Q^a(s)$. Although the gain of the entries in $Q^c(s)$ are small, they nonetheless can have a significant effect on the dynamics of the IFFL.

In Figure 3.3 we plot the time-lapse response of $y_2(t)$ and $y_3(t)$ for varying parameter values of $k_{d,2}$. The $k_{d,2}$ parameter is a Michaelis constant that determines the effective competitiveness of substrate x_2 in binding with C_0. As $k_{d,2}$ increases, the competitiveness of substrate x_2 is diminished, relative to the competitiveness of x_1 and x_3. Attenuating $k_{d,2}$ can be viewed as similar to swapping out a strong LVA marker Clp-XP degradation with a weaker LVA marker on the species x_2. So far, in the experimental literature, there are only three known LVA markers that confer varying binding affinities to their associated protein. In our simulation, we consider five potential values for $k_{d,2}$: 500, 1625, 2750, 3875, and $5000\,\mu M$ corresponding to five artificial LVA markers of varying strengths.

Notice that as we decrease the effective competitiveness of y_2 for Clp-XP, this also coincides with an increased ζ_2 crosstalk magnitude. Here, we have computed $\zeta_2 = \int_0^t y_2^c(t) - y_2^a(t)$. We find that $|\zeta_2|$ increases as $k_{d,2}$ increases. In Figure 3.3B-D, ζ_2 is plotted as a percentage of maximum absolute change across all values of $k_{d,2}$.

We see that the time-lapse response of $y_2(t)$ increases monotonically for all t as the crosstalk $\zeta_2(t)$ increases. This is consistent with biological intuition, since an increase in competition for resource loading (or a decreased ability for y_2 to compete for enzymes) results in prolonged lifetimes of each individual y_2 (TetR-YFP) protein. This in turn results in

higher repression levels of y_3 in the incoherent feedforward loop. Increased competition for Clp-XP from substrates y_3 and y_1 have the effect of damping y_3 dynamics and reinforcing the pulsatile response of the IFFL. The crosstalk in this circuit thus has the effect of effectively strengthening the negative feedback of y_2 on y_3, encouraging the downward transient after $t \approxeq 0.75$ hours. Our network analysis shows we can improve the robustness of an IFFL's pulse by attenuating the relative binding affinity of the repressor to its protease.

In general, crosstalk effects do not necessarily reinforce the feedback architecture of a biocircuit. This underscores the importance of having techniques for quantifying crosstalk in a synthetic gene network and validating that designed interactions are dominant over crosstalk interactions. In the next two sections, we illustrate these concepts with experimental systems implemented in vitro and in vivo.

3.4 Identifying the Dynamical Structure of A Genelet Repressilator From Experimental Data

The best illustration of dynamical structure reconstruction is one that involves experimental data. We take as a first test case the synthetic genelet repressilator developed by Kim and Winfree [49]. The genelet repressilator consists of three DNA switches that repress one another through indirect sequestration. Specifically, each DNA switch transcribes its mRNA product only when its activator strand binds to complete its T7 RNA polymerase promoter sequence. The mRNA product produced from each DNA switch, in turn, acts as an inhibitor to the downstream switch by binding to the downstream switch's DNA activator molecule. Thus, by sequestering the DNA activator from completing the T7 RNA polymerase promoter region, the mRNA product of the upstream switch inhibits activation of the downstream switch. Figure 3.5A shows the mechanistic design of the genelet switch.

The genelet switch relies heavily on RNaseH to degrade any activator-mRNA inhibitor complexes. Without degradation, the binding of activator to mRNA inhibitor is much faster

than unbinding and so sequestration is effectively irreversible. Thus, in order for the repressilator to function properly, RNaseH must degrade its target substrates sufficiently fast. If RNaseH is saturated with high levels of a particular substrate, this slows the degradation of other substrates, creating a crosstalk interaction between competing DNA-RNA complexes.

By performing network reconstruction on the genelet repressilator, we can determine how much crosstalk exists in the biocircuit. To reconstruct $Q(s)$ and $P(s)$, we performed a single experiment with three perturbations applied in series. To perturb each switch we pipetted a small perturbative concentration of DNA inhibitor (a DNA analogue of RNA inhibitor). Since DNA is not degradable in a T7 expression system, it effectively acts as a step input since it binds to DNA activator and does not degrade. In this way, our perturbation design ensures sufficiency of excitation and independent perturbation of each activator (and downstream switch), thereby satisfying the conditions of Theorem 2 in [37] and Theorem in [58]

A detailed model of the repressilator can be found in the supplement of [49]. Since the derivation is lengthy, it suffices to write the idealized dynamical structure function $Q^a(s)$ of this system, corresponding to the detailed model provided in Supplementary Section 1.6. The structure is obtained by linearizing the system, transforming into the Laplace domain, eliminating hidden variables to obtain the following:

$$\begin{bmatrix} 0 & 0 & Q^a_{13}(s) \\ Q^a_{21}(s) & 0 & 0 \\ 0 & Q^a_{32}(s) & 0 \end{bmatrix}$$

reflecting the cyclic structure of the system. Though the parameters of $Q^a(s)$ are unknown, we know that for every entry where $Q^a_{ij}(s) \equiv 0$, estimating corresponding entry in $Q^c(s)$ from experimental gives a functional description of the crosstalk present in the network. The experimental data used to fit $Q^c(s)$ and $P^c(s)$ are plotted in Figure 3.6, along with their respective fits. For each row i of $Q^c(s)$, we use $Y_j, j \neq i$ and U_i as inputs and Y_i as the output for a direct MIMO $p \times 1$ transfer function estimation problem. The impulse response

for the convolution kernel $Q(t)$ of the reconstructed $Q(s)$ is plotted in Figure 3.7.

If we compute the corresponding $\mathcal{H}_\infty$ gain of each entry in $Q_{ij}(s)$ and scale by the maximum gain, we obtain

$$\begin{pmatrix} 0 & 0.07 & \mathbf{0.73} \\ \mathbf{1.0} & 0 & 0.3 \\ 0.053 & \mathbf{0.17} & 0 \end{pmatrix}.$$

We see significant crosstalk on the edge $Q_{23}(s)$ and minor crosstalk from entries $Q_{31}(s)$ and $Q_{12}(s)$. This crosstalk need not occur simultaneously, since the $\mathcal{H}_\infty$ gain calculates the worst-case or maximum gain over all possible frequencies. With the exception of $Q_{23}(s)$, all other crosstalk entries have strictly smaller $\mathcal{H}_\infty$ gain than the designed edge. Examining the impulse response of the convolution kernel confirms these observations; the crosstalk edge $Q_{23}(t)$ has a larger impulse response than *designed* edge$Q_{32}(t)$.

There is also a gain imbalance between the *designed* edges $Q_{32}(s)$, $Q_{13}(s)$ and $Q_{21}(s)$. In order for the oscillator to perform properly, it needs to have approximately the same gain along each edge in the network. Having applied our network reconstruction algorithm, this allows us to identify design-level criteria for improving the oscillator. In particular, we can increase the gain of the edge in $Q_{32}(s)$ by adjusting the binding affinity of the activator DNA with its inhibitor RNA, or by increasing the concentration of the corresponding downstream switch T_{23}. This design insight is not obvious when perusing the experimental trajectories of each switch in Figure 3.6. Inferring dynamical structure functions yields a mesoscopic view of system interactions — enough detail to pinpoint the source of failure at the component level, but abstracted enough to avoid the ill-posed nature of full state-space realization problems.

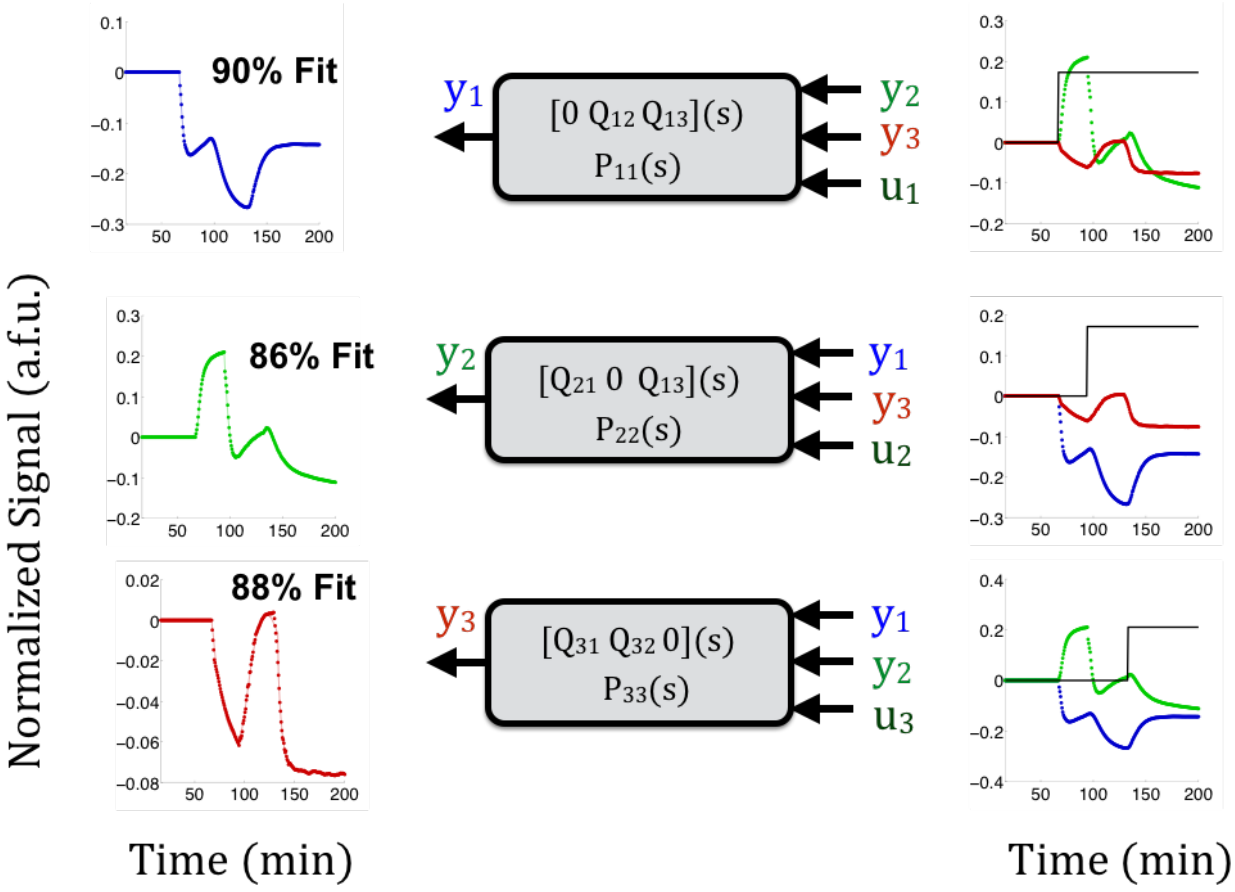

Figure 3.6: Time-series experimental data from *in vitro* network perturbation experiments of a T7 RNAP genelet repressilator. Three outputs are measured simultaneously, y_1, y_2, and y_3, corresponding to DNA switches T_{31}, T_{12} and T_{23}. DNA homologues of the RNA inhibitors rI_j $j = 1, 2, 3$ are injected at small concentrations to provide a step input perturbation to the corresponding component Y_j in the genelet circuit.

3.5 Using Network Reconstruction to Prototype and Validate a Novel Event Detector From Experimental Data

Network reconstruction can provide critical information to prototype and validate novel synthetic biocircuits [73]. Network reconstruction enables a detailed understanding of how synthetic biocircuit components are interacting with one another; when a biocircuit performs suboptimally, a reconstructed network model highlights parts of the biocircuit that need to be redesigned. Thus, while traditional troubleshooting approaches involve exhaustive part-by-part optimization to achieve global functionality, network reconstruction enables model-directed approach to troubleshooting.

As a proof of concept, our goal was to prototype a completely novel transcription-based event detector biocircuit. Event detectors are useful because of their ability to perform temporal logic. Making temporal logic decisions enable applications such as programmed differentiation, where the goal is to perform some operation based on a combinatorial and temporal sequences of events that dictate cell fate.

So far there are two demonstrations of temporal logic gates: 1) a temporal logic gate that differentiates start times of two chemical outputs [43] and 2) a molecular counter that counts the number of sequential pulses of inducers [30]. Both event detectors use serine integrases to perform irreversible recombination. The advantage of an integrase-based approach is the persistent nature of DNA-based memory. At the same time, the drawback is that these biocircuits function as a one-time use device. Both the molecular counter and temporal logic gate cannot reset once they have been triggered.

In contrast, transcription based event detectors use proteins instead of DNA to encode a memory state. The advantage is that proteins are non-permanent, since they are diluted through cell growth or can be tagged for degradation. On the other hand, maintaining protein state over multiple generations is metabolically expensive and the dynamics of the circuit

can become sensitive to production and growth phase of the cells. Therefore, a transcription based event detector biocircuit must be designed with precise timing, balance of production rates, and carefully tuned gain of each transcriptional regulator. This provides a perfect use case for our network reconstruction algorithm.

Our transcriptional event detector consists of two constitutively expressed relay genes, AraC and LasR, that transmit the arrival of two distinct induction events (arabinose and HSL) to relay output promoters pBAD and pLas respectively. To record these induction events historically, the output of each relay gene is coupled to one of two combinatorial promoters (pBAD-Lac or pLas-Tet) in a toggle switch. Each combinatorial promoter implements NIMPLY logic, e.g. pBAD-Lac (pLas-Tet) expresses TetR (LacI) only when arabinose (HSL) and AraC (LasR) are present and LacI (TetR) is absent. Thus, when one analyte (e.g. arabinose) arrives, it triggers latching of the toggle switch only if the toggle switch is unlatched to begin with or the prior latching protein state has been diluted out. The relay outputs thus transmit the *current or recent* induction event state while the toggle switch maintains the *historical* induction event state. Thus, depending on the order of arrival of each inducer, we obtain different biocircuit states. Figure 3.8 details the genetic elements in the event detector biocircuit and the designed component interaction network.

An idealized model for the system (assuming no crosstalk) is written as

$$
\begin{aligned}
\dot{x}_1 &= \rho_1 m_1 - \delta_p x_1, \\
\dot{x}_2 &= \rho_2 m_2 - \delta_p x_2, \\
\dot{x}_3 &= \rho_3 m_3 - \delta_p x_3, \\
\dot{x}_4 &= \rho_4 m_4 - \delta_p x_4, \\
\dot{m}_1 &= \frac{k_1(k_l + u_5/k_{M,u5})}{(1 + u_5/k_{M,u5})} + u_1 - \delta_m m_1, \\
\dot{m}_2 &= \frac{k_2(k_l + u_5/k_{M,u5})}{(1 + x_3/k_{M,3} + u_5/k_{M,u5})} + u_2 - \delta_m m_2, \\
\dot{m}_3 &= \frac{k_3(k_l + u_6/k_{M,u6})}{(1 + x_2/k_{M,2} + u_6/k_{M,u6})} + u_3 - \delta_m m_3, \\
\dot{m}_4 &= \frac{k_4(k_l + u_6/k_{M,u6})}{(1 + u_6/k_{M,u6})} + u_4 - \delta_m m_4, \\
\mathbf{y} &= \begin{bmatrix} \mathbf{I}_{4\times 4} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^T & \mathbf{m}^T \end{bmatrix}^T
\end{aligned}
\tag{3.12}
$$

where the measured outputs of the system are $y_i = x_i, i = 1, ..., 4$, ρ_i is the translation rate of m_i into x_i, δ_p is the effective dilution rate of $x_i, i = 1, ..., 4$ and k_i is the catalytic transcription rate for m_i, $u_1, ..., u_4$ are DNA inputs to perturb $m_1, .., m_4$ and u_5 and u_6 are arabinose and HSL, respectively. The dynamical structure function for this system is calculated by linearizing the system about a nominal operating point, (x_0, m_0), taking a Laplace transform and solving out the hidden variables $m_1, ..., m_4$. We present a simplified case here, assuming algebraic symmetry of the parameters $k_i = k, \rho_i = \rho, k_{M,i} = k_M$ as it does not qualitatively change the structure of $(Q(s), P(s))$. We obtain:

$$
Q^a(s) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & Q_{23}(s) & 0 \\ 0 & Q_{32}(s) & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}
$$

$$P^a(s) = \begin{bmatrix} P_{11}(s) & 0 & 0 & 0 \\ 0 & P_{22}(s) & 0 & 0 \\ 0 & 0 & P_{33}(s) & \\ 0 & 0 & 0 & P_{44}(s) \end{bmatrix}$$

where $P_{ii}(s) = \rho/(\delta_m + s)(\delta_p + s)$ for $i = 1, .., 4$ and

$$Q_{23}(s) = \frac{-k\rho(k_l + u_5/k_M)}{k_M(\delta_m + s)(\delta_p + s)(u_5/k_M + x_3/k_M + 1)^2}$$
$$Q_{32}(s) = \frac{-k\rho(k_l + u_6/k_M)}{k_M(\delta_m + s)(\delta_p + s)(u_6/k_M + x_2/k_M + 1)^2}$$

It is important to note that in the absence of protein degradation, $Q_{23}(s)$ and $Q_{32}(s)$ can be approximated with first-order SISO transfer functions. This observation will be validated empirically in the sequel.

These expressions for $Q(s)$ and $P(s)$ are of the idealized dynamical structure function of the *alternative* system. This is the intended network structure of the event detector, which will hold in the absence of crosstalk. Depending on the abundance of transcription factors such as LacI, TetR, and AraC, as well as commonly shared transcriptional and translational proteins, the *actual* dynamical structure function $Q^c(s)$ may be fully-connected or possess undesired interactions. This raises two important questions: 1) under what structural constraints on $Q^c(s)$ can we achieve robust performance with the event detector, and 2) when the event detector fails, how is this failure characterized by the dynamical structure function?

To answer these questions, we first constructed this event detector *in vitro* in the TX-TL system [73], demonstrating coarse functionality with four fluorescent reporters. The *in vitro* event detector appeared to show differentiated outputs, depending on whether arabinose or HSL arrived first. However, in *in vivo* tests, induction with HSL at 1 μM seemed to override any prior latching from arabinose. When we attenuated HSL concentration from 1 μM to 1 nM, the circuit maintained its latched state (Figure 3.9). What was the explanation for this

concentration dependent difference in performance?

To diagnose the cause, we performed *in vitro* network reconstruction experiments on the biocircuit at 1 nM and 1 μM HSL induction concentrations. Each reconstruction experiment consisted of perturbing a single gene in the event detector; this entailed pipetting an additional 1.5 nM of plasmid containing the gene of interest into a TX-TL reaction mix [73], in addition to a premixed solution containing 2 nM of each gene in the event detector biocircuit.

We measured the output of the pBAD, pBAD-Lac and pTet-Las promoters using pBAD-CFP, pBAD-Lac YFP, and pTet-Las RFP reporter genes respectively. To measure the output of the pLas promoter, we simultaneously tested a variant where the pBAD-CFP gene was replaced with a pLas-CFP gene. All TX-TL experiments were performed as bulk reactions; the data was background subtracted using a negative control reaction containing the unperturbed event detector, smoothed with a standard moving-window filter, and normalized in each channel by the maximum signal achieved in the 1 nM condition.

Since the perturbations to each gene were DNA based, we knew that addition of DNA corresponded to direct perturbation of the target gene, ensuring $P(s)$ is diagonal. This feature of TX-TL based network inference is particularly useful. In contrast to *in vivo* where perturbation experiments are limited to spikes (or transient pulses) chemical inducers, TX-TL perturbation experiments enable direct addition of DNA corresponding to the target gene. These direct perturbations effectuate a step input since once plasmid DNA is suddenly added, it persists and doesn't degrade, which guarantees sufficiency of excitation conditions for model identifiability [1, 37].

The natural expression curve of a TX-TL experiment mirrors the dynamics of an unstable system (see Figures 3.10 and 3.12). There are three reasons for this. Firstly, since TX-TL extract is harvested from log-phase cells (where Clp-XP protease expression is attenuated), protein degradation rates are significantly lower. Secondly, proteins that are marked for degradation typically suffer from enzyme loading effects; thus we elected not to tag our

proteins for degradation. Thirdly, TX-TL bulk reactions do not simulate cell growth and protein dilution. All these factors make it easy to characterize production in TX-TL reactions, but in the absence of protein degradation and dilution. This presents a challenge, since the effect of small input perturbations can be masked numerically by ever-increasing protein expression levels.

The usual approach to this challenge is to identify unstable transfer functions using standard system identification routines [58]. However, the negative feedback of the system is inherently masked by the dynamics of large unstable poles. We propose here another method, suited to the often unstable dynamics of biocircuits monitored in bulk reactions (both *in vivo* and *in vitro*). The key insight is that certain perturbations, e.g. DNA-based or inducer-based perturbations, can be modeled as perturbations to the initial state $x_p(0) = x_n(0) + u$. The dynamics of the unperturbed system can be written as

$$\dot{x}_n = f(x_n), \quad x_n(0) = x_0$$

and the dynamics of the perturbed system can be written as

$$\dot{x}_p = f(x_p), \quad x_p(0) = x_n(0) + u.$$

What insight about the dynamics of f can be gained if we consider the difference of the two trajectories? Specifically, what insight about the local dynamics of f, i.e. the Jacobian $\mathbf{A} = \mathbf{Df}(x_0)$ can we gain, from considering the difference of the two trajectories? Define $z = x_p(t, x_p(0)) - x_n(t, x_n(0)$ and note that if $u = 0$, then $z \equiv 0$ for all t. Under the conditions of the Hartman-Grobman Theorem in Chapter 2.8 of [77], we can approximate $z(t)$ as

$$\begin{aligned} x_p(t) - x_n(t) &\approxeq e^{At} x_p(0) - e^{At} x_n(0) \\ &= e^{At} \left(x_p(0) - x_n(0) \right) \end{aligned} \tag{3.13}$$

Thus, $\dot{z} \cong Az + Bu$ where $B = \text{diag}\,(\delta(t), ..., \delta(t))$,$\delta(t)$ is the Dirac-delta function and $u = z(0)$. The network dynamical structure $Q_z(s)$ is identical to that of the original system, $Q_{x_n}(s)$ since they are both defined with A. In our subsequent analysis, any reconstruction of $Q_z(s) = Q_{x_n}(s)$ will be performed using the difference of perturbed and nominal trajectories.

To reconstruct each row $(Q_i(s), P_i(s))$ we used Y_i as output data and $Y_j, j \neq i$ and U_i as input data to fit a 1 by 4 transfer function. We used a model order of $n = 1$ poles for each transfer function, motivated by two observations: 1) measurements were collected at the translational level, which by time-scale separation arguments could be approximated well by a first order transfer function and 2) higher order transfer functions tended to introduce artifact dynamics that decreased the quality of fit.

