Learning from Disorder and Noise in Physical Biology

Taylor Emil Firman
University of Denver

Follow this and additional works at: https://digitalcommons.du.edu/etd
Part of the Biological and Chemical Physics Commons, and the Biophysics Commons

Recommended Citation
https://digitalcommons.du.edu/etd/1465

This Dissertation is brought to you for free and open access by the Graduate Studies at Digital Commons @ DU. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Digital Commons @ DU. For more information, please contact jennifer.cox@du.edu,dig-commons@du.edu.
Learning from disorder and noise in
Physical Biology

A Dissertation
Presented to
the Faculty of Natural Sciences and Mathematics
University of Denver

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

by
Taylor Firman
June 2018
Advisor: Dr. Kingshuk Ghosh
Stochasticity, disorder, and noise play crucial roles in the functioning of many biological systems over many different length scales. On the molecular scale, most proteins are envisioned as pristinely folded structures, but intrinsically disordered proteins (IDPs) have no such folded state and still serve distinct purposes within the cell. At the scale of gene regulation, realistic in vivo conditions produce stochastic fluctuations in gene expression that can lead to advantageous bet-hedging strategies, but can be difficult to characterize using a deterministic framework. Even at the organismal scale, germband extension (GBE) in Drosophila melanogaster embryos systematically elongates the epithelial tissue using cell intercalation, but leaves cells in highly heterogeneous geometries. Throughout this work, we will demonstrate that these characteristics are not just arbitrary artifacts to be glossed over, but are actually very intentional frameworks that are harnessed by the respective systems to their own advantage.

In some cases, they can also be harnessed by researchers to better characterize or even control the system through various biophysical techniques. In the case of IDPs, we will introduce an analytical model that can predict the conformational size of these disordered proteins and identify specific “hot spots” in their sequences that hold significant influence over the shape (and therefore function) of each protein. To better understand stochastic gene expression, the power of stochastic inference methods will be put on display, specifically methods using modeling systems based on the principle of Maximum Caliber. These require no direct knowledge about the archi-
architecture of the underlying genetic network and make quantitative predictions using the entire content of experimentally realistic time-series data. Finally, we will break down the process of GBE using node-based Monte Carlo simulations to show that while anisotropic tension is enough to qualitatively reproduce convergent extension, competing active extension mechanisms must be introduced in each cell to achieve heterogeneous cell configurations and quantitative agreement with experiment. These studies will collectively demonstrate that randomness and fluctuation do not always imply disarray and intractibility, but instead can convey adaptability and possibility. As such, these characteristics should be embraced by the field of biophysics.
Acknowledgements

First and foremost, I would like to thank my advisor, Dr. Kingshuk Ghosh, for his guidance throughout my graduate school career. Thank you for teaching me, thank you for arguing with me, and thank you for your patience throughout these years. To my labmates, Dr. Lucas Sawle, Jon Huihui, and Stephen Wedekind, as well as all of the students in the Molecular and Cellular Biophysics program at DU, thank you for your camaraderie and for making me a better scientist. To our collaborators here at the University of Denver, Dr. Dinah Loerke, Dr. Todd Blankenship, and both of their respective labs, thank you for being so formative in my time her and for enriching our research together. I would also like to thank our collaborators outside of DU: Dr. Gábor Balázi, Dr. Brian Munsky, and Dr. Steve Pressé for their invaluable contributions to our MaxCal research; Dr. Adam de Graff, Dr. David Minde, Dr. Geoffrey Lynn, Dr. Alex Holehouse, and Dr. Andreas Vitalis for their insight into IDPs; Dr. Jeff Cameron, Dr. Kristin Moore, and Dr. Sean Shaheen for guiding the application of MaxCal to polyploidy. Thank you to Ben Fotovich at the High Performance Computing facility of DU for his extensive help in computation.

To Alan Thorndike, thank you for inspiring me to pursue science and for teaching me to look for the eigenvectors in life rather than the coordinates. To my partner, Ashley, you are my rock and inspiration. Thank you for putting up with my antics over the course of this experience and thank you for being the amazing woman that you are. To my parents, Rich and Denise, and my siblings, Katie, Sean, Jake, Angela, and Joel, I love you all. It’s been a long six years and I’m not entirely sure how I got lucky enough to be here, but it has a lot to do with you guys. Thanks for standing by me. And to the next generation of Firmans, my nieces and nephews, thank you for brightening our lives and here’s hoping one of you becomes a scientist.
# Table of Contents

Acknowledgements ........................................ iv
List of Tables ............................................. viii
List of Figures ............................................ ix

I Introduction ........................................... 1

II Disorder on the Molecular Level ..................... 10
   1 Size Prediction of Disordered Proteins ............... 11
      1.1 Materials and Methods ............................ 15
      1.1.1 High-Throughput Model for IDPs ................. 15
      1.1.2 CAMPARI Simulation Procedure ................. 18
      1.2 Results ........................................... 19
      1.2.1 Sequence-specific theory is consistent with all-atom simulations of IDP sequences .......... 19
      1.2.2 DisProt database has a heterogeneous size distribution .... 20
      1.2.3 Synthetic sequences yield similar trends .......... 24
      1.2.4 Phosphorylation can have “hot spots” for modulating conformation ............................. 27
      1.3 Conclusion .......................................... 31

III Stochasticity in Genetic and Biochemical Networks 33
   2 Competition in Complexation Reactions .............. 34
      2.1 Materials and Methods ............................. 37
      2.1.1 Stochastic model for two competing reactions ........ 37
      2.1.2 Exact equilibrium distribution ...................... 40
      2.1.3 Equations of motion for different moments ........... 40
      2.2 Results and Discussion ............................ 42
      2.2.1 Quantifying equilibrium fluctuations and correlations ........ 42
      2.2.2 Determination of dynamical quantities ............. 54
<table>
<thead>
<tr>
<th>Section</th>
<th>Content</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3</td>
<td>Conclusion</td>
<td>56</td>
</tr>
<tr>
<td>3</td>
<td>Maximum Caliber Applied to Auto-Activation Networks</td>
<td>59</td>
</tr>
<tr>
<td>3.1</td>
<td>Materials and Methods</td>
<td>61</td>
</tr>
<tr>
<td>3.1.1</td>
<td>Generating synthetic data for an auto-activating circuit</td>
<td>61</td>
</tr>
<tr>
<td>3.1.2</td>
<td>Maximum Caliber model for auto-activating circuit</td>
<td>63</td>
</tr>
<tr>
<td>3.1.3</td>
<td>Parameter estimation via maximum likelihood</td>
<td>66</td>
</tr>
<tr>
<td>3.1.4</td>
<td>Dealing with experimental data</td>
<td>68</td>
</tr>
<tr>
<td>3.2</td>
<td>Results and Discussion</td>
<td>70</td>
</tr>
<tr>
<td>3.2.1</td>
<td>MaxCal accurately infers underlying rate parameters</td>
<td>70</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Distributions predicted from MaxCal agree well with data</td>
<td>71</td>
</tr>
<tr>
<td>3.2.3</td>
<td>MaxCal provides an effective feedback parameter for the circuit</td>
<td>74</td>
</tr>
<tr>
<td>3.2.4</td>
<td>MaxCal can be applied when dealing with noisy fluorescence trajectories</td>
<td>76</td>
</tr>
<tr>
<td>3.3</td>
<td>Conclusion</td>
<td>76</td>
</tr>
<tr>
<td>4</td>
<td>Maximum Caliber Applied to Toggle Switch Networks</td>
<td>79</td>
</tr>
<tr>
<td>4.1</td>
<td>Materials and Methods</td>
<td>80</td>
</tr>
<tr>
<td>4.1.1</td>
<td>Generating synthetic data for TS</td>
<td>80</td>
</tr>
<tr>
<td>4.1.2</td>
<td>MaxCal model for TS</td>
<td>82</td>
</tr>
<tr>
<td>4.1.3</td>
<td>Estimating parameters from stochastic trajectories in M1</td>
<td>85</td>
</tr>
<tr>
<td>4.1.4</td>
<td>Estimating parameters from stochastic trajectories in M3</td>
<td>86</td>
</tr>
<tr>
<td>4.1.5</td>
<td>Creating and analyzing synthetic fluorescence trajectories for M2 and M4</td>
<td>87</td>
</tr>
<tr>
<td>4.2</td>
<td>Results and Discussion</td>
<td>89</td>
</tr>
<tr>
<td>4.2.1</td>
<td>MaxCal accurately infers underlying rates and observables for TS</td>
<td>89</td>
</tr>
<tr>
<td>4.2.2</td>
<td>MaxCal accurately infers underlying rates and observables for TS using fluorescence trajectories of both proteins (M2)</td>
<td>92</td>
</tr>
<tr>
<td>4.2.3</td>
<td>MaxCal accurately infers underlying rates and observables for TS using only one protein trajectory (M3)</td>
<td>93</td>
</tr>
<tr>
<td>4.2.4</td>
<td>MaxCal accurately infers underlying rates and observables for TS using only one fluorescence trajectory (M4)</td>
<td>95</td>
</tr>
<tr>
<td>4.3</td>
<td>Conclusion</td>
<td>97</td>
</tr>
<tr>
<td>5</td>
<td>Maximum Caliber Applied to Repressilator Networks</td>
<td>99</td>
</tr>
<tr>
<td>5.1</td>
<td>Materials and Methods</td>
<td>100</td>
</tr>
<tr>
<td>5.1.1</td>
<td>Gillespie Reaction Network</td>
<td>100</td>
</tr>
<tr>
<td>5.1.2</td>
<td>MaxCal Model Descriptions</td>
<td>101</td>
</tr>
<tr>
<td>5.1.3</td>
<td>Predictions and Metrics Used to Assess MaxCal Representations</td>
<td>108</td>
</tr>
<tr>
<td>5.1.4</td>
<td>Inclusion of Fluorescence</td>
<td>113</td>
</tr>
<tr>
<td>5.2</td>
<td>Results and Discussion</td>
<td>114</td>
</tr>
<tr>
<td>5.2.1</td>
<td>MaxCal Models Using Protein Number</td>
<td>114</td>
</tr>
</tbody>
</table>
IV Noise on a Multicellular Level

6 Germband Extension Modeling

6.1 Materials and Methods

6.1.1 Vertex-based model

6.1.2 Monte-Carlo-like energy minimization protocol

6.1.3 Imaging protocol and grid initialization

6.1.4 Metrics for comparison of model to experiment

6.2 Results and Discussion

6.2.1 ‘Passive’ line tension model fails to reproduce data during extension.

6.2.2 Softer conditions during GBE do not resolve discrepancies between model and experiment.

6.2.3 Presence of ‘active’ extension mechanism is a plausible explanation based on experimental observations.

6.3 Conclusion

V Concluding Remarks

Bibliography

Appendices

A Derivation of Renormalized Kuhn Length

B Simple Approximation for \( \langle m_1 \rangle \) and \( \langle (\Delta m_1)^2 \rangle \)

C Application of Finite State Projection to Maximum Caliber

D Application of MaxCal to Alternate Auto-Activation Circuits

E Detailed Protocols for Cell Grid Simulations
List of Tables

1.1 Parameters best representing coil-globule transition of all-atom simulations. ................................................. 21
1.2 Sequences of wt and phosphorylated species of P0A8H9. .......... 28
3.1 Comparison of true rates and predicted rates using MaxCal. ........ 71
3.2 Effective rates from fluorescence trajectories. ........................ 77
4.1 Comparison of true and predicted rates/metrics from MaxCal using both trajectories (M1 & M2). ............................. 90
4.2 Comparison of true and predicted rates/metrics from MaxCal using both trajectories (M3 & M4). ............................. 94
5.1 Reaction rates used to generate synthetic trajectories. ............... 102
5.2 Comparison of experimental and MaxCal predicted rates and statistics from protein number based models. ...................... 117
5.3 Comparison of experimental and MaxCal predicted rates and statistics from fluorescence based models. ...................... 120
D.1 Comparison of true rates and predicted rates using MaxCal on alternate self-promotion model. ......................... 179
E.1 Parameter values used in cell grid simulations. ....................... 182
List of Figures

1.1 Theory recapitulates temperature-dependent coil-globule transitions of all-atom simulations. .................................................. 20
1.2 Distribution of protein size, quantified by $x$ and $\bar{x}$, for IDPs in DisProt. 22
1.3 Predicted heat maps of protein conformations in DisProt. .............. 24
1.4 Predicted heat maps of protein conformations in DisProt accounting for parameter variation. ............................................. 24
1.5 Predicted heat maps of synthetic sequence conformations. .............. 26
1.6 Predicted heat maps of synthetic sequence conformations accounting for parameter variation. ............................................. 26
1.7 Theoretical prediction agree well with all-atom simulation of wt and phosphorylated species of P0A8H9. ................................. 29

2.1 Fanofactor phase diagrams for different complexation conditions. .... 43
2.2 Probability distribution comparison under competitive and non-competitive conditions. .................................................. 46
2.3 Phase diagrams of variance differences under competitive and non-competitive conditions. .................................................. 47
2.4 Nonzero relative noise for fixed $M$ and relatively infinite $N_1$, $N_2$. .... 48
2.5 Nonzero relative noise for fixed $N_1$ and relatively infinite $M$, $N_2$. .... 49
2.6 Higher competition levels amplifies negative correlation. .................. 51
2.7 Inferred equilibrium constants can vary widely in competitive conditions. 53
2.8 Time evolution comparison between prediction and simulation. ....... 54

3.1 Positive feedback circuit. ................................................... 63
3.2 Predicted distributions agree well with the “true” distributions. ........ 72
3.3 MaxCal can capture distributions under varying degrees of feedback. 75
3.4 Predicted distributions from fluorescence trajectories with 200% noise using PFNC. ....................................................... 77

4.1 Predicted distributions agree well with the “true” distributions for M2. 92
4.2 Predicted distributions agree well with the “true” distributions for M2. 93
4.3 Predicted distributions agree well with the “true” distributions for M3. 95
4.4 Predicted distributions agree well with the “true” distributions for M4. 96
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>MaxCal reproduces basic characteristics of repressilator circuit.</td>
</tr>
<tr>
<td>5.2</td>
<td>Primary behaviors of Model 1 agree well with simulation.</td>
</tr>
<tr>
<td>5.3</td>
<td>Three-dimensional behaviors of protein number based MaxCal models agree well with simulation.</td>
</tr>
<tr>
<td>5.4</td>
<td>Primary behaviors of Model 2 agree well with simulation.</td>
</tr>
<tr>
<td>5.5</td>
<td>Primary behaviors of Model 3 agree well with simulation.</td>
</tr>
<tr>
<td>5.6</td>
<td>Primary behaviors of Model 4 agree well with simulation.</td>
</tr>
<tr>
<td>5.7</td>
<td>Primary behaviors of Model 5 agree well with simulation.</td>
</tr>
<tr>
<td>5.8</td>
<td>Three-dimensional behaviors of fluorescence based MaxCal models agree well with simulation.</td>
</tr>
<tr>
<td>5.9</td>
<td>Primary behaviors of Model 6 agree well with simulation.</td>
</tr>
<tr>
<td>5.10</td>
<td>Primary behaviors of Model 7 agree well with simulation.</td>
</tr>
<tr>
<td>5.11</td>
<td>Primary behaviors of Model 8 agree well with simulation.</td>
</tr>
<tr>
<td>6.1</td>
<td>Cell grid samples and metric schematics.</td>
</tr>
<tr>
<td>6.2</td>
<td>‘Passive’ model produces accurate convergence, but misrepresents extension.</td>
</tr>
<tr>
<td>6.3</td>
<td>Softer conditions are not enough to resolve discrepancies.</td>
</tr>
<tr>
<td>6.4</td>
<td>Area dependency on the number of sides of a cell.</td>
</tr>
<tr>
<td>6.5</td>
<td>Examples of final cell configurations from all models.</td>
</tr>
<tr>
<td>6.6</td>
<td>Active extension mechanism reproduces experimental metrics.</td>
</tr>
<tr>
<td>6.7</td>
<td>Active extension mechanism resolves energetic discrepancies from experiment.</td>
</tr>
<tr>
<td>D.1</td>
<td>Predicted distributions for alternate model agree well with the “true” distributions.</td>
</tr>
<tr>
<td>E.1</td>
<td>Sample calculation of tissue extension metric using experimental images.</td>
</tr>
</tbody>
</table>
Part I

Introduction
Stochasticity and disorder are frequently overlooked in the field of biology. Too often, biological systems are portrayed as orderly to help ourselves understand them. Proteins are pictured in terms of well-defined three-dimensional structures with each residue contributing in the function of the protein.\textsuperscript{1} Gene expression is laid out into organized reaction diagrams with each reactant playing an important role in the system that cannot be ignored.\textsuperscript{2} Multicellular tissues are expected to work together in synchronous harmony in order to produce perfectly regular shapes and patterns.\textsuperscript{3} In actuality, the biological realm is a rather tumultuous one. At the microscopic level, consider intrinsically disordered proteins (IDPs), an entire class of proteins that lack stable “folded” conformations.\textsuperscript{4} At the cellular level, consider how two genetically identical cells can produce vastly different behavioral phenotypes purely due to stochastic gene expression.\textsuperscript{5} At a tissue level, consider the highly heterogeneous morphologies produced by epithelial cells during germband extension (GBE) in \textit{Drosophila melanogaster} embryos.\textsuperscript{6} However, this stochasticity should not automatically be categorized as a bad thing. IDPs serve important biological functions ranging from transcriptional regulation and DNA condensation to cell differentiation and possibly membrane-less organelle formation.\textsuperscript{4,7,8} The bifurcation of behavioral phenotypes in isogenic cells can actually allow a cell colony to diversify and effectively “hedge its bets” against drastic environmental changes.\textsuperscript{9} The heterogeneous morphologies of GBE may be the result of competing active extension mechanisms within each cell that are necessary to produce sufficient elongation for proper physiology.\textsuperscript{10} Not only are stochasticity, disorder, and noise present in biology, they are indispensable to some of its core operations. This work will demonstrate numerous examples across different length scales of biology where stochasticity, disorder, and noise are harnessed by the system in question. In some cases, they can also be
harnessed by researchers to better characterize or even control the system through various biophysical techniques, both experimental and computational.

**Disorder on the Molecular Level**

Beginning at a molecular view, we will present an analytical model that can be used to predict the general conformation of intrinsically disordered proteins. This protein classification is of particular (and relatively newfound) interest because it challenges the long-held ”structure-function” paradigm. Despite not having what is traditionally considered to be a “folded” state, IDPs have well-defined *distributions* of conformations, where some shapes will be more probable than others, but they are constantly transitioning between them. These specific but dynamic distributions help to facilitate myriad behaviors including the spatial search of DNA by many transcription factors as well as the identification of misfolded proteins for repair or degradation.