The impulse response of each entry $Q_{ij}(s)$ is plotted in Figures 3.11 (for 1 nM HSL induction response) and 3.13 (for 1 μM HSL induction response). At 1 nM HSL, the balance between LacI and TetR repression gain is relatively balanced (within one order of magnitude). Calculating the $||\mathcal{H}||_\infty$ gain of each entry $Q_{ij}(s)$ and normalizing by the maximum entry, we obtain the following normalized gain matrix:

$$\begin{pmatrix} 0 & 0.027 & 2.2 \cdot 10^{-3} & 2.9 \cdot 10^{-3} \\ 2.6 \cdot 10^{-3} & 0 & \mathbf{0.43} & 0.01 \\ 0.022 & \mathbf{1.0} & 0 & 2.9 \cdot 10^{-4} \\ 4.9 \cdot 10^{-3} & 0.14 & 8.2 \cdot 10^{-3} & 0 \end{pmatrix}$$

The largest gain occurs in entry $Q_{32}(s)$, reflecting the amplitude of the impulse response achieved by the convolution kernel $Q_{32}(t)$ in Figure 3.11. Ideally, in the 1 nM induction condition, only $Q_{23}(s)$ and $Q_{32}(s)$ are non-zero and implement mutual repression of Y_2 and Y_3. With the exception of $Q_{42}(s)$, this is essentially true since all other entries in $Q_{ij}(s)$ have gain two to three orders of magnitude below $||Q_{23}(s)||_\infty$ and $||Q_{32}(s)||_\infty$. From Figure 3.11, we see that Q_{23} and Q_{32} both implement negative repression as well, verifying that the circuit is functioning as intended.

Notice that while the amplitude of the gain is balanced, in the presence of 1 nM, Q_{23} has persistent negative repression for all t, corresponding to a pole at $s = 0$. This is because the reconstruction experiments are performed in the presence of 1 nM HSL, which favors TetR expression over LacI through the course of the experiment. However, the important observation is that the gain is balanced. This provides a physical explanation for why the event detector biocircuit maintains its latched state in the presence of 1 nM HSL.

Performing the same analysis with a background concentration of $1\mu M$ HSL, we obtain a dramatically different dynamical network structure. Using the data in Figure 3.12 to reconstruct $Q(s)$, we see that the feedback between $Q_{23}(s)$ and $Q_{32}(s)$ is no longer balanced in Figure 3.13. This becomes clear when we calculate the $\mathcal{H}_\infty$ gain of each $Q_{ij}(s)$ to obtain the normalized gain matrix:

$$\begin{pmatrix} 0 & 1.7 \cdot 10^{-4} & 7.1 \cdot 10^{-3} & 0 \\ 1.1 \cdot 10^{-3} & 0 & \mathbf{1.0} & 1.8 \cdot 10^{-4} \\ 0 & 0 & 0 & 0 \\ 7.5 \cdot 10^{-4} & 2.6 \cdot 10^{-4} & 2.9 \cdot 10^{-4} & 0 \end{pmatrix}.$$

We remark that in this case, entry Q_{31}, Q_{32}, Q_{34} appeared as numerical noise, with polynomial coefficients less than 1.5×10^{-3}, which we thresholded out in the matrix above. Plotting the $\mathcal{H}_2$ gain of each Q_{ij} and normalizing by the gain matrix, we see the most significant of these entries is Q_{34}

$$\begin{pmatrix} 0 & 1.7 \cdot 10^{-4} & 7.1 \cdot 10^{-3} & 1.3 \cdot 10^{-5} \\ 1.1 \cdot 10^{-3} & 0 & 1.0 & 1.8 \cdot 10^{-4} \\ 1.9 \cdot 10^{-6} & 4.0 \cdot 10^{-6} & 0 & 0.073 \\ 7.5 \cdot 10^{-4} & 2.6 \cdot 10^{-4} & 2.9 \cdot 10^{-4} & 0 \end{pmatrix}$$

which reveals that Q_{34} achieves 7% of the maximum gain achieved by Q_{23}. Most importantly, we see that 1 μM induction of HSL essentially abolishes the negative repression from Y_3 to Y_2 (LacI). Both the $\mathcal{H}_\infty$ and $\mathcal{H}_2$ gain imbalance between Q_{32} and Q_{23} are more than 5

orders of magnitude. Thus, when induced with $1\mu M$ HSL, any memory of LacI dominance is immediately overrided by repression with TetR. Designing a *robust* event detector requires the balance of gain between both genes in the toggle switch to be invariant to varying concentrations of the input concentrations (HSL and arabinose).

3.6 Conclusion

The dynamical structure function models the dependencies among measured states. It is a flexible representation of network structure that naturally adapts to the constraints imposed by experimental measurement. Since identifiability conditions of the dynamical structure function have been well characterized, appropriate experimental design can ensure that the process of network reconstruction produces a sensible answer.

Most importantly, network reconstruction of the dynamical structure function can be used to validate the intended network design of a synthetic biological system. In specific cases, where orthogonality between two chemical species is intended, the entries in a reconstructed dynamical structure function provide a direct estimate of crosstalk or interference between the two species of interest. More generally, the dynamical structure function allows us to characterize the operational or active network and study the relationships between environmental parameters, active network dynamics, and biocircuit performance. We have integrated theory, simulation, and experiments to demonstrate that dynamical structure functions can be a powerful tool for understanding, engineering, and validating novel synthetic gene networks.

3.7 Experimental Methods

All plasmids were constructed using either Golden Gate assembly [24] or Gibson isothermal assembly [33] in *E. coli*. Plasmids were sequence verified in JM109 cloning strains and transformed into MG1655ΔLacI, courtesy of R. J. Krom and the J. J. Collins as a two-

plasmid system with kanamycin and chloremphenicol selection. All *in vivo* experiments were carried out with $n = 2$ replicates using MatriPlates (Brook Life Science Systems MGB096-1-2-LG-L) 96 square-well glass bottom plates at 29 C in a H1 Synergy Biotek plate reader using 430/470, 505/535/ and 580/610 nm excitation/emission wavelengths. Cell density was quantified with optical density at 600 nm.

For *in vitro* experiments, all genelet repressilator reconstruction experiments were carried out at 37 C in a Horiba Spectrofluoremeter with 1 minute readout times, using Rhodamine Green, TYE 563 and Texas Red flourophores with 10 nm monochromamtor excitation and emission bands centered at 502/527, 549/563, and 585/615 nm respectively. TX-TL experiments were performed using extract prepared according to the methods described in [73, 91]. All network reconstruction reactions were performed using 10 μL reaction volumes in Nunc 384 square well glass-bottom plates (ThermoFisher Scientific Cat. No. 142761) at 29 C to reflect *in vivo* conditions.

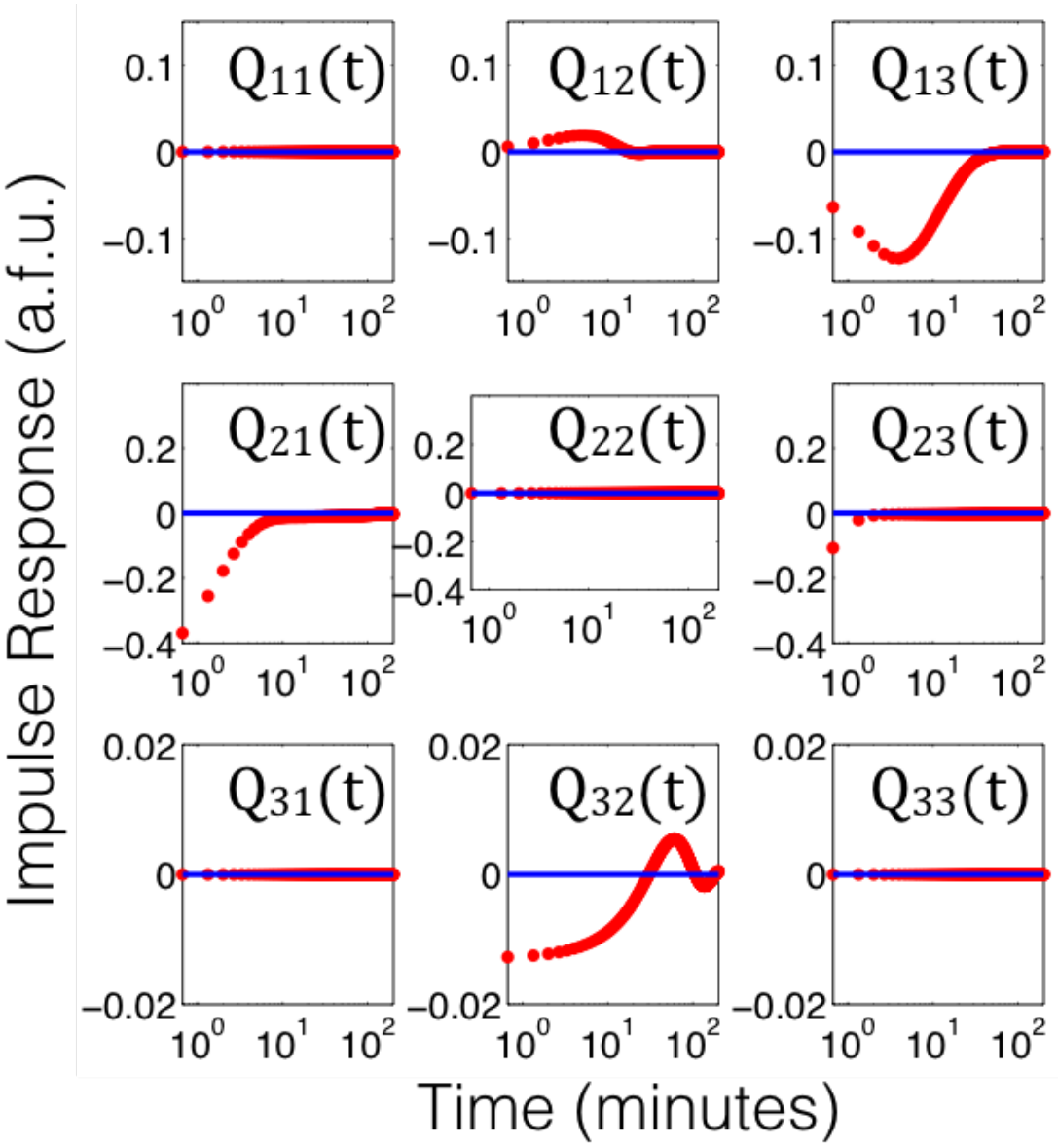

Figure 3.7: Impulse response of the estimated convolution kernel $Q(t)$ matrix. $Q(s)$ is estimated directly from experimental data, transformed into the frequency domain, and simulated in time for $t = 0$ to $t = 300$ minutes.

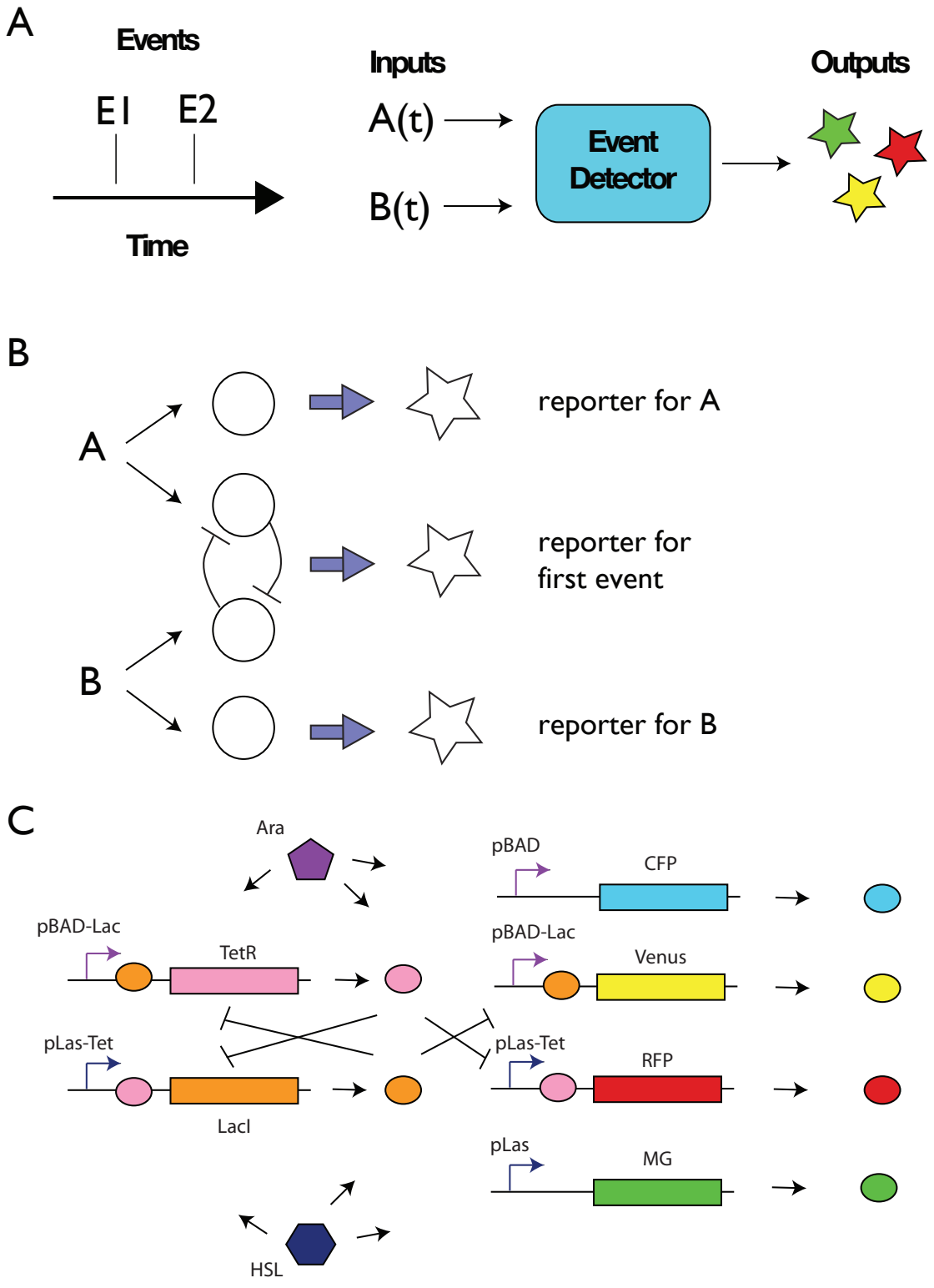

Figure 3.8: (**A**) Left: We design an event detector to determine the identity and relative ordering of two events E_1 and E_2 occurring within a finite time horizon. (**B**) A schematic showing the logic of the circuit for the event detector. Arrival of event type A triggers transient reporter for A (top) and latching of the toggle in a A-dominant state as a memory state. Similarly, arrival of event type B triggers transient reporter for B (bottom) and latching of the toggle in a B-dominant state as a memory state. (**C**) A diagram showing the synthetic biocircuit parts used to implement the network architecture in (**B**).

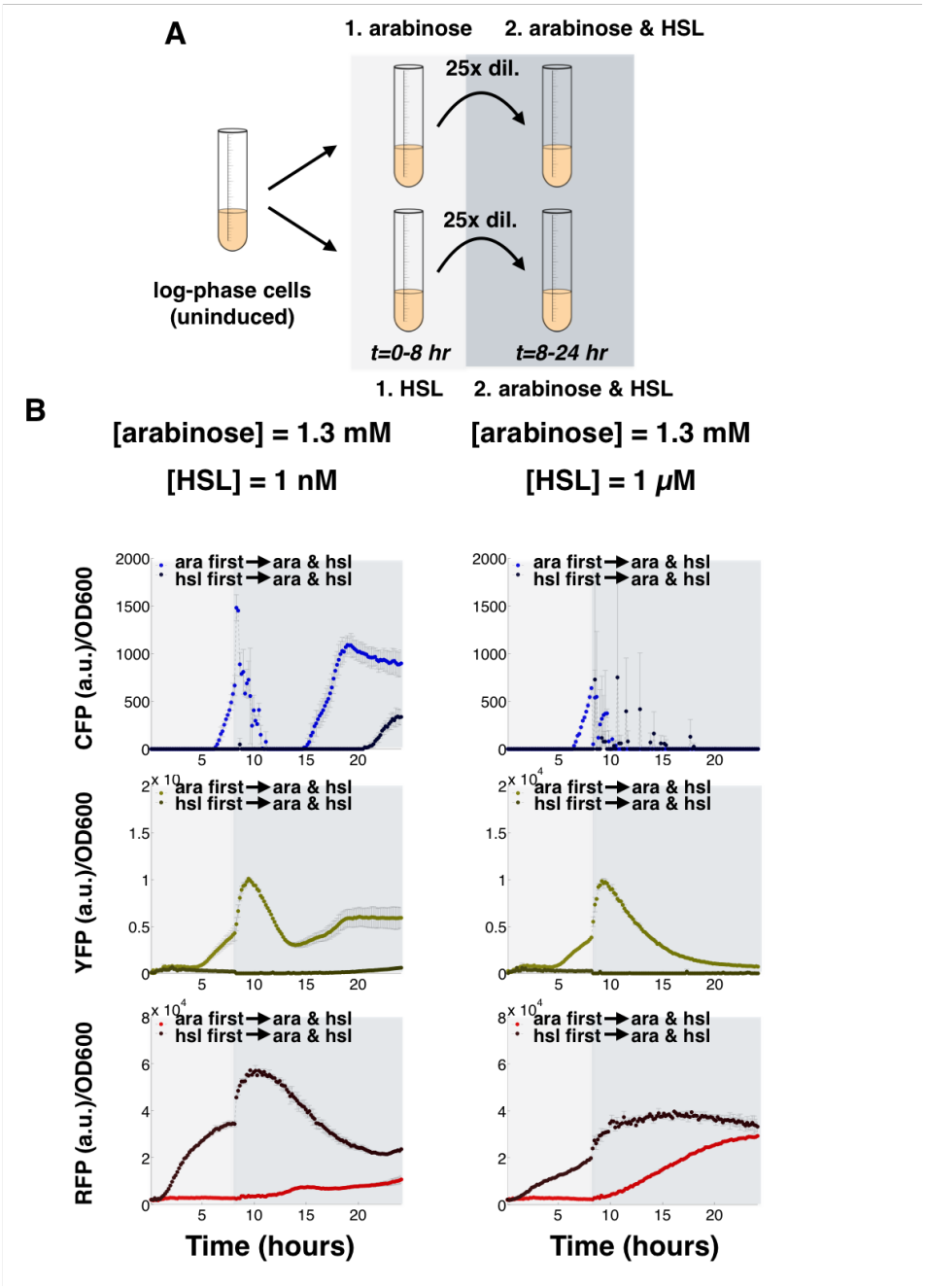

Figure 3.9: *In vivo* plate reader experiments, testing the temporal logic properties of the event detector. Notice that at 1 nM HSL, the event detector functions properly, expressing different levels of CFP, YFP and RFP depending on the the order of arrival of arabinose and HSL. At 1 μM HSL, the temporal logic properties of the event detector completely are abolished.

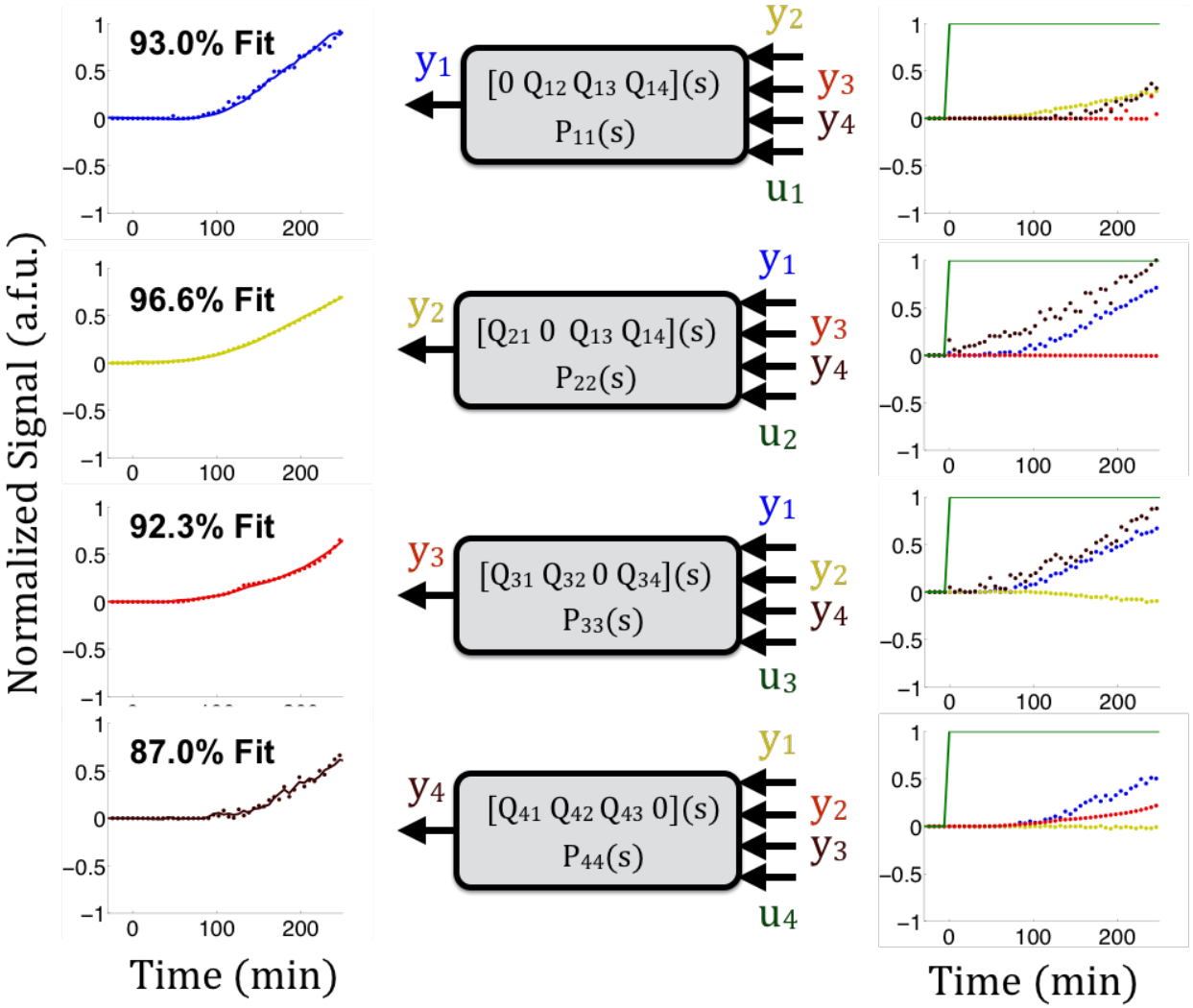

Figure 3.10: Time-series data of the event detector from network reconstruction experiments in the TX-TL system with 1 nM HSL.

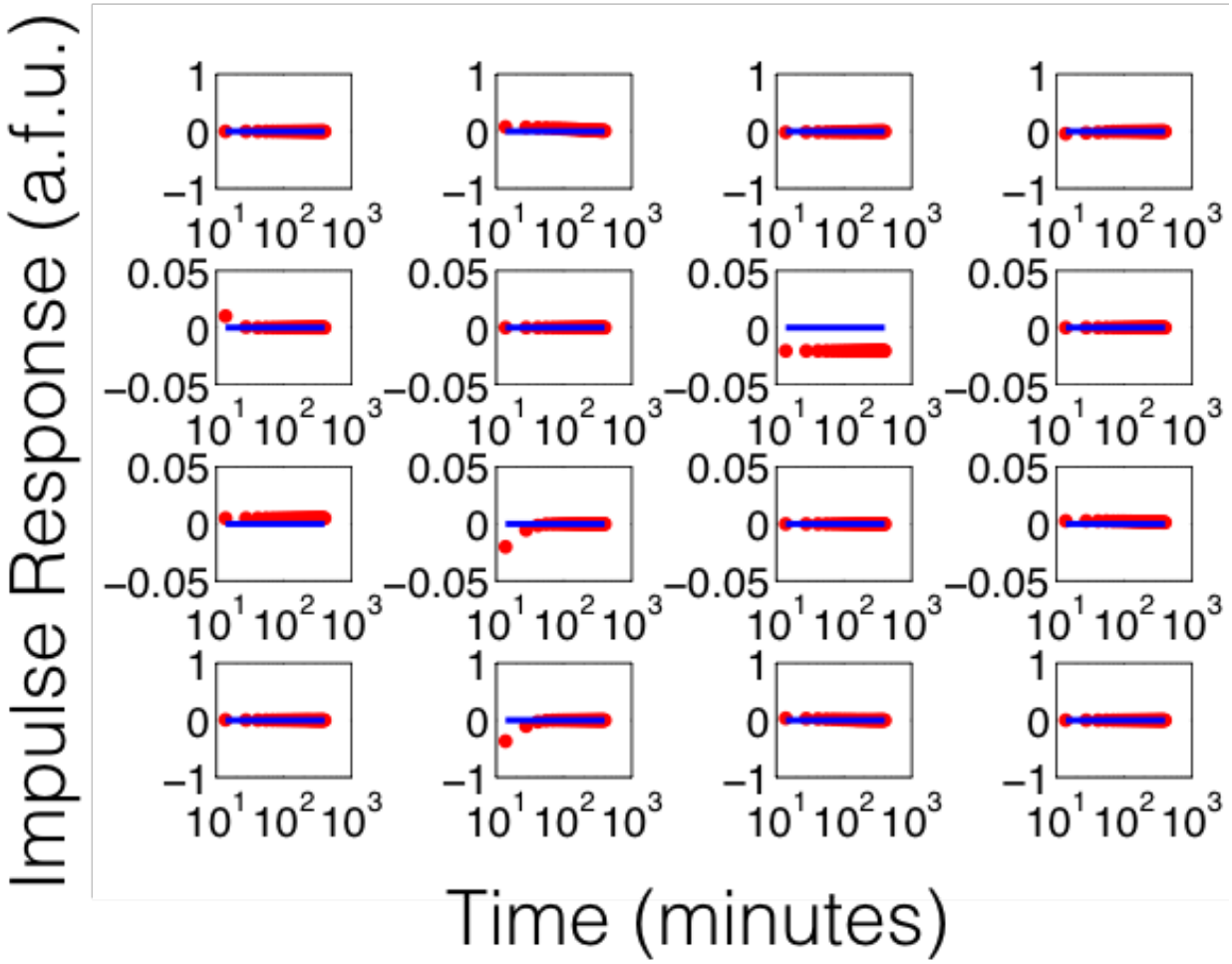

Figure 3.11: Impulse response of the estimated convolution kernel $Q(t)$ matrix when the event detector biocircuit is induced with 1 nM HSL. $Q(s)$ is estimated directly from experimental data, transformed into the frequency domain, and simulated in time, plotted in log-scale to make fast transients dynamics more visible.

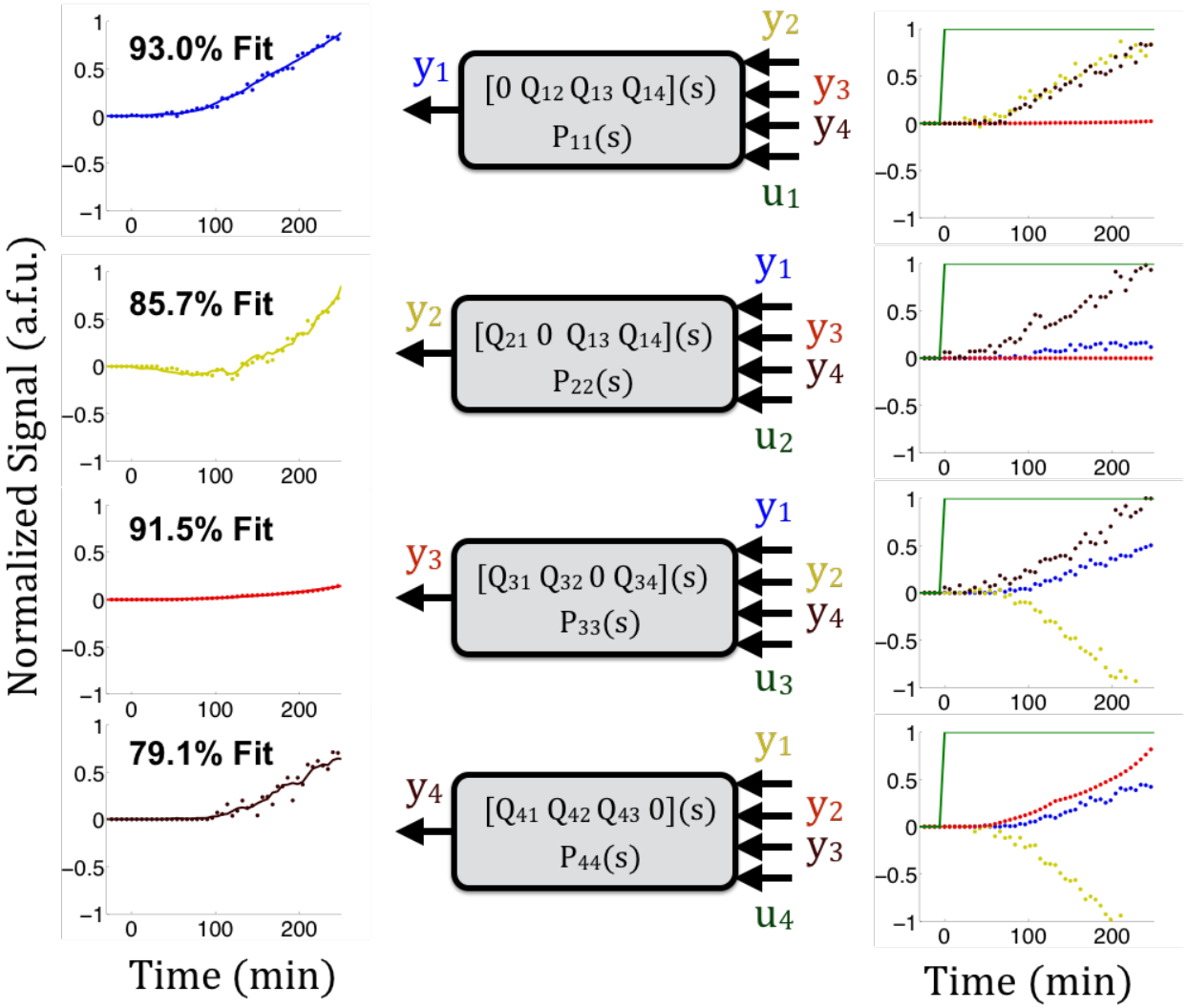

Figure 3.12: Time-series data of the event detector from network reconstruction experiments in the TX-TL system with 1 μM HSL.

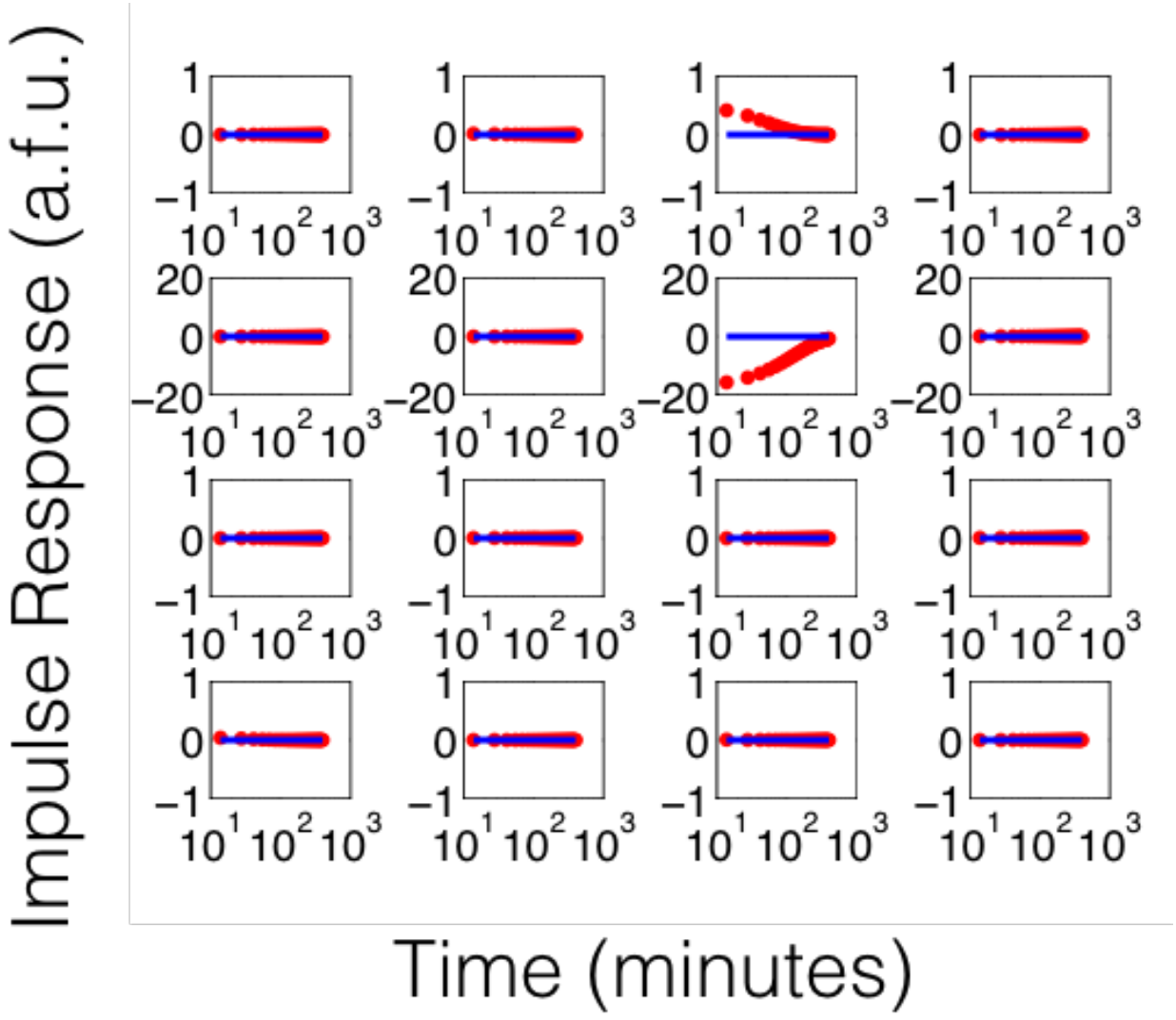

Figure 3.13: Impulse response of the estimated convolution kernel $Q(t)$ matrix when the event detector biocircuit is induced with 1 μM HSL. $Q(s)$ is estimated directly from experimental data, transformed into the frequency domain, and simulated in time, plotted in log-scale to make fast transient dynamics more visible.

Chapter 4

Analysis of Context Effects with Dynamical Structure Functions

4.1 Background

One of the primary goals of synthetic biology is to manipulate and synthesize novel biochemical devices to achieve an objective. For *in vivo* applications, this often means exploiting the resources available in a host organism. For example, the authors in [90] combine the technology of combinatorial promoters with the native transcription and translation machinery in *E. coli* to achieve robust oscillation. In [22], the authors took advantage of native protein folding machinery and plasmid replication proteins to monitor a low copy number oscillator using GFP expressed on a high copy number plasmid — by so doing, they were able to determine that low copy number gene expression was stochastic. More recently, the authors in [26] take advantage of the clustered regularly interspaced short palindromic repeats (CRISPR) pathway and endoribonucleases to achieve polycistronic transcriptional and translational expression. In each scenario, these synthetic technologies utilize the resources available in a cell to achieve their objective, whether those resources are available in abundance or scarcity. Thus, it is becoming clear that the successful implementation of synthetic designs requires understanding the context of host chassis and resources [9].

Substantial experimental work has been done experimentally that indicates resources can be scarce in the cell. In [27], the authors demonstrate that too many LVA-tagged

components in a synthetic circuit causes saturation of Clp-XP protease, creating coupling between originally orthogonal components. They show these coupling interactions are strong enough to destroy the robustness of the oscillator from [90].

On the modeling side, substantial work has been done to quantify both theoretically and empirically the scarcity of resources and the impact it can have on synthetic biocircuit dynamics. The scarcity of resources often leads to unintentional coupling, referred to as retroactivity [96], crosstalk [103], loading effects [85], etc. Strategies for attenuating this crosstalk have been proposed in [45], [96], and [103]. There is ongoing work about how to scale these strategies for larger systems, where multiple components may be subject to retroactivity and a limit exists for the number of independent inputs for attenuating retroactivity.

Our goal in this thesis is complementary to the work in [45, 85, 96]: we seek to understand the effects of resource crosstalk on synthetic circuit performance, specifically how resource crosstalk can lead to right half plane zeros in the local dynamics of a transcription-translation system about an equilibrium point. In addition, our work should be viewed as complementary to the analysis on glycolysis systems, as it considers yet another scenario where right half plane zeros play a role in limiting system performance in biological systems [11]. In general, it is well known in that right half plane zeros can impose fundamental limitations on system performance, see [44] for example. Since resource limitations are not necessarily modeled in every situation, it is important to understand what kinds of scenarios involving resource limitations can lead to a right half plan zero. With that understanding, we can design biocircuits to avoid these scenarios or design them to ensure the effect of these right-half plane zero is small.

One goal in synthetic biology is to design well-behaved and robust biological engineered modules. Consequentially, such modules must have a well-behaved and robust input-output response. As a first step, we will restrict our attention to systems that tend towards a locally asymptotically stable equilibrium point and study their input-output response with

the transfer function. Further, as these modules may be comprised of multiple submodules; studying the intput-output dynamics and the zero dynamics at the level of the state-space realization may be challenging. Thus, we will use transfer functions to characterize the precise conditions under which right half plane zeros can arise from resource limitations.

We organize the subsequent sections as follows: in Section 4.2 we develop a motivating example system to illustrate how crosstalk interactions can lead to a non-minimum phase transfer function. In Section 4.3 we show that a simple network motif plays the primary role in introducing right half plane zeros — we present both a motivating example and a generalizing result that characterizes the parametric and functional conditions under which right half plane zeros exist. Next, in Section 4.5, we apply the results of Section 4.3 to identify a general class of transcription-translation systems that have a right half plane zero under certain parametric conditions. Finally, in Section 4.6 we show how degradation and ribosomal loading in certain types of synthetic biocircuits can have the potential to introduce a right half plane zero and adversely affect the master stress response of a host *E. coli* cell.

4.2 Example: Resource Limitations in a Signal Cascade

Whether a synthetic biocircuit is implemented *in vivo* or *in vitro*, if it utilizes transcriptional or translational machinery, e.g. NTP, ATP, polymerases, σ-factors, ribosomes, or degrades using shared degradation enzymes, e.g. ribonucleases or proteases, then it has the potential to saturate or overload these resource molecules and interfere with other processes in the system. These saturation effects can lead to sequestration of enzymes from critical processes that would otherwise function normally. These sequestration effects are the basis of resource mediated-crosstalk interactions in a system. When these crosstalk interactions are substantial, they can lead to unwanted coupling between otherwise orthogonal processes. Further, when that coupling augments the existing network of interactions to form a certain type

of network motif, then right half plane zeros can appear in the transfer function. To gain intuition let us consider an example system that uses a ubiquitous network motif: the signal cascade.

Let x_O and x_I be two proteins in a signal cascade network translated from mRNA molecules m_O and m_I respectively. Let u be an input (e.g. an inducer or allosteric activator) that activates the cascade via x_I. Suppose that x_I represses expression of mRNA transcript m_O. If expressed, m_O translates to x_O as the final output protein of the cascade. We suppose that the production and degradation of x_O and x_I can be described as Michaelis-Menten functions (without competitive effects). We write the model for this system as follows:

$$\begin{aligned}
\dot{m}_I &= \alpha_I - \delta m_I \\
\dot{x}_I &= R_I \frac{m_I/k_{M,I}}{1 + m_I/k_{M,I}} - D_I \frac{x_I/\kappa_{M,I}}{1 + x_I/\kappa_{M,I}} + k_{IU} u \\
\dot{m}_O &= \alpha_O + \frac{k^r_{cat}}{k_{M,OI} + x_I} - \delta m_O \\
\dot{x}_O &= R_O \frac{m_O/k_{M,O}}{1 + m_O/k_{M,O}} - D_O \frac{x_O/\kappa_{M,O}}{1 + x_O/\kappa_{M,O}} \\
y &= x_O
\end{aligned} \tag{4.1}$$

Here we have attempted to capture a signal cascade in the simplest possible terms. In doing so, we acknowledge that we have omitted the usual Hill functions that describe transcriptional activation and ignored the intricate processes behind the production of mRNA, including isomerization, strand elongation, fall-off, etc.. Our goal is to capture the essence of the network structure and relationships between states in this signal cascade, but with minimal complexity so that when we add modeling terms to describe resource competition, the introduction of a right half plane zero will be transparent. After linearizing, we compute

the transfer function as

$$G(s) = \frac{-k_{IU}\frac{R_O/k_{M,O}}{\left(1+m_{O,e}/k_{M,O}\right)^2}\frac{k^r_{cat}}{(1+x_{I,e}/k_{M,OI})^2}}{(s+\delta)\left(s-\frac{-D_I/\kappa_{M,I}}{(1+x_{I,e}/\kappa_{M,I})^2}\right)\left(s+\frac{D_O/\kappa_{M,O}}{(1+x_{O,e}/\kappa_{M,O})^2}\right)}.$$

Clearly, $G(s)$ has no right half plane zeros, so the system is minimum phase. Now consider a model that incorporates the effects of substrates competing for the same resources. In particular, we will suppose that m_I and m_O compete for the same ribosomes R to translate x_I and x_O respectively and that x_O and x_I compete for the same degradation enzymes D. We write it as follows:

$$\begin{aligned}
\dot{m}_I &= \alpha_I - \delta m_I \\
\dot{x}_I &= \frac{R\frac{m_I}{k_{M,I}}}{1+\frac{m_I}{k_{M,I}}+\frac{m_O}{k_{M,O}}} - \frac{D\frac{x_I}{\kappa_{M,I}}}{1+\frac{x_I}{\kappa_{M,I}}+\frac{x_O}{\kappa_{M,O}}} + k_{IU}u \\
\dot{m}_O &= \alpha_O + \frac{k^r_{cat}}{k_{M,OI}+x_I} - \delta m_O \\
\dot{x}_O &= \frac{R\frac{m_O}{k_{M,O}}}{1+\frac{m_O}{k_{M,O}}+\frac{m_I}{k_{M,I}}} - \frac{D\frac{x_O}{\kappa_{M,O}}}{1+\frac{x_O}{\kappa_{M,O}}+\frac{x_I}{\kappa_{M,I}}} \\
y &= x_O.
\end{aligned} \tag{4.2}$$

The linearized system is of the form

$$\begin{bmatrix} -\delta & 0 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & a_{24} \\ 0 & -k_{OI} & -\delta & 0 \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}, \tag{4.3}$$

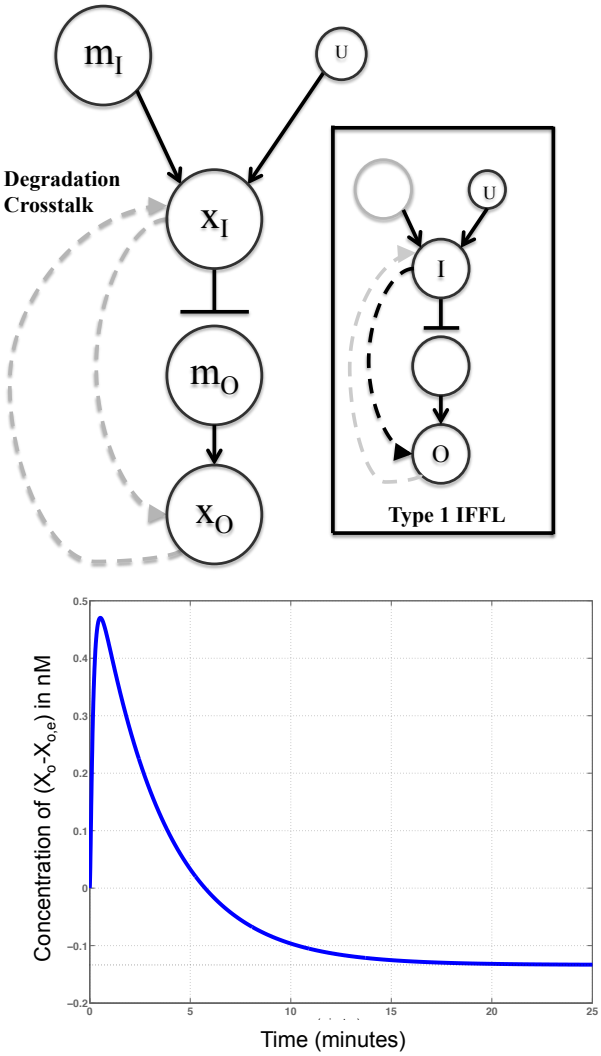

Figure 4.1: (Top) An illustration of the signal cascade system (4.2) including sequestration interactions from degradation enzyme loading. The actual cascade (not including resource limitation effects) has the same structure as system (4.1). the input u upregulates x_I, which subsequently represses expression of the output gene's mRNA m_O. However, x_I sequesters degradation enzyme from x_O, which has the effect of increasing x_O concentration. Thus, we draw an effective (positive) arrow from x_I to x_O. Similarly, x_O sequesters degradation enzyme from x_I so we draw a (positive) arrow from x_O to x_I. In particular, one of the crosstalk edges introduces a Type I incoherent feedforward loop into the system (see inset). (Bottom) The step response of the linearization of system (4.2) is plotted here. The transfer function is now non-minimum phase with right half plane zero $z = 0.0012$. Parameter values for the simulation were $R_I = R_O = 1.51$ nM/s, $D_I = D_O = 2$ nM/s, $k_{M,I} = 20$ nM, $k_{M,O} = 40$ nM , $\kappa_{M,I} = 3.1$ nM, $\kappa_{M,O} = 2.9$ nM, $\alpha_I = .002$ nM/s, $\alpha_O = .001$ nM/s, $\delta = 0.005$/s, $k_{IU} = 10^{-7}$ /s and $k_{OI} = 625 \times 10^{-9}$ /s..