In light of this realization, this leads us to a rather industrious question: given a particular sequence of amino acids, is it possible to analytically predict general characteristics of these distributions? To start, conformation size is largely governed by electric charge patterning. Previous studies have addressed the point that a higher net charge leads to more expanded, rod-like conformations. To improve on this coarse-grain observation, other researchers mapped out the entire phase-space of positive ($f_+$) and negative ($f_-$) charge concentration to generally assess where broad categories of conformations are located. However, the field is still in need of a more objective and quantitative method for predicting IDP conformation size. For instance, what about the case of two proteins with the same net charge and same charge concentrations, but different patterning of those charges? Can we predict
whether these proteins will have markedly different sizes or reside in the same coil-like or globule-like state?

To answer this question, we will introduce an analytical model based on first-order principles that can not only differentiate between these two proteins but also provide a direct prediction of the size of each protein. This model will first be applied to a set of well-characterized proteins to demonstrate its accuracy compared to simulated standards. Then, we will move to the entire collection of IDPs in the DisProt database to predict previously unseen trends across the entire phase space of charge concentration using realistic protein sequences. Such widespread, high-throughput analysis is impossible with the detailed but time-consuming methods of all-atom simulations. This will finally lead us to select interesting and illustrative examples where the addition of charge via phosphorylation at different locations will result in two proteins with identical charge concentrations but different sequencing, and as a result, vastly different conformations.

**Stochasticity in Genetic and Biochemical Networks**

Moving up in scale, we will then analyze the dynamics of the complex networks governing gene expression and feedback within those networks. Starting out simple, we will consider the nontrivial dynamics of competing complexation reactions. These reactions are fairly ubiquitous within the context of biophysics, but too often they are thought of in isolation and in terms of averages, completely ignoring the influence of any outside interactions or any stochastic fluctuations in the number of complexes. When considering complexation from an *in vitro* standpoint, averages are a perfectly suitable description since any sort of fluctuation becomes insignificant when considering numbers on the order of Avogadro’s number. However, switching gears to an *in vivo* standpoint, it is a common occurrence for a cell to have less than
100 copies of a particular protein, rendering these fluctuations much more consequen-
tial. Moreover, reactions hardly ever work alone within the busy context of the cell. Multiple reactants will regularly compete over shared resources, producing a much different outcome than if they operated in isolation.²⁵

In our analysis, we will demonstrate that mass-action methods actually make incorrect predictions about the behavior of complexation reactions in the case of small protein numbers.²⁶ Furthermore, the inclusion of a competing reactant will amplify the stochasticity of the system compared to a lone complexation, even producing nonzero relative noise when certain reactants are infinite. This will illustrate that mass-action based methods ignoring fluctuation are insufficient when describing small number scenarios. In addition, well-established stochastic methods like Gillespie algorithms²⁷,²⁸ must have the entire picture in order to provide any sort of valuable insight as to the dynamics of the system. Typical experiments can only measure limited numbers of species and the “bottom-up” mentalities associated with these algorithms require an unrealistic amount of information.²⁹,³⁰

Considering these factors, we turn to a “top-down” approach that infers meaningful metrics and predictions directly from experimental time-series data without making any assumptions as to architecture of the underlying genetic network. Specifically, we will utilize the principle of Maximum Caliber (MaxCal).³¹–³⁴ As the dynamic analog of maximum entropy, MaxCal tries to maximize the path entropy, or caliber, of the model while constraining some of its average behaviors to match experiment. To exhibit the power and utility of this concept, we will apply MaxCal to a series of common patterns in genetic networks, or motifs,³⁵–³⁷ gradually increasing in complexity as we progress.

To start, our simplest model will evaluate a single-gene auto-activating circuit.³⁸ This self-promotion leads to a bimodal “all-or-nothing” behavior where the protein of
interest can switch between being present in large numbers and not being present at all. While the genetics remain the same between these states, vastly different cellular behaviors can arise from these two disparate expression levels. This switch-like behavior allows groups of cells to execute the “bet-hedging” strategy mentioned earlier to prevent the collective from being wiped out by extreme environmental shifts while ensuring that the majority achieves optimal growth. Using in silico protein number trajectories as experimental input, we will apply our MaxCal model\textsuperscript{39} as if this auto-activation circuit is hidden from the view of researchers. By utilizing the full extent of the information contained in the input trajectories, MaxCal is able to objectively and quantitatively predict numerous behaviors of the genetic circuit used to create these simulated trajectories, including protein production/degradation rates, dwell times in either state, and protein number distributions. It is even able to extract effective feedback metrics that are unavailable using other stochastic methods. Furthermore, we will demonstrate that our methodology is easily transferable to trajectories presented in fluorescence rather than protein number, a much more realistic expectation from experiment.

Stepping up in complexity, MaxCal will then be applied to the two-gene toggle switch motif.\textsuperscript{40–42} In this circuitry, both proteins mutually repress the expression of their counterpart, instigating the same “all-or-nothing” response where one protein species can only be present in large numbers while the other is barely present and vice versa. With the similarity in these switch-like behaviors, the application to this motif addresses concerns about MaxCal as to whether it can handle systems with multiple reactants involved. Again using synthetic protein number and fluorescence trajectories, we will demonstrate that MaxCal is able to perform with similar accuracy when considering higher order systems, even when additional species (including mRNA) are hidden from the view of researchers.\textsuperscript{43} Previous studies have performed similar pre-
liminary research on this model, but our work establishes a more objective protocol using experimentally realistic datasets.

Finally, the most extensive version of our MaxCal gene expression model will be introduced to characterize the three-gene oscillatory motif known as a repressilator. This synthetic gene network mimics the periodic behaviors of circadian rhythms through circular repression where the presence of one protein represses the production of the next, producing very regular oscillation in expression levels of all three species. Not only does this genetic network once again increase the number of reactants to monitor, it also presents an entirely different dynamical property of oscillation, making it the most difficult test for MaxCal yet. By harnessing the computing efficiency of GPU programming and a discretized form of Finite State Projection, we will confirm that MaxCal can effectively and accurately reproduce the complex behaviors of the repressilator circuit from both protein number and fluorescence trajectories. The predictive power of our model even extends to include such detailed statistics as oscillatory periods, peak amplitudes, and three-dimensional protein number distributions.

In the context of gene regulation, mass-action approaches appeal to researchers because of their simplicity, but the data reduction inherent to these methods renders them ineffective when describing noisy small number scenarios, typical in vivo conditions. On the other hand, many stochastic methods account for this noise, but the level of detail necessary for them to function makes them both computationally unruly and experimentally unrealistic. All three of the applications of MaxCal above will convey that the “top-down” approach of MaxCal provides a compromise between these two mentalities by providing comprehensible macroscopic predictions while accounting for stochastic fluctuation and by harnessing the full extent of experimental data while remaining computationally realistic.
Noise on a Multicellular Level

In the final section, we will escalate to the multicellular scale of organized tissues, a regime not exempt from the effects of stochasticity. We will analyze the morphological process of GBE in *Drosophila melanogaster* embryos\(^{50,51}\) using a stochastic Monte Carlo simulation method. During development, the epithelial tissue of these embryos goes through what’s called convergent extension\(^{52}\) (CE) where the tissue *converges* along the dorsal-ventral axis and *extends* along the anterior-posterior axis. As a conserved process across a large percentage of species, CE creates the elongated body shape we are used to seeing in not just fruit flies, but a number of popular model organisms including the ascidian notochord,\(^{53,54}\) the chick primitive streak,\(^{55}\) and various organ systems.\(^{56-62}\) What makes this process particularly interesting is that during this elongation, cell division is largely absent and it is purely the result of cells intercalating and reorganizing amongst themselves.

How does this type of wide-spread coordination occur? It was originally thought that an anisotropic distribution of the tension-generating protein, myosin, was enough to initiate the process of contraction,\(^{63-65}\) and from there, the counteracting response of volume conservation and membrane elasticity would be enough to expand the tissue in the perpendicular direction.\(^{66}\) Our model will demonstrate that while these components are enough to generate the broad, qualitative behaviors of basic CE, the exact, quantitative behaviors seen in experimental tissues require additional mechanisms. Most notably, final tissue configurations are much more heterogeneous and disordered compared to the uniformly hexagonal grids generated by simulation. However, in the same vein as the active extension mechanism driven by myosin, it is conceivable that an active extension mechanism could also be present,\(^{10}\) possibly caused by targeted recruitment of the adhesive E-cadherin proteins. By including such a mechanism into the energetic considerations of our model, we will show that
not only is this a viable explanation, but a probable one based on the level of quantitative agreement across multiple metrics including cellular shape factor, overall tissue extension, and geometrical heterogeneity.

With these illustrative examples, this work aims to convince researchers that stochasticity, disorder, and noise are unavoidable within the context of biology, nor should they always be avoided. From microscopic to macroscopic length scales, we will establish archetypal instances where stochasticity can be embraced by science and scientist alike.
Part II

Disorder on the Molecular Level
Chapter 1

Size Prediction of Disordered Proteins

Intrinsically disordered proteins (or regions, collectively termed as IDP here) lack well-defined, stable structures and participate in important biological functions such as transcriptional regulation, cellular differentiation, and DNA condensation. These functions often rely on the molecular recognition features of IDPs that facilitate formation of multi-protein fuzzy complexes or spatial search by transcription factors. However, the ability to form such complexes – using fly casting or monkey bar mechanisms, for example – strongly depends on the conformational statistics of IDPs. Conformational features of IDPs are also important in identifying misfolded protein substrates for repair or degradation. Thus, IDPs challenge the traditional structure-function paradigm of globular proteins and require us to rethink function in terms of conformational statistics in the disordered ensemble of proteins. Unfolded state ensembles of regular proteins – capable of folding into well-defined structures – are also important in understanding traditional problems in protein science. Specifically, it has been proposed that thermophilic proteins may have more compact
unfolded states compared to their mesophilic orthologs.\textsuperscript{16,75,76} Recent work has also shown that thermophiles may have lower interaction energies in their unfolded state compared to mesophiles.\textsuperscript{77}

In parallel to the emerging role of IDPs in biological function, studies on the formation of membrane-less organelles and their physical properties are fundamentally changing our views of cellular biology.\textsuperscript{8,78,79} These organelles are liquid-like, capable of fusing and dripping, and their existence spans diverse systems including \textit{C. elegans} embryos, fungi, and mammalian cells.\textsuperscript{78,80–87} \textit{In vitro} experiments have also reported liquid-liquid phase separation using solutions of proteins.\textsuperscript{79,87–91} Together, \textit{in vitro} and \textit{in vivo} experiments have demonstrated the important role of IDPs in the formation of these liquid-like organelles.\textsuperscript{8,91} Simple sets of interactions among proteins may hold the clue to the formation of these droplets.\textsuperscript{8,92,93} Conformations of single-chain IDPs may relate to multi-chain physics, giving us insights into phase separation mechanisms. Mean-field homopolymer theories show that single-chain conformation as well as multi-chain solution properties can be described using the same interaction parameter $\chi$, capturing the chemical mismatch between the solvent and the solute. Recent work of Lin and Chan\textsuperscript{94} has explicitly shown that single-chain conformational properties (such as radius of gyration) are correlated with the critical temperature of phase separation in heteropolymers.\textsuperscript{95,96}

Clearly, there is a growing need to understand the conformational properties of IDPs,\textsuperscript{20} with the central question looming: how do molecular interactions encoded in the protein sequence dictate IDP conformations? Experiments from Schuler and colleagues as well as all-atom simulations by Pappu and coworkers have shown the important role of charge in IDP conformations.\textsuperscript{12,13,18,19,97,98} Consistent with this, properties of IDPs can be modulated by pH and salt.\textsuperscript{12,13,99,100} It is also interesting to note that phosphorylation sites in proteins are often in the disordered regions and
can be used to modulate the conformational ensemble.\textsuperscript{101–104} Phosphorylation sites can modify signaling responses from continuous to switch-like.\textsuperscript{102} Recent work has also shown that phosphorylation-induced folding can offer novel mechanisms of biological regulation using IDPs.\textsuperscript{105} Similarly, phase separation and aggregation can be controlled by phosphorylation.\textsuperscript{106} Phosphorylation of serine-rich regions of disordered proteins is crucial in the regulation of RNA granules in \textit{C. elegans}.\textsuperscript{83} Phosphorylation of the N-terminal region of human HP1\textalpha{} promotes phase-separated droplets and serves as a mechanism for gene silencing by heterochromatin.\textsuperscript{87}

\textit{How exactly does protein charge govern IDP conformation?} The simplest metric of net charge per residue (NCPR) has been used to describe IDPs.\textsuperscript{17,18} Although successful for some proteins, net charge is too coarse of a description; multiple sequences can have the same net charge with varying degrees of positive and negative charges. Recognizing this deficiency, Pappu and colleagues have presented a novel classification scheme (diagram-of-states) of IDPs based on the fraction of both positive ($f_+$) and negative charges ($f_-$)\textsuperscript{18,19} and inferred composition-function relationship in IDPs.\textsuperscript{97} For example, transcriptional repression activity of gene silencing proteins depends on the placement of the linker region in the $f_+, f_-$ diagram-of-states. The boundaries in this diagram-of-states also provide thumb rules for conformational statistics, e.g. if a protein will be in the coil or globule phase. However, the degree of swelling or the degree of compactness cannot be derived from this diagram-of-states. Furthermore, the above classification is insensitive to variation in size due to different placement of charges (positive and negative) in sequences with identical $f_+$ and $f_-$. Coarse-grained models of Srivastava and Muthukumar\textsuperscript{107} and all-atom simulations of Das and Pappu\textsuperscript{19} have shown that charge patterning – even for a fixed charge composition – can cause significant variation in conformation. Das and Pappu\textsuperscript{19} proposed a novel patterning metric to provide insights to these changes. Subsequently, analyti-
cal theory has provided rationale for these changes and introduced a novel sequence charge decoration (SCD) metric to distinguish these conformations.\textsuperscript{16} It is also important to note that there are regions in the $f_+, f_-$ diagram-of-states where proteins cannot be primarily classified as coil-like or globule-like.

In light of these developments, new questions emerge: Can we go beyond $f_+$ and $f_-$ to define the conformational space of IDPs? How strong are the variations across the $f_+, f_-$ phase space? Better yet, how strong are the variations for proteins in the same region of $f_+, f_-$ phase space? Even ignoring variation in hydrophobicity, what fraction of proteins in DisProt are coil-like and globule-like? Within coil-like proteins, how swollen are they? Said differently, what is the distribution of the degree of compactness or swelling with respect to a Flory Random Coil (FRC) for IDPs in the DisProt database? Furthermore, how do these conformations vary with changes in solution conditions such as temperature, particularly with respect to coil-globule transition? While some of these issues have been addressed with simulations,\textsuperscript{19,107–109} there is no analytical theory to understand these and apply them at a large scale to gain further insights. We address these questions by providing a comprehensive theory to compute IDP conformation as a function of sequence charge patterning and the excluded volume parameter (repulsive or attractive as temperature is varied). Existing theories for coil-globule transition are for homopolymers and are not applicable to delineate subtle variations due to charge patterning with constant composition. Earlier work of Sawle and Ghosh\textsuperscript{16} developed a heteropolymer theory to include sequence specificity in modeling coil-like conformations of proteins, but it was unable to describe collapsed globules. Here, we bridge the two. The high-throughput nature of the theory allows us to predict the conformation of all proteins in the DisProt database and yields multiple novel insights. Simultaneously, the theory shows that subtle variations in charge patterning, either due to mutation or post-translational modification,
can alter conformations. These insights and tools can be further used to design IDP sequences of desired conformations to test against single-molecule experiments that probe IDP conformations and are becoming more and more accessible.\textsuperscript{98,103,110}

1.1 Materials and Methods

1.1.1 High-Throughput Model for IDPs

We begin with the Hamiltonian ($H_t$) for a polymer chain in the presence of inter-monomer excluded volume, electrostatic interaction, and a three-body repulsive term given by:\textsuperscript{111}

$$\beta H_t = \frac{3}{2l} \int_0^L \! ds \left( \frac{dR(s)}{ds} \right)^2 + \frac{l_b}{l^2} \int_0^L \! ds \int_0^s \! ds' q(s)q(s') \exp \left( -\frac{\kappa |R(s) - R(s')|}{|R(s) - R(s')|} \right)$$

$$+ l \int_0^L \! ds \int_0^s \! ds' \omega(s, s') \delta[R(s) - R(s')]$$

$$+ \omega_3 l^3 \int_0^L \! ds \int_0^s \! ds' \int_0^{s''} \! ds'' \delta[R(s) - R(s')] \delta[R(s') - R(s'')]$$.

where $s$ is the contour length variable of the backbone, $R(s)$ is the position vector at $s$, $L$ is the total contour length with $l$ denoting the Kuhn length, $q(s)$ is the charge (unitless) at $s$, $\kappa$ is the inverse Debye length, the Bjerrum length is $l_b = 7.2$ Å$\,(298/T)$ with $T$ as the absolute temperature, $\omega(s, s')$ is the excluded volume interaction parameter between $s$ and $s'$, $\omega_3$ is the three-body repulsion parameter, $\delta$ is the Dirac delta function, and $k_b$ is Boltzmann’s constant with $\beta = 1/(k_bT)$. The three-body term is essential to describe the globule phase and coil-globule transition.\textsuperscript{112–114}

For simplicity, it has been assumed to be independent of the types of amino acids involved in the interaction. Using a variational approach similar to the one introduced by Muthukumar,\textsuperscript{115–118} the free energy ($\beta F$) of a chain as a function of the chain
expansion parameter $x$, sequence charge decoration $Q$, three-body repulsion $B$, and excluded volume interaction $\Omega$ (either attractive or repulsive) is given by

$$\beta F(x) = \frac{3}{2}(x - \ln x) + \left(\frac{3}{2\pi}\right)^{3/2} \frac{\Omega}{x^{3/2}} + \omega_3 \left(\frac{3}{2\pi}\right)^3 \frac{B}{2x^3} + \frac{l_b}{l} \frac{Q}{x^{1/2}} \sqrt{\frac{6}{\pi}}. \quad (1.2)$$

The chain expansion parameter is explicitly defined as $x = \left(\frac{R_{ee}}{R_{ee,frc}}\right)^2$, with $R_{ee}$ being the end-to-end distance of the protein of interest and $R_{ee,frc}$ being the end-to-end distance in the FRC limit (in the absence of any interaction). Switching from continuous to discrete notation, $\Omega$, $Q$, and $B$ are given respectively by

$$\Omega = \frac{1}{N} \sum_{m=2}^{N} \sum_{n=1}^{m-1} \omega_{m,n} (m - n)^{-1/2};$$

$$Q = \frac{1}{N} \sum_{m=2}^{N} \sum_{n=1}^{m-1} q_m q_n (m - n)^{1/2};$$

$$B = \frac{1}{N} \sum_{p=3}^{N} \sum_{m=2}^{p-1} \sum_{n=1}^{m-1} \frac{(p - n)}{[(p - m)(m - n)]^{3/2}},$$

where $N$ is the number of monomers in the protein of interest, $\omega_{m,n}$ is the excluded volume interaction parameter between the $m^{th}$ and $n^{th}$ monomer, and $q_i$ is the charge of the $i^{th}$ monomer. For the subsequent discussion, we ignore the sequence specificity of the excluded volume term and approximate its temperature dependence as $\omega_{m,n} = \omega = v(1 - \Theta/T)$, with $v$ controlling the magnitude of the excluded volume contributions and $\Theta$ being a reference temperature below which it is attractive and above which it is repulsive. Variations of this temperature dependence, as may be needed due to more complex temperature-dependent hydrophobic effects, can also be readily implemented. At this point, several limits are worth noticing for consistency checks. Equations 1.2 and 1.3 immediately recover the fully-ionized polyelectrolyte limit ($q_m = q_n = \pm 1$) of Muthukumar and colleagues. However, the
details of the sequence patterning are now embedded in the first two equalities in equation 1.3 enabling us to model heteropolymers. Furthermore, $Q$ is identical to the sequence charge decoration (SCD) metric introduced in our earlier work\textsuperscript{16} Its implication and usage will be discussed later in the context of IDP conformations. The free energy will be minimized with respect to $x$ numerically for arbitrary temperatures and sequence decorations to predict IDP conformations, i.e. $x$, for a given choice of $\omega_3$, $v$, and $\Theta$.