where $k_{OI} = k^r_{cat}/\left(1 + x_{I,e}/k_{M,OI}\right)^2$,

$$
\begin{aligned}
a_{21} &= \frac{\frac{R}{k_{M,I}}(1+\frac{m_{O,e}}{k_{M,O}})}{(1+\frac{m_{I,e}}{K_{M,I}}+\frac{m_{O,e}}{K_{M,O}})^2}, \; a_{41} = \frac{-\frac{R}{K_{M,I}}(\frac{m_{O,e}}{K_{M,O}})}{(1+\frac{m_{I,e}}{k_{M,I}}+\frac{m_{O,e}}{k_{M,O}})^2}, \\
a_{22} &= \frac{\frac{-D}{\kappa_{M,I}}(1+\frac{x_{O,e}}{\kappa_{M,O}})}{(1+\frac{x_{I,e}}{\kappa_{M,I}}+\frac{x_{O,e}}{\kappa_{M,O}})^2} \; , \; a_{42} = \frac{\frac{D}{\kappa_{M,I}}\frac{x_{O,e}}{\kappa_{M,O}}}{(1+\frac{x_{I,e}}{\kappa_{M,I}}+\frac{x_{O,e}}{\kappa_{M,O}})^2}, \\
a_{23} &= \frac{-\frac{R}{k_{M,O}}(\frac{m_{I,e}}{k_{M,I}})}{(1+\frac{m_{I,e}}{k_{M,I}}+\frac{m_{O,e}}{k_{M,O}})^2} \; , \; a_{43} = \frac{\frac{R}{k_{M,O}}(1+\frac{m_{I,e}}{k_{M,I}})}{(1+\frac{m_{I,e}}{k_{M,I}}+\frac{m_{O,e}}{k_{M,O}})^2}, \\
a_{24} &= \frac{\frac{D}{\kappa_{M,O}}\frac{x_{I,e}}{\kappa_{M,I}}}{(1+\frac{x_{I,e}}{\kappa_{M,I}}+\frac{x_{O,e}}{\kappa_{M,O}})^2} \; , \; a_{44} = \frac{\frac{-D}{\kappa_{M,O}}(1+\frac{x_{I,e}}{\kappa_{M,I}})}{(1+\frac{x_{I,e}}{\kappa_{M,I}}+\frac{x_{O,e}}{\kappa_{M,O}})^2},
\end{aligned}
\tag{4.4}
$$

and $B = k_{IU}\mathbf{e_2}$ and $C = \mathbf{e_4}^T$, where $\mathbf{e}_i$ denotes the i^{th} standard basis vector. We then have

$$G(s) = \frac{(a_{42}k_{IU})s - a_{43}k_{OI}k_{IU} + a_{42}\delta k_{IU}}{D(s)}$$

with characteristic polynomial $D(s) = s^3 - (a_{22}+a_{44}-\delta)s^2 - (a_{24}a_{42} - a_{22}a_{44} + (a_{22}+a_{44})\delta - a_{23}k_{OI})s + a_{22}a_{44}\delta - a_{22}a_{42}\delta - a_{23}a_{43}k_{OI} + a_{24}a_{43}k_{OI}$. With the appropriate constraints on the balance between degradation and production, the poles will lie in the left half plane. However, notice that $G(s)$ has a zero at

$$z = \frac{a_{43}k_{OI}}{a_{42}} - \delta, \tag{4.5}$$

and since $\delta > 0$, the zero is in the right half plane if the ratio $(a_{43}k_{OI})/a_{42}$ is sufficiently large. The coefficient a_{42} describes how x_I impacts x_O via indirect competition for the limited degradation enzyme. As the amount of x_I at equilibrium increases, a_{42} grows smaller and less x_O is degraded, since the degradation enzymes become more likely to be bound to x_I substrate. However, through the signal cascade, x_I inhibits x_O with effective gain $a_{43}k_{OI}$. As the amount of x_I at equilibrium increases, a_{42} can grow small enough so that $a_{43}k_{OI}/a_{42}$ approaches δ from the right hand side, resulting in a small (slow) right half plane zero which

place stronger constraints on the controller. Figure 4.1 shows the step response for system (4.2) with a particularly slow right half plane zero.

Notice that if the signal cascade was designed so that x_I activated x_O, then the term $-k_{OI}$ would be replaced with k_{OI} and the zero would be

$$z = -(\frac{a_{43}k_{OI}}{a_{42}} + \delta) < 0. \tag{4.6}$$

Thus, it appears that the incoherence, or opposing dynamics of 1) x_I repressing x_O and 2) x_I "promoting" the abundance of x_O by saturating degradation enzyme, is necessary to produce a right half plane zero. If the signal cascade is designed so that x_I activates x_O, then the incoherent feedforward loop becomes a coherent feedforward loop and the right half plane zero disappears. It is the incoherent feedforward loop that makes $G(s)$ non-minimum phase; thus the next section focuses on characterizing how and when incoherent feedforward loops produce right half plane zeros in $G(s)$.

4.3 Analysis of the Incoherent Feedforward Loop Network Motif with Dynamical Structure Theory

In this section, we characterize how right half plane zeros arise in incoherent feedforward loops (IFFLs). Incoherent feedforward loops are a common network motif in natural biological systems [36, 62] and they have been proposed recently as a design motif for achieving robust adaption in synthetic biocircuits [48]. We show in this section that IFFLs can also produce right half plane zeros under certain parametric conditions.

We first illustrate this fact with a simple example to provide intuition for the general

result. Consider the following linear two state model for a feedforward loop:

$$\begin{aligned} \dot{x}_I &= -\delta_I x_I + k_{IU} u \\ \dot{x}_O &= -\delta_O x_O + k_{OI} x_I + k_{OU} u \\ y &= x_O \end{aligned} \tag{4.7}$$

The transfer function for this system is

$$G(s) = \frac{k_{OU} s + \delta_I k_{OU} + k_{OI} k_{IU}}{(s + \delta_I)(s + \delta_O)}$$

and has a zero at

$$z = -\left(\frac{k_{OI} k_{IU}}{k_{OU}} + \delta_I \right).$$

Notice the similarity between z here and the zero in equation (4.6). The feedforward loop is coherent (incoherent) whenever the sign of $k_{OI}k_{IU}$ is the same as (opposite of) the sign of k_{OU}. This condition succinctly characterizes all four types of incoherent feedforward loops and all four types of coherent feedforward loops. In the nonlinear setting, such a succinct characterization may be hard to find, but as our analysis pertains to transfer functions, this condition will suffice for determining if a feedforward loop is incoherent or coherent.

Since δ_I represents a degradation rate for x_I, then $\delta_I > 0$ and the potential for $z > 0$ exists only when

$$\frac{k_{OI} k_{IU}}{k_{OU}} < 0$$

and δ_I small enough. Notice that if we separate the B matrix as

$$B = B_I + B_D = \begin{bmatrix} k_{IU} \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ k_{OU} \end{bmatrix}$$

we can decompose the transfer function $G(s)$ as

$$\begin{aligned} G(s) &\equiv G_I(s) + G_D(s) \\ &\equiv C(sI - A)^{-1}B_I + C(sI - A)^{-1}B_D \\ &= \frac{k_{OI}k_{IU}}{(s+\delta_I)(s+\delta_O)} + \frac{k_{OU}}{(s+\delta_O)}, \end{aligned}$$

where $G_I(s)$ represents the transfer function describing dynamics of the feedforward loop from U to the output X_O that requires the intermediate state I and $G_D(s)$ is the transfer function that describes the dynamics of the feedforward loop that go directly from U to X_O.

Viewed this way, we see that the feedforward loop transfer function $G(s)$ has a right half plane zero when the gain of $G_I(s)$ has opposite sign of the gain of $G_D(s)$, i.e. the two modes are incoherent, and the gain of $G_I(s)$ is sufficiently larger than the gain of $G_D(s)$. The two transfer functions capture the dynamics of the two pathways for controlling X_O. When those pathways are incoherent and the gain of the pathway with an intermediate state (i.e. the slower pathway) is sufficiently large, then the step response $G(s)$ is temporarily dominated by $G_D(s)$ since $G_I(s)$ must integrate one state before the signal propagates to x_O. The result is a transient consistent with the sign of the gain of $G_D(s)$. However, in the long run, $G_I(s)$ dominates the dynamics of $G(s)$ since the gain of $G_I(s)$ is larger, resulting in convergence to steady state in a direction opposite the initial transient driven by $G_D(s)$. These dynamics are a direct consequence of the structure of the incoherent feedforward loop. The incoherent feedforward loop thus yields structural intuition into the characteristic inverse step response observed for a non-minimum phase SISO transfer function with a single right half plane zero.

In general, biological systems possess many states and potentially many feedforward loops embedded in a single network. Moreover, the location of the input and the placement of the reporter molecule, i.e. the output of the system, play a key role in determining if a feedforward loop even exists between the input and the output. This is consistent with classical examples of non-minimum phase systems; sensor placement relative to actuator

location can make all the difference in eliminating a right half plane zero. [44]. If a feedforward loop exists, there may be several and in particular, the sequestration effects of resource loading (e.g. ribosomes, polymerases, transcription factors, ribonucleases, proteases, shared metabolic enzymes) may result in additional feedforward loops that were not included in the designed or natural system. If multiple feedforward loops are present, then it is critical to determine what the overall dominant or 'net' feedforward loop is, and whether it is incoherent or coherent. Typically, the model for a multi-state transcription-translation system, when considering the local dynamics about an equilibrium point, can be approximated with a linear time-invariant state space realization.

However, it is not easy to determine the existence of a incoherent feedforward loop at first glance from the state-space realization (hence we use transfer function calculus to calculate right half plane zeros instead of Rosenbrock matrices). Often, it is easier to consider a candidate intermediate node x_I and the output node x_O in the network and ask if there is an *effective* incoherent feedforward loop in the system. In this scenario, it would be useful to find a simpler representation of system structure that embeds the dynamics of unnecessary intermediate states as open loop transfer functions and describes the overall effect of u on x_I, x_I on x_O, and u on x_O. The dynamical structure function [38] is a convenient representation of structure that has this property. We develop the following lemma, based on the techniques in [38].

Lemma 1. *Consider the system*

$$\begin{aligned} \dot{z} &= Az + Bu \\ y &= \begin{bmatrix} c & 0 \end{bmatrix} z \end{aligned} \tag{4.8}$$

where $z = \begin{bmatrix} x_O & x_I & x^T \end{bmatrix}^T$, $x_O(t), x_I(t) \in \mathbb{R}$ *for all* t, $x(t) \in \mathbb{R}^{n-2}$ *for all* t, $A \in \mathbb{R}^{n \times n}, B \in$

$\mathbb{R}^n, c \in \mathbb{R}, C \in \mathbb{R}^{1\times n}$. *Then the system can be expressed as*

$$\begin{bmatrix} sX_O(s) \\ sX_I(s) \end{bmatrix} = \begin{bmatrix} W_{OO}(s) & W_{OI}(s) \\ W_{IO}(s) & W_{II}(s) \end{bmatrix} \begin{bmatrix} X_O(s) \\ X_I(s) \end{bmatrix} + \begin{bmatrix} V_O(s) \\ V_I(s) \end{bmatrix} U(s) \tag{4.9}$$

where $W_{OI}(s)$ is a transfer function describing the open loop dynamics from $X_I(s)$ to $X_O(s)$ involving only the states in $X(s)$ (excluding $X_I(s)$ and $X_O(s)$), $W_{OO}(s)$ is a transfer function describing self-regulatory open loop dynamics of X_O that involve only states in $X(s)$, $V_O(s)$ is the open loop transfer function from U to X_O describing dynamics that involve only states in $X(s)$, etc.

Proof. Observing the partitioning in the state vector $z = \begin{bmatrix} x_O & x_I & x^T \end{bmatrix}^T$, we can write state space equation matrices in block form as:

$$\begin{bmatrix} \dot{x_O} \\ \dot{x_I} \\ \dot{x} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \begin{bmatrix} x_O \\ x_I \\ x \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} u \tag{4.10}$$
$$y = cx_O + \begin{bmatrix} 0 \end{bmatrix} x = cx_O$$

Assuming $X(0) = 0$, we take Laplace transforms and solve for $X(s)$ in the third row to obtain

$$X(s) = (sI - A_{33})^{-1} \left(\begin{bmatrix} A_{31} & A_{32} \end{bmatrix} \begin{bmatrix} X_O(s) \\ X_I(s) \end{bmatrix} + B_3 U(s) \right),$$

and noting that $(sI - A_{33})^{-1}$ exists almost everywhere on $\mathbb{C}$, substituting this expression for $X(S)$ results in

$$\begin{bmatrix} sX_O(s) \\ sX_I(s) \end{bmatrix} = \begin{bmatrix} W_{OO}(s) & W_{OI}(s) \\ W_{IO}(s) & W_{II}(s) \end{bmatrix} \begin{bmatrix} X_O(s) \\ X_I(s) \end{bmatrix} + \begin{bmatrix} V_O(s) \\ V_I(s) \end{bmatrix} U(s),$$

where

$$\begin{aligned} W_{OO}(s) &= A_{11} + A_{13}(sI - A_{33})^{-1}A_{31}, \\ W_{OI}(s) &= A_{12} + A_{13}(sI - A_{33})^{-1}A_{32}, \\ W_{IO}(s) &= A_{21} + A_{23}(sI - A_{33})^{-1}A_{31}, \\ W_{II}(s) &= A_{22} + A_{23}(sI - A_{33})^{-1}A_{32}, \\ V_O(s) &= B_1 + A_{13}(sI - A_{33})^{-1}B_3, \\ V_I(s) &= B_2 + A_{23}(sI - A_{33})^{-1}B_3. \end{aligned} \tag{4.11}$$

□

This lemma shows that an arbitrary state space realization can be used to compute the open loop (proper) transfer functions describing the relationships between the output state x_O, intermediate x_I and input u. Notice that the form of equation (4.9) resembles the form of system (4.7); immediately, the question arises if the findings in the prequel generalize to transfer functions. The next result answers this question:

Theorem 1. *Suppose the system (4.8) is asymptotically stable. Suppose, for all $x \in \mathbb{R}_{\geq 0}$ we have that*

$$\left(\frac{W_{OI}(x)V_I(x)}{V_O(x)}\right) \leq 0 \text{ and } W_{II}(x) \geq \left(\frac{W_{OI}(x)V_I(x)}{V_O(x)}\right).$$

Further, if

$$W_{II}(s) - \frac{W_{OI}(s)V_I(s)}{V_O(s)} = k(s + D) + f_p(s),$$

where $0 \leq k < 1, D \in \mathbb{R}$ and $f_p(s)$ is a proper transfer function, then the transfer function of system (4.8) has at least one zero z in the closed right half plane of $\mathbb{C}$. Moreover, z is a nonnegative real number.

Proof. Let the set of nonnegative real numbers be denoted as X. Define $f(s) = W_{II}(s) - \frac{W_{OI}(s)V_I(s)}{V_O(s)}$. After some algebra, the transfer function of the system can written as

$$G(s) = \frac{s - f(s)}{(s - W_{OO}(s))(s - W_{II}(s)) - W_{OI}W_{IO}}$$

To show $G(s)$ has at least one zero in X, it suffices to show that $x - f(x)$ has a root $z \in X$. Since $\frac{W_{OI}(x)V_I(x)}{V_O(x)} \leq 0$ and $W_{II}(x) \geq \frac{W_{OI}(x)V_I(x)}{V_O(x)}$, then $f(x) \geq 0$ for all $x \in X$. Since the system is asymptotically stable, this implies that $f(x)$ has no right half plane poles in X. Define $p(x) = x - f(x) = x - (k(x + D) + f_p(x))$; clearly $p(x)$ is continuous since $f(x)$ has no poles in X. Notice that $p(0) = -f(0)$. If $f(0) = 0$, then we are done. If $f(0) \neq 0$, then $f(0) > 0$ which implies $p(0) < 0$. Next, write $p(x) = (1 - k)x - kD - f_p(x)$. Since $f_p(x)$ is proper, it is globally bounded on X. Denote

$$M = \max\{\sup_{x \in X} f_p(x), |kD|\}.$$

Then if $x > (M + kD)/(1 - k)$, $p(x) > 0$ and by continuity of $p(x)$ and the intermediate value theorem, $p(z) = 0$ for some $z \in X$. □

Remark 6. *The constraint that $f(x)$ be at least relative degree -1 can be interpreted as a constraint on the structure of the system (4.8). Since $W_{II}(s)$ is either proper or strictly proper, the improperness of $f(x)$ can only arise from the ratio $\frac{W_{OI}(x)V_I(x)}{V_O(x)}$ being improper. Since $W_{OI}(x)$ and $V_I(x)$ are proper, again, the only way that the ratio is improper is if $V_O(x)$ is strictly proper. When $V_O(x)$ has relative degree 0, then the $f(x)$ has relative degree 0, when $V_O(x)$ has relative degree one, then $f(x)$ has relative degree -1, and so forth. Thus, the constraint on $V_O(x)$ is that it possesses direct feedthrough from u to x_O (relative degree 0), or that there is effectively at most one integrator from u to x_O (relative degree 1).*

Remark 7. *The condition that $f(x) \geq 0$ for all $x \in X$ can be viewed as a constraint on the zeros of $f(x)$. Since the poles of $f(x)$ all lie in the open left half plane, $f(x)$ can never have a negative denominator. Therefore, to ensure that $f(x) > 0$ for all $x \in X$, any right half plane zero in $f(x)$ must have even algebraic multiplicity.*

4.4 Autocatalyic Systems

The autocatalytic system is a classic system where the end product of the system is also a catalyst for one or more of the reactions. Autocatalytic systems have been studied extensively, especially in the context of glycolysis [11]. In fact, the authors of [11] demonstrate that the existence of a right half plane zero in the glycolytic pathway of *S. cerevisiae* is the source of oscillatory behavior and a fundamental constraint on ATP dynamics. As an extension of their analysis, we pose a nonlinear model for an autocatalytic pathway with n chemical species and show that in the linearized model, autocatalysis produces an incoherent feedforward loop from the input of the pathway to the end product molecule. We consider the following Michaelis-Menten model for an autocatalytic system:

$$\begin{aligned}
\dot{x}_1 &= E^p \frac{x_p^{k_p} u}{k_{M,P} + x_p^{k_p}} - E^1 \frac{x_1^{k_1}}{k_{M,1} + x_1^{k_1}} \\
\dot{x}_2 &= E^1 \frac{x_1^{k_1}}{k_{M,1} + x_1^{k_1}} - E^2 \frac{x_2^{k_2}}{k_{M,2} + x_2^{k_2}} \\
\vdots &= \qquad \vdots \\
\dot{x}_{n-1} &= E^{n-2} \frac{x_{n-2}^{k_{n-2}}}{k_{M,n-2} + x_{n-2}^{k_{n-2}}} - E^{n-1} \frac{x_{n-1}^{k_{n-1}}}{k_{M,n-1} + x_{n-1}^{k_{n-1}}} \\
\dot{x}_p &= n_p \frac{E^{n-1} x_{n-1}^{k_{n-1}}}{k_{M,n-1} + x_{n-1}^{k_{n-1}}} - \frac{n_c E^p x_p^{k_p} u}{k_{M,P} + x_p^{k_p}} - \delta_C
\end{aligned} \tag{4.12}$$

where $E^p, E^2, ..., E^n$ are dedicated metabolic or housekeeping enzymes that convert x_p to x_2, x_2 to x_3, etc. We use the parameters n_c to denote the number of x_p molecules consumed to produce x_2 and n_p to denote the number of x_p molecules produced from precursor chemical species (or metabolite) x_n. The terms $k_{M,j}$, $j = p, 2, .., n$ are Michaelis constants for the n different enzymatic conversion reactions. Linearizing the system, the (A, B) and C matrices

have the following form:

$$
\begin{aligned}
A &= \begin{bmatrix} -a_{11} & 0 & \dots & 0 & a_{1p} \\ a_{11} & -a_{22} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & -a_{n-1,n-1} & 0 \\ 0 & 0 & \dots & n_p a_{n-1,n-1} & -n_c a_{p,p} \end{bmatrix} \\
B &= \begin{bmatrix} k_{1u} & 0 & \dots & 0 & -k_{pu} \end{bmatrix}^T \\
C &= \begin{bmatrix} 0 & 0 & \dots & 0 & 1 \end{bmatrix}
\end{aligned}
$$

where for $j = 1, .., n-1$ the entries of A can be written as

$$
a_{jj} = a_{j,j}(\mathbf{x_e}) = E^j \frac{k_{M,j} x_{j,e}^{k_{j-1}}}{\left(k_{M,j} + x_{j,e}^{k_j}\right)^2}
$$

$$
\begin{aligned}
a_{p,p}(\mathbf{x_e}, u_e) &= a_{1,p}(\mathbf{x_e}, u_e) = E^p u_e \frac{k_{M,p} x_{p,e}^{k_{p-1}}}{\left(k_{M,p} + x_{p,e}^{k_p}\right)^2} \\
k_{pu}(\mathbf{x_e}) &= n_c k_{1u}(\mathbf{x_e}) = n_c E^p \frac{x_{p,e}^{k_p}}{\left(k_{M,p} + x_{p,e}^{k_p}\right)}
\end{aligned}
$$

Taking Laplace transforms, we can inductively solve out $x_2, .., x_{n-1}$ and write the system in the form of equation (4.9), taking $x_O = x_p$ and $x_I = x_1$ to get the following terms for (W, V)

$$
W_{OI}(s) = \prod_{j=2}^{n-1} \frac{a_{j-1,j-1}}{(s + a_{j,j})}(\mathbf{x_e})
$$

$$\left[\begin{array}{ccc|ccccc} -\delta_1 & 0 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & -\delta_2 & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \frac{R^{\text{tot}}}{K_{M,1}}\frac{\left(\sum_{j\neq 1}\frac{m_{j,e}}{K_{M,j}}+1\right)}{\left(1+\sum_j \frac{m_{j,e}}{K_{M,j}}\right)^2} & -\frac{R^{\text{tot}}}{K_{M,2}}\frac{m_{1,e}}{K_{M,1}}\frac{1}{\left(1+\sum_j \frac{m_{j,e}}{K_{M,j}}\right)^2} & -\delta & 0 & \dots & 0 & -\frac{R^{\text{tot}}}{K_{M,3}}\frac{m_{1,e}}{K_{M,1}}\frac{1}{\left(1+\sum_j \frac{m_{j,e}}{K_{M,j}}\right)^2} & \dots & -\frac{R^{\text{tot}}}{K_{M,n}}\frac{m_{1,e}}{K_{M,1}}\frac{1}{\left(1+\sum_j \frac{m_{j,e}}{K_{M,j}}\right)^2} \\ \hline -\frac{R^{\text{tot}}}{K_{M,1}}\frac{m_{2,e}}{K_{M,2}}\frac{1}{\left(1+\sum_j \frac{m_{j,e}}{K_{M,j}}\right)^2} & \frac{R^{\text{tot}}}{K_{M,2}}\frac{\left(1+\sum_{j\neq 2}\frac{m_{j,e}}{K_{M,j}}\right)}{\left(1+\sum_j \frac{m_{j,e}}{K_{M,j}}\right)^2} & 0 & -\delta & \dots & 0 & -\frac{R^{\text{tot}}}{K_{M,3}}\frac{m_{2,e}}{K_{M,2}}\frac{1}{\left(1+\sum_j \frac{m_{j,e}}{K_{M,j}}\right)^2} & \dots & -\frac{R^{\text{tot}}}{K_{M,n}}\frac{m_{2,e}}{K_{M,2}}\frac{1}{\left(1+\sum_j \frac{m_{j,e}}{K_{M,j}}\right)^2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -\frac{R^{\text{tot}}}{K_{M,1}}\frac{m_{n,e}}{K_{M,n}}\frac{1}{\left(1+\sum_j \frac{m_{j,e}}{K_{M,j}}\right)^2} & -\frac{R^{\text{tot}}}{K_{M,2}}\frac{m_{n,e}}{K_{M,n}}\frac{1}{\left(1+\sum_j \frac{m_{j,e}}{K_{M,j}}\right)^2} & 0 & 0 & \dots & -\delta & -\frac{R^{\text{tot}}}{K_{M,3}}\frac{m_{n,e}}{K_{M,n}}\frac{1}{\left(1+\sum_j \frac{m_{j,e}}{K_{M,j}}\right)^2} & \dots & -\frac{R^{\text{tot}}}{K_{M,n}}\frac{\left(1+\sum_{j\neq n}\frac{m_j}{K_{M,j}}\right)}{\left(1+\sum_j \frac{m_{j,e}}{K_{M,j}}\right)^2} \\ 0 & 0 & 0 & 0 & \dots & 0 & -\delta_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & \dots & -\delta_n \end{array}\right] \tag{4.13}$$

$W_{OO} = -n_c a_{p,p}(\mathbf{x_e}), W_{IO} = a_{1,p}(\mathbf{x_e}), W_{II} = -a_{1,1}(\mathbf{x_e}), V_O = -k_{pu} = -n_c k_{1u}$, and $V_I = k_{1u}$. By Theorem 1, since $\frac{W_{OI}(x)V_I}{V_O} < 0$ for all $x \in \mathbb{R}_{\geq 0}$, then if

$$-a_{11}(\mathbf{x_e}) < n_c \prod_{j=2}^{n-1} \frac{a_{j-1,j-1}(\mathbf{x_e})}{s + a_{j,j}},$$

the transfer function from u to x_p, the product molecule that is consumed to be produced in the autocatalytic system, has a positive real zero.

4.5 Input-Coupled Systems

We now consider a general class of transcriptional and translational systems, comprised of at least two orthogonal genes. We suppose these two genes are activated by a small input molecule u_1 and thus refer to this type of system as an input-coupled system. The model is

written as

$$
\begin{aligned}
\dot{m}_1 &= \beta_1 - \delta_1 m_1 + K_{m_1,u} u_1 \\
\dot{m}_2 &= \beta_2 - \delta_2 m_2 + K_{m_2,u} u_1 \\
\dot{p}_1 &= R^{\text{tot}} \frac{m_1/K_{M,1}}{1 + \sum_{j=1}^{n} m_j/K_{M,j}} - \delta p_1 \\
\vdots &= \qquad \vdots \\
\dot{p}_n &= R^{\text{tot}} \frac{m_n/K_{M,n}}{1 + \sum_{j=1}^{n} m_j/K_{M,j}} - \delta p_n \\
\dot{m}_3 &= \beta_3 - \delta_3 m_3 \\
\vdots &= \qquad \vdots \\
\dot{m}_n &= \beta_n - \delta_n m_n.
\end{aligned}
$$

Let us examine what happens if we introduce only ribosomal loading on translation of mRNA into protein. Here we have assumed that binding of the ribosomal machinery to the ribosome binding site of an mRNA molecule happens much faster and that the ribosomal translational complex satisfies a Michaelis-Menten assumption, i.e. it reaches steady state much faster than $m_1, ..m_n, p_1, ..., p_n$.

To investigate the existence of a right half plane zero, we calculate the Jacobian in equation (4.13) for some nominal equilibrium point x_e, and define the block elements of the the Jacobian J as

$$
J \equiv \left[\begin{array}{c|c} A_{11}(x_e) & A_{12}(x_e) \\ \hline A_{21}(x_e) & A_{22}(x_e) \end{array}\right]
$$

with

$$
B = \begin{bmatrix} K_{m_1,u} & K_{m_2,u} & 0 & \dots & 0 \end{bmatrix}^T
$$

and

$$C = \begin{bmatrix} 0 & 0 & 1 & 0 \dots & 0 \end{bmatrix}.$$

Notice that the signed Boolean structure of A_{11} is identical to the signed Boolean structure of the A matrix in the IFFL system (4.15). Following the pattern discovered in the above example, if we suppose

$$\frac{\frac{m_{j,e}}{K_{M,j}}}{\left(\frac{m_{2,e}}{K_{M,2}}\right)^2} = O(\epsilon) \text{ for } j \neq 2 \; ; \; n = O(1) \tag{4.14}$$

then a direct application of the Woodbury matrix identity, allows us to write the transfer function as $G(s)$

$$\begin{aligned}
&= C \left(\left[\begin{array}{c|c} sI - A_{11}(x_e) & -A_{12}(x_e) \\ \hline -A_{21}(x_e) & sI - A_{22}(x_e) \end{array} \right]^{-1} \right) B \\
&= C \left[\begin{array}{c|c} (sI - A_{11} - A_{12}(sI - A_{22})^{-1} A_{21})^{-1} & \star \\ \hline \star & \star \end{array} \right] B \\
&= \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}^T \left((sI - A_{11}) - A_{12}(sI - A_{22})^{-1} A_{21} \right)^{-1} \begin{bmatrix} K_{m_1,u} \\ K_{m_2,u} \\ 0 \end{bmatrix}
\end{aligned}$$

and since $1 >> \epsilon$, if we pull out $(m_2/K_{M,2})^2$ from the denominator in A_{12}, A_{21}, and A_{22} it is easy to see that A_{12} is $O(\epsilon)$ and A_{21}, A_{22} is at most $O(1)$, thus implying that

$$\left((sI - A_{11}) - A_{12}(sI - A_{22})^{-1} A_{21} \right)^{-1} \approxeq (sI - A_{11})^{-1}.$$

Next, note that the signed Boolean structure of (A_{11}, B_1) is of the form

$$\left(\begin{bmatrix} -\delta_1 & 0 & 0 \\ 0 & -\delta_2 & 0 \\ K_{p_1,m_1} & -K_{p_1,m_2} & -\delta \end{bmatrix}, \begin{bmatrix} K_{m_1u} \\ K_{m_2u} \\ 0 \end{bmatrix} \right) \tag{4.15}$$

Permuting the states, so that p_1 is x_O and m_2 is x_I and $(u - u_e)$ is u, and applying the equations in (4.11) we get $W_{II}(s) = -\delta_2, V_I(s) = K_{m_2u}, W_{OO}(s) = -\delta$ and

$$\begin{aligned} W_{OI}(s) = & \quad -K_{p_1,m_2} \\ V_O(s) = & \quad \frac{K_{p_1,m_1}K_{m_1u}}{s+\delta_1}, \end{aligned}$$

with

$$f(x) = -\delta_2 - \left(-\frac{K_{p_1,m_2}K_{m_2,u}}{K_{p_1,m_1}K_{m_1,u}}(s+\delta_1) \right).$$

In this case, notice that the incoherence in K_{p_1,m_2} and K_{p_1,m_1} determines the sign of $\frac{W_{OI}(x)V_I(x)}{V_O(x)} \leq 0$ Also, $f(x)$ has relative degree -1 and the condition that $0 \leq K < 1$ implies that

$$K_{p_1,m_2}K_{m_2,u} < K_{p_1,m_1}K_{m_1,u} \tag{4.16}$$

and if

$$\left(\frac{K_{p_1,m_2}K_{m_2,u}}{K_{p_1,m_1}K_{m_1,u}/\delta_1} \right) > \delta_2 \tag{4.17}$$

then $f(x) \geq 0$ for all nonnegative real x and by Theorem 1 the system will have a right half plane zero.

In this particular class of systems, we assume degradation is not substantially saturated, i.e. we can approximate each degradation rate as linear. The reader will find that if an input-coupled system is posed with degradation crosstalk as the sole source of crosstalk, the system will have the potential to possess a right half plane zero only if 1) there is a down-regulation of one gene by another, and 2) a third gene dominates use of the degradation enzymes, so much that it sequesters the enzymes from the first or second. The key is the introduction of

an incoherent feedforward loop in the system. When there are multiple sources of resource-mediated crosstalk, again, multiple feedforward loops will be present and a right half plane zero may be present if the dominant feedforward loop is an incoherent feedforward loop.

4.6 Clp-XP Loading and Implications on the σ^{38} (RpoS) Regulated Stress Response

In this section we show how loading effects introduced by two competing pathways: 1) a pathway that is introduced synthetically with strong production gain and degradation gain and 2) the stress response pathway regulated by the master stress response regulator σ^{38} (RpoS) results in an incoherent feedforward loop with the potential for a right half plane zero. Specifically, we consider the effects of adding a high copy number gene that is engineered to have an LVA tag [27], a standard modification tag added to proteins to tune degradation rates. However, since Clp-XP degrades σ^{38} and any LVA-tagged molecule, when LVA-tagged proteins are produced in high quantity by a high copy number gene, the result is a sudden increase in Clp-XP degradable proteins which can lead to Clp-XP saturation [29]. Clp-XP regulates σ^{38} concentration, so if enough Clp-XP is sequestered, the result is that the effective lifetime of a σ^{38} molecule is extended. Furthermore, σ^{38} is the master stress response (up)regulator, an increase in its lifetime results in activation of critical stress response genes. These stress response genes can have adverse effects on cell metabolism, unnecessarily tax transcriptional and translational machinery (e.g. HPI and HPII catalases [59] which convert toxic hydrogen peroxide molecules into hydrogen and water), or in the worst case, induce cell lysis (e.g. the protein entericidin which induces cell lysis [6]).

While there are certain scenarios where inducing cell death may be the goal of a synthetic circuit, it is often the goal to engineer biocircuits that do not adversely impact the health of its host or minimally perturb the activity of host housekeeping genes. Therefore, it is important to understand whether such a synthetic circuit can adversely affect the cell's

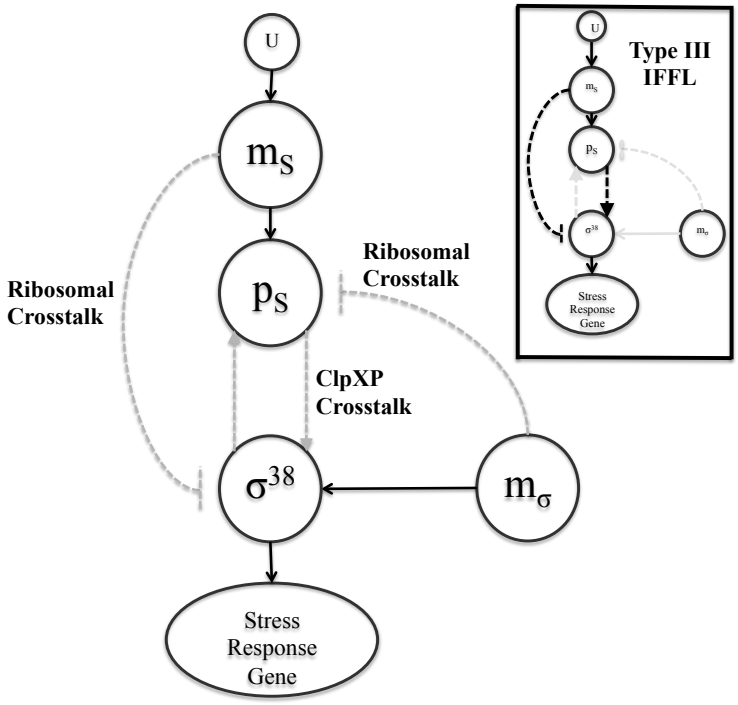

Figure 4.2: A diagram illustrating the interactions between chemical species in system (4.18). A synthetic gene is induced by a small molecule u, resulting in expression of mRNA molecule m_S. m_S translates into LVA-tagged protein p_S-LVA. The m_S mRNA molecules sequester ribosomes from m_σ, the mRNA transcript for σ^{38} — this creates a crosstalk interaction where m_S effectively down regulates σ^{38} expression (similarly, m_σ down regulates p_S expression, but only weakly when the synthetic gene is at high copy number). The p_S protein sequesters Clp-XP from σ^{38}, — this creates a crosstalk interaction where p_S in effect extends the lifetime of σ^{38}, which we indicate with an up-regulation arrow from p_S to σ^{38}. The inset highlights the Type III incoherent feedforward loop [62] is introduced via ribosomal and Clp-XP sequestration interactions. The relevant crosstalk interaction arrows in the IFFL are drawn as dotted and a darker color than the other crosstalk interactions.

ability to regulate its stress response. A schematic illustrating the interactions of the circuit, including the indirect crosstalk interactions (dotted) is shown in Figure 4.2. We model the system as follows:

$$\begin{aligned}
\dot{m}_S &= P^{\text{tot}} k^p_{\text{cat}} \frac{N \frac{D_S}{K_{M,D}}}{1 + N\frac{D_S}{K_{M,D}} + \frac{D_\sigma}{K_{M,\sigma_D}}} - \delta_S m_s + k_{su} u, \\
\dot{m}_\sigma &= P^{\text{tot}} k^p_{\text{cat}} \frac{\frac{D_\sigma}{K_{M,\sigma_D}}}{1 + N\frac{D_S}{K_{M,D}} + \frac{D_\sigma}{K_{M,\sigma_D}}} - \delta_\sigma m_\sigma, \\
\dot{p_S} &= k_{\text{cat}} \frac{R^{\text{tot}} \frac{m_s}{K_{M,s}}}{1 + \frac{m_s}{K_{M,S}} + \frac{m_\sigma}{K_{M,\sigma}}} - \kappa_{\text{cat}} \frac{C^{\text{tot}} \frac{P_S}{\kappa_{M,S}}}{1 + \frac{P_s}{\kappa_{M,S}} + \frac{\sigma^{38}}{\kappa_{M,\sigma}}}, \\
\dot{\sigma^{38}} &= k_{\text{cat}} \frac{R^{\text{tot}} \frac{m_\sigma}{K_{M,\sigma}}}{1 + \frac{m_\sigma}{K_{M,\sigma}} + \frac{m_S}{K_{M,S}}} - \kappa_{\text{cat}} \frac{C^{\text{tot}} \frac{\sigma^{38}}{\kappa_{M,\sigma}}}{1 + \frac{P_s}{\kappa_{M,S}} + \frac{\sigma^{38}}{\kappa_{M,\sigma}}}, \\
\dot{x}_{stress} &= \alpha \frac{\sigma^{38}}{k_M + \sigma^{38}} - \delta_x x_{stress}\,.
\end{aligned} \tag{4.18}$$

The Jacobian of the system has the following form:

$$\begin{bmatrix}
-a_{11} & 0 & 0 & 0 & 0 \\
0 & -a_{22} & 0 & 0 & 0 \\
a_{31} & -a_{32} & -a_{33} & a_{34} & 0 \\
-a_{41} & a_{42} & a_{43} & -a_{44} & 0 \\
0 & 0 & 0 & a_{54} & -a_{55}
\end{bmatrix},$$

with $B = \begin{bmatrix} k_{su} & 0 & 0 & 0 & 0 \end{bmatrix}^T$, where a_{ij} are computed in the usual fashion. Notice, the definition of the transfer function depends on which state we choose as our output. Since we are considering the copy number of our circuit to be particularly large, e.g. if the circuit was implemented on a high copy plasmid, then our concern is how drawing from the resources of the cell affects a critical survival mechanism — the stress response pathway. Thus, we are interested in how inducing our synthetic pathway with u affects production of stress response protein x_{stress}. Here x_{stress} can be interpreted as any of the proteins typically (positively)

regulated by σ^{38}, e.g. Thus, if u renders the cell unable to respond to stress, or worse yet, indirectly activates the stress response, this could lead to poor performance of the synthetic circuit or in the worst case, destruction of the host via cell lysis.

Computing the transfer function gives

$$G(s) = \frac{(-a_{41}a_{54}k_{su})\left(s - \frac{a_{43}a_{31}}{a_{41}} + a_{33}\right)}{D(s)}$$

where $D(s) = (s+a_{55})(a_{33}a_{44} - a_{34}a_{43} + a_{33}s + a_{44}s + s^2)(s+a_{11})$ and the zero can be written simply as

$$z = \frac{a_{54}a_{43}a_{31}k_{su}}{a_{54}a_{41}k_{su}} - a_{33} = \frac{k_{\text{cat}}C^{\text{tot}}\left(\frac{\sigma_e^{38}}{\kappa_{M,\sigma}}\frac{K_{M,\sigma}}{m_\sigma^e} - 1\right)}{\left(1 + \frac{p_s^e}{\kappa_{M,s}} + \frac{\sigma_e^{38}}{\kappa_{M,\sigma}}\right)^2}$$

which is positive if $\frac{K_{M,\sigma}}{m_\sigma^e} - \frac{\kappa_{M,\sigma}}{\sigma_e^{38}} > 0$. Examining the first expression for z, we see that the system has a right half plane zero if the effective gain of "up-regulation" of σ^{38} via Clp-XP saturation ($a_{43}a_{31}k_{su}$), normalized by the effective gain of "down-regulation" of σ^{38} via ribosomal loading ($a_{41}k_{su}$) is sufficiently large, specifically to exceed the rate of degradation of p_S (a_{33}). When overall up-regulation of σ^{38} only slightly exceeds degradation of σ^{38}, the result is a particularly slow right half plane zero.

Copy number of the synthetic circuit also plays a role in determine the size of z. When z is positive and small, increasing the copy number N in system (4.18) increases m_s^e which results in an increase in p_s^e. Notice that increasing p_s^e also results in a decrease in σ_e^{38}. If $z > 0$, then increasing copy number N drives z towards 0, resulting in a slower settling time and larger amplitude of the inverse transient. We consider the average copy numbers N of 4 standard replication origins that are used to carry synthetic circuits in *E. coli* and plot the step response as a function of N (Figure 4.3).

Recall that one goal in synthetic biological design is to implement synthetic pathways in biology that do not jeopardize the health of the cell or unintentionally activate unnecessary pathways. When we use Clp-XP to mediate degradation in our synthetic circuit, a right

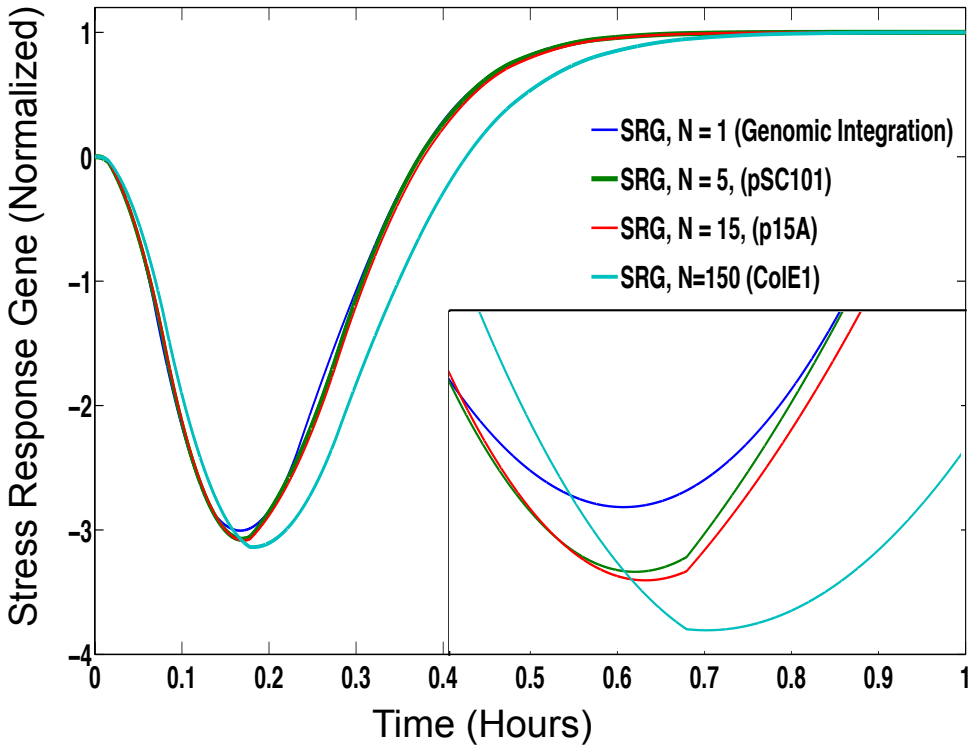

Figure 4.3: Normalized step response curves of stress response gene x_{stress} for the nonlinear system (4.18) plotted as a function of time and the synthetic gene copy number N. We selected average values representative of low, medium and high copy numbers of plasmid (displayed with a standard ORI of that copy number). These curves were generated using the following parameters: $P^{\text{tot}} = 200$ nM, $R^{\text{tot}} = 80$ nM, $C^{\text{tot}} = 75$ nM, $D_s = D_\sigma = 1$ nM, $k^p_{\text{cat}} = .009$ /s, $k_{\text{cat}} = \kappa_{\text{cat}} = .0002$ /s, $K_{M,S} = K_{M,\sigma} = K_{M,D} = K_{M,\sigma_D} = 500$ nM , $\kappa_{M,S} = \kappa_{M,\sigma} = 3$ nM, $\delta_S = \delta_\sigma = \delta_x = 5 \times 10^{-3}$ /s, $\alpha = 10$ nM/s, $k_M = 30$ nM, $k_{su} = .0001$ /s. A step of 1 nM was used. (Inset) A magnified view of the trough in the inverse transient of the step response.

half plane zero is introduced into the system, resulting in significant coupling to the stress response genes of the cell. Further, induction of our synthetic circuit results in 1) a transient (inverse) dynamic wherein the cell loses its ability to respond to stress, 2) the eventual up-regulation of stress response genes whether or not there actually is environmental stress. Both of these outcomes can be deleterious to the cell.

In presenting this example, our purpose is to simply increase understanding of the *potentially* adverse consequences of introducing LVA-tagged molecules on a high copy number biocircuit. We do not claim that right half plane zeros will always exist for such biocircuits,

as the existence of the right half plane zero is dependent on the equilibrium values of the chemical species involved, the Michaelis-Menten constants, and the amount of ribosomes, polymerases, and Clp-XP in each cell. Thus, our results should be viewed as an additional consideration when using Clp-XP to control degradation rates.

4.7 Conclusion

In this part of the thesis we reviewed the principle of resource loading and examined its effects on a signal cascade as a general motif within synthetic and natural signaling networks. We showed that saturation and competition of degradation enzymes produces unintended crosstalk interactions. Those crosstalk interactions introduced an *effective* incoherent feedforward loop, and under certain parametric conditions, a right half plane zero in the local dynamics of an equilibrium point. We analyzed the incoherent feedforward loop using a simple example and derived sufficient conditions for a multi-dimensional SISO system to have an incoherent feedforward loop and additionally, a right half plane zero. We then applied this result to derive parametric conditions under which a class of transcription-translation systems and a synthetic system leveraging LVA degradation technology would have a right half plane zero. We stress that cells always deal with finite resources [82] and as shown in [27], expressing just two genes was already enough to completely saturate ClpXp using typical promoters (pTet and pAra). It is thus likely that for any reasonably sized circuit, the resources for either production or degradation machinery will be saturated. Therefore, the issue of characterizing how right half plane zeros arise from resource limitations is an important design consideration, especially as synthetic biologists begin to address the challenge of assembling more complicated biocircuits as well as integrated systems consisting of multiple biocircuit parts.