For a further consistency check, it is instructive to compare the free energy formalism presented above with an alternate but equivalent approach (see Appendix A for details) used in our earlier work\textsuperscript{16} on heteropolymer conformations without the three-body term. This alternate formalism\textsuperscript{16,120–123} is based on the idea that the full Hamiltonian (equation 1.1) can be approximated by a renormalized Hamiltonian ($H_r$) with only the connectivity term (the first term in the Hamiltonian; see equation A.1 in Appendix A) where the Kuhn length $l$ is replaced by a renormalized Kuhn length ($l_r$) that contains all the details of the interaction. The equation for the renormalized Kuhn length $l_r$ is determined by demanding that (see equations A.2 and A.3 in Appendix A for the rational)

\begin{equation}
\langle R_{cc}^2(H_t - H_r) \rangle_r = \langle R_{cc}^2 \rangle_r \langle (H_t - H_r) \rangle_r, \quad (1.4)
\end{equation}

where the ensemble average $\langle ... \rangle_r$ denotes averages over the renormalized Hamiltonian.

Using this formalism, the equation for $l_r$, and hence $x$, is derived (see Appendix A for the details of the derivation) as

\begin{equation}
x^{3/2} \left( 1 - \frac{1}{x} \right) = \left( \frac{3}{2\pi} \right)^{3/2} \frac{\Omega}{x} + \frac{l_b}{l} \sqrt{\frac{2}{3\pi}} Q + \omega_3 \left( \frac{3}{2\pi} \right)^3 \frac{B}{x^{5/2}}, \quad (1.5)
\end{equation}
where $\Omega$, $Q$, and $B$ are defined by equation 1.3. It is instructive to note that setting $d\beta F(x)/dx = 0$ using equation 1.2 yields equation 1.5. However, equation 1.2 can give us distributions and help us choose the minimum when there is bistability. On the other hand, for some special parameter choices (specifically near abrupt coil-globule transitions), using equation 1.5 will produce three solutions (one local minima, one global minima and one local maxima being unstable). The global minima can be chosen by drawing the tie line (for coil-globule coexistence) using Maxwell equal area construction and will give a result identical to numerical minimization of equation 1.2. Thus, both formalisms are exactly equivalent, but we use equation 1.2 for the ease of implementation. Note that by setting $\omega_3 = 0$, we recover our earlier result (see equation 13 in Sawle and Ghosh\textsuperscript{16}) for the end-to-end distance of a heteropolymer (with explicit charge decoration) in the zero salt limit without the three-body term. Our result also reproduces equation 5.5 of Muthukumar\textsuperscript{114} describing neutral homopolymers with the two-body and three-body term. Furthermore, setting $\omega_2 = 0$, $\omega_3 = 0$, and $q_m = q_n = 1$, we recover equation 3.26 (zero salt limit) of Muthukumar.\textsuperscript{118} We also recover the same functional form of Ha and Thirumalai\textsuperscript{121} in the polyelectrolyte limit (assuming uniform charge) in the presence of two-body and three-body interactions. To reiterate, Equations 1.2 and 1.3 (or equation 1.5 alternately) are the key results of the paper, which will be numerically minimized to compute $x$, and hence chain conformation, for arbitrary temperatures and sequence decorations to predict IDP conformations.

### 1.1.2 CAMPARI Simulation Procedure

We benchmark the theoretical model described above by performing all-atom Monte Carlo simulation using CAMPARI for IDPs with arbitrary sequences.\textsuperscript{124,125} Temperature-dependent conformations of IDPs were generated by carrying out twelve
independent simulations, each utilizing thermal replica exchange over fifteen temperatures: 280K, 286K, 293K, 298K, 305K, 315K, 325K, 335K, 350K, 365K, 380K, 395K, 410K, 430K, and 450K. Each simulation ran for $1.4 \times 10^7$ Monte Carlo steps with the first $1.5 \times 10^6$ steps eliminated as equilibration and a pdb file was generated every $4.0 \times 10^3$ steps. Energy considerations, Monte Carlo move sets, and replica swapping procedures are identical to that of Das and Pappu. Since our model is only concerned with zero salt conditions, the simulated system only consists of the protein of interest along with neutralizing Na\(^+\) and Cl\(^-\) ions and a spherical boundary with a radius of 400 Å was used. To represent phosphorylation in simulation and theory, the residue being phosphorylated is replaced by glutamic acid (E). This closely mimics the effect of phosphorylation in our coarse-grained model and has been used in previous studies to represent phosphorylation using CAMPARI.

1.2 Results

1.2.1 Sequence-specific theory is consistent with all-atom simulations of IDP sequences

All-atom simulation by Das and Pappu\(^{19}\) has quantitatively demonstrated temperature-dependent conformation changes in IDPs. We test the ability of our theory to capture coil-globule transition by fitting the simulations of five different IDP sequences, selected by Das and Pappu to best represent a range of weak polyampholytes. Figure 1.1 shows that the theoretical model captures these transitions well with the respective parameter values reported in Table 1.1. It is also instructive to note that by setting $\omega_3 = 0$, equation 1.2 yields earlier work\(^{16}\) which was further tested against all-atom simulation data of the radius of gyration for thirty different
variants of Glu-Lys sequences at room temperature. This provides additional benchmarks for the theory, but the absence of the three-body interaction (i.e. $\omega_3 = 0$) does not allow modeling the globule state. Hence, it could not be used to model the data presented in Figure 1.1.

Figure 1.1: Theory recapitulates temperature-dependent coil-globule transitions of all-atom simulations. Analytical theory (solid line) based on equation 1.2 and equation 1.3 was fitted to temperature-dependent simulated data (solid circle) of Das and Pappu using CAMPARI for five different disordered sequences (DisProt ID mentioned in the figure). For simulation results, $x$ was calculated using $R_G^2 \approx R_{ee}^2/6$, and for theoretical results, the best fit parameter values ($\omega_3$, $v$, and $\Theta$) are listed in Table 1.1.

1.2.2 DisProt database has a heterogeneous size distribution

Next, we predict the distribution of protein sizes of the naturally occurring IDPs in the DisProt database using the charge decoration in their primary sequence. DisProt is a database curating experimentally classified disordered proteins and regions. For the rest of this study, we will only consider the full sequences of disor-
Table 1.1: Parameters best representing coil-globule transition of all-atom simulations. The first five rows report parameter values for five proteins (names listed in column 1) extracted by using least-squared-error fitting procedures applied to the all-atom simulations of Das and Pappu (shown in Figure 1.1). The last row reports the average and standard deviation of these parameters.

<table>
<thead>
<tr>
<th>Protein</th>
<th>$\omega_3$</th>
<th>$v$</th>
<th>$\Theta$ (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DP00246</td>
<td>0.038</td>
<td>1.33</td>
<td>297</td>
</tr>
<tr>
<td>DP00011</td>
<td>0.140</td>
<td>2.19</td>
<td>362</td>
</tr>
<tr>
<td>DP00438</td>
<td>0.080</td>
<td>1.83</td>
<td>353</td>
</tr>
<tr>
<td>DP00436</td>
<td>0.084</td>
<td>0.78</td>
<td>401</td>
</tr>
<tr>
<td>DP00348</td>
<td>0.148</td>
<td>1.51</td>
<td>465</td>
</tr>
<tr>
<td>Avg. $\pm$ St. Dev.</td>
<td>0.098 $\pm$ 0.041</td>
<td>1.53 $\pm$ 0.47</td>
<td>376 $\pm$ 56</td>
</tr>
</tbody>
</table>

As of 11/2/2017, DisProt had 803 IDPs of which we used 727 IDPs by excluding all proteins containing unknown or irregular amino acids as well as those with more than 1000 amino acids for computational ease. We assign $q = 1$ for lysine and arginine, and $q = -1$ for glutamic acid and aspartic acid. In the absence of the exact values of the parameters $\omega_3$, $v$, and $\Theta$, we used an average of the values reported in Table 1.1 (see row six). We set $T = 300K$ to represent the DisProt size distribution near room temperature. The predicted distribution of $x$ demonstrates that proteins in the DisProt database have a wide variation in conformational size. The majority of proteins (77%) are predicted to be in the globule state (i.e. $x < 1$) with dimensions smaller than FRC (see blue curve in Figure 1.2).

Conformational properties of IDPs have been broadly classified in terms of $f_+$ and $f_-$ in five key regions: R1) weak polyampholytes and polyelectrolytes, R2) region between R1 and R3, R3) strong polymapholytes, and R4-R5) strong polyelectrolytes. These regions have been associated with typical conformations such as coil, globule, hairpin, etc. (see Figure 7 in for details). Proteins in R4 and R5 generally behave
as expected since theories for polyelectrolytes are well-developed. However, more than 99% of the proteins in DisProt reside in R1, R2, and R3 (specifically $f_+, f_- < 0.34$) and it is not clear how protein size varies within these regions. Therefore, the subsequent discussion will only focus on R1, R2, and R3. For a quantitative understanding of size heterogeneity based on charge content, we subdivide R1, R2, and R3 into a smaller grid spacing of $f_+$ and $f_-$ with $\Delta f_+ = \Delta f_- = 0.02$ and compute the average of predicted $x$ values at $T = 300$K for all proteins within a given bin. We denote this bin-dependent protein-averaged size as $x(f_+, f_-)$ and this is plotted as a heat map in Figure 1.3A with darker colors representing higher $x(f_+, f_-)$ values. Along the diagonal of this heat map (i.e. $f_+ \approx f_-$), $x$ values are usually small, while moving away from the diagonal (and increasing the net charge of the protein) produces much larger protein conformations. This highlights inter-bin variations in protein size and is displayed most prominently in R2 and at the leading edge of R3. It is important to note that the regions of small $x$ (shown with light color) do not reach
unphysically low values (such as zero), consistent with the left end of the distribution in Figure 1.2. Next, we quantify intra-bin variation expected to arise from variations in charge patterning among sequences with similar $f_+$ and $f_-$. We do this by computing the standard deviation in $x$ (defined as $\sigma_x(f_+, f_-)$) at $T = 300K$ among all the proteins that belong in a given bin. The resulting values are plotted as a heat map in Figure 1.3B, with darker colors representing higher intra-bin standard deviation. A similar pattern is recognized here with smaller standard deviations appearing along the diagonal and larger standard deviations (and therefore higher sequence specificity) further away from the diagonal. The degree of intra- and inter-bin variation is further illustrated by computing the fraction ($\phi$) of proteins that are coil-like (defined by $x > 1$) in a given bin. If $\phi$ is close to unity or zero, either all proteins are coil-like or all proteins are globule-like, suggesting that intra-bin variation is low. Conversely, $\phi \approx 0.5$ suggests that there exists a healthy mixture of coil-like and globule-like proteins, implying the highest intra-bin variation. As seen in Figure 1.3C, $\phi$ is small along the diagonal and approaches unity much further away from the diagonal, but there are non-trivial ‘sweet spots’ between these extremes where $\phi \approx 0.5$. It is in these locations where sequence specificity strongly influences size and where small modifications such as phosphorylation or mutation will have the largest impact, a topic discussed later in more detail.

The predictions in Figure 1.3 were made using the average values of the parameters from Table 1.1, but again, this ignores fluctuation in $\omega_3$, $v$, and $\Theta$. As before, we now include variation in $\omega_3$, $v$, and $\Theta$ to calculate the average of the weighted values of $x$ as $\bar{x}(f_+, f_-)$, the standard deviation of $x$ as $\sigma_x(f_+, f_-)$, and the resulting coil fraction as $\bar{\phi}(f_+, f_-)$. We notice the trends are very similar to the unweighted results (compare Figure 1.3 with Figure 1.4).
Figure 1.3: Predicted heat maps of protein conformations in DisProt. Panel (A) shows the average $x$ for DisProt proteins with positive ($f_+$) and negative charge fractions ($f_-$) corresponding to their bin location, panel (B) shows the standard deviation in $x$ for a given bin, and panel (C) shows the fraction ($\phi$) of IDPs in the coil state for a given bin. Averages, standard deviations, and fractions were calculated using all the proteins in a given bin. Blue bins correspond to locations in the $f_+, f_-$ phase space that contain fewer than two proteins. Black lines delineate regions R1, R2, and R3 as defined in Das and Pappu, PNAS, 110, 13392 (2013).

Figure 1.4: Predicted heat maps of protein conformations in DisProt accounting for parameter variation. Panel (A) shows the average $\overline{x}$ for DisProt proteins with positive ($f_+$) and negative charge fractions ($f_-$) corresponding to their bin location, panel (B) shows the standard deviation in $\overline{x}$ for a given bin, and panel (C) shows the fraction ($\overline{\phi}$) of IDPs in the coil state ($\overline{x} > 1$) for a given bin. Blue bins correspond to locations in the $f_+, f_-$ phase space that contain fewer than two proteins. Black lines delineate regions R1, R2, and R3 as defined in Das and Pappu, PNAS, 110, 13392 (2013).

1.2.3 Synthetic sequences yield similar trends

To further understand the origin of these changes and to establish that these variations are due to charge patterning rather than chain length variation or paucity of DisProt proteins, we performed a similar analysis with synthetic sequences. For each bin in $f_+, f_-$ phase space, we create a polypeptide of 100 amino acids using a
three letter code: lysine (K), aspartic acid (E), and alanine (A). Enough lysine and
aspartic acid residues are used to produce the corresponding \( f_+ \) and \( f_- \) respectively,
with alanine residues filling in the rest to mimic neutral amino acids. The sequence is
then randomly shuffled \( 10^4 \) times, keeping the same composition of K, E, and A, and
yielding sufficient variations in sequence patterning. Next, we calculate the values of
\( x \) for each random sequence. The heat maps in Figure 1.5 report the average (panel
A) and standard deviation (panel B) of the \( x \) values as well as the corresponding \( \phi \)
values (panel C) for each \( f_+ \), \( f_- \) pair. The averages and standard deviations were
calculated over all the sequences within a given bin. We notice the average of \( x \)
remains low along the diagonal (\( f_+ \approx f_- \)) and gradually increases away from the
diagonal, consistent with our observation with the DisProt database. The standard
deviation, however, shows a previously unseen trend. Variance remains low along the
diagonal and begins to increase away from it, but at a certain point, it reaches a
maximum and steadily decreases. This maximum is due to the extreme sensitivity of
coil-globule transitions on charge patterning for sequences that reside in these ‘sweet
spots’ of the charge composition parameter space. Away from this region, sequences
higher in net charge will predominantly be in the coil-like state regardless of sequence
patterning. Similarly on the other extreme, anything lower in net charge (near the
\( f_+ \approx f_- \) region) will be primarily globule-like independent of specific patterning.
However, near the maximum variance, a protein’s likelihood of being in the coil or
globule state strongly depends on the specifics of the charge patterning in addition
to charge composition. While this behavior does not contradict our findings with
the DisProt database, the trend could not be fully appreciated in DisProt due to
an insufficient number of proteins far away from the diagonal where the ‘sweet spot’
region exists. The region of maximum variance also corresponds to \( \phi \) values around
0.5, consistent with our interpretation of equal likelihood of coil and globule. The
above findings were generated using the fixed average values of $\omega_3$, $v$, and $\Theta$ listed in row six of Table 1.1. As before, the weighted calculations of $\bar{x}$, $\sigma$, and $\bar{\phi}$ accounting for parameter variation were performed (see Figure 1.6). Changes were minimal and the same interpretations hold.

Figure 1.5: **Predicted heat maps of synthetic sequence conformations.** Panel (A) shows the average $x$ for 100AA synthetic proteins generated using K, E, and A to match the positive ($f_+$) and negative charge fractions ($f_-$) corresponding to their bin location, panel (B) shows the standard deviation in $x$ for a given bin, and panel (C) shows the fraction ($\phi$) of proteins in the coil state ($x > 1$) for a given bin. Black lines delineate regions R1, R2, and R3 as defined in Das and Pappu, PNAS, 110, 13392 (2013).

Figure 1.6: **Predicted heat maps of synthetic sequence conformations accounting for parameter variation.** Panel (A) shows the average $\bar{x}$ for 100AA synthetic proteins with positive ($f_+$) and negative charge fractions ($f_-$) corresponding to their bin location, panel (B) shows the variance in $\bar{x}$ for a given bin, and panel (C) shows the fraction ($\bar{\phi}$) of proteins in the coil state ($\bar{x} > 1$) for a given bin. Black lines delineate regions R1, R2, and R3 as defined in Das and Pappu, PNAS, 110, 13392 (2013).
1.2.4 Phosphorylation can have “hot spots” for modulating conformation

The applications above illustrate variation among unmodified wild-type (wt) proteins in the DisProt database. The proposed model can be further applied to predict changes in conformation when the wt protein’s charge patterning is altered due to post-translational modification (PTM), such as phosphorylation. Phosphorylation typically modifies amino acids serine (S), threonine (T), and tyrosine (Y) by adding a negatively charged phosphate group, consequently changing the charge composition and patterning. We notice that it is possible to make drastic changes in conformation by phosphorylating even as few as only two sites. Interestingly, not all phosphorylation sites have the same change and the extent of change in size depends on the specific site. As an illustration, we present results for the IDP P0A8H9 (UniProt ID). The wt protein shows coil-globule transition with temperature as evidenced by all-atom simulation using CAMPARI (blue circles in Figure 1.7A). Using a least-squared-error fitting procedure, we extract representative parameter values of $\omega_3 = 0.096$, $v = 1.47$, and $\Theta = 380.0\text{K}$ to best fit our model (solid blue line in Figure 1.7A) to the simulated data. Next, we use these parameters to predict the temperature-dependent conformations for two post-translationally modified sequences that have each undergone phosphorylation at two residues (see Table 1.2). The sequence P0A8H9$_{S54S56}$ is phosphorylated at residue locations 54 and 56, while the alternate sequence P0A8H9$_{S2T15}$ is phosphorylated at residue locations 2 and 15. The predicted differences in size between these two sequences as a function of temperature can be significant (see solid lines in Figure 1.7B). To further validate our theoretical prediction, we carried out all-atom simulations for these two sequences (red and green circles in Figure 1.7B). We again notice a significant difference between the two sequences, although the mag-
magnitude of difference differs from the theoretical prediction. The differences between theoretical and simulated values are attributed to changes in the global parameter values ($\omega_3$, $v$, $\Theta$) upon phosphorylation. As a further check, we fitted the simulated profiles of the two phosphorylated species and found these parameters to be $\omega_3 = 0.180$, $v = 2.90$, and $\Theta = 399.1\text{K}$ for $\text{P0A8H9}_{\text{S2T15}}$, and $\omega_3 = 0.063$, $v = 0.85$, and $\Theta = 448.2\text{K}$ for $\text{P0A8H9}_{\text{S54S56}}$. The degree of change in the values of these global parameters will depend on the parent sequence and phosphorylation sites.

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>P0A8H9&lt;sub&gt;WT&lt;/sub&gt;</td>
<td>MSETITVNCPTCGKTVVVWGEISPFRPFCSKRCQ</td>
</tr>
<tr>
<td></td>
<td>LIDLGEWAAEEKRIPSSGLSESDDWSEEPKQ</td>
</tr>
<tr>
<td>P0A8H9&lt;sub&gt;S54S56&lt;/sub&gt;</td>
<td>MSETITVNCPTCGKTVVVWGEISPFRPFCSKRCQ</td>
</tr>
<tr>
<td></td>
<td>LIDLGEWAAEEKRIPSSGDL&lt;sub&gt;E&lt;/sub&gt;EEDWSEEPKQ</td>
</tr>
<tr>
<td>P0A8H9&lt;sub&gt;S2T15&lt;/sub&gt;</td>
<td>MEETITVNCPTCG&lt;sub&gt;K&lt;/sub&gt;EVVVWGEISPFRPFCSKRCQ</td>
</tr>
<tr>
<td></td>
<td>LIDLGEWAAEEKRIPSSGDLSESDDWSEEPKQ</td>
</tr>
</tbody>
</table>

*Table 1.2: Sequences of wt and phosphorylated species of P0A8H9.* Phosphorylation locations (shown in bold) were selected in order to maximize the difference in SCD (equation 1.6) between the two phosphorylated species.

Predicted and simulated difference between the two sequences is noticeable even around room temperature. This is remarkable given that the two phosphorylated sequences have identical charge composition ($f_+$ and $f_-$) and nearly identical patterning except for the two specific sites of phosphorylation. This finding conveys two key concepts: i) IDP conformations can vary drastically with minimal perturbation (as little as two phosphorylated residues), and ii) there are certain hotspots of phosphorylation that can trigger significant changes in conformation. This observation may yield further insights into experimental findings that suggest IDPs utilize phosphorylation to achieve graded or switch-like responses\textsuperscript{102} or to control phase separation.\textsuperscript{106}
Figure 1.7: Theoretical prediction agree well with all-atom simulation of wt and phosphorylated species of P0A8H9. Panel (A) shows temperature-dependent conformation (x) change of P0A8H9WT using all-atom simulation (blue circles) that were used to fit theory (solid line) to select values of $\omega_3 = 0.096$, $v = 1.47$, and $\Theta = 380$K. Panel (B) shows the predicted temperature-dependent conformations (x) of P0A8H9S54S56 (green line) and P0A8H9S2T15 (red line) using these parameters and their comparison against all-atom simulations (green and red circles).

The extent of conformation changes due to phosphorylation – drastic or modest – crucially depends on the complex coupling between charge patterning and temperature noticed by Srivastava and Muthukumar (see Figure 2 in\textsuperscript{107}), or broadly speaking, sequence (charge and hydrophobic) details and temperature.\textsuperscript{19,108,109} Our theory reveals the complexity of the coupling in an analytical framework just using a coarse-grained model for protein charge. If room temperature is in the vicinity of the coil-globule transition temperature of an IDP sequence, slight alteration in the charge patterning can alter the effective intra-chain interaction and consequently cause drastic changes in conformation. In contrast, if the protein – at room temperature – is away from the transition point, the impact of charge patterning on the conformation will be modest or negligible. While the observed changes have been quantified by tuning the temperature to alter the effective excluded volume ($\omega$) parameter, similar changes can be envisioned due to local changes in the solution properties that are
likely to occur inside the cytoplasm due to significant inter-compartment heterogeneity inside a cell.\textsuperscript{128}

It is also important to note that there are counter examples, such as wild type Ash1 and its phosphorylated version (Ash1p) who minimally differ in conformation.\textsuperscript{126} Our finding is not in contradiction with this observation. As discussed above, if proteins are away from the transition point, the change in size upon phosphorylation can be minimal. Furthermore, as seen in the case of Ash1 and Ash1p, the similarities can be due to inter-segment statistics that contribute to overall $R_g$. Our present theory is only for the end-to-end distance ($R_{ee}$) which is not directly related to $R_g$ for heteropolymer chains.\textsuperscript{129–131} More accurate estimation of $R_g$ would require calculation of inter-residue distances as well.\textsuperscript{16} Nevertheless, these insights to either amplify conformational changes for function or buffer effects of large changes in charge patterning provide first steps in designing synthetic sequences and possibly even understanding allostery in IDPs.\textsuperscript{132,133}

**Sequence Charge Decoration (SCD) metric can identify phosphorylation sites that induce maximal change**

*How does one choose proteins and sites of phosphorylation that may reveal such drastic changes upon phosphorylation?* Analyzing equation 1.2 and equation 1.3, we note the effect of electrostatics is purely governed by $Q$, which is equivalent to Sequence Charge Decoration (SCD), earlier introduced by Sawle and Ghosh.\textsuperscript{16} Specifically, SCD is defined as

$$SCD = \frac{1}{N} \sum_{m=2}^{N} \sum_{n=1}^{m-1} q_m q_n (m-n)^{1/2}.$$ \hspace{1cm} (1.6)
The role of SCD has also been explored in modeling unfolded protein states and understanding thermophilic adaptation\textsuperscript{16,76} as well as polyampholytic phase behavior.\textsuperscript{94} In fact, this metric was used to identify the example above (P0A8H9) and the two specific choices of phosphorylation that yield maximal difference. Using this metric, we can now readily identify multiple proteins that can have significant changes in conformation upon phosphorylation. The choice of proteins and the phosphorylation sites can be further investigated by using appropriate values of the parameters $\omega_3$, $v$, and $\Theta$ – either using the average (in row six of Table 1.1) as a first guess or all-atom simulation for a more accurate estimation of the parameters – to quantify expected changes in size. The example above is only one protein among several others that were predicted from our theory. However, we chose to study the protein P0A8H9 due to its small size (65 amino acids), making it amenable to all-atom simulation in a reasonable time.

1.3 Conclusion

In summary, we present a theoretical method to compute temperature-dependent conformations of polymers with specific distributions of charges. The coupling between charge patterning and temperature allows us to explore novel ways to manipulate heteropolymer conformation. We provide a new diagram-of-states using sequence-conformation relation for Intrinsically Disordered Proteins (IDP) in the DisProt database by predicting their size distribution. Our calculation reveals significant variation among proteins with similar charge composition, therefore attributing it only to charge patterning. Similarly, we show that post-translational modifications that alter protein charge, such as phosphorylation, can be used to control protein size as well. With temperature regulation and subtle changes in charge location (as
few as two phosphorylation sites), drastic differences in protein size can be realized. Predicted changes can be guided by a simple sequence charge decoration (SCD) metric using only the location of charges and chain length. The initial guess for the sequences can be further explored by computing $x$ using either the average IDP parameters or simulation data when available. These predicted changes can be further tested by additional all-atom simulations or experiments. We hope the theory presented here will motivate new experiments using single molecule techniques to explore IDP conformations and possibly reveal new insights into IDP function.
Part III

Stochasticity in Genetic and Biochemical Networks
Chapter 2

Competition in Complexation Reactions

Complexation reactions are ubiquitous in biology. The most common examples are biochemical reactions where enzymes and substrates bind to facilitate the conversion of substrates to products. While the majority of these enzyme-substrate reactions in biochemical pathways involve small molecule substrates, it is also possible to envision complexation between two macromolecules. Such is the case during translation and transcription which requires complexation between nucleic acids and proteins. Other relevant biological examples include multimerization of proteins leading to functionally beneficial supra-molecular complexes. Similarly, formation of toxic protein aggregates from natively folded or unfolded proteins also involve complexation. These reactions have received considerable attention starting with the most celebrated Michealis-Menten kinetics where enzyme molecules speed up conversion of substrates to products by forming enzyme-substrate complexes as an intermediate. However, it should be remembered that these reactions seldom work in isolation. On the contrary, multiple reactions are often coupled by sharing the same
resource. After all, it is the concerted and coupled network of complex biochemical reactions that is at the very heart of systems and synthetic biology. Coupling between different species is present in different forms — circuits involving feedback are the primary example of such direct interactions. However, sharing a common resource can also mediate indirect modes of communication between two reactants. Perhaps the most common example is competitive binding of two different substrates on the same enzyme. Broadly speaking, competition leads to numerous intriguing biological phenomena such as ultrasensitivity in cellular signalling, temporal cooperativity, and promotion or suppression of oscillatory robustness of biochemical networks, to name a few.

While competition is natural in biology, the role of small numbers presents another intriguing aspect of biology. For example, copies of proteins that participate in biochemical reactions could be as low as a few tens to hundreds inside a cell. Compartmentalization inside cells makes the copy number problem even more prominent. This is in strong contrast to standard in-vitro chemical reactions involving a large number of reactant molecules where fluctuations are negligible and averages are a perfectly reasonable description of the system. Traditional Michaelis-Menten treatment of enzyme kinetics, using laws of mass action, works within this framework providing information about averages. However, due to small copy numbers of proteins inside of a cell, one has to separate results of standard in-vitro reactions from their in-vivo counterparts. Clearly, an unavoidable consequence of such small numbers is the high level of fluctuation, implying averages alone may not be a good prescription of the system. Small number fluctuations in gene expression has already shown the need to adopt statistical physics principles to quantify results beyond simple averages. Single molecule biophysics is the extreme example of small numbers. Example problems such as RNA folding, DNA translocation,
single enzyme molecule processivity,\textsuperscript{144} and motion of molecular motors\textsuperscript{22} highlight the importance of studying the entire distribution instead of just the average. These examples raise the natural question: How sensible is it to think in terms of averages if noise is comparable to the mean\textsuperscript{145,146} or higher? As a consequence, care must be taken to employ traditional approaches in reaction network analysis that are formulated only in terms of averages, such as mass action laws.

The roles of competition and stochasticity, individually, have been the subject of major research interest in recent times. However, the combined effect\textsuperscript{147} of the two remained largely unexplored in spite of its prevalence throughout biology. The most common and unavoidable example is when few copies of an enzyme molecule act on multiple substrates in a biochemical pathway. Workhorse protein molecules called kinases, involved in signal transduction, often bind multiple different proteins to transmit signals.\textsuperscript{148} Rationing of a limited number of ribosomes among many competing mRNA’s during protein expression is another example where stochasticity and competition are inherently linked. In a recent work, it was shown that such systems can exhibit correlated resonance mediated via competition.\textsuperscript{147}

Motivated by the prevalence of these examples throughout biology, we have chosen to investigate the simple system of two complexation reactions, \(A + B \leftrightarrow AB\) and \(A + C \leftrightarrow AC\). Thus, reactants \(B\) and \(C\) share the same species \(A\), mediating a possible layer of indirect interaction between the two despite not being directly coupled. Our earlier work\textsuperscript{149} highlighted the importance of stochasticity in a single complexation reaction such as \(A + B \leftrightarrow AB\). We demonstrated that even when particle numbers are not too low, of the order of 20-40, the effect of noise is important. In the present work, we extend this study in the context of competition. Several interesting questions emerge: i) How does noise in one reactant (say \(C\)) influence the noise in the other (\(B\))? ii) Does the competition enhance or suppress noise when
compared to the single complexation reaction? iii) If so, how is it possible to dial down or dial up such effects by engineering the second reaction?

In the example considered, we assume the total (free and complexed) number of A, B, and C molecules are fixed due to tight regulation inside a cell. This assumption is reasonable when the rates of complexation and dissociation are faster than any other reactions A, B, and C molecules are involved in. Statistical mechanics of gene expression and repression is a typical example, where binding/unbinding rates of transcription factors can be faster than the production and degradation rates of the activator/repressor protein molecules and the gene.22,23 Steady state assumption is commonplace in several other biological examples as well, such as Michaelis-Menten enzyme kinetics.22,24 Thus we use the fixed copy number approximation to motivate the simplest toy model as a conceptual framework to study the role of competition that is typical in biology. However, lifting the assumption of steady state will be interesting and has been left for future studies.

Below, we first outline the master equation formalism employed to describe the system. Results and their implications are discussed in the following section.

## 2.1 Materials and Methods

### 2.1.1 Stochastic model for two competing reactions

We consider the reaction $A + B \leftrightarrow AB$ (herein denoted as reaction 1) and $A + C \leftrightarrow AC$ (denoted as reaction 2) for a system having a total of $M$ number of $A$ molecules, $N_1$ number of $B$ molecules, and $N_2$ number of $C$ molecules. The stochastic variables are the numbers of complexes formed, i.e. the number of molecules of $A$ that participate to form $AB$ (denoted by $m_1$) and the number that participate to form $AC$ (denoted by $m_2$). $P(m_1, m_2; t)$ is the joint probability distribution of having $m_1$
number of $AB$ complexes and $m_2$ number of $AC$ complexes at time $t$. Let $k_{1f}$ and $k_{1b}$ denote the first reaction’s forward and reverse rates respectively, while $k_{2f}$ and $k_{2b}$ represent that of the second. Using master equation formalism, the time evolution of the probability distribution is

$$
\frac{dP(m_1, m_2; t)}{dt} = -(k_{1b}m_1 + k_{2b}m_2)P(m_1, m_2; t) - (k_{1f}(M - m_1 - m_2)(N_1 - m_1)
+ k_{2f}(M - m_1 - m_2)(N_2 - m_2))P(m_1, m_2; t)
+ k_{1f}(M - m_1 - m_2 + 1)(N_1 - m_1 + 1)P(m_1 - 1, m_2; t)
+ k_{2f}(M - m_1 - m_2 + 1)(N_2 - m_2 + 1)P(m_1, m_2 - 1; t)
+ k_{1b}(m_1 + 1)P(m_1 + 1, m_2; t) + k_{2b}(m_2 + 1)P(m_1, m_2 + 1; t) \tag{2.1}
$$

It should be noted that the equation above can be derived using the principle of Maximum Caliber (MaxCal) as well. This is not surprising given the equivalence between MaxCal and Markov process. Master equations for two state systems and single bimolecular complexation reactions using MaxCal have been derived and systematically analyzed. We choose not to discuss such details here because our present goal is the quantitative analysis of the role of noise rather than the formalism. An equivalent description of the system can be obtained by performing Kinetic Monte Carlo simulation using transition probabilities (between $t$ and $t + \Delta t$).
defined as,

\[
W(m_1, m_2, t + \Delta t; m_1, m_2, t) = 1 - (k_{1b}m_1 + k_{2b}m_2) \\
+ k_{1f}(M - m_1 - m_2)(N_1 - m_1) \\
+ k_{2f}(M - m_1 - m_2)(N_2 - m_2)\Delta t
\]

\[
W(m_1, m_2, t + \Delta t; m_1 - 1, m_2, t) = k_{1f}(M - m_1 - m_2 + 1)(N_1 - m_1 + 1)\Delta t
\]

\[
W(m_1, m_2, t + \Delta t; m_1, m_2 - 1, t) = k_{2f}(M - m_1 - m_2 + 1)(N_2 - m_2 + 1)\Delta t
\]

\[
W(m_1, m_2, t + \Delta t; m_1 + 1, m_2, t) = k_{1b}(m_1 + 1)\Delta t
\]

\[
W(m_1, m_2, t + \Delta t; m_1, m_2 + 1, t) = k_{2b}(m_2 + 1)\Delta t
\]  

(2.2)

where \(W(m_1, m_2, t + \Delta t; m_1, m_2, t)\) is the probability of keeping the number of \(AB\) and \(AC\) complexes (\(m_1\) and \(m_2\)) unchanged; \(W(m_1, m_2, t + \Delta t; m_1 - 1, m_2, t)\) is the probability of increasing the number of \(AB\) complexes from \(m_1 - 1\) to \(m_1\); \(W(m_1, m_2, t + \Delta t; m_1, m_2 - 1, t)\) is the probability of increasing the number of \(AC\) complexes from \(m_2 - 1\) to \(m_2\); \(W(m_1, m_2, t + \Delta t; m_1 + 1, m_2, t)\) is the probability of decreasing the number of \(AB\) complexes from \(m_1 + 1\) to \(m_1\); and \(W(m_1, m_2, t + \Delta t; m_1, m_2 + 1, t)\) is the probability of decreasing the number of \(AC\) complexes from \(m_2 + 1\) to \(m_2\). All the transitions are between time \(t\) and \(t + \Delta t\), with \(\Delta t\) being small enough to ensure that there is only one transition in the short time interval. Similar simulations have been performed earlier for other discrete systems.\(^{31,149}\) Gillespie simulation\(^{28}\) can be used to simulate the time evolution of the same system as well.
2.1.2 Exact equilibrium distribution

The equilibrium distribution $P(m_1, m_2)$ is derived in the limit $dP(m_1, m_2; t)/dt = 0$ and is given by

$$P(m_1, m_2) = \frac{M!}{(M - m_1 - m_2)!} \frac{N_1!}{(N_1 - m_1)!} \frac{N_2!}{(N_2 - m_2)!} \frac{1}{m_1! m_2!} \frac{x_1^{m_1} x_2^{m_2}}{Z},$$

(2.3)

where $x_1 = k_{1f}/k_{1b}$, $x_2 = k_{2f}/k_{2b}$ (herein referred to as reaction constants), and $Z$ is the partition function ensuring normalization. This can be immediately verified by noticing that $P(m_1, m_2)$ satisfies equation 2.1 when $dP(m_1, m_2; t)/dt = 0$. Similar combinatorial expressions for equilibrium probability distributions have been derived for other biochemical reactions.\textsuperscript{136,149,153}

2.1.3 Equations of motion for different moments

While the exact time evolution of the entire probability distribution is described by equation 2.1, it is possible to derive equations of motion for different moments as well. Multiplying both sides by $m_1$ and summing over all possible values of $m_1$, we get the equation of motion for the average of $m_1$,

$$\frac{d\langle m_1 \rangle}{dt} = -k_{1b}\langle m_1 \rangle + k_{1f}\langle (M - m_1 - m_2)(N_1 - m_1) \rangle.$$

(2.4)

Similarly, the equation of motion for the average of $m_2$ can be derived as

$$\frac{d\langle m_2 \rangle}{dt} = -k_{2b}\langle m_2 \rangle + k_{2f}\langle (M - m_1 - m_2)(N_2 - m_2) \rangle.$$

(2.5)
Multiplying equation 2.1 by \( m_1 m_2 \) and summing over all values of \( m_1 \) and \( m_2 \), the equation of motion for \( \langle m_1 m_2 \rangle \) becomes

\[
\frac{d\langle m_1 m_2 \rangle}{dt} = k_{1f}\langle (M - m_1 - m_2)(N_1 - m_1)m_2 \rangle + k_{2f}\langle (M - m_1 - m_2)(N_2 - m_2)m_1 \rangle - (k_{1b} + k_{2b})\langle m_1 m_2 \rangle
\]

(2.6)

Next, we multiply equation 2.1 by \( m_1^2 \) and sum over all values of \( m_1 \) to obtain the equation of motion for \( \langle m_1^2 \rangle \),

\[
\frac{d\langle m_1^2 \rangle}{dt} = k_{1f}\langle (M - m_1 - m_2)(N_1 - m_1) \rangle + 2k_{1f}\langle (M - m_1 - m_2)(N_1 - m_1)m_1 \rangle - 2k_{1b}\langle m_1^2 \rangle + k_{1b}\langle m_1 \rangle
\]

(2.7)

Similarly, the equation of motion for \( \langle m_2^2 \rangle \) is

\[
\frac{d\langle m_2^2 \rangle}{dt} = k_{2f}\langle (M - m_1 - m_2)(N_2 - m_2) \rangle + 2k_{2f}\langle (M - m_1 - m_2)(N_2 - m_2)m_2 \rangle - 2k_{2b}\langle m_2^2 \rangle + k_{2b}\langle m_2 \rangle
\]

(2.8)

It should be noted that these equations do not form a closed set and progressively involve higher order moments. However, within a truncation scheme (described in section 2.2.2), the set can be made finite and the time development of various moments of \( m_1 \) and \( m_2 \) can be computed.
2.2 Results and Discussion

2.2.1 Quantifying equilibrium fluctuations and correlations

From equation 2.3, we compute different moments of the stochastic variable $m_1$ (number of $AB$ complexes) and present the analysis and implication of these results below.