Chapter 5

Modeling Stochastic Environmental Context Perturbations on Synthetic Gene Networks

5.1 Background

Cell to cell variability in gene expression [22] is a property of small volume, small copy number biochemical systems. From a controls standpoint, this variability imposes fundamental constraints on feedback performance [56] and create a challenge in designing circuits that must function around a specific operating point. Classic studies of synthetic oscillators [21] reveal that variable gene expression leads to variable oscillator phases, desynchronization, variable amplitude etc. However, recent strategies using combinatorial promoter architectures provide hope that the design of a robust oscillator [90] is possible. However, when the same oscillator was exposed to loading effects in [27], oscillation disappeared entirely in some cells while other cells produced slow irregular oscillations. Despite using a stochastic model to account for cell-to-cell variability, the stochastic model used in [90] could not account for environmental disturbances.

Perhaps the most widely accepted stochastic model for biochemical systems is the chemical master equation, a special instance of the forward Chapman Kolmogorov equation [47]. In [35], the author shows that the chemical master equation is an exact model for a well-mixed, thermally equilibrated gas-phase system. Typically, when used to model biochemical

systems in liquid phase, it is deemed a "mesoscopic description" of dynamics, as it is considered an intermediate representation between the microscopic representation of molecular dynamics and the macroscopic representation of a mass action kinetics model. Thus, in theory, the chemical master equation contains the necessary information to capture the randomness of molecules colliding and moving in a well-mixed volume as well as an appropriate level of abstraction to escape the analytical burden of simulating physical trajectories and collisions of individual molecules in the system.

However, finding an analytical solution for the chemical master equation is generally difficult, if not impossible, as it is typically infinite dimensional in the state-space [47]. Except in special instances where models are amenable to generating function approaches for exact solutions [80, 92] or where conservation laws enable finite bounds on the state-space [68], exact solutions for the master equation are difficult to obtain in closed form analytical expressions. Two alternatives exist to address this problem: 1) simulation using techniques such as τ-leaping, hybrid approaches, time scale separation approaches, or 2) reducing the model to a simpler or tractable form, e.g. using the finite state space projection algorithm [68], the sliding window abstraction approach [41], as well as spectral methods using basis functions to expand and approximate probability densities [18, 23, 72].

One of the outstanding challenges in modeling stochastic biochemical systems is the problem of accounting for system complexity—in a single cell there are millions of biomolecules present at any point in the cell cycle and many are often neglected in models but are critical to system function, e.g. ribosomes, RNAP, tRNA, σ-factors. Additionally, there is strong evidence to suggest that host, compositional, spatial and functional context, often ignored in synthetic and systems biology models, play a role in regulating gene expression [**?** **?**]. Global variables such as these are typically unaccounted for in stochastic models, yet they impact the dynamics of the biological systems studied.

In control theory, it is standard to include modeling terms that account for environmental disturbances [44] — often the disturbances are considered bounded and controllers are

subsequently designed to be robust to disturbances contained within those bounds. Such a perspective may be valuable when complemented with recent advances in experimental techniques employing optogenetic *ex vivo* control based on *in silico* models to regulate *in vivo* gene expression in cells [64, 93]. Once a notion of environmental disturbance is formulated, we can begin to probe the robustness of a particular *ex vivo* controller with respect to environmental perturbations in a stochastic modeling framework.

Toward this end, in this work we develop an approach for capturing environmental disturbances using the chemical master equation. We view our efforts as supplementary to the model reduction results of [18, 23, 72], as their techniques can be applied in concert with our own approaches or in a stepwise approach. Our results are complementary to the results in [42, 92], where total output noise is decomposed into system-extrinsic noise and system-intrinsic noise. Here we consider decomposition at the level of system *dynamics* as opposed to system outputs, as ultimately our goal is a framework for designing synthetic systems with dynamics robust to bounded environmental disturbances. Additionally, we seek to develop a framework that enables exclusion of any system variables that add unwanted model complexity but that do not substantially enrich the dynamical behavior of the system. Therefore, our aim is to develop models that account for environmental disturbances, but only those that substantially impact the dynamics of the system.

This chapter is organized as follows. In Section 5.2 we introduce notation, define the concept of a chemical reaction system, and the review the classical chemical master equation: a dynamical model for describing the reaction system dynamics. In Section 5.3, we introduce the notion of a system-environment decomposition on a chemical reaction system preparatory for our main result. In Section 5.5 and 5.6, we derive the main result, an additive decomposition of (plant) system dynamics into two terms: the first being a description of the evolving intrinsic uncertainty in the system and the second being a description of the disturbance that extrinsic uncertainty can have on the intrinsic state. We conclude in Section 5.7 with two simple examples of environmental disturbance, a model describing loading

effects between two genes and a model describing antibiotic perturbation to transcription and translation rates.

5.2 Preliminaries: The Chemical Master Equation

In this section, we introduce the mathematical framework and notation for our analysis. We begin by reviewing the concept of a chemical reaction system. Since we ultimately seek to decompose this global system into a specific system and its environment, we will refer to it as the global chemical reaction system.

Definition 6. *Define $\mathcal{C} = (\mathcal{S}, \mathcal{R})$ to be a global chemical reaction system with $\mathcal{S} = \{S_1, .., S_N\}$ being a set containing all N chemical species in the global chemical reaction system. Let $\mathcal{R} = \{R_1, ..., R_M\}$ be a set enumerating all M reactions in $\mathcal{C}$.*

Remark 8. *The elements of the set $\mathcal{R}$ are reactions. Mathematically, a reaction $R_j \in \mathcal{R}$ is defined based on a species set $\mathcal{S}$ and can be thought of as an ordered 4-tuple of sets $R_j = (\{c_1, .., c_k\}, \{S_1, .., S_k\}, \{d_1, .., d_n\}, \{P_1, ..., P_n\})$, where $c_1, .., c_k, d_1, .., d_n \in \mathbb{N}$, $S_1, .., S_k, P_1, .., P_n \in \mathcal{S}$. The first set of R specifies the stoichiometry of the reactants, the second set the list of reactants, the third set the stoichiometry of the products, and the fourth set the list of the products. Typically, we will follow convention and express the reaction R_j as $c_1S_1 + \cdots + c_kS_k \to d_1P_1 + \cdots + d_nP_n$, and not as an ordered 4-tuple. The above convention can be viewed as an implicit reference to the underlying mathematical object that defines the reaction R_j: an ordered 4-tuple of sets.*

The global chemical reaction system is thus a list of all potential chemical species and chemical reactions occurring in a relevant biological chassis, e.g. a cell, an *in vitro* test tube, vesicle, etc. In principle, the size of $\mathcal{S}$ and $\mathcal{R}$ are very large, since it must include all possible partial products of transcription, i.e. aborted transcripts, background molecules critical for metabolism, intermediate metabolites, etc. Most biological models exclude the complexity

found in the global chemical reaction system, as it contents are mostly unknown, in addition to being computationally and analytically intractable.

We restrict our attention to global chemical reaction systems whose contents are well stirred, in a fixed volume, and at a constant temperature. Under these conditions, we define $X(t)$ to be a vector of copy numbers, with $X_i(t)$ being the copy number of S_i at time $t, i = 1, .., N$. We suppose that for each reaction $R_j \in \mathcal{R}$ there exists a propensity function $w_j(X(t))$ that characterizes the probability of reaction R_j firing in time interval dt as $w_j(X(t))dt$ [35]. We note this is an assumption, rather than a consequence as in [35] since $\mathcal{C}$ is not necessarily a gas-phase system. We define the stoichiometric transition vector for each reaction R_j as $\xi_j = \begin{bmatrix} \nu_1 & \dots & \nu_N \end{bmatrix}^T$, where ν_k describes the stoichiometric change in X_k during reaction R_j. Thus, if $X = x_o$ before R_j fires, then $X = x_o + \xi_j$ after R_j has fired. Further, with some abuse of notation, we will suppose that if $X = (Y, Z)$, then $\xi_j[Y]$ denotes the subvector of ξ_j that records the stoichiometric change of Y. The chemical master equation of the system $\mathcal{C}$ is then given as

$$\begin{aligned} \frac{d}{dt}P(X(t)|X(t_o)) = &\sum_{j=1}^{M} w_j(X(t) - \xi_j)P(X(t) - \xi_j|X(t_o)) \\ &- P(X(t)|X(t_o))\sum_{j=1}^{M} w_j(X(t)) \end{aligned} \tag{5.1}$$

The chemical master equation specifies the evolution of the joint probability mass function of $X(t)$. Since $X(t)$ is a vector of species copy numbers, its entries take on nonnegative integer values. We refer to the set of values that $X(t)$ can take as the configuration space.

5.3 Decomposition of the Global Chemical Reaction System

Now that we have a way of describing the global chemical reaction system $\mathcal{C}$, we can consider its relationship to a system of interest. This system may coincide with all the measurable chemical species in the global chemical reaction system, a select set of genes under study and their associated transcriptional and translational products, or even a set of chemical species that are associated with a synthetic biocircuit. Our representation of this system should thus be flexible, as it may require the inclusion of specific reporter molecules and their precursor mRNA transcripts, or include only a single chemical species, corresponding to an inducible and measurable protein. The only constraint we impose is that all its chemical species are within $\mathcal{S}$.

Definition 7. *Let $S_1, .., S_n \in \mathcal{S}$ be a list of relevant chemical species. We define the chemical reaction system*

$$\mathcal{S}^p \equiv (S^p, R^p)$$

associated with this list of species and refer to this as our system or plant, where $S^p \equiv \{S_1, .., S_n\}$ and $R^p = \{R_j \in \mathcal{R} | \text{all products and reactants of } R_j \text{ are in } S^p\}$.

Notice in defining such a system in the global chemical reaction system, we assume knowledge of a pre-specified list of chemical species $S_1, .., S_n$. This list of chemical species then determines the list of reactions intrinsic to this system, as they do not require the presence of chemical species outside the system to function. Alternatively, we could proceed by defining a list of relevant reactions and subsequently impose that all products and reactants associated with those reactions be the list of species for our system. However, a reaction set defined in that manner may not include all self-contained reactions of chemical species in the system, as there may be other chemical reactions that only involve elements of S^p. Finally, we use this particular approach as it is typical to think of biological systems first

as a collection of chemical species and subsequently enumerate the list of relevant reactions. We define the environmental chemical reaction system as follows:

Definition 8. *Define the chemical reaction system* $\mathcal{S}^e = (S^e, R^e)$ *as the environmental system, where*

$$S^e \equiv \mathcal{S} \backslash S^p \ , \ R^e \equiv \mathcal{R} \backslash R^p$$

We will suppose X is ordered so that $X = (X^p, X^e)^T$, i.e. the first n elements specify the copy number of the species in S^p while the last $N-n$ elements specify the species in S^e. Viewing the chemical master equation as a state-space model with $P(X^p(t), X^e(t)|X^p(t_o), X^e(t_o))$, we will refer to $P(X^p(t))$ as the state of the system and $P(X^e(t))$ as the state of the environmental system. Finally, we denote the number of reactions in R^p and R^e as m_p and m_e respectively.

5.4 An Additive Decomposition of the Chemical Master Equation

Our goal is achieve a representation of the chemical master equation that captures only state of the system $\mathcal{S}^p$, $P(X^p(t))$, how it evolves over time and how the environmental system impacts that evolution. Ideally, we would like to write a decomposition of the form

$$\frac{d}{dt} P(X^p(t)) = f(X^p(t)) + g(X^p(t), X^e(t)). \tag{5.2}$$

Such a representation would allow us to include in $f(X^p(t))$ any dynamics that are relevant to the system, e.g. for design or parameter estimation purposes, while the environmental disturbance term $g(X^e(t))$ would act as a perturbation or disturbance to the nominal system's trajectory. As the derivation of the decomposition is long, we divide it into two parts: the first part evaluates the consequences of decomposing the species set $\mathcal{S}$ of the global chemical reaction system and the second part evaluates the consequences of decomposing the reaction

set $\mathcal{R}$ of the global chemical reaction system.

5.5 Part I : Leveraging the Partition on Chemical Species

The primary obstacle to achieving a decomposition of the form (5.2) is that the chemical master equation (5.1) describes the evolution of the joint probability mass function $P(X^p(t), X^e(t)|X^p(t_o), X^e(t_o))$. Typically, to separate $P(X^p(t))$ from

$$P(X^p(t), X^e(t), X^p(t_o), X^e(t_o))$$

requires an assumption of independence between the stochastic processes $X^p(t)$ and $X^e(t)$. This is a strong assumption, one that contradicts the very purpose of our analysis: to understand how the environmental state affects system dynamics.

Alternatively, we consider averaging out the effects of $X^e(t)$, i.e. marginalizing the joint probability mass to obtain the marginal in $X^p(t)$. Rather than laboriously analyzing the effect of individual sample trajectories of $X^e(t)$ on $X^p(t)$, this approach has the advantage of describing the average effect of the distribution of sample trajectories $X^e(t)$ on $X^p(t)$. First, we write the chemical master equation to include the decomposition of the global chemical species set $\mathcal{S}$:

$$\begin{aligned}\frac{d}{dt}P\left(\begin{bmatrix}X^p(t)\\X^e(t)\end{bmatrix}\right) = \sum_{j=1}^{M} w_j \left(\begin{bmatrix}X^p(t)\\X^e(t)\end{bmatrix} - \xi_j\right) P\left(\begin{bmatrix}X^p(t)\\X^e(t)\end{bmatrix} - \xi_j\right) \\ - w_j \left(\begin{bmatrix}X^p(t)\\X^e(t)\end{bmatrix}\right) P\left(\begin{bmatrix}X^p(t)\\X^e(t)\end{bmatrix}\right).\end{aligned}$$

Here we have suppressed the convention of carrying the initial condition as a conditioning argument in each probability mass function, as it will make the derivation easier to read. With some abuse of notation, we write the argument of the probability mass function as X^p or X^e, which will be an abbreviation for the probability mass function actually evaluated at a point

(x, y) in the configuration space, i.e. $P(X^p(t) = x, X^e(t) = y)$. If we use $P(X^p(t), X^e(t))$ to refer to the *probability mass function*, we will explicitly say so. The same notation will hold true for conditional and marginal probability density functions. Let $\mathcal{S}(X^e)$ denote the set of values that X^e can assume in the configuration space. If we sum over $\mathcal{S}(X^e)$, the left hand side becomes

$$\begin{aligned}
&\sum_{\mathcal{S}(X^e)} \frac{d}{dt} P\left(\begin{bmatrix} X^p(t) \\ X^e(t) \end{bmatrix}\right) = \frac{d}{dt} \sum_{\mathcal{S}(X^e)} P\left(\begin{bmatrix} X^p(t) \\ X^e(t) \end{bmatrix}\right) \\
&= \frac{d}{dt} \sum_{\mathcal{S}(X^e)} P\left(X^e(t) | X^p(t)\right) P\left(X^p(t)\right) \\
&= \frac{d}{dt} P(X^p(t)) \sum_{\mathcal{S}(X^e)} P\left(X^e(t) | X^p(t)\right) \\
&= \frac{d}{dt} P\left(X^p(t)\right).
\end{aligned}$$

The first equality holds due to uniform convergence of the sum

$$\sum_{S(X^e)} P(X^e(t) | X^p(t), X^p(t_o), X^e(t_o)).$$

The second and third equality holds from the law of conditioning. In the last equality, we use the fact that the *probability mass function* $P(X^e(t)|X^p(t), X^p(t_o), X^e(t_o))$ when summed over all values of X^e in the configuration space is unity. We now address the right hand side of the chemical master equation. Summing over $\mathcal{S}(X^e)$ and conditioning on $X^p(t)$ gives

$$\begin{aligned}
&\sum_{\mathcal{S}(X^e)} \Big[\sum_{j=1}^{M} w_j \left(\begin{bmatrix} X^p(t) \\ X^e(t) \end{bmatrix} - \xi_j\right) P\left(\begin{bmatrix} X^p(t) \\ X^e(t) \end{bmatrix} - \xi_j\right) \\
&- w_j \left(\begin{bmatrix} X^p(t) \\ X^e(t) \end{bmatrix}\right) P\left(\begin{bmatrix} X^p(t) \\ X^e(t) \end{bmatrix}\right) \Big]
\end{aligned}$$

$$
\begin{aligned}
=&\sum_{j=1}^{M}\sum_{\mathcal{S}(X^e)} w_j \left(\begin{bmatrix} X^p(t) \\ X^e(t) \end{bmatrix} - \xi_j\right) f_c(X^p, X^e) f_m(X^p) \\
&-\sum_{j=1}^{M}\sum_{\mathcal{S}(X^e)} w_j \left(\begin{bmatrix} X^p(t) \\ X^e(t) \end{bmatrix}\right) P\left(X^e(t)|X^p(t)\right) P\left(X^p(t)\right)
\end{aligned}
$$

where the conditional and marginal probability mass functions are written as

$$
\begin{aligned}
f_c(\cdot) &= P\left(X^e(t) - \xi_j\left[X^e\right] | X^p(t) - \xi_j\left[X^p\right]\right) \\
f_m(\cdot) &= P(X^p(t) - \xi_j\left[X^p\right]).
\end{aligned}
$$

For each $j = 1, .., M$ we pull out the marginal of $X^p(t)$ and summing over $\mathcal{S}(X^e)$ gives

$$
\begin{aligned}
\frac{d}{dt}P(X^p(t)) = &\sum_{j=1}^{M} P(X^p(t) - \xi_j\left[X^p\right])\alpha_j(X^p(t) - \xi_j\left[X^p\right]) \\
&-\sum_{j=1}^{M} P\left(X^p(t)\right)\alpha_j(X^p(t))
\end{aligned}
$$

where

$$
\alpha_j(X^p(t)) \equiv \sum_{\mathcal{S}(X^e)} w_j \left(\begin{bmatrix} X^p(t) \\ X^e(t) \end{bmatrix}\right) P\left(X^e(t)|X^p(t)\right).
$$

In summary, the preceding equations marginalize the master equation of the joint probability mass function, leveraging the decomposition of X into X^p and X^e. We consider $\alpha_j(X^p(t))$ as the averaged propensity functions for the system $\mathcal{S}$, since they can also be expressed as

$$
\alpha_j(X^p(t)) = \mathbb{E}_{X^e(t)|X^p(t)}\left[w_j(X^p(t), X^e(t))\right].
$$

Writing out the marginalized master equation, we have

$$\frac{d}{dt}P(X^p(t)) = \sum_{j=1}^{M} \alpha_j(X^p(t) - \xi_j\,[X^p])P(X^p(t) - \xi_j\,[X^p]) - \sum_{j=1}^{M} \alpha_j(X^p(t))P\,(X^p(t)) \tag{5.3}$$

and note that the averaged propensity functions $\alpha_j(X^p(t))$ specify the probability that reaction R_j will happen in the time interval $[t, t+dt]$, $\alpha_j(X^p(t)dt$, averaged over all possible values of $X^e(t)$. The decomposition of the species set $\mathcal{S} = S^p \cup S^e$ thus produces a representation of the chemical master equation that describes only the evolution of the marginal density $P(X^p(t))$.

5.6 Part II: Leveraging the Partition on the Chemical Reactions

If we now incorporate the decomposition on the reaction set $\mathcal{R}$, we can also rewrite the propensity functions in terms of the m_p reactions that only involve chemical species in S^p and m_e reactions involving system or environmental species. Notice the term $\sum_{j=1}^{M} \alpha_j(X^p(t) - \xi_j\,[X^p])P(X^p(t) - \xi_j\,[X^p])$ can be written as

$$\begin{aligned} &\sum_{j=1}^{m} \alpha_j(X^p(t) - \xi_j\,[X^p])P\,(X^p(t) - \xi_j\,[X^p]) \\ =&\sum_{j=1}^{m_p} \alpha_j(X^p(t) - \xi_j\,[X^p])P\,(X^p(t) - \xi_j\,[X^p]) \\ +&\sum_{j=1}^{m_e} \alpha_j(X^p(t) - \xi_j\,[X^p])P\,(X^p(t) - \xi_j\,[X^p]) \end{aligned}$$

and since the first m_p reactions do not involve X^e, we can write for each reaction $j = 1, ..m_p$ the associated propensity function for those reactions as $w_j(X^p(t), X^e(t)) = w_j(X^p(t))$ and

so we can further write $\sum_{j=1}^{M} \alpha_j(X^p(t) - \xi_j[X^p])P(X^p(t) - \xi_j[X^p])$ as

$$
\begin{aligned}
&= \sum_{j=1}^{m_p} P\left(X^p(t) - \xi_j[X^p]\right) \times \\
&\sum_{\mathcal{S}(X^e)} w_j\left(X^p(t) - \xi_j[X^p]\right) P\left(X^e(t)|X^p(t)\right) \\
&+ \sum_{j=1}^{m_e} P\left(X^p(t) - \xi_j[X^p]\right) \sum_{\mathcal{S}(X^e)} \alpha_j \left(\begin{bmatrix} X^p(t) \\ X^e(t) \end{bmatrix} - \xi_j \right) \\
&= \sum_{j=1}^{m_p} w_j\left(X^p(t) - \xi_j[X^p]\right) \times \\
&P\left(X^p(t) - \xi_j[X^p]\right) \sum_{\mathcal{S}(X^e)} P\left(X^e(t)|X^p(t)\right) \\
&+ \sum_{j=1}^{m_e} P\left(X^p(t) - \xi_j[X^p]\right) \sum_{\mathcal{S}(X^e)} \alpha_j \left(\begin{bmatrix} X^p(t) \\ X^e(t) \end{bmatrix} - \xi_j \right) \\
&= \sum_{j=1}^{m_p} w_j\left(X^p(t) - \xi_j[X^p]\right) P\left(X^p(t) - \xi_j[X^p]\right) (1) \\
&+ \sum_{j=1}^{m_e} \alpha_j(X^p(t) - \xi_j[X^p]) P\left(X^p(t) - \xi_j[X^p]\right)
\end{aligned}
$$

with a similar derivation holding for $\xi_j \equiv 0$, thus implying that the marginalized chemical master equation, or state-space model for $P(X^p(t))$, becomes:

$$
\begin{aligned}
&\frac{d}{dt}P\left(X^p(t)\right) \\
&= \sum_{j=1}^{m_p} w_j\left(X^p(t) - \xi_j\left[X^p\right]\right) P\left(X^p(t) - \xi_j\left[X^p\right]\right) \\
&- \sum_{j=1}^{m_p} w_j\left(X^p(t)\right) P\left(X^p(t)\right) \\
&+ \sum_{j=1}^{m_e} \alpha_j \left(\begin{bmatrix} X^p(t) \\ X^e(t) \end{bmatrix} - \xi_j\right) P\left(X^p(t) - \xi_j\left[X^p\right]\right) \\
&- \sum_{j=1}^{m_e} \alpha_j \left(\begin{bmatrix} X^p(t) \\ X^e(t) \end{bmatrix}\right) P\left(X^p(t)\right) \\
&\equiv f^P(P(X^p(t))) + f^E(P(X^e(t)|X^p(t)), P(X^p(t))) \qquad (5.4)
\end{aligned}
$$

where '$\equiv$' indicates that we define f^P and f^E as

$$
\begin{aligned}
f^P(\cdot) &= \sum_{j=1}^{m_p} w_j\left(X^p(t) - \xi_j\left[X^p\right]\right) P\left(X^p(t) - \xi_j\left[X^p\right]\right) \\
&\qquad - \sum_{j=1}^{m_p} w_j\left(X^p(t)\right) P\left(X^p(t)\right), \\
f^E(\cdot) &= \sum_{j=1}^{m_e} \alpha_j \left(\begin{bmatrix} X^p(t) \\ X^e(t) \end{bmatrix} - \xi_j\right) P\left(X^p(t) - \xi_j\left[X^p\right]\right) \\
&\qquad - \sum_{j=1}^{m_e} \alpha_j \left(\begin{bmatrix} X^p(t) \\ X^e(t) \end{bmatrix}\right) P\left(X^p(t)\right).
\end{aligned}
$$

To summarize, we first imposed a decomposition on the chemical species of the global chemical reaction system to obtain two subsystems: the system of interest and its environment. Second, we marginalized the chemical master equation to obtain a master equation that described only the time-evolution of the state of the system $P(X^p(t))$ using averaged propensity functions $\alpha_j(X^p(t))$. Finally, we imposed knowledge about the dependencies of the reactions and this resulted in a simple additive decomposition of the marginalized dynamics.

$$\frac{d}{dt}P\left(X^p(t)\right) = f^P(P(X^p(t))) + f^E(P(X^e(t)|X^p(t)), P(X^p(t)))$$

Notice this decomposition depends on two functionals: f^P which depends only on the state of the system $P(X^p(t))$ and f^E which depends on the state of the environmental system $P(X^e|X^p(t))$ and the state of the system $P(X^p(t))$. Since our derivation began from the chemical master equation (5.1) of the state of the global chemical reaction system $\mathcal{S}$ and since we have not imposed any additional assumptions—only using the normalization property of a probability mass function and conditioning arguments—the decomposition is exactly consistent with the dynamics of the global chemical reaction system.