**Coefficient of variation can be high**

We construct a quantity $f_1$ as the ratio of the average and standard deviation of $m_1$ as

$$f_1(M, N_1, N_2, x_1, x_2) = \frac{\sqrt{\langle m_1^2 \rangle - \langle m_1 \rangle^2}}{\langle m_1 \rangle} = \frac{\sqrt{\langle (\Delta m_1)^2 \rangle}}{\langle m_1 \rangle}$$

(2.9)

The estimates of $f_1$ provide a quantitative measure of stochasticity in the system and has been used to estimate stochasticity in different biological systems.\(^{139}\) It is surprising to note that fluctuations are significant even with reactant quantities greater than a few (of the order of $50 – 100$) and with reaction constants $(x_1, x_2)$ throughout a wide range of values (between $10^{-2}$ and $10^8$). It is possible to achieve $f_1$ as high as $150\%$ or greater in parts of the phase diagrams shown (see Figure 2.1). These regions correspond to the reactant quantities and reaction constants that are typical in many biological reactions\(^{22,138,139}\) (e.g. binding of oxygen to myoglobin:\(^{154}\) $\Delta \varepsilon \approx -7.04k_BT$, $x \approx 1.1 \times 10^3$; binding of NtrC to DNA:\(^{155}\) $\Delta \varepsilon \approx -17.47k_BT$, $x \approx 3.9 \times 10^7$). Thus, the notion of averages can be misleading in these examples due to the broad variance of the distribution. Parameters for the figures throughout this study were chosen to represent either regions of extremity or regions of typical values throughout biology.
Figure 2.1: Fanofactor phase diagrams for different complexation conditions. Phase Diagrams of $f_1$ for a) $M = 50, x_1 = 1,000, x_2 = 20,000$, and b) $N_1 = 50, x_1 = 1,000, x_2 = 20,000$. Different colors denote different $f_1$ percentages (blue: $50\% < f_1 < 75\%$, green: $75\% < f_1 < 100\%$, red: $100\% < f_1 < 150\%$, black: $f_1 > 150\%$).

Simpler approximate analytical result for fluctuation

While equation 2.3 provides the full and exact distribution, we intend to provide a simpler but approximate relation here by approximating equation 2.3 with a Gaussian distribution. The average ($\langle m_1 \rangle = \overline{m}_1$) and variance ($\langle (\Delta m_1)^2 \rangle$) of the distribution can be determined from

$$x_1 = \frac{\overline{m}_1}{(M - \overline{m}_1 - \overline{m}_2)(N_1 - \overline{m}_1)}; \quad x_2 = \frac{\overline{m}_2}{(M - \overline{m}_1 - \overline{m}_2)(N_2 - \overline{m}_2)}$$

(2.10)

and,

$$\langle (\Delta m_1)^2 \rangle = \sigma_1^2 = \frac{P_2}{P_1 P_2 - Q^2}$$

(2.11)

where $P_1$, $P_2$, and $Q$ are defined by equation B.6 in Appendix B. Details of the derivation can also be found in Appendix B.

Estimates of standard deviation predicted using the equation above are usually within 5% of the exact results given moderate conditions ($M, N_1, N_2 > 25; 1,000 <$
$x_1, x_2 < 200,000$; figure not shown) but are highly overestimated in the specific conditions of high $x_1$, low $x_2$, low $M$, and $N_1 + N_2 \approx M$. Thus, equation 2.11 provides a simple but reliable estimate of the noise of the system for a broad range of values of $M$, $N_1$, $N_2$, $x_1$, and $x_2$ that can be readily used in experimental systems. These results also help us understand several asymptotic limits and gain valuable insights (discussed below). It is interesting to note that the average using this approximation is the same as the mass action law prediction.

**Competition enhances fluctuation**

Figure 2.1 indicates that an increase in $N_2$ enhances fluctuation in the first reaction ($f_1$). This can be rationalized by realizing that the additional number of $C$ molecules compete for the shared resource $A$. The competition in turn creates a shortage of the available number of $A$ molecules to react with $B$. The lowering of the available number ($M_{av}$) of $A$ molecules is associated with increased fluctuation in the first reaction. This implies $f_1$ would be higher for the two reaction case when compared to the single reaction (non-competitive case) having the same $x_1$, $M$ and $N_1$. While this observation is important and can be employed to alter fluctuations in the first reaction, it is not surprising due to an effective reduction in the available number of $A$ molecules to react with $B$, causing $M_{av} < M$. However, perhaps a more relevant question would be: How does the fluctuation compare with the single reaction scenario that has the same reduced number ($M_{av}$) of $A$ molecules instead of $M$? Thus we provide a direct comparison by considering two cases: i) fluctuations in the first reaction, within the competitive reaction scheme, for fixed values of $N_1, N_2, M, x_1$, and $x_2$, and ii) fluctuations in the single reaction scheme having the same $N_1, x_1$, but a reduced value of $M$ given by $M' = M - \langle m_2 \rangle$, where $\langle m_2 \rangle$ is the average number of complexes formed by the second reaction in case (i). This way, we ensure the single
reaction in case (ii) has access to the same number of $A$ molecules, on average, as it would in the presence of the second reaction. This allows a more direct comparison of noise in the first reaction both in the absence and in the presence of the second reaction. We illustrate this by comparing the distribution of $m_1$ under conditions (i) (solid lines in Figure 2.2) and (ii) (dashed lines in Figure 2.2), defined above. We have chosen values of $M$, $N_1$, and $N_2$ in the tens to hundreds range, corresponding to the typical number of proteins inside a cellular compartment. To simplify the scenario while keeping realistic values, $x_1$ and $x_2$ were assumed to be equal at $x_1 = x_2 = 1000$, a typical value of reaction constants observed in biology.\textsuperscript{22,154} When $N_2$ is sufficiently low, as expected, the two distributions are very similar (not shown). However, with the addition of a comparable number of $C$ molecules driving the second reaction, we notice the width in the competitive reaction scenario (case i) is significantly larger compared to the single reaction (case ii). However, the averages remained very similar due to the constraint imposed by reducing the number of $A$ molecules to $M'$ (see Figure 2.2). The single reaction case is limited by the hard constraints on the total number of molecules (either $A$ or $B$) exhibiting the sharp drop in the probability. However, coupling with a second reaction allows a bigger spread due to a much “softer” condition at the expense of the fluctuations in the second reaction. This is essentially how the particle baths of different species ($B$ and $C$) couple to effectively enhance fluctuations. Fluctuations in the number of free reactants $B$ would be the same as fluctuations in the number of $AB$ complexes. Fluctuations in the free reactant $A$ is higher than the fluctuation in the $AB$ complexes under competition. Thus the claim of fluctuation enhancement under competition holds for both non-complexed and complexed species.

Next, we quantitatively explore this noise enhancement in the entire phase space. We denote $\sigma_1^2$ as the variance for the two reaction scheme (case i), and $(\sigma'_1)^2$ as the
Figure 2.2: **Probability distribution comparison under competitive and non-competitive conditions.** Distributions under competitive (case i) conditions (solid line) and non-competitive (case ii) conditions (dashed line) with $M = 50$, $N_1 = N_2 = 500$ and $x_1 = x_2 = 1,000$ (case (i): $\langle m_1 \rangle = 25.0$, $\sigma_1 = 3.45$; case (ii): $\langle m'_1 \rangle = 25.0$, $\sigma'_1 = 0.01$).

We explicitly compute the phase diagram of a new metric, $\eta = (\sigma_1^2 - (\sigma'_1)^2)$, for different reactant quantities and reaction constants. By subtracting $(\sigma'_1)^2$, we isolate the extrinsic noise due to the presence of the second reaction from the inherent noise in the single reaction. To facilitate the calculation of factorial functions, we choose $M'$ to be an integer less than or equal to $M - \langle m_2 \rangle$, therefore overestimating noise in the second scenario. In spite of the overestimation in $(\sigma'_1)^2$, Figure 2.3A and 2.3B indicate that the difference is positive and high in most of the regions of the phase space that are biologically relevant.

Selected regions of the phase space where the difference is small can be classified in three regimes (blue in Figure 2.3A,B): a) $N_1$ is sufficiently large compared to $N_2$ and $M$, restricting $m_1$ to the deterministic limit of $M$ and mimicking single reaction scenario; b) $N_2$ is sufficiently larger than $N_1$ and $M$ is fixed, forcing $m_1$ to approach 0, consequently reducing the width of the distribution; c) $M$ is sufficiently large com-
Figure 2.3: Phase diagrams of variance differences under competitive and non-competitive conditions. Phase diagrams of $\eta = \sigma^2_1 - (\sigma'_1)^2$ for a) $M = 50$, $x_1 = 1000$, $x_2 = 1000$, and b) $N_1 = 50$, $x_1 = 1000$, $x_2 = 1000$. Different colors denote different $\eta$ values (blue: $2.0 < \eta < 4.0$, green: $4.0 < \eta < 6.0$, red: $6.0 < \eta < 8.0$, black: $\eta > 8.0$).

Compared to both $N_2$ and $N_1$ nullifying the effect of competition. The first two scenarios depict the situation where one reaction overwhelms the other, forcing $m_1$ to approach either $M$ or 0, while the regime of large $M$ can be associated with “weak-competition” due to a relative abundance of the shared resource. With the exception of such extreme regions, the major part of the phase diagram shows significant enhancement in variance due to competition. Moreover, we notice that a proportional increase of $N_1$ and $N_2$ (in Figure 2.3A) or $M$ and $N_2$ (in Figure 2.3B) enhances width. This is in contrast to the general expectation that large numbers reduce fluctuations. We explore this effect and provide an alternate mathematical argument to further augment our claim that competition enhances noise. From equation 2.11, we notice that the amount of variance can be significantly high if $P_1 P_2 \simeq Q^2$. We consider a particularly interesting limiting case where $N_1$ and $N_2$ are infinitely large, maintaining a fixed
stoichiometry of \( N_2 = \beta N_1 \), and \( M \) is finite. In this limit, equation 2.11 yields,

\[
\sigma_1^2 = \frac{M \beta x_1 x_2}{(x_1 + \beta x_2)^2} \quad \text{and} \quad f_1 = \sqrt{\frac{\beta x_2}{M x_1}} \tag{2.12}
\]

The coefficient of variation \((f_1)\) is non-zero and can be made indefinitely large by increasing \( \beta \) i.e. the relative proportion of \( N_2 \) compared to \( N_1 \). This is in sharp contrast to \( \beta = 0 \) (no competition), when the variance is zero and the coefficient of variation is zero. Non-zero and high values of \( f_1 \) indicate that analyzing such competing biochemical reactions using averages and mass action laws can be misleading even if competing species are present in large amounts. Figure 2.4 demonstrates this effect by calculating the coefficient of variation from the exact distribution (equation 2.3).

![Figure 2.4: Nonzero relative noise for fixed \( M \) and relatively infinite \( N_1, N_2 \). Plots of \( f_1 \) and \( \sigma_1^2 \) as a function of \( \beta = N_2/N_1 \) with \( x_1 = x_2 = 1,000 \) and varying values of \( M \) (blue lines: \( M = 5 \); green lines: \( M = 10 \); red lines: \( M = 50 \)). The maximum value between \( N_1 \) and \( N_2 \) is 10,000 and the other is derived based on the value of \( \beta \) (e.g. for \( \beta = 0.5 \), \( N_2 = 5,000 \) and \( N_1 = 10,000 \); for \( \beta = 2 \), \( N_2 = 10,000 \) and \( N_1 = 5,000 \)). Another similarly interesting regime emerges when \( M \) and \( N_2 \) are infinitely large, maintaining a fixed stoichiometry \( M/N_2 = \alpha \), and \( N_1 \) is finite. We focus on the condition \( \alpha < 1 \). Under this condition \( M \) is relatively less compared to \( N_2 \), and it
is possible to explore the effect of competition. Using equation 2.11, the asymptotic results are:

\[ \sigma_1^2 = \frac{N_1 x_1 x_2 \alpha (1 - \alpha)}{[(1 - \alpha) x_2 + \alpha x_1]^2} \] and \[ f_1 = \sqrt{\frac{(1 - \alpha) x_2}{N_1 \alpha x_1}} \] (2.13)

We explore this in Figure 2.5 using the exact distribution. Despite the number of reactants being infinite, a proper balance between them can produce non-zero fluctuation in the presence of competition. Moreover, it is possible to indefinitely increase the coefficient of variation when \( N_1 \) is fixed and \( M/N_2 \) is very low. On the other hand, when \( \alpha > 1 \), it is not surprising that the competition is negligible due to the relative abundance of the shared resource, hence \( f_1 = 0 \), recovering the result of the pure single reaction.

Figure 2.5: **Nonzero relative noise for fixed \( N_1 \) and relatively infinite \( M, N_2 \).** Plots of \( f_1 \) and \( \sigma_1^2 \) as a function of \( \alpha = M/N_2 \) with \( x_1 = x_2 = 1,000 \) and varying values of \( N_1 \) (blue lines: \( N_1 = 5 \); green lines: \( N_1 = 25 \); red lines: \( N_1 = 50 \)). The maximum value between \( M \) and \( N_2 \) is 10,000 and the other is derived based on the value of \( \alpha \) (e.g. for \( \alpha = 0.5 \), \( M = 5,000 \) and \( N_2 = 10,000 \); for \( \alpha = 2 \), \( M = 10,000 \) and \( N_2 = 5,000 \)).

It is interesting to note, when \( \alpha < 1 \), fluctuations in the number of the free (non complexed) \( A \) and \( B \) molecules are also non zero. The coefficient of variation \( f_A \) (for
the free $A$ molecules) and $f_B$ (for the free $B$ molecules) are given by

$$f_A = \sqrt{\frac{(1 - \alpha)x_2}{\alpha}}; \quad f_B = \sqrt{\frac{\alpha x_1}{(1 - \alpha)N_1 x_2}}. \quad (2.14)$$

On the other hand when $\alpha \geq 1$, $f_B$ diverges. In the regime $\alpha > 1$, the coefficient of variation for the free $C$ molecules is finite and is given by

$$f_C = \sqrt{x_2(\alpha - 1)} \quad (2.15)$$

Thus, the different lines of argument above establish the significant effect of noise enhancement in the limit of strong competition. Similar effects of noise enhancement have been observed in the context of gene expression.\textsuperscript{156} Since such coupled biochemical reactions are inherent in biology, this effect should be realized in many systems. These high fluctuations indicate an evolutionary need to adopt other complex machinery such as feedback to control noise when it is not desirable. The effect of competition in feedback regulated circuits has already shown interesting dynamical properties.\textsuperscript{25} Conversely, it can be argued that the enhancement of noise under competitive schemes may be beneficial in biology due to several noise induced phenomena such as stochastic resonance,\textsuperscript{157,158} stochastic focusing,\textsuperscript{159,160} stochastic bistability/bifurcation, and decision making.\textsuperscript{41,42,44,136,161–163} Although more complicated than the example presented here, it should be remembered that the exclusive toggle switch has similar characteristics by sharing the same plasmid.\textsuperscript{41,42,44}

**Competition promotes correlation**

Another interesting consequence of shared resources would be a degree of correlation established between $B$ and $C$. We explore this effect between the two types of
complexes, $AB$ and $AC$, by computing their correlation $R^2$,

$$R^2 = \frac{\langle m_1m_2 \rangle - \langle m_1 \rangle \langle m_2 \rangle}{\sigma_1 \sigma_2}.$$  

(2.16)

Figure 2.6: **Higher competition levels amplifies negative correlation.** Plot of $R^2$ as a function of $N_2$ for a fixed value of $M = 50$ with a) $x_1 = x_2 = 1$ and b) $x_1 = x_2 = 1,000$. The blue line represents $N_1 = 2$, the green line $N_1 = 10$, the red line $N_1 = 20$, and the black line $N_1 = 50$.

In the absence of correlation, the metric would be zero. We notice negative correlation between the two species due to competition mediated via a common resource (Figure 2.6). Moreover, we notice the degree of correlation between the two species can be increased by increasing $N_2$ or $N_1$ or both while keeping $M$ fixed. Under these conditions, the two species strongly compete with each other for the shared resource, explaining the strong negative correlation. This is consistent with a previous observation on correlation resonance in the context of competing reactions including protein degradation and production.\textsuperscript{147} It is also interesting to notice that under strong binding conditions (high values of $x_1$ and $x_2$), the degree of correlation undergoes an all-or-none transition from zero to negative unity when $N_2$ is increased for fixed val-
ues of \(M\) and \(N_1\) (see Figure 2.6B). This resembles cooperative transitions observed in numerous other biological settings.\textsuperscript{136,164,165}

**Equilibrium constants can be highly heterogeneous**

A quantity of common interest in biochemical reactions is the equilibrium constant, \(K_1\) for the first reaction and \(K_2\) for the second reaction. These are defined as the ratios of the number of complexes to the product of the numbers of free reactants. Due to the stochastic nature of the number of complexes, \(m_1\) and \(m_2\), equilibrium constants \(K_1\) and \(K_2\) will be stochastic as well and are defined as

\[
K_1 = \frac{m_1}{(M - m_1 - m_2)(N_1 - m_1)}; \quad K_2 = \frac{m_2}{(M - m_1 - m_2)(N_2 - m_2)}. \tag{2.17}
\]

The joint distribution of \(K_1\) and \(K_2\) was computed from the distribution of \(m_1\) and \(m_2\) (equation 2.3) and is shown in Figure 2.7. These distributions show a broad variance, rendering the notion of one single reaction constant questionable. Traditional chemical kinetics will only depict one point in this two dimensional space of reaction constants. Values for Figure 2.7 (specifically \(x_1 = x_2 = 100\)) were chosen to mimic a biological example with a dissociation constant of \(K_d \approx 15\, \text{mM}\).\textsuperscript{22}

**Mass action prediction deviates from the exact average**

Mass action predictions of the number of complexes can be an overestimate or underestimate of the true average (large time limit in Figure 2.8). We further explored the difference between the exact average (\(\langle m_1 \rangle\)) and the mass action prediction in the parameter space of reactant quantities \((M, N_1, N_2)\) and reaction constants \((x_1, x_2)\). The deviations are maximum, as expected, when reactant quantities are extremely low, of the order of unity. However, we find the difference to be as high as 15%
Figure 2.7: **Inferred equilibrium constants can vary widely in competitive conditions.** Joint distribution of $K_1$ and $K_2$ for $M = 10,000$, $N_1 = N_2 = 5,000$ and $x_1 = x_2 = 100$. Different color points denote different probabilities (blue: $P(K_1, K_2) < 0.25%$; red: $0.25% < P(K_1, K_2) < 3%$; black: $P(K_1, K_2) > 3%$) and the green square denotes the coordinates of $\langle K_1 \rangle = 145.9$ and $\langle K_2 \rangle = 145.9$; $K_1 = K_2 = 100$ would be the single value predicted by mass action laws.

Even under realistic conditions: $M, N_1, N_2 > 20$; $1,000 < x_1, x_2 < 200,000$. An important consequence of this is a potential danger if the process is executed in reverse – inferring the rate constants using mass action laws (equation 2.10) by measuring the average values of $m_1$ and $m_2$ in noisy biological experiments. The relative percentage difference between the true and predicted values of $x_1$ is extremely amplified with high $x_1$, low $x_2$, low $M$, and $N_1 + N_2 \approx M$ in order to exaggerate a smaller denominator in equation 2.10. However, even outside of these specific conditions, the error can be as high as 25% under realistic conditions. This high level of error indicates that in small number situations, in-vivo conditions, it is imperative to consider the role of noise when determining the true equilibrium constant of a reaction.
2.2.2 Determination of dynamical quantities

We perform Kinetic Monte Carlo simulation using the transition rates given in equation 2.2 to simulate several time evolved trajectories starting with a given initial condition. Time evolution of the average number of complexes ($\langle m_1(t) \rangle$) and the standard deviation ($\sqrt{\langle (m_1(t))^2 \rangle}$) is calculated by gathering statistics from these trajectories. Figure 2.8 shows the simulation result. From the early time behavior, it is evident that standard deviation can be higher than the average itself indicating that the application of mass action laws should be used with caution, consistent with our discussion above.

Figure 2.8: Time evolution comparison between prediction and simulation. Time evolution of the average (Monte Carlo simulation = black circles; analytical solution with fluctuation = black line, analytical mass action solution = green line) and standard deviation (Monte Carlo simulation = red circles; analytical solution with fluctuation = red line) of the number of complexes $m_1$ (left panel) and $m_2$ (right panel) with $M = 5$, $N_1 = 10$, $N_2 = 5$, $k_{1f} = 0.01$, $k_{1b} = 1$, $k_{2f} = 1$, $k_{2b} = 0.5$, and initial conditions of $m_1(t = 0) = 0$, and $m_2(t = 0) = 0$. Equilibrium values are: $\langle m_1 \rangle_{\text{exact}} = 0.11$, $\langle m_1 \rangle_{\text{massaction}} = 0.13$, $\sigma_{1,\text{exact}} = 0.32$, $\langle m_2 \rangle_{\text{exact}} = 3.8$, $\langle m_2 \rangle_{\text{massaction}} = 3.6$, $\sigma_{2,\text{exact}} = 0.8$.
Mean-field time evolution equations fail to predict correct averages

Equation 2.4 and 2.5 can be simplified assuming $\langle m_1^2(t) \rangle = \langle m_1(t) \rangle^2$, $\langle m_2^2(t) \rangle = \langle m_2(t) \rangle^2$, and $\langle m_1(t)m_2(t) \rangle = \langle m_1(t) \rangle \langle m_2(t) \rangle$. Under these mean-field approximations, we neglect fluctuations in the stochastic variables $m_1$, $m_2$ as well as their correlations. These approximations lead to

$$
\frac{d\langle m_1 \rangle}{dt} = -k_{1b}\langle m_1 \rangle + k_{1f}(M - \langle m_1 \rangle - \langle m_2 \rangle)(N_1 - \langle m_1 \rangle)
$$

$$
\frac{d\langle m_2 \rangle}{dt} = -k_{2b}\langle m_2 \rangle + k_{2f}(M - \langle m_1 \rangle - \langle m_2 \rangle)(N_2 - \langle m_2 \rangle).
$$

It should be noted that these are identical to evolution equations derived using mass action laws in traditional chemical kinetics. The solution of these mean-field average equations deviate from the actual averages obtained from Monte Carlo simulation (see Figure 2.8). Thus it is possible to realize the breakdown of mass action laws due to strong fluctuations in such non-linear systems.

**Modified time evolution equations predict correct averages and fluctuations**

As illustrated above, neglecting fluctuations is a poor approximation for predicting time evolution of the average. Here, we propose a better approximation by including fluctuations. The exact time dependence of the fluctuation and average would require solving equation 2.4, 2.5, 2.6, 2.7, and 2.8 simultaneously. However, as mentioned earlier, these do not form a closed set of equations. We make the following
approximation to truncate the set of equations:

\[
\begin{align*}
\langle m_1^3(t) \rangle &= 3\langle m_1^2(t) \rangle \langle m_1(t) \rangle - 2\langle m_1(t) \rangle^3 \\
\langle m_2^3(t) \rangle &= 3\langle m_2^2(t) \rangle \langle m_2(t) \rangle - 2\langle m_2(t) \rangle^3 \\
\langle m_1^2(t)m_2(t) \rangle &= 2\langle m_1(t)m_2(t) \rangle \langle m_1(t) \rangle \\
&\quad - 2\langle m_1(t) \rangle^2 \langle m_2(t) \rangle + \langle m_1^2(t) \rangle \langle m_2(t) \rangle \\
\langle m_1(t)m_2^2(t) \rangle &= 2\langle m_1(t)m_2(t) \rangle \langle m_2(t) \rangle \\
&\quad - 2\langle m_2(t) \rangle^2 \langle m_1(t) \rangle + \langle m_2^2(t) \rangle \langle m_1(t) \rangle
\end{align*}
\] (2.19)

For a single complexation reaction, it has been already demonstrated that higher order moments beyond the second moment can be assumed to be zero for predicting time evolution.\textsuperscript{149} We numerically solve these sets of equations and compare with the exact results from Monte Carlo simulation (see Figure 2.8). Inclusion of fluctuations improves the prediction over mean-field mass action equations. The set of approximations in equation 2.19 effectively assumes the distribution to be symmetric around the mean. Although this is not accurate, it is less severe of an approximation than the mass action laws (equation 2.18) that neglect fluctuations altogether. The agreement with the exact result thus implies that neglecting fluctuations is a much worse approximation than neglecting the skewness of the distribution when predicting the time evolved average and variance.

2.3 Conclusion

We studied the equilibrium distribution and time evolution of different moments for the chemical reactions \( A + B \leftrightarrow AB \) and \( A + C \leftrightarrow AC \) to explicitly study the effect of competition in noisy biochemical reactions. We make several key observations: i)
Equilibrium fluctuations (Figure 2.1) can be significant when numbers of reactant molecules are of the order of tens to hundreds, typical in biology. ii) The degree of noise in the first reaction can be altered by tuning the number of reactant molecules in the second reaction, while keeping everything else fixed. This is due to the absorption effect provided by the second reaction that reduces the available number of shared resources for the first reaction. As a result, fluctuation increases (Figure 2.1).

iii) We notice the same enhancement of noise holds true even when we isolate the effect of the reduced number of available reactant molecules under the competitive case. We do so by calculating the variance for the single reaction (in the absence of competition) having a renormalized number of $A$ molecules, $M' = M - \langle m_2 \rangle$, where $M$ is the total number of $A$ molecules in the competitive case and $\langle m_2 \rangle$ is the average number of $A$ molecules taken away by the second reaction. Under these conditions, we again notice fluctuations are significantly higher in the presence of the second reaction although the averages can be the same (Figure 2.2). We present detailed quantitative analysis of this noise enhancement in the broad parameter space that could be biologically relevant.

iv) With finite resources (fixed $M$), fluctuations can be non-zero even when the number of $B$ molecules is infinite, provided there is an infinite number of $C$ molecules such that $N_2/N_1$ is finite. In fact, the noise can be indefinitely amplified when the ratio is increased (equation 2.12). A similar noise enhancement is seen when $N_1$ is held fixed but both $M$ and $N_2$ are infinite such that $M < N_2$. This is in contrast to the single reaction scheme that has zero fluctuations in the limit of one of the reactants being infinite. This could be of serious consequence in biochemical networks, where even if substrates are present in high concentrations, neglecting variance could be misleading due to enhanced fluctuations resulting from resource sharing. Thus, analysis based on mass action formalism should be applied with care. These regimes exhibit non-zero fluctuations in some of the free (non com-
plexed) reactants as well. v) We find competition enhances correlation between two species (Figure 2.6) when resources become limited, consistent with effects seen in other biochemical networks. vi) We report numerical errors in predicting averages using mass action laws by comparing against exact results. The errors have been identified both in the equilibrium value and the time evolution of the average. Our alternate analytical formulation to compute the time development of averages and variances provides excellent agreement with the exact result. We believe that the detailed quantitative study presented here using the tools of chemical reactions and stochastic physics, will motivate future researchers to consider the effect of noise and stochasticity in both theoretical and experimental efforts, especially in the context of competing reactions.
Chapter 3

Maximum Caliber Applied to Auto-Activation Networks

Biological function is largely dictated by gene networks that control protein expression in single cells. Understanding details of these networks and consequently building quantitative models is essential to control gene expression and ultimately regulate cellular dynamics. However, model development has been limited due to the lack of information about the complex web of interactions (including feedback regulation) that defines these networks. Typical experiments only provide partial information by measuring the expression levels of one or two proteins of interest using fluorescent tags, much less than the actual number of entities (mRNAs, promoters, nucleotides, amino acids) involved in the process of gene expression. This problem of partial information is a key challenge for model building. Although the number of species monitored is limited, experimental read-outs contain crucial information by recording the entire time trajectory of fluctuating protein expression levels. The stochastic nature of the trajectories is due to small copy numbers of molecules involved in these reactions. The details of noise statistics encode the
details of network architecture. This provides a potentially useful avenue to infer details of network architecture by analyzing noisy protein expression levels.\textsuperscript{30,171–175} In spite of realizing the power of this approach,\textsuperscript{171,172,174,175} such efforts are still in their infancy. Existing models are either too simple, with limited single-cell-level predictive power, or too detailed, requiring too many unknown parameters.\textsuperscript{29} The most common stochastic approaches first define sets of reaction networks to be simulated using a Gillespie algorithm\textsuperscript{27} or related methods and then fit different observables to determine the corresponding reaction rate parameters. A major drawback of these methods is that they are ‘bottom-up’ and require detailed knowledge of the underlying reaction network. This is particularly challenging when networks involve feedback, a common feature in many natural networks and synthetic biology. It is currently impossible to test many of these ad-hoc assumptions independently. Furthermore, these approaches can involve too many parameters that can fit the same data with multiple models, creating additional challenges for efficient parameter estimation.\textsuperscript{30} The challenges of having too many parameters are also problematic for circuit design\textsuperscript{5,176–187} as it requires ways to efficiently explore parameter space to test different models, thus demanding models with the least possible amount of parameters.

To circumvent these obstacles, we propose a ‘top-down’ approach in modeling these networks. We use the principle of Maximum Caliber (MaxCal) to model stochastic trajectories with minimal information. We show the application of MaxCal on a simple auto-activating circuit, a common motif in many biological circuits.\textsuperscript{38} MaxCal maximizes path entropy subject to constraints, similar to maximum entropy on state space, and directly works with path trajectories. This makes MaxCal directly applicable to experimentally measured time trajectories of protein numbers. We establish the methodology on synthetic data generated using Gillespie simulations\textsuperscript{27} of a known auto-activating circuit. This trajectory data serves as the input data –
a proxy for experimental data – to MaxCal. The minimal model of MaxCal is then applied to the raw trajectory statistics in conjunction with maximum likelihood (ML) to determine representative parameters for the model. These parameters can predict other statistics of the data and quantitatively infer several underlying physical variables that are not visible otherwise. In the following section, we first describe the synthetic circuit and the generation of \textit{in silico} data that mimic experimental data. Next, we introduce MaxCal and its specific application to this circuits. We show how MaxCal along with ML can be used to infer model parameters and make predictions. Comparing these predictions against the known model allows us to benchmark the predictive capabilities of MaxCal. Finally, we present how the methodology can be applied when the input data is not in protein number but in arbitrary fluorescence, a common challenge in interpreting experimental data.

\section{3.1 Materials and Methods}

\subsection{3.1.1 Generating synthetic data for an auto-activating circuit}

Considering the complexity of natural networks with many unknown or incompletely understood interactions, synthetic biologists are building mimics of frequently-occurring parts of bigger networks, called network motifs.\textsuperscript{35–37} One natural network motif with important biological function that has inspired the design of many synthetic gene circuits is feedback regulation. Previous work\textsuperscript{44} has demonstrated the application of MaxCal on double-negative (overall positive) feedback circuits where two genes mutually repress each other, commonly referred to as a toggle switch circuit.\textsuperscript{40,42} Here we consider a positive feedback circuit where a single gene auto-activates itself. As a proof of concept, we apply MaxCal to synthetic data generated \textit{in silico} using models for which the underlying parameters are known. This will serve as a proxy
for experimental data and provide us with a gold standard to which we can compare when demonstrating how well MaxCal performs given stochastic trajectories.

While most models ignore the presence of mRNA for simplicity, ours will explicitly account for it when generating synthetic data. This will illustrate that for the purposes of our inference methods, additional hidden species like mRNA will not affect the accuracy of our metrics and predictions. Among several models of auto-activation in different biological contexts,\textsuperscript{161,188–194} we will utilize the one below, similar to the one studied by Elston et al,\textsuperscript{195} to generate stochastic synthetic data that will serve to mimic experimental time traces:

\begin{equation}
\alpha \xrightarrow{g} \alpha + a ; \quad a \xrightarrow{d} \emptyset ; \quad a \xrightarrow{p} a + A ; \quad A \xrightarrow{r} \emptyset \\
A + A \xrightarrow{f_d}{\text{kd}} A_2 ; \quad \alpha + A_2 \xrightarrow{f_p}{b_p} \alpha^* ; \quad \alpha^* \xrightarrow{g^*}{\text{kd}} \alpha^* + a
\end{equation}

where some generic mRNA $a$ is created from its corresponding gene $\alpha$ at a rate of $g$ and is degraded at a rate of $d$. This mRNA produces the protein it encodes ($A$) at a rate of $p$. Protein $A$ can then degrade at a rate of $r$ or dimerize into $A_2$ with forward and backward rates of $f_d$ and $b_d$ respectively. To incorporate feedback, $A_2$ binds and unbinds from its promoter at rates of $f_p$ and $b_p$ respectively, sending the promoter into and out of its activated state, $\alpha^*$. In this activated state, $\alpha^*$ creates mRNA $a$ at a much faster rate $g^*$, capturing the essentials of a positive feedback mechanism. Rates are chosen to produce switching times that are representative of experiments\textsuperscript{38} while maintaining synthesis and degradation rates of both protein and mRNA in the realm of typical rates.\textsuperscript{22} A Gillespie algorithm\textsuperscript{27} was applied to this network to generate the stochastic trajectories of protein ($A$) levels shown in Figure 3.1A and three major features are worth noting: i) two clearly separated high and low states, ii) a large amount of fluctuation within each state, and iii) stochastic switching between the
two states. In the next section, we will first attempt to reproduce these three basic features in MaxCal using as simple of a framework as possible.

![Figure 3.1](image)

**Figure 3.1:** Positive feedback circuit. (A) Typical time trace of the number of proteins in a self-promotion circuit using the reaction scheme in equation 3.1 ($g = 0.05 \text{ s}^{-1}$, $g^* = 0.5 \text{ s}^{-1}$, $d = 0.2 \text{ s}^{-1}$, $p = 0.02 \text{ s}^{-1}$, $r = 1.0 \times 10^{-3} \text{ s}^{-1}$, $f_d = 5.0 \times 10^{-3} \text{ s}^{-1}$, $b_d = 50.0 \text{ s}^{-1}$, $f_p = 6.0 \times 10^{-3} \text{ s}^{-1}$, $b_p = 3.0 \times 10^{-5} \text{ s}^{-1}$, assuming intrinsic time unit is in seconds). Data is recorded every 300 s. (B) Typical time trace of the number of proteins in the minimal model of self-promotion using MaxCal ($h_a = -0.512$, $h_A = 0.585$, $K_{A\alpha} = 0.0298$, $M = 15$, $\Delta t = 300 \text{ s}$).