Also, notice that the term $f^P(P(X^p(t)))$ depends on the exact propensity functions $w_j(X^p(t))$ in its definition. Thus, if $S^e = \emptyset$ or $R^e = \emptyset$, the term $f^P(P(X^p(t)))$ is precisely the right half-side of the chemical master equation (5.1). This is important for several reasons: 1) the term $f^P(X^p(t))$ can be viewed as the *complete* dynamics for a simple model system involving only the system variables, see the supplementary information of [42, 80, 92] for examples, 2) recognizing that $f^P(X^p(t))$ describes a simple model system's dynamics, it may be possible to omit any species in S^p that make the simplified model $\dot{P}(X^p(t)) = f^P(X^p(t))$ intractable or to introduce additional species from S^e to ensure the presence of conservation laws, potentially making the configuration space of S^p finite.

5.7 Using the System-Environment Decomposition to Model Environmental Disturbances: Examples

Ribosomal loading between two genes

We now consider an example system to illustrate how our decomposition enables modeling of environmental disturbances. We suppose the system of interest consists of $n = 2$ chemical

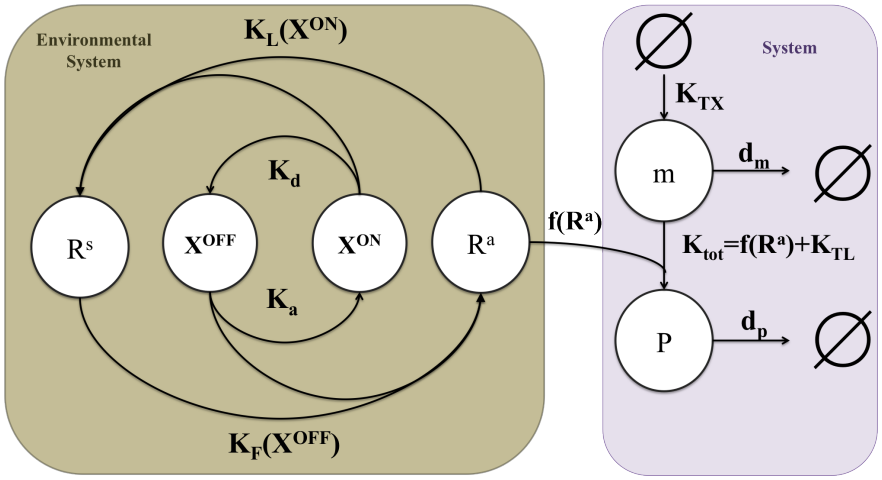

Figure 5.1: A schematic illustrating the interactions between chemical species in the system (5.5). mRNA and protein are produced constitutively, while ribosome in the abundant state R^a are able to augment production. A competing gene X sequesters ribosome away from m, facilitating R's transition from the abundant state R^a to the scarce state R^s. A confluence of two arrows indicates that either k_L depends on X^{ON}, k_F depends on X^{OFF}, or K_{tot} depends on $f(R)$.

species, S_1 is an mRNA m which encodes the protein p. There are $m_p = 4$ system reactions:

$$\emptyset \xrightarrow{k_{TX}} m, \qquad m \xrightarrow{k_{TL}} m + p,$$
$$m \xrightarrow{\delta_m} \emptyset, \qquad p \xrightarrow{\delta_p} \emptyset$$

A diagram illustrating the structure of this simple system is shown in Figure 5.1. We will assume that basal expression of mRNA is very low in the absence of an environmental cue (e.g. transcriptional machinery is scarce) which results in a small basal transcription rate k_{TX}. Furthermore, we will assume that in the absence of a separate environmental cue (e.g. ribosomal and translation machinery is scarce), the rate of translation k_{TL} is quite small. Finally, since the rates represent weak or basal expression, we suppose that k_{TX} and k_{TL} are zero-order rates that do not depend on the actual concentration of mRNA or protein (i.e. they are rate limited by RNAP and ribosome counts). We suppose that the degradation rates do depend on the copy number of m and p. We write the dynamics of the chemical master equation for the isolated (or toy) system as $\dot{P}(p, m, t) = f^P(P(X^p, t))$ where $f^P(P(X^p, t))$

equals

$$
\begin{aligned}
f^P(P(X^p,t)) &= k_{TL}P(p-1,m,t) + k_{TX}P(p,m-1,t) \\
&+ \delta_m(m+1)P(p,m+1,t) + \delta_p(p+1)P(p+1,m,t) \\
&- k_{TL}P(p,m,t) - k_{TX}P(p,m,t) - (\delta_m m + \delta_p p)P(p,m,t)
\end{aligned} \tag{5.5}
$$

The solution for this system is obtained by computing the probability generating function $F(z_1,z_2) = \sum_{m,p} z_1^p z_2^m P(p,m,t)$. Transforming the system (5.5) we obtain

$$
\begin{aligned}
\frac{\partial F}{\partial t} &= (k_{TL}z_1 + \delta_p - \delta_p)\frac{\partial F}{\partial z_1} + (k_{TX}z_2 + \delta_m - \delta_m z_2)\frac{\partial F}{\partial z_2} \\
&\quad + (-k_{TL} - k_{TX})F(z_1,z_2,t).
\end{aligned}
$$

By applying the method of characteristics, we obtain a closed form expression for the probability generating function $F(z_1,z_2,t)$:

$$
\begin{aligned}
& e^{-(k_{TX}+k_{TL})t} \sum_{m,p\in\mathbb{Z}\geq 0} \left((z_1+\delta_p)e^{(d_p-k_{TL})t} - \delta_p\right)^p \\
& \times \left((z_2+\delta_m)e^{(\delta_m-k_{TX})t} - \delta_m\right)^m P(p,m,0)
\end{aligned}
$$

from which we can calculate the probability mass function to be written as follows:

$$
\begin{aligned}
& P(p=k_1, m=k_2, t) = e^{-(k_{TX}+k_{TL})t}\frac{1}{k_1!k_2!}\times \\
& \sum_{m,p\in\mathbb{Z}\geq 0} \Bigg(\prod_{j=0,k_1\leq p}^{k_1-1} (p-j) f_1(k_1,t) \left(\delta_p e^{-(k_{TL}-\delta_p)t} - \delta_p\right)^{p-k_1} \\
& \prod_{i=0,k_2\leq m}^{k_2-1} (m-i) f_2(k_2,t) \left(\delta_m e^{-(k_{TX}-\delta_m)t} - \delta_m\right)^{m-k_2} \\
& \times P(p,m,0)\Bigg)
\end{aligned} \tag{5.6}
$$

and $f_1(k_1, t) = e^{-(k_{TL}-\delta_p)k_1 t}, f_2(k_2, t) = e^{-(k_{TX}-\delta_m)k_2 t}$ As our system has been chosen to be relatively simple, i.e. reflecting simplified models that typically exclude environmental species from the list of chemical species, the solution to (5.5) is a closed form analytical solution. Notice that the configuration space is not necessarily finite, so we sum over the positive integers.

Our goal is to now modify the system dynamics using $f^E(P(X^e(t)))$ to explore the effect of ribosomal loading. Because the complexity of the system may introduce nonlinearities into the generation function, our approach will be to simulate the perturbed system and compare it to the isolated system. We suppose the environmental system contains the ribosomal species R necessary to increase translation rates, but that a species X has the ability to sequester ribosomes away from translating m to p. We will suppose that ribosomes can assume two states: either abundant or scarce and we denote them as R^a and R^s, respectively. When X is in the X^{ON} state, it facilitates the conversion of R^a to R^s (and vice versa when X is on the X^{OFF} state). We denote the environmental reactions as follows:

$$\begin{aligned} R^s &\xrightarrow{k_{F(X)}} R^a, & X^{OFF} &\xrightarrow{k_a} X^{ON}, \\ R^a &\xrightarrow{k_L(X)} R^s, & X^{ON} &\xrightarrow{k_d} X^{OFF}, \\ R^a + m &\xrightarrow{k_{TLF}} R^a + p + m, \end{aligned} \tag{5.7}$$

with the last reaction denoting enhanced translation rates of m due to the abundance of ribosomes.

We suppose that X is a gene regulated by some external input and switches randomly between its off and on states, independent of the current state of R. We assume its expression to be strong, so that the dynamics of m and p do not impact its transition rates. We then

write the solution for X as $p(X,t) = e^{At}P(X,0)$ where

$$A = \begin{bmatrix} -k_d & k_a \\ k_d & -k_a \end{bmatrix},$$

We suppose that $P(X(0)) = \begin{bmatrix} k_a & k_d \end{bmatrix}^T$ where $k_a + k_d = 1$. Under these assumptions, we can write

$$P(X(t)) = e^{At} \begin{bmatrix} k_a \\ k_d \end{bmatrix} = \begin{bmatrix} k_a \\ k_d \end{bmatrix}$$

Further, substituting and conditioning with $P(X(t))$ gives us the following linear equation.

$$\frac{d}{dt} \begin{bmatrix} P(R^a,t) \\ P(R^s,t) \end{bmatrix} = \begin{bmatrix} -k_a k_L & k_d k_F \\ k_a k_L & -k_d k_F \end{bmatrix} \tag{5.8}$$

The solution can be substituted into $f^E(P(X^e|p,m,t), P(X^p|t))$, allowing us to write it as

$$-k_{TLF}P(R^a,t)P(p,m,t) + k_{TLF}P(R^a,t)P(p-1,m,t).$$

A simulation of the system is plotted in Figure 5.2; we see that protein expression is the highest when the probability that gene X stays off is close to 1. The reason is that when X is on, the amount of free ribosomes decrease (sequestration of ribosomes by X) and the amount of p produced is less.

To summarize, in this example we have posed a simple approach for capturing the effects of enzyme sequestration or loading effects [96]. We showed that the state of the protein and mRNA of our system can be strongly influenced by the state of X, which is a chemical species that does not directly interact with m or p. Thus, ribosomal loading can lead to indirect interactions between chemical species, even in a chemical master equation modeling framework.

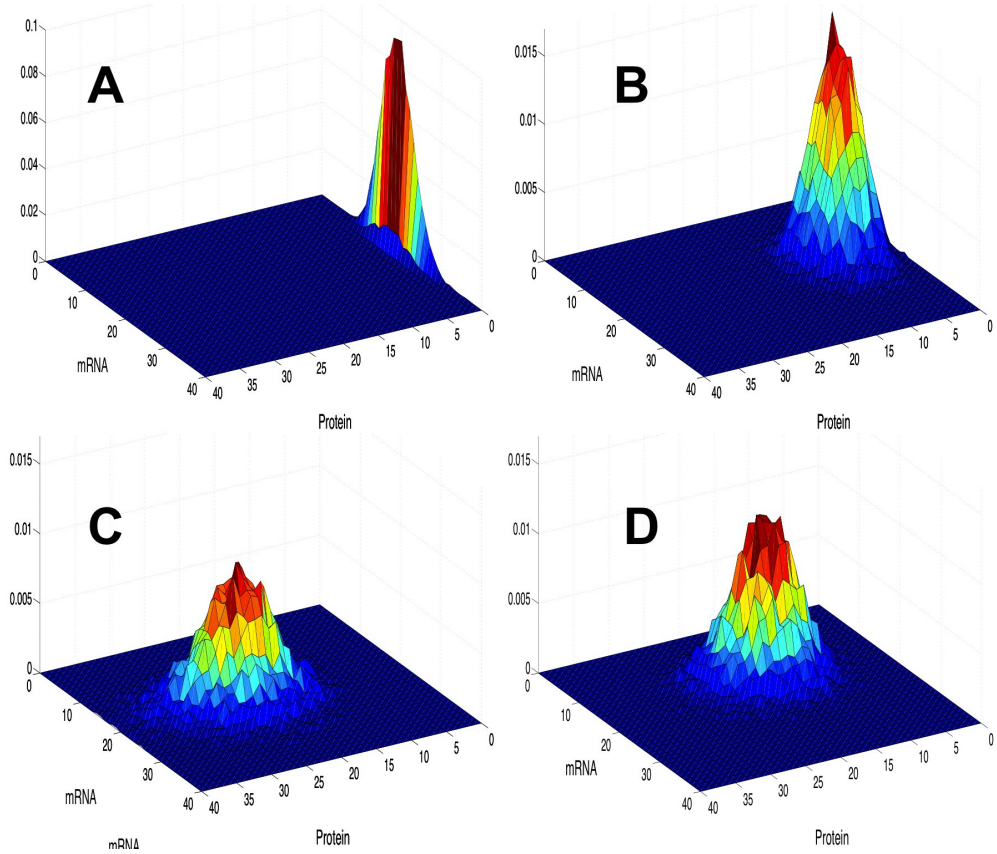

Figure 5.2: (A) The unperturbed system (5.7) plotted at $t = 13$ minutes. Notice that weak basal expression of protein produces a distribution that reflects high mRNA average copy number but low average protein copy number. Here $k_a = k_d = 0$. (B) The system simulated with ribosomal loading effects and a gene X that is in X^{ON} state with high probability ($k_a = .9$) Notice the significant reduction in protein expression when compared against C or D. (C) The system simulated with X in the off state ($k_d = .9$) with high probability and in the on state with low probability ($k_a = 1$). Consequently, the system has significantly higher concentrations of protein than in plots A or B. (D) The system (5.7) when X has one half probability of being on and one half probability of being off. All simulations were performed in MATLAB using the Gillespie Stochastic Simulation Algorithm. Common parameters used for all four simulations were $k_F = 0.6$ /s , $k_L = 0.4$/s, $k_{TX} = 0.052$ /s , $k_{TLF} = 0.4$ /s, $\delta_m = 1.4 \times 10^{-3}$ s , $\delta_p = .015$ s and $m(0), p(0) \sim$(Poiss(7)))

Stochastic switch with antibiotic attenuation

We now examine a particular approach for modeling the effect of antibiotics on a system. We suppose the system carries no resistance for two antibiotics and that these two antibiotics, when perturbing the system, can reduce the rate of transcription and translation respectively. The system is composed of an mRNA and a protein, whose expression is controlled by an upstream binary oscillator X.

We denote the two states of the binary oscillator as X^H and X^L. When in the high state, transcription and translation of m and p occurs using the same chemical reactions as in Example 1; however, we denote the high state propensity coefficient of transcription as k_{TXH} and the high state propensity coefficient of translation K_{TXL}. In the low state, the reaction structure is the same again, but this time with the low transcription and translation reaction propensity coefficients k_{TXL} and k_{TLL}. A diagram of the system is shown in Figure 5.3. Let $P_H(p, m, t)$ be the probability mass function for a system with $X = X^H$ and $P_L(p, m, t)$ be the probability mass function for a system with $X = X^L$. Notice that P_H (and P_L) can be obtained by a direct application of the solution generated using the method of probability generating functions from Example 1, evaluated with $k_{TX} = k_{TXH}$ and $k_{TL} = k_{TLH}$ ($k_{TX} = k_{TLH}$ and $k_{TL} = k_{TLL}$). Thus, we can calculate $P(p, m, t)$ as

$$\begin{aligned} P(p, m, t) &= P(p, m, t|X^H)P(X^H, t) + P(p, m, t|X^L)P(X^L, t) \\ &= P_H(p, m, t)P(X^H, t) + P_L(p, m, t)P(X^L, t) \end{aligned}$$

We suppose that $P(X, t) = P(X) = \frac{1}{2}$ is stationary and we model it as a Bernoulli random variable with parameter $p = .5$, i.e. the oscillator is unbiased. With these assumptions, we can calculate the solution of the system $P(p, m, x, t)$ and in particular, the marginal $P(p, m, t)$. Without any environmental disturbances, the distribution $P(p, m, t)$ has a bi-

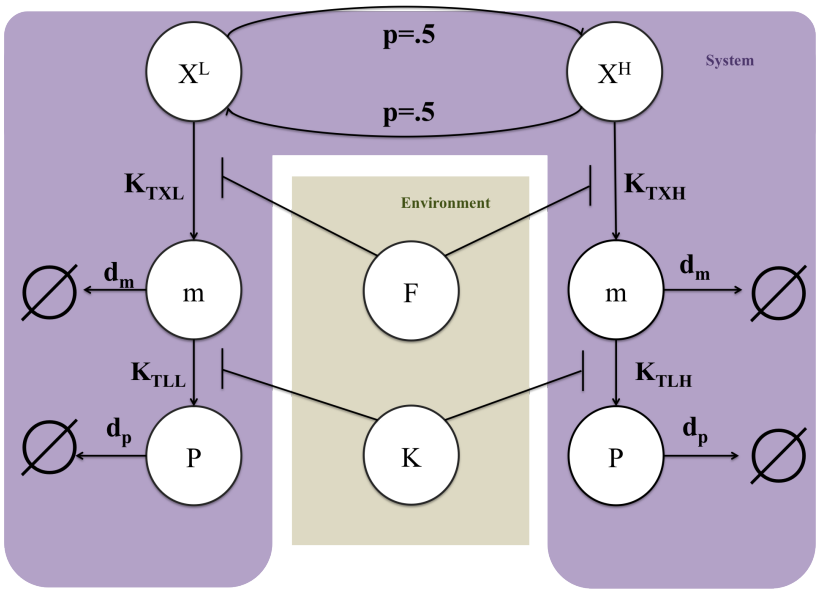

Figure 5.3: A schematic illustrating the reaction channels and interactions between chemical species in Example 2. F and K are antibiotics that attenuate transcription and translation rates respectively. X is a binary oscillator switching back and forth between a low and a high state with no transition bias. In the high state, m and p are produced with faster transcription and translation rates. At the low state, m and p are produced at much slower rates.

modal distribution, see Figure 5.4A. However, let us now introduce an environmental system to add disturbance to the dynamics of the system. In particular, we suppose there are two types of antibiotic added to the system. The first, which we denote as K, can be thought of as an antibiotic that disrupts ribosomal activity (e.g. kanamycin, streptomycin, chloromphenicol). The second, which we denote as F, can be viewed as an antibiotic that disrupts the transcription process (e.g. rafimycin). Accordingly, we suppose their effect on transcriptional and translation reactions has an overall negative effect. In particular, we suppose that their reactions are of the following form:

$$F \xrightarrow{-\Delta_{TX}F} m, \qquad K + m \xrightarrow{-\Delta_{TL}K} P \tag{5.9}$$

That is, regardless if $X = X^H$ or $X = X^L$, the antibiotic K decreases the rate of translation as a function of K while the antibiotic F slows the rate of transcription as a function of F. Let us assume that K and F have independent distributions to describe their copy number.

Besides this assumption, let us suppose that we do not know the distribution. We can write the environmental disturbance $f^E(P(X^e(t)|X^p(t))$ as

$$\begin{aligned}
&\sum_{x=0}^{\infty} -\Delta_{TX} x P(F = x|p, m, t) P(p, m-1, t) \\
&+ \sum_{y=0}^{\infty} -\Delta_{TL} y P(K = y|p, m, t) P(p, m-1, t) \\
&- \left(\sum_{x=0}^{\infty} -\Delta_{TX} x P(F = x|p, m, t) P(p, m, t) \right) \\
&- \left(\sum_{y=0}^{\infty} -\Delta_{TL} y P(K = y|p, m, t) P(p, m, t) \right)
\end{aligned}$$

In this scenario, we have an analytically tractable model for our system but no clear expression for the conditional distribution of the environment $P(X^E|p, m, t)$. Hence there is no way to compute or simulate $P(X^e|p, m, t)$. However, we can justify using a particular distribution by the principle of maximum entropy, which specifies the functional form of the distribution if there are constraints on the moments of $P(X^e|p, m, t)$. Certainly, we can assume that the mean value of K and F are both finite. If so, then from Theorem 5.7 in [14] we then can write

$$P(X^e = x|p, m, t) = Cr^x$$

where $C = \frac{1}{\mu_{X^e}}, r = \frac{\mu_{X^e}}{\mu_{X^e}+1}$ and $X^e = F$ or K. Further, if we suppose that μ_F and μ_K are given (or estimated using empirical measurements), then we get that

$$\begin{aligned}
f^E(\cdot) = &-\Delta_{TX} \mathbb{E}_{F|p,m,t}[F] P(p, m-1, t) \\
&- \Delta_{TL} \mathbb{E}_{K|p,m,t}[K] P(p, m-1, t) \\
&+ \Delta_{TX} \mathbb{E}_{F|p,m,t}[F] P(p, m, t) \\
&+ \Delta_{TL} \mathbb{E}_{K|p,m,t}[K] P(p, m, t)
\end{aligned}$$

Notice that this expression for f^E leads to a closed form solution of $P(X^p, t) = P(p, m, t)$ in this case. Writing down the expression for $P_H(p, m, t)$ and $P_L(p, m, t)$ the reader will see that f^E has the effect of perturbing the transcription and translation rates of the original system to be

$$k_{TX}^{pet} = k_{TX} - \Delta_{TX}\mathbb{E}_{F|p,m,t}\left[F\right],$$
$$k_{TL}^{pet} = k_{TL} - \Delta_{TL}\mathbb{E}_{K|p,m,t}\left[K\right],$$

where k_{TX} and k_{TL} can be replaced with k_{TXH}, k_{TXL}, k_{TLH}, k_{TLL} respectively to obtain $P_H(p, m, t)$ and $P_L(p, m, t)$ as a function of the perturbed rates. The final solution is then calculated as before, as

$$\begin{aligned} P(p, m, t) &= P(p, m, t|X^H)P(X^H, t) + P(p, m, t|X^L)P(X^L, t) \\ &= P_H(p, m, t)P(X^H, t) + P_L(p, m, t)P(X^L, t) \end{aligned}$$

In Figure 5.4, we plot the results of our simulation. When we perturb just with antibiotic F, the mean of the mRNA decreases while the mean of the protein remains approximately the same (the second peak remains and bimodality is not abolished). When we perturb the system with just K, there is a decrease in translation rates and bimodality disappears. Finally, when we perturb with F and K at the same time, both protein and mRNA copy number decrease as expected and we lose bimodality.