### 3.1.2 Maximum Caliber model for auto-activating circuit

Maximum caliber is a variational principle that gives a prescription to infer dynamics by maximizing the path entropy,\textsuperscript{34,44,140,149,151,152,196–202} or *caliber*, subject to known constraints enforced via Lagrange multipliers. For the gene circuit of interest, there are three minimal constraints that must be in place: i) protein synthesis, ii) protein degradation, and iii) auto-activation/positive feedback. We enforce the first two by restricting the average number of proteins that are created in a discrete time interval ($\Delta t$) as well as the average number of proteins that are destroyed.\textsuperscript{34,44} To do this, we define $\ell_\alpha$ as the production state variable which describes the number of proteins that are created in the time interval and ranges as integer values between zero and some predefined maximum value ($M$), i.e. $0 \leq \ell_\alpha \leq M$. We also define $\ell_A$ as the
degradation state variable which describes the number of previously existing proteins that still exist at the end of the time interval. Clearly, $\ell_A$ ranges as integer values between zero and the number of proteins present at the beginning of the time interval ($N_A$), i.e. $0 \leq \ell_A \leq N_A$. The corresponding Lagrange multipliers for these two constraints are $h_\alpha$ and $h_A$ and the probability of observing a particular combination of $\ell_\alpha$ and $\ell_A$ is defined as $P_{\ell_\alpha, \ell_A}$. Next, we implement the constraint of positive feedback, the idea that a high number of proteins ($N_A$) should positively correlate with the production of $A$. This is done by introducing a third Lagrange multiplier, $K_{A\alpha}$, that enforces a coupling between protein production and the presence of proteins by constraining the average of $\ell_\alpha \ell_A$. This is the lowest order term in the coupling of these two variables that must be imposed to capture the essence of feedback. As will be shown in chapter 4, similar arguments are used to build models to describe negative feedback in toggle switch circuitry.\textsuperscript{44} The four basic ingredients of the model, described above, yield the caliber as

\begin{equation}
C = -\sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} P_{\ell_\alpha, \ell_A} \log P_{\ell_\alpha, \ell_A} + h_\alpha \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} \ell_\alpha P_{\ell_\alpha, \ell_A} \\
+ h_A \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} \ell_A P_{\ell_\alpha, \ell_A} + K_{A\alpha} \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} \ell_\alpha \ell_A P_{\ell_\alpha, \ell_A},
\end{equation}

and the corresponding caliber-maximized path probabilities are

\begin{equation}
P_{\ell_\alpha, \ell_A} = Q^{-1} \left( \frac{N_A}{\ell_A} \right) \exp(h_\alpha \ell_\alpha + h_A \ell_A + K_{A\alpha} \ell_\alpha \ell_A) ; \\
Q = \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} \left( \frac{N_A}{\ell_A} \right) \exp(h_\alpha \ell_\alpha + h_A \ell_A + K_{A\alpha} \ell_\alpha \ell_A).
\end{equation}

With this path probability distribution, stochastic trajectories are generated using a Monte Carlo method to select a path for each time point. The system then creates
and destroys the number of proteins corresponding to the $\ell_\alpha$ and $\ell_A$ of the selected path and the time of the system advances by the predetermined $\Delta t$. A quick search of the parameter phase space ($h_\alpha$, $h_A$, $K_{A\alpha}$, $M$) reveals that even with just these four parameters, bimodal behaviors like the ones seen in the self-promotion circuits of the previous section (characterized with nine parameters in equation 3.1) can be reproduced (see Figure 3.1B). For efficient and accurate computation, protein number probability distributions are generated using a discrete-time version of Finite State Projection (FSP) described in Appendix C, where each state represents a different number of proteins present in the system up to some predefined maximum, $N_{\text{max}}$. This method is needed to provide a systematic way to truncate the infinite phase space of possible states since protein number does not have an upper bound. FSP provides a rigorous self-consistent approach to ensure the truncation error is within a pre-determined error-bound (see Appendix C for exact application).

Furthermore, the state variables $\ell_\alpha$ and $\ell_A$ directly relate to effective protein synthesis and degradation rates analogous to the reaction rates in equation 3.1. Specifically,

$$p_{\text{eff}} = p(a)_{N_L} \approx p\left(\frac{g}{d}\right) \approx \frac{\langle \ell_\alpha \rangle_{N_L}}{\Delta t}, \quad p_{\text{eff}}^* = p(a)_{N_H} \approx p\left(\frac{g^*}{d}\right) \approx \frac{\langle \ell_\alpha \rangle_{N_H}}{\Delta t},$$

$$r(N) = \frac{N - \langle \ell_A \rangle_N}{N\Delta t}, \quad r \approx \sum_N P_{\text{eq}}(N)r(N),$$

where $\langle \cdots \rangle_i$ represents the average of a quantity of interest when $N_A = i$, $N_L$ is the peak of the protein number distribution in the low state, $N_H$ is the peak in the high state, and $P_{\text{eq}}(N)$ is the probability of having $N$ proteins within the system at relative equilibrium (calculated via discrete-time FSP). Furthermore, we can extract an effective feedback metric $F_{A\alpha}$ in terms of the Pearson correlation coefficient between
\( \ell_\alpha \) and \( \ell_A \):

\[
F_{A\alpha} = \frac{\langle \ell_\alpha \ell_A \rangle_{\text{tot}} - \langle \ell_\alpha \rangle_{\text{tot}} \langle \ell_A \rangle_{\text{tot}}}{\sqrt{\langle (\ell_\alpha^2)_{\text{tot}} \rangle_{\text{tot}} - \langle \ell_\alpha \rangle_{\text{tot}}^2} \sqrt{\langle (\ell_A^2)_{\text{tot}} \rangle_{\text{tot}} - \langle \ell_A \rangle_{\text{tot}}^2}}, \tag{3.5}
\]

\[
\langle X \rangle_{\text{tot}} = \sum_{N_A=0}^\infty \sum_{\ell_\alpha=0}^M \sum_{\ell_A=0}^{N_A} XP_{eq}(N_A)P_{\ell_\alpha,\ell_A},
\]

where \( X \) represents any combination of the stochastic variables of interest used in the first part of the equation. We anticipate \( 0 < F < 1 \) for positive feedback, but it is not restricted to this range.

### 3.1.3 Parameter estimation via maximum likelihood

The exercise above ensures that the minimal model of MaxCal with only four parameters is capable of producing the general features of a bimodal system. Next, we proceed to benchmark the performance of the model quantitatively when given a particular stochastic trajectory to characterize. This will allow us to learn about quantitative details of the underlying network by decoding information hidden in the noisy raw trajectory. For example, we may be interested in inferring the effective synthesis/degradation rates or the degree of feedback \( (F_{A\alpha}) \), quantities that are not directly available from the raw experimental trajectory. Below, we provide the framework to quantitatively infer these specific characteristics of a network from the stochastic trajectory.

Consider an experimentally observed trajectory of sufficiently long time \( T \) expressed in the units of the typical time scale \( (\Delta t) \) used for sampling the data. In this intrinsic time unit \( (\Delta t) \), we have \( T + 1 \) frames at which protein number has been recorded. Now consider a particular transition between two subsequent frames, say \( t \) and \( t + 1 \), in which the protein number changed from \( i \) to \( j \). We denote the proba-
bility of this one-step (single-frame) transition as $P(j, t + 1; i, t)$ which is abbreviated as $P_{i \rightarrow j}$. These one-step transition probabilities can be determined from MaxCal as

$$P_{i \rightarrow j} = \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{i} \delta(\ell_\alpha + \ell_A - j) P_{\ell_\alpha, \ell_A},$$  \hspace{1cm} (3.6)$$

where $\delta$ is the Dirac delta function, and $P_{\ell_\alpha, \ell_A}$ are functions of the Lagrange multipliers, described by equation 3.3. The likelihood ($\mathcal{L}$) of observing the experimental trajectory given a specific set of MaxCal parameters ($h_\alpha$, $h_A$, $K_{A\alpha}$, $M$) can then be calculated as

$$\mathcal{L} = \prod_{t=0}^{T-1} P(N_{t+1}, t + 1; N_t, t) = \prod_{\{i \rightarrow j\}} P^{\omega_{i \rightarrow j}}_{i \rightarrow j}.$$  \hspace{1cm} (3.7)$$

where $N_t$ is the number of proteins present in frame $t$, $\omega_{i \rightarrow j}$ is the total number of $i \rightarrow j$ one-step transitions, and the second product is over all possible transitions between different values of $i$ and $j$. As outlined above (equation 3.6), $P_{i \rightarrow j}$’s are determined using MaxCal, hence the likelihood is a function of $h_\alpha$, $h_A$, $K_{A\alpha}$ and $M$. Thus, we can maximize the likelihood of the trajectory to select $h_\alpha$, $h_A$, $K_{A\alpha}$, $M$.

However, experiments (and our Gillespie simulations) have no upper limit on production analogous to $M$ in MaxCal. Rare fluctuations leading to unusually large jumps in protein number (greater than $M$) in one time step will severely penalize the likelihood of parameter values that are otherwise most likely. This discontinuous jump in likelihood will erroneously eliminate the most likely set of parameters. We avoid this problem by calculating transition probabilities over multiple intervals ($m$ frames) for a given set of MaxCal parameters using FSP (see Appendix C). We denote the probability of a multi-step (multiple-frame) transition as $P(j, t + m; i, t)$,
abbreviated as $P_{(i \rightarrow j), m}$. This slightly modifies our likelihood function $\mathcal{L}$ as

$$\mathcal{L} = \prod_{n=1}^{T/m} P(N_{t+m}, t+m; N_t, t = m(n-1)) = \prod_{\{i \rightarrow j\}} P_{(i \rightarrow j), m}^{\omega_{(i \rightarrow j), m}} ,$$

(3.8)

where $\lfloor/\rfloor$ is the standard floor division operation and $\omega_{(i \rightarrow j), m}$ is the total number of $i \rightarrow j$ transitions over $m$ frames. An objective choice of $m$ can be provided by using the average residence time (in frames) in the high and low state. However, our result is not sensitive to the choice of $m$ and is robust for a range of values around the typical value.

### 3.1.4 Dealing with experimental data

While the procedure above is applicable to synthetic data in terms of protein number, typical experimental read-outs are in arbitrary fluorescence units. Furthermore, the amount of fluorescence measured per protein is noisy and requires one to de-convolute fluorescence fluctuations from protein number fluctuations. To mimic typical experimental readouts with these challenges, we use the same synthetic data from the auto-activating circuit, but “corrupt” them to create in silico fluorescence trajectories that are likely to be observed in an experiment. We assume the probability distribution of fluorescence intensity ($I$) measured per protein to be a Gaussian distribution\textsuperscript{139,203,204} centered at $f_0$ with a standard deviation of $\sigma$, i.e. $\langle I \rangle = f_0$ and $\langle I^2 \rangle - \langle I \rangle^2 = \sigma^2$. With this assumption, the fluorescence measured from $N$ proteins would follow a probability distribution that is a convolution of $N$ protein fluorescence distributions leading to a Gaussian distribution with mean $Nf_0$ and variance $N\sigma^2$. To “corrupt” simulated trajectories of protein numbers, we select a fluorescence for each time point from this distribution where the mean and variance depends on the protein number $N$. While the procedure described here assumes the fluorescence per
protein follows a Gaussian distribution, we can use a similar approach for Gamma distributions\textsuperscript{205,206} as well.

With this ‘synthetic fluorescence trajectory’ closely mimicking realistic experimental situations, we propose a strategy of deconvolution in which the fluorescence fluctuation is included when calculating the likelihood of a set of MaxCal parameters. In this approach (termed Parallel Fluorescence-to-Number Conversion, or PFNC), we assume the average and variance of the fluorescence per protein distribution are known (i.e. both \( f_0 \) and \( \sigma \) are given), possibly obtained via low-intensity photobleaching experiments.\textsuperscript{207–219} With this information, we can incorporate the fluorescence distribution into the likelihood function (equation 3.8), modifying it to

\[
\mathcal{L} = \prod_{n=1}^{T/m} \left( \sum_{N_t} \sum_{N_{t+m}} \Phi(N_t | f_t) P(N_{t+m}, t + m; N_t, t = m(n-1)) \Phi(N_{t+m} | f_{t+m}) \right)
\]

where \( f_t \) is the fluorescence at frame \( t \), and \( \Phi(N_t | f_t) \) is the conditional probability that \( N_t \) proteins are present given a fluorescence measurement of \( f_t \). These conditional probabilities are calculated via Bayes’ theorem as

\[
\Phi(N | f) = \frac{P_g(f | N) P_{eq}(N)}{P_{tot}(f)}
\]

where \( P_g(f | N) \) is the Gaussian fluorescence distribution with an average and standard deviation of \( N f_0 \) and \( \sqrt{N}\sigma \) as mentioned above, \( P_{eq}(N) \) is the protein number distribution at relative equilibrium calculated via discrete-time FSP,\textsuperscript{49} and \( P_{tot}(f) \) is the fluorescence probability distribution over the entire trace. The probability \( P(N_{t+m}, t + m; N_t, t) \) is determined as above using MaxCal and is a function of the Lagrange multipliers and \( M \). The new likelihood function (equation 3.9) is then maximized to determine \( h_\alpha, h_A, K_{A\alpha}, \) and \( M \).
3.2 Results and Discussion

3.2.1 MaxCal accurately infers underlying rate parameters

Using the procedures described above, we determine $h_\alpha$, $h_A$, $K_{A\alpha}$ and $M$ for a given stochastic trajectory in terms of either protein number or fluorescence readout. These parameters fully specify the minimal MaxCal model and are capable of making multiple predictions, such as the underlying rate parameters. Effective values for the underlying production and degradation rates can be predicted using the average value of the production and degradation state variables respectively (see equation 3.4). To see how well these inferred rates compare to the true values, we applied our inference method to input trajectories that are approximately 2000 frames long with an intrinsic sampling rate of five minutes ($\Delta t = 300$ s), equivalent to trajectories of seven days. Furthermore, we used 100 such trajectories, equivalent to tracking protein numbers in 100 cells. These numbers were chosen to closely match typical experimental conditions. In order to quantify the variance of the effective rate estimates, we apply our method to ten different sets of these simulations and present the average and standard deviation of the ten sets of predicted rates. Using simulations from the reaction rates listed in Figure 3.1, the predicted values compare well against the “true” values used to generate the synthetic data (see Table 3.1). The robustness of the prediction was further tested by creating synthetic data using different values of $g$, $g^*$, and $r$, and similar accuracies were produced. In addition, the inference scheme was applied to alternate model of positive feedbacks – different from equation 3.1 – to generate the synthetic data, and again, the inferred rates matched well with input values (see Appendix D for details). However, it is important to realize that equation 3.4 is only an approximation to infer the intrinsic production and degradation rates. Thus, it is possible to have deviations between the inferred and true rates – higher than the
ones reported in Table 3.1 – while MaxCal captures the temporal statistics well (e.g. fluctuations in the high/low states and transitions between states).

<table>
<thead>
<tr>
<th></th>
<th>True Values</th>
<th>Inferred Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{\text{eff}}$ (s$^{-1}$)</td>
<td>$5.0 \times 10^{-3}$</td>
<td>$5.6 \pm 0.3 \times 10^{-3}$</td>
</tr>
<tr>
<td>$p_{\text{eff}}^*$ (s$^{-1}$)</td>
<td>$50.0 \times 10^{-3}$</td>
<td>$42.2 \pm 1.8 \times 10^{-3}$</td>
</tr>
<tr>
<td>$r$ (s$^{-1}$)</td>
<td>$1.0 \times 10^{-3}$</td>
<td>$0.93 \pm 0.05 \times 10^{-3}$</td>
</tr>
<tr>
<td>$\tau_{L\rightarrow H}$ (s)</td>
<td>$73.4 \times 10^3$</td>
<td>$93.8 \pm 81.6 \times 10^3$</td>
</tr>
<tr>
<td>$\tau_{H\rightarrow L}$ (s)</td>
<td>$82.7 \times 10^3$</td>
<td>$99.4 \pm 86.8 \times 10^3$</td>
</tr>
<tr>
<td>$S_I$ (bits)</td>
<td>8.95</td>
<td>9.11 \pm 0.05</td>
</tr>
<tr>
<td>$S_h$ (bits)</td>
<td>9.54</td>
<td>9.00 \pm 0.04</td>
</tr>
<tr>
<td>$S_l$ (bits)</td>
<td>6.68</td>
<td>7.67 \pm 0.06</td>
</tr>
<tr>
<td>$S_{cg}$ (bits)</td>
<td>1.03</td>
<td>1.02 \pm 0.01</td>
</tr>
</tbody>
</table>

Table 3.1: Comparison of true rates and predicted rates using MaxCal. The first column reports “true” underlying protein synthesis and degradation rates used to create synthetic input data ($d = 0.2$ s$^{-1}$, $p = 0.02$ s$^{-1}$, $f_d = 5.0 \times 10^{-3}$ s$^{-1}$, $b_d = 50.0$ s$^{-1}$, $f_p = 6.0 \times 10^{-3}$ s$^{-1}$, $b_p = 3.0 \times 10^{-5}$ s$^{-1}$), average residence times in the high and low states, and corresponding path informational entropies. Synthetic input data was recorded at $\Delta t = 300$s. The second column reports the average and standard deviation of the same quantities of interest, but extracted using the MaxCal model on ten sets of synthetic data, each consisting of 100 trajectories of 7 days.

3.2.2 Distributions predicted from MaxCal agree well with data

For a more detailed demonstration of how well MaxCal describes data, we further compared MaxCal predicted distributions to that of the input data (generated from the reaction network in equation 3.1). Figure 3.2A shows that the protein number distribution predicted from MaxCal agrees well with the input data from either model in that the locations and widths of the two peaks are comparable between the two
approaches. Next, we compare the distribution of dwell times predicted by MaxCal to that obtained from the synthetic data from both models. The agreement for the shape of the distribution and the average dwell times in the low and high state (see Figure 3.2B, 3.2C and Table 3.1) are reasonable.

Figure 3.2: Predicted distributions agree well with the “true” distributions. (A) Protein number probability distributions from synthetic input trajectories (blue) and predicted MaxCal trajectories (red). (B) Low state and (C) high state residence time probability distributions for synthetic input trajectories (blue) and predicted MaxCal trajectories (red). Underlying Gillespie reaction rates are the same as those used in Table 3.1 and the extracted MaxCal parameters used are a representative example from the ten sets extracted to make Table 3.1.

The comparisons between “true” and predicted values for multiple observables show that the minimal model of MaxCal with only four parameters can make reasonable predictions for data generated with more complex models (with seven parameters). To further quantify the quality of the parameter extraction and performance of our minimal model against the actual model with more parameters, we compare the informational content in the ‘synthetic’ Gillespie trajectories and trajectories generated by MaxCal using these parameters. We compute path informational entropy as

\[ S_I = - \sum_{i,j} P_i P_{i \rightarrow j} \log_2(P_i P_{i \rightarrow j}), \]  

\[ (3.11) \]
where $P_i$ is the probability of having $i$ proteins in the system and $P_{i \rightarrow j}$ is the probability of transitioning from $i$ proteins to $j$ proteins after a single frame. If our MaxCal model is too simple and cannot adequately capture the dynamics of the Gillespie trajectories used, its $S_I$ will be notably different from that of the Gillespie model. We find that the MaxCal model selected by ML has only a 1.8% difference in path informational entropy compared to the ‘synthetic’ input data from Gillespie simulations (see Table 3.1). This provides quantitative verification that the minimal constraints used in equation 3.2 are sufficient to describe the auto-activating circuit modeled here.