Our example thus illustrates a simple way of modeling the effects of antibiotics on transcription and translation. It does not require complete knowledge about the distribution of the antibiotics but it does require some estimate on the parameters for μ_F, μ_K, and Δ_{TX}, Δ_{TL}.

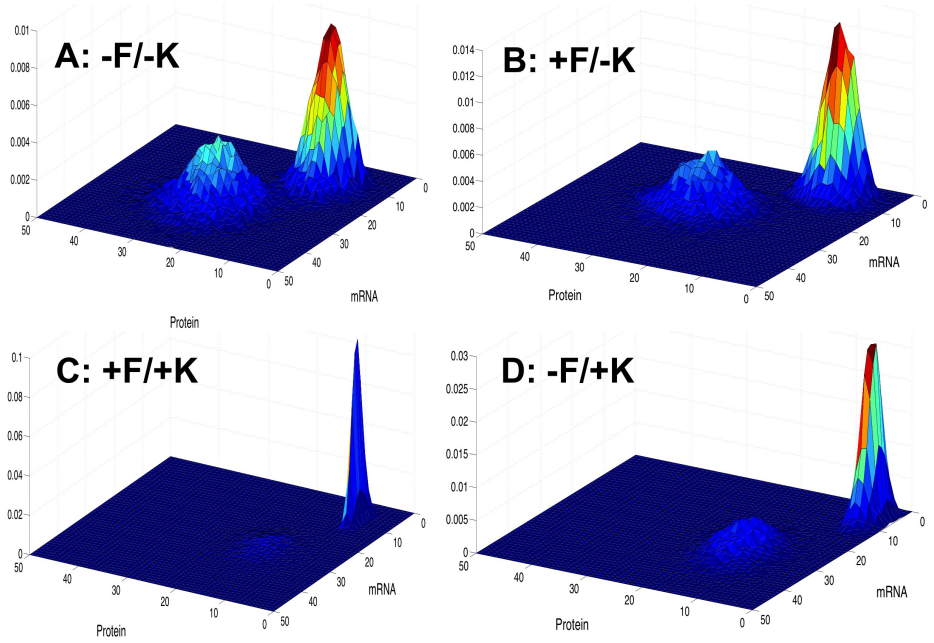

Figure 5.4: A: The unperturbed system is bimodal, with a large peak at low mRNA and low protein count and a smaller peak at high protein and high mRNA count. B: Antibiotic F reduces transcription, thus reducing average mRNA count. (compare to A) C: Antibiotics F and K reduce transcription and translation and abolish bimodality of the system. D: Antibiotic K reduces translation, abolishing bimodality but leaves a strong peak at lower mRNA copy numbers.

5.8 Conclusions

In this chapter we derived a decomposition of the chemical master equation into an additive sum of two terms: the first describes the dynamics of a system of interest, the second has the interpretation of the averaged environmental disturbance or more precisely, averaged propensity functions for all reactions involving environmental species. We illustrated the use of this decomposition to model two types of environmental effects: 1) the effect of ribosomal loading from an orthogonal gene with high (or low) demand for the ribosomes in a cell, 2) the effect of antibiotics on a bimodal system with unknown environmental distribution. We approximated the latter environmental effect by using a maximum entropy distribution to show that antibiotics directly perturb the transcription and translation rates of the system, scaled by the mean of the antibiotic copy number distribution.

Chapter 6

Future Work

Broadly speaking, my future research will center on two themes: 1) engineering complex synthetic biological networks and 2) robustness analysis of large-scale cyberphysical systems. The common thread in each of these research thrusts will be a focus on bringing to bear nonlinear and stochastic techniques of controller synthesis and closed-loop analysis to engineer robust networks.

As a next step in engineering more complex synthetic biological systems, I would like to explore approaches that involve distributed control. Distributed control systems benefit from the reduced power-to-load ratios per device. Natural biological systems already leverage this paradigm in microbial consortia. My goal here will be to engineer complex heterogeneous communities of microbes that perform diversified functions.

Successful distributed control strategies requires device modularity and communication standards. Thus, a supporting research direction will be development of standardized biological devices, parts, models of parts, and communication protocols to interconnect devices. The challenge here is that device abstraction is not well-defined. As this thesis has shown, there is context interference at multiple levels of biocircuit design. Developing appropriate models and conceptual schemes for abstracting synthetic biological devices will be an important research challenge.

Context interference, or more broadly, system interdependencies enforced by hidden or unmeasured states are a defining aspect in modern critical infrastructure and cyberphysical

systems. With the digital and information revolution, our world is becoming increasingly interconnected and interdependent. What are best practices, both in terms of control and robust design, for engineering interconnected large-scale complex networks?

Two key research questions arise in addressing this question, model reduction and model representation. As we have shown in our thesis, linear systems provide abstracted representations for network structure through dynamical structure functions. However, the critical infrastructure networks that shape the energy and economic landscape of our world interface with social networks and varying degrees of systemic uncertainty. Therefore, developing reduced order models and structural representations for nonlinear stochastic networks are research directions which I plan to pursue. Further, I would like to understand how to use reduced-order models to gain insight into what features of comprising networks give rise to self-organized criticality and system fragility.

Bibliography

[1] J. Adebayo, et al. (2012). 'Dynamical structure function identifiability conditions enabling signal structure reconstruction'. In *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pp. 4635–4641. IEEE.

[2] J. R. Babendure, et al. (2003). 'Aptamers switch on fluorescence of triphenylmethane dyes'. *Journal of the American Chemical Society* **125**(48):14716–14717.

[3] V. L. Balke & J. Gralla (1987). 'Changes in the linking number of supercoiled DNA accompany growth transitions in Escherichia coli.'. *Journal of Bacteriology* **169**(10):4499–4506.

[4] J. M. Berg, et al. (2002). *Biochemistry*. Freeman and Company: New York.

[5] L. Bintu, et al. (2005). 'Transcriptional regulation by the numbers: models'. *Current Opinion in Genetics & Development* **15**(2):116–124.

[6] R. E. Bishop, et al. (1998). 'The entericidin locus of Escherichia coli and its implications for programmed bacterial cell death'. *Journal of Molecular Biology* **280**(4):583 – 596.

[7] H. Bremer & P. P. Dennis (1996). *Modulation of chemical composition and other parameters of the cell by growth rate.*, chap. 97. Springer.

[8] H. Buc & W. R. McClure (1985). 'Kinetics of open complex formation between Escherichia coli RNA polymerase and the lac UV5 promoter. Evidence for a sequential mechanism involving three steps'. *Biochemistry* **24**(11):2712–2723.

[9] S. Cardinale & A. Arkin (2012). 'Contextualizing context for synthetic biology–identifying causes of failure of synthetic biological systems'. *Biotechnology Journal* .

[10] F. Ceroni, et al. (2015). 'Quantifying cellular capacity identifies gene expression designs with reduced burden'. *Nature methods* **12**(5):415–418.

[11] F. Chandra., et al. (2011). 'Glycolytic Oscillations and Limits on Robust Efficiency'. *Science* **333**(6039):187–192.

[12] J. Chappell, et al. (2013). 'Validation of an entirely in vitro approach for rapid prototyping of DNA regulatory elements for synthetic biology'. *Nucleic Acids Research* **41**(5):3471–3481.

[13] S. Chong, et al. (2014). 'Mechanism of transcriptional bursting in bacteria'. *Cell* **158**(2):314–326.

[14] K. Conrad (2004). 'Probability distributions and maximum entropy'. *Expository Paper* **6**.

[15] R. S. Cox, et al. (2007). 'Programming gene expression with combinatorial promoters'. *Molecular Systems Biology* **3**(1):145.

[16] J. H. Davis, et al. (2011). 'Design, construction and characterization of a set of insulated bacterial promoters'. *Nucleic Acids Research* **39**(3):1131–1141.

[17] D. Del Vecchio, et al. (2008). 'Modular cell biology: retroactivity and insulation'. *Molecular Systems Biology* **4**(1):161.

[18] P. Deuflhard, et al. (2008). 'Adaptive discrete Galerkin methods applied to the chemical master equation'. *SIAM Journal on Scientific Computing* **30**(6):2990–3011.

[19] M. Drolet (2006). 'Growth inhibition mediated by excess negative supercoiling: the interplay between transcription elongation, R-loop formation and DNA topology'. *Molecular Microbiology* **59**(3):723–730.

[20] A. D. Edelstein, et al. (2014). 'Advanced methods of microscope control using μManager software'. *Journal of Biological Methods* **1**(2).

[21] M. B. Elowitz & S. Leibler (2000). 'A synthetic oscillatory network of transcriptional regulators'. *Nature* **403**(6767):335–338.

[22] M. B. Elowitz, et al. (2002). 'Stochastic Gene Expression in a Single Cell'. *Science* **297**(5584):1183–1186.

[23] S. Engblom (2009). 'Spectral approximation of solutions to the chemical master equation'. *Journal of computational and applied mathematics* **229**(1):208–221.

[24] C. Engler, et al. (2009). 'Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes'. *PloS one* **4**(5):e5553.

[25] C. Engler, et al. (2008). 'A one pot, one step, precision cloning method with high throughput capability'. *PloS ONE* **3**(11):e3647.

[26] L. Q. et al (2012). 'RNA processing enables predictable programming of gene expression'. *Nature Biotechnology* **30**(10):1002–1006.

[27] N. C. et al (2011). 'Queueing up for enzymatic processing: correlated signaling through coupled degradation'. *Molecular Systems Biology* **7**(561).

[28] G. S. Filonov, et al. (2015). 'In-gel imaging of RNA processing using Broccoli reveals optimal aptamer expression strategies'. *Chemistry & Biology* **22**(5):649–660.

[29] Å. Fredriksson, et al. (2007). 'Decline in ribosomal fidelity contributes to the accumulation and stabilization of the master stress response regulator σS upon carbon starvation'. *Genes & development* **21**(7):862–874.

[30] A. E. Friedland, et al. (2009). 'Synthetic gene networks that count'. *science* **324**(5931):1199–1202.

[31] T. S. Gardner, et al. (2000). 'Construction of a genetic toggle switch in Escherichia coli'. *Letters to Nature* **403**.

[32] A. J. Gates & L. M. Rocha (2015). 'Control of complex networks requires both structure and dynamics'. *arXiv preprint arXiv:1509.08409* .

[33] D. G. Gibson, et al. (2010). 'Creation of a bacterial cell controlled by a chemically synthesized genome'. *science* **329**(5987):52–56.

[34] D. G. Gibson, et al. (2009). 'Enzymatic assembly of DNA molecules up to several hundred kilobases'. *Nature Methods* **6**(5):343–345.

[35] D. T. Gillespie (1992). 'A rigorous derivation of the chemical master equation'. *Physica A: Statistical Mechanics and its Applications* **188**(13):404 – 425.

[36] L. Goentoro & M. W. Kirschner (2009). 'Evidence that fold-change, and not absolute level, of β-catenin dictates Wnt signaling'. *Molecular cell* **36**(5):872–884.

[37] J. Gonçalves & S. Warnick (2008). 'Necessary and sufficient conditions for dynamical structure reconstruction of LTI networks'. *Automatic Control, IEEE Transactions on* **53**(7):1670–1674.

[38] J. Gonçalves, et al. (2007). 'Dynamical Structure Functions for the Reverse Engineering of LTI Networks'. *IEEE Transactions of Automatic Control, 2007* .

[39] A. Gyorgy, et al. (2015). 'Isocost lines describe the cellular economy of genetic circuits'. *Biophysical Journal* **109**(3):639–646.

[40] K. Y. Han, et al. (2013). 'Understanding the Photophysics of the Spinach–DFHBI RNA Aptamer–Fluorogen Complex To Improve Live-Cell RNA Imaging'. *Journal of the American Chemical Society* **135**(50):19033–19038.

[41] T. A. Henzinger, et al. (2009). 'Sliding window abstraction for infinite Markov chains'. In *In Proc. CAV, volume 5643 of LNCS*, pp. 337–352. Springer.

[42] A. Hilfinger & J. Paulsson (2011). 'Separating intrinsic from extrinsic fluctuations in dynamic biological systems'. *Proceedings of the National Academy of Sciences* **108**(29):12167–12172.

[43] V. Hsiao, et al. (2015). 'A population-based temporal logic gate for timing and recording chemical events'. *bioRxiv* p. 029967.

[44] K. Z. J. Doyle, K. Glover (1996). *Robust and Optimal Control*. Prentice Hall, Englewood Cliffs, N.J.

[45] S. Jayanthi & D. D. Vecchio (2012). 'Tuning Genetic Clocks Employing DNA Binding Sites'. *PLoS ONE* **7**(7):e41019.

[46] T. Kalisky, et al. (2007). 'Cost–benefit theory and optimal design of gene regulation functions'. *Physical Biology* **4**(4):229.

[47] N. V. Kampen (2007). *Stochastic processes in physics and chemistry*. North Holland.

[48] J. Kim & R. M. Murray (2011). 'Analysis and design of a synthetic transcriptional network for exact adaptation'. In *Biomedical Circuits and Systems Conference (BioCAS), 2011 IEEE*, pp. 345–348. IEEE.

[49] J. Kim & E. Winfree (2011). 'Synthetic in vitro transcriptional oscillators'. *Molecular systems biology* **7**(1):465.

[50] H. Kobayashi, et al. (2004). 'Programmable cells: interfacing natural and engineered gene networks'. *Proceedings of the National Academy of Sciences of the United States of America* **101**(22):8414–8419.

[51] J. O. Korbel, et al. (2004). 'Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs'. *Nature Biotechnology* **22**(7):911–917.

[52] S. Kosuri, et al. (2013). 'Composability of regulatory sequences controlling transcription and translation in Escherichia coli'. *Proceedings of the National Academy of Sciences* **110**(34):14024–14029.

[53] M. H. Larson, et al. (2008). 'Applied force reveals mechanistic and energetic details of transcription termination'. *Cell* **132**(6):971–982.

[54] M. E. Lee, et al. (2015). 'A highly characterized yeast toolkit for modular, multipart assembly'. *ACS Synthetic Biology* .

[55] T. S. Lee, et al. (2011). 'BglBrick vectors and datasheets: a synthetic biology platform for gene expression'. *Journal of Biological Engineering* **5**(1):1–14.

[56] I. Lestas, et al. (Jan.). 'Noise in Gene Regulatory Networks'. *Automatic Control, IEEE Transactions on* **53**(Special Issue):189–200.

[57] L. F. Liu & J. C. Wang (1987). 'Supercoiling of the DNA template during transcription'. *Proceedings of the National Academy of Sciences* **84**(20):7024–7027.

[58] L. Ljung (1999). *System Identification–Theory for the User*. Prentice Hall.

[59] P. C. Loewen, et al. (1998). 'Regulation of the rpoS regulon of Escherichia coli'. *Canadian Journal of Microbiology* **44**:707–717.

[60] R. Lutz & H. Bujard (1997). 'Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements'. *Nucleic Acids Research* **25**(6):1203–1210.

[61] T. Maier, et al. (2011). 'Quantification of mRNA and protein and integration with protein turnover in a bacterium'. *Molecular Systems Biology* **7**(1):511.

[62] S. Mangan & U. Alon (2003). 'Structure and function of the feed-forward loop network motif'. *Proceedings of the National Academy of Sciences* **100**(21):11980–11985.

[63] S. Meyer, et al. (2014). 'Torsion-Mediated Interaction between Adjacent Genes'. *PLoS Comput Biol* **10**(9):e1003785.

[64] A. Milias-Argeitis, et al. (2011). 'In silico feedback for in vivo regulation of a gene expression circuit'. *Nature Biotechnology* .

[65] R. Milo, et al. (2010). 'BioNumbersthe database of key numbers in molecular and cell biology'. *Nucleic acids research* **38**(suppl 1):D750–D753.

[66] D. Mishra, et al. (2014). 'A load driver device for engineering modularity in biological networks'. *Nature Biotechnology* **32**(12):1268–1275.

[67] T. S. Moon, et al. (2012). 'Genetic programs constructed from layered logic gates in single cells'. *Nature* **491**(7423):249–253.

[68] B. Munsky & M. Khammash (2006). 'The finite state projection algorithm for the solution of the chemical master equation'. *The Journal of Chemical Physics* **124**:044104.

[69] V. K. Mutalik, et al. (2013a). 'Precise and reliable gene expression via standard transcription and translation initiation elements'. *Nature Methods* **10**(4):354–360.

[70] V. K. Mutalik, et al. (2013b). 'Quantitative estimation of activity and quality for collections of functional genetic elements'. *Nature methods* **10**(4):347–353.

[71] H. C. Nelson & R. T. Sauer (1985). 'Lambda repressor mutations that increase the affinity and specificity of operator binding'. *Cell* **42**(2):549–558.

[72] M. Nip, et al. (2012). 'A spectral methods-based solution of the Chemical Master Equation for gene regulatory networks'. In *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pp. 5354–5360. IEEE.

[73] V. Noireaux, et al. (2003). 'Principles of cell-free genetic circuit assembly'. *Proceedings of the National Academy of Sciences* **100**(22):12672–12677.

[74] M. L. Opel & G. Hatfield (2001). 'DNA supercoiling-dependent transcriptional coupling between the divergently transcribed promoters of the ilvYC operon of Escherichia coli is proportional to promoter strengths and transcript lengths'. *Molecular Microbiology* **39**(1):191–198.

[75] Z.-A. Ouafa, et al. (2012). 'The nucleoid-associated proteins H-NS and FIS modulate the DNA supercoiling response of the pel genes, the major virulence factors in the plant pathogen bacterium Dickeya dadantii'. *Nucleic Acids Research* **40**(10):4306–4319.

[76] J. S. Paige, et al. (2011). 'RNA mimics of green fluorescent protein'. *Science* **333**(6042):642–646.

[77] L. Perko (2013). *Differential equations and dynamical systems*, vol. 7. Springer Science & Business Media.

[78] A. R. Rahmouni & R. D. Wells (1992). 'Direct evidence for the effect of transcription on local DNA supercoiling *in vivo*'. *Journal of Molecular Biology* **223**(1):131–144.

[79] K. Y. Rhee, et al. (1999). 'Transcriptional coupling between the divergent promoters of a prototypic LysR-type regulatory system, the ilvYC operon of Escherichia coli'. *Proceedings of the National Academy of Sciences* **96**(25):14294–14299.

[80] D. R. Rigney (1979). 'Stochastic model of constitutive protein levels in growing and dividing bacterial cells'. *Journal of Theoretical Biology* **76**(4):453 – 480.

[81] M. A. Savageau (2001). 'Design principles for elementary gene circuits: Elements, methods, and examples'. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **11**(1):142–159.

[82] M. Scott, et al. (2010). 'Interdependence of Cell Growth and Gene Expression: Origins and Consequences'. *Science* **330**(6007):1099–1102.

[83] K. E. Shearwin, et al. (2005). 'Transcriptional interference–a crash course'. *TRENDS in Genetics* **21**(6):339–345.

[84] J. Shin & V. Noireaux (2012). 'An E. coli cell-free expression toolbox: application to synthetic gene circuits and artificial cells'. *ACS Synthetic Biology* **1**(1):29–41.

[85] D. Siegal-Gaskins, et al. (2013). 'Biomolecular resource utilization in elementary cell-free gene circuits'. *The Proceedings of the IEEE American Control Conference, to appear* .

[86] D. Siegal-Gaskins, et al. (2014). 'Gene circuit performance characterization and resource usage in a cell-free breadboard'. *ACS Synthetic Biology* **3**(6):416–425.

[87] M. J. Smanski, et al. (2014). 'Functional optimization of gene clusters by combinatorial design and assembly'. *Nature biotechnology* .

[88] L. Sommerlade, et al. (2009). 'Estimating causal dependencies in networks of nonlinear stochastic dynamical systems'. *Physical Review E* **80**(5):051128.

[89] B. C. Stanton, et al. (2014). 'Genomic mining of prokaryotic repressors for orthogonal logic gates'. *Nature Chemical Biology* **10**(2):99–105.

[90] J. Stricker, et al. (2008). 'A fast, robust and tunable synthetic gene oscillator'. *Nature* **456**(7221):516–519.

[91] Z. Z. Sun, et al. (2013). 'Linear DNA for rapid prototyping of synthetic biological circuits in an Escherichia coli based TX-TL cell-free system'. *ACS Synthetic Biology* .

[92] P. S. Swain, et al. (2002). 'Intrinsic and extrinsic contributions to stochasticity in gene expression'. *Proceedings of the National Academy of Sciences* **99**(20):12795–12800.

[93] J. J. Tabor, et al. (2011). 'Multichromatic Control of Gene Expression in Escherichia coli'. *Journal of Molecular Biology* **405**(2):315 – 324.

[94] Z. Tuza, et al. (2013). 'An *in silico* modeling toolbox for rapid prototyping of circuits in a biomolecular "breadboard system'. In *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, pp. 1404–1410. IEEE.

[95] J. J. Tyson, et al. (2003). 'Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell'. *Current opinion in cell biology* **15**(2):221–231.

[96] D. D. Vecchio, et al. (2008). 'Modular cell biology : retroactivity and insulation'. *Molecular Systems Biology* **4**:161.

[97] J.-W. Veening, et al. (2004). 'Visualization of differential gene expression by improved cyan fluorescent protein and yellow fluorescent protein production in Bacillus subtilis'. *Applied and environmental microbiology* **70**(11):6809–6815.

[98] E. Weber, et al. (2011). 'A Modular cloning system for standardized assembly of multigene constructs'. *PLoS ONE* **6**(2).

[99] D. M. Wolf & A. P. Arkin (2003). 'Motifs, modules and games in bacteria'. *Current opinion in microbiology* **6**(2):125–134.

[100] X. S. Xie, et al. (2008). 'Single-molecule approach to molecular biology in living bacterial cells'. *Annual Reviews of Biophysics* **37**:417–444.

[101] E. Yeung, et al. (2013). 'Resource competition as a source of non-minimum phase behavior in transcription-translation systems'. In *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, pp. 4060–4067. IEEE.

[102] E. Yeung, et al. (2012a). 'Quantifying crosstalk in biochemical systems'. In *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pp. 5528–5535. IEEE.

[103] E. Yeung, et al. (2012b). 'Quantifying Crosstalk in Biochemical Systems'. *The Proceedings of the 49th IEEE Conference on Decision and Control* .

[104] J. W. Young, et al. (2012). 'Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy'. *Nature Protocols* **7**(1):80–88.

[105] Y. Yuan, et al. (2011). 'Robust dynamical network structure reconstruction'. *Automatica* **47**(6):1230–1235.

[106] J. Zhang, et al. (2002). 'Creating new fluorescent probes for cell biology'. *Nature Reviews Molecular Cell Biology* **3**(12):906–918.

[107] K. Zhou, et al. (1996). *Robust and Optimal Control*. Prentice Hall, N. J.