The overall path entropy has contributions from three types of fluctuations: i) within the high state, ii) within the low state, and iii) transitions between the high and low states. In order to further explore how MaxCal generated path entropy captures details of these fluctuations, we compute three additional path entropies: $S_h$, $S_l$, and $S_{cg}$. $S_h$ and $S_l$ are computed in the same fashion as $S_I$, but only consider parts of the trajectory in the high state and low state respectively. To assign parts of a trajectory to the low and high state, the locations of the low and high state peaks are used as thresholds ($N = 5$ and $N = 50$ in the case of the Gillespie distribution (blue) of Figure 3.2A). Once the protein level is less (greater) than or equal to the lower (upper) threshold, the system is considered to be in the low (high) state. It then remains in that state until it reaches the opposite threshold. To measure $S_{cg}$, the trajectory is first coarse-grained into a binary trajectory between the low state ($N_{cg} = 0$) and the high state ($N_{cg} = 1$). $S_{cg}$ is then calculated in the same manner as equation 3.11. We find that MaxCal generated estimates of $S_h$ and $S_{cg}$ are in excellent agreement with the input data, while $S_l$ differs by slightly less than 15% from the input (see Table 3.1). The analysis above provides a quantitative measure of performance for MaxCal with a given set of constraints. These measures can be further used to determine the need for incorporating higher order combinations of
the state variables to the caliber function (equation 3.2, e.g. \( \langle \ell \alpha \ell^2 \rangle_A \), \( \langle \ell^2 \alpha \ell \rangle_A \) etc.) to develop models of higher complexity.\textsuperscript{197}

### 3.2.3 MaxCal provides an effective feedback parameter for the circuit

We also extract the effective feedback parameter, \( F_{A\alpha} \), using equation 3.6. As a demonstration of its usefulness, if we compared the MaxCal parameters extracted from experimental traces with varying concentrations of inducer,\textsuperscript{38} \( F_{A\alpha} \) would be expected to vary proportionately, representing the degree of coupling between the production of \( A \) and concentration of \( A \). To mimic the effect of varying inducer concentrations, we generated synthetic data with higher or lower promoter binding rates, \( f_p \), to effectively increase or decrease the amount of self promotion in the system. Next, we applied our MaxCal framework to these trajectories with different levels of self promotion. Figure 3.3A-C shows that MaxCal reproduces comparable protein number distributions regardless of the degree of self-promotion, and as seen in the titles above each panel, the effective feedback \( F_{A\alpha} \) changes accordingly.

Estimating the effective feedback parameter can be important, as it determines the onset of bimodality from unimodality as well as the relative population in the high and low states. Bimodal protein distributions and stochastic switching between the two states often dictate phenotypic variability, a characteristic of bet-hedging strategies used by microbes to evade stress such as antibiotic.\textsuperscript{9,38,221} Consequently, different strains that have evolved under different selection pressures may differentially tune their level of feedback.\textsuperscript{222} Similarly, it may be interesting to see whether strains using ‘resistance’ or ‘tolerance’ mechanisms to evade antibiotics\textsuperscript{223} evolve their feedback parameters differently. Applying MaxCal on experimental trajectories of different

74
Figure 3.3: **MaxCal can capture distributions under varying degrees of feedback.** Protein number probability distributions from synthetic input trajectories (blue) and predicted MaxCal trajectories (red) for different levels of self-promotion, specifically (A) $f_p = 0.5 \times 10^{-3} \text{ s}^{-1}$, (B) $f_p = 2.0 \times 10^{-3} \text{ s}^{-1}$, and (C) $f_p = 6.0 \times 10^{-3} \text{ s}^{-1}$. All other underlying Gillespie reaction rates are the same as those used in Table 3.1.

strains evolved under different conditions to infer these feedback parameters can give us further insights into evolvability and selection. Similarly, this metric can be useful when describing circuits with negative feedback as well.

The ability to extract an effective feedback parameter is a special feature of MaxCal that provides a coarse-grain description of feedback. This is in contrast to traditional parameterization schemes that invoke auxiliary species and multiple reactions involving many parameters to describe feedback. As a result, MaxCal can provide a model with less parameters compared to traditional bottom-up approaches. This is true even when describing circuits with multiple species beyond the single-gene expression circuit used in this study.\textsuperscript{34,44} The success of MaxCal presented here demonstrates that hidden species such as RNA do not significantly alter the predictive power of MaxCal. It is also important to note that MaxCal is exactly equivalent to the master equation when describing systems without feedback, e.g. biochemical cycles where states interconvert amongst themselves.\textsuperscript{32,34,152}
3.2.4 MaxCal can be applied when dealing with noisy fluorescence trajectories

The results above illustrate applicability of MaxCal when experimental trajectories are expressed in protein number fluctuations. We now proceed to demonstrate applicability of MaxCal when data is reported in noisy fluorescence trajectories instead of protein number trajectories. We use the PFNC methodology described earlier to infer the underlying model from the noisy data. Figure 3.4 and Table 3.2 show the performance of these strategies tested against “corrupted” synthetic data and PFNC infers rates with reasonable accuracy even when noise is as high as 200% (see bottom rows in Table 3.2). Success of PFNC is further demonstrated by comparing ‘true’ and predicted distributions of protein numbers and dwell times (see Figure 3.4) at this level of noise. PFNC performs at this level due to the incorporation of fluorescence fluctuation within its ML procedure. While the above results were extracted from data using a Gaussian fluorescence distribution, we carried out similar exercises using a Gamma distribution for the fluorescence per protein\cite{205,206} and similar accuracies were produced. This highlights the need for carrying out controlled photobleaching experiments to learn about the average as well as the noise in the fluorescence per protein distribution to faithfully infer underlying dynamics. In summary, the exercise above demonstrates broad applicability of MaxCal, even when experimental data is not in protein number but in fluorescence with high fluctuation.

3.3 Conclusion

We use the principle of Maximum Caliber (MaxCal) – akin to the principle of maximum entropy applied to describe path probabilities – to model protein number fluctuations as observed in genetic circuits. We demonstrate the application of Max-
Table 3.2: **Effective rates from fluorescence trajectories.** The first row reports “true” underlying protein synthesis and degradation rates used to create synthetic input data (same rates and conditions as Table 3.1). The second row reports the average and standard deviation of MaxCal-inferred rates when trajectories are in protein number. Rows 3-6 report extracted rates for synthetically corrupted trajectories generated using different levels of noise in fluorescence per protein compared to the average (indicated in column 1).

<table>
<thead>
<tr>
<th>Noise</th>
<th>$p_{\text{eff}}$ (s$^{-1}$)</th>
<th>$p'_{\text{eff}}$ (s$^{-1}$)</th>
<th>$r$ (s$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>$5.0 \times 10^{-3}$</td>
<td>$50.0 \times 10^{-3}$</td>
<td>$1.0 \times 10^{-3}$</td>
</tr>
<tr>
<td>0%</td>
<td>$5.6 \pm 0.3 \times 10^{-3}$</td>
<td>$42.2 \pm 1.8 \times 10^{-3}$</td>
<td>$0.93 \pm 0.05 \times 10^{-3}$</td>
</tr>
<tr>
<td>50%</td>
<td>$5.2 \pm 0.2 \times 10^{-3}$</td>
<td>$41.0 \pm 2.0 \times 10^{-3}$</td>
<td>$0.91 \pm 0.05 \times 10^{-3}$</td>
</tr>
<tr>
<td>100%</td>
<td>$5.5 \pm 0.3 \times 10^{-3}$</td>
<td>$47.8 \pm 2.6 \times 10^{-3}$</td>
<td>$1.08 \pm 0.06 \times 10^{-3}$</td>
</tr>
<tr>
<td>150%</td>
<td>$5.9 \pm 0.3 \times 10^{-3}$</td>
<td>$54.8 \pm 3.9 \times 10^{-3}$</td>
<td>$1.25 \pm 0.08 \times 10^{-3}$</td>
</tr>
<tr>
<td>200%</td>
<td>$5.9 \pm 0.3 \times 10^{-3}$</td>
<td>$58.2 \pm 3.6 \times 10^{-3}$</td>
<td>$1.35 \pm 0.07 \times 10^{-3}$</td>
</tr>
</tbody>
</table>

Figure 3.4: **Predicted distributions from fluorescence trajectories with 200% noise using PFNC.** (A) Protein number probability distributions from ‘true’ synthetic input trajectories (blue) and MaxCal trajectories using PFNC inference strategy (red). Comparisons between ‘true’ and PFNC are also shown for (B) low state and (C) high state residence time probability distributions. Underlying Gillespie reaction rates are the same as those used in Table 3.2.

Cal in a positive feedback circuit, a common motif in many naturally occurring and synthetic circuits. Specifically, we consider a single-gene auto-activating circuit where a minimal model based on MaxCal was developed with three physical constraints: protein synthesis, protein degradation, and positive feedback. Through this analysis,
we make four key conclusions. First, the minimal model is capable of producing the switch-like behavior of the circuit. Second, the model shows its usefulness to quantitatively infer underlying parameters. To mimic raw data from experiment, synthetic data was generated using a Gillespie algorithm with a known reaction network model to produce trajectories of fluctuating protein numbers. MaxCal correctly infers underlying rates when compared to the ‘known’ values. Furthermore, MaxCal-predicted distributions agree well with the ones derived from the input data. Third, MaxCal provides an effective feedback parameter to characterize these circuits that can be useful for circuit design as well as analysis of differently evolved strains. Finally, we show how similar methods can be applied when the raw trajectory is in fluorescence rather than protein number, a typical attribute of experimental data. We demonstrate this by “corrupting” the same synthetic protein number trajectories with Gaussian fluctuation to create noisy fluorescence trajectories. Even in the regime of high fluorescence noise, we show that the integrated approach of PFNC can infer underlying rates and distributions of observables by including both MaxCal-generated transition probabilities and fluorescence fluctuation in a model’s likelihood. The method presented here demonstrates the potential application of MaxCal on broader problems in gene networks involving feedback, even when data is presented in fluorescence.
Chapter 4

Maximum Caliber Applied to
Toggle Switch Networks

In the previous chapter, we successfully demonstrated the predictive capabilities of MaxCal on a single-gene auto-activating circuit with synthetic time traces generated from a known reaction network with five species: activated and inactivated promoter, mRNA, protein monomer, and protein dimer. We showed that it is possible to infer underlying rate parameters with the stochastic trajectory of only one of these species, specifically the information about the protein monomer. New questions now emerge: how likely are we to succeed with MaxCal when the number of hidden species increases further? For example, can we describe more complex circuits with multiple genes expressing multiple proteins that interact with each other? To test the applicability of MaxCal to systems with increasing numbers of hidden species that dictate the dynamics of the protein of interest, we generate stochastic time trajectories for four specific models. The first model (M1) simulates positive feedback in a two-gene mutually repressing circuit – known as a toggle switch (TS)\textsuperscript{40,42,44} – with mRNA dynamics included and provides the fluctuating time traces of both proteins for use in
the MaxCal inference machinery. In the second model (M2), we test MaxCal’s ability to infer underlying rate parameters when the stochastic trajectories in M1 are given in fluorescence rather than protein number, mimicking experimental conditions. In the third model (M3), we use the same TS circuit but with the temporal information of only one protein (expressed in protein number). Thus, M3 strongly tests the scalability of MaxCal inference by using the same TS circuit but providing information on only one of the proteins with six other species remaining hidden. In the fourth model (M4), we further challenge MaxCal by giving it the stochastic trajectory from M3 in fluorescence rather than protein number. The underlying parameters used to generate these synthetic time traces, thereby considered as “true” values, are compared against MaxCal inferred parameters for benchmarking. We find that in all four cases, MaxCal performs reasonably well. This demonstrates the potential of MaxCal to infer information about network details that are not directly measurable in experiments, even when there are many hidden species compared to available data and even when they are further convolved with fluorescence noise. In the next section, we describe the details of the underlying reaction networks as well as our methodology for different models.

4.1 Materials and Methods

4.1.1 Generating synthetic data for TS

Synthetic biologists are building small motifs to achieve specific functions and test different hypotheses.\textsuperscript{35–38,181,186,224,225} One such motif is the two-gene mutually repressing TS circuit, and in the following models, we test MaxCal’s inferential power with this circuit. In a previous work,\textsuperscript{44} we showed that MaxCal can capture fluctuations in TS using synthetic data. However, we neglected the role of mRNA in
the generation of synthetic data and ignored the impact of experimentally relevant sampling times ($\Delta t$) on our inference methods. Similar to M1, we now explicitly incorporate mRNA using the following reaction scheme:\textsuperscript{42}

\begin{align}
\alpha & \xrightarrow{g} \alpha + a ; \quad a \xrightarrow{d} \emptyset \\
a & \xrightarrow{p} a + A ; \quad A \xrightarrow{r} \emptyset \\
\alpha + A & \xrightarrow{f_b} \alpha^* ; \quad \alpha^* \xrightarrow{g} \alpha^* + a ; \quad \alpha^* \xrightarrow{g^*} \alpha^* + b \\
\alpha & \xrightarrow{g} \alpha + b ; \quad b \xrightarrow{d} \emptyset \\
b & \xrightarrow{p} b + B ; \quad B \xrightarrow{r} \emptyset \\
\alpha + B & \xrightarrow{f_b} \alpha' ; \quad \alpha' \xrightarrow{g^*} \alpha' + a ; \quad \alpha' \xrightarrow{g} \alpha' + b
\end{align}

where generic mRNA’s $a$ and $b$ are created from the gene $\alpha$ (at different loci) at a rate of $g$ and are degraded at a rate of $d$. These mRNAs produce their respective proteins $A$ and $B$ at a rate of $p$ while $A$ and $B$ degrade at a rate of $r$. To incorporate feedback, either protein can bind and unbind the promoter site at rates of $f$ and $b$ respectively, sending the promoter into and out of different expression states ($\alpha^*$ for $A$ and $\alpha'$ for $B$). In $\alpha^*$, the gene creates mRNA $b$ at a much slower rate $g^*$ while $\alpha'$ creates $a$ at the slower rate $g^*$, capturing the essentials of mutual repression. Rates are again chosen to produce representative switching times\textsuperscript{38} while maintaining realistic synthesis and degradation rates.\textsuperscript{22} A Gillespie algorithm\textsuperscript{27} is then used to generate stochastic trajectories of protein levels that exhibit bistability. M1, M2, M3, and M4 use the same underlying circuit to generate the stochastic time traces, but M1 and M2 provide time trajectories of both proteins $A$ and $B$ to MaxCal for inference. To further challenge MaxCal, M3 and M4 provide the noisy protein number/fluorescence trajectory of only one protein to mimic a situation where only one of the proteins may be tagged.
4.1.2 MaxCal model for TS

For TS, two genes express two different proteins, and consequently, we extend our MaxCal model from chapter 3 to two proteins. For the time interval $\Delta t$, we utilize two production/birth variables, $\ell_\alpha$ and $\ell_\beta$, ranging from zero to $M$ for protein $A$ and $B$ respectively. Likewise, we define $\ell_A$ and $\ell_B$ to denote the number of preexisting $A$ and $B$ molecules respectively that remain after the time interval $\Delta t$. Accordingly, $0 \leq \ell_A \leq N_A$ and $0 \leq \ell_B \leq N_B$ with $N_A$ and $N_B$ being the initial number of $A$ and $B$ proteins present respectively. Lagrange multipliers $h_\alpha$, $h_\beta$, $h_A$, and $h_B$ are used to constrain the average values of $\ell_\alpha$, $\ell_\beta$, $\ell_A$, and $\ell_B$. However, we assume symmetry in $A$ and $B$ by setting $h_\alpha = h_\beta$ and $h_A = h_B$. Next, the constraint of mutual repression is imposed by coupling the production variable of $A$ with the degradation variable of $B$ and vice versa. Specifically, we constrain the average of $\ell_\alpha \ell_B$ and $\ell_\beta \ell_A$ using the corresponding Lagrange multiplier $K_{A\beta}$, identical to earlier work on TS using MaxCal. However, in contrast to the previous model, we now add auto-feedback by further constraining the average of $\ell_\alpha \ell_A$ and $\ell_\beta \ell_B$ with the Lagrange multiplier $K_{A\alpha}$. The rationale for this additional term in the Hamiltonian is twofold: i) mutual repression implies auto-activation. We intend to infer model parameters for experimentally relevant values of sampling time $\Delta t$ (on the order of minutes), and for such relatively large time intervals, one is likely to witness the impact of auto-activation. ii) Perhaps more importantly, the four constraints of $\langle \ell_\alpha \ell_B \rangle$, $\langle \ell_\beta \ell_A \rangle$, $\langle \ell_\alpha \ell_A \rangle$, and $\langle \ell_\beta \ell_B \rangle$ now provide a complete description that includes all possible cross-talks between the birth and degradation variables up to the second order and avoids arbitrarily assigning $K_{A\alpha} = 0$. We intend to let the data select values of $K_{A\alpha}$ within a model that is self-consistent up to the second order. This minimal model of repression

82
and promotion together yield the caliber as

\[ C = - \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} \sum_{\ell_\beta=0}^{M} \sum_{\ell_B=0}^{N_B} P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B} \log P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B} + \frac{h_\alpha}{M} \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} \sum_{\ell_\beta=0}^{M} \sum_{\ell_B=0}^{N_B} (\ell_\alpha + \ell_\beta) P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B} + \frac{h_A}{M} \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} \sum_{\ell_\beta=0}^{M} \sum_{\ell_B=0}^{N_B} (\ell_A + \ell_B) P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B} + K_{A\alpha} \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} \sum_{\ell_\beta=0}^{M} \sum_{\ell_B=0}^{N_B} (\ell_\alpha \ell_A + \ell_\beta \ell_B) P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B} + K_{A\beta} \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} \sum_{\ell_\beta=0}^{M} \sum_{\ell_B=0}^{N_B} (\ell_\beta \ell_A + \ell_\alpha \ell_B) P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B}, \]

where \( P_{\ell_\alpha,\ell_\beta,\ell_A,\ell_B} \) is the probability of obtaining a particular combination of \( \ell_\alpha, \ell_\beta, \ell_A, \) and \( \ell_B \). Upon maximizing the caliber, the path probability \( P_{\ell_\alpha,\ell_\beta,\ell_A,\ell_B} \) becomes

\[ P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B} = Q^{-1} \binom{N_A}{\ell_A} \binom{N_B}{\ell_B} \exp[h_\alpha(\ell_\alpha + \ell_\beta) + h_A(\ell_A + \ell_B) + K_{A\alpha}(\ell_\alpha \ell_A + \ell_\beta \ell_B) + K_{A\beta}(\ell_\beta \ell_A + \ell_\alpha \ell_B)]. \]

\[ Q = \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} \sum_{\ell_\beta=0}^{M} \sum_{\ell_B=0}^{N_B} \binom{N_A}{\ell_A} \binom{N_B}{\ell_B} \exp[h_\alpha(\ell_\alpha + \ell_\beta) + h_A(\ell_A + \ell_B) + K_{A\alpha}(\ell_\alpha \ell_A + \ell_\beta \ell_B) + K_{A\beta}(\ell_\beta \ell_A + \ell_\alpha \ell_B)]. \]

Similar to equation 3.4 for the auto-activating circuit, we can approximate effective protein synthesis rates (unrepressed \( p_{\text{eff}} \) and repressed \( p^*_{\text{eff}} \)) and protein degradation
rates \( r \) for the TS circuit in terms of MaxCal variables:

\[
p_{\text{eff}} = p(a)_{N_H,N_L} = p(b)_{N_L,N_H} \approx P \left( \frac{g^*}{d} \right) \approx \frac{\langle \ell_\alpha \rangle_{N_H,N_L}}{\Delta t} = \frac{\langle \ell_\beta \rangle_{N_L,N_H}}{\Delta t},
\]

\[
p_{\text{eff}}^* = p(a)_{N_L,N_H} = p(b)_{N_H,N_L} \approx P \left( \frac{g^*}{d} \right) \approx \frac{\langle \ell_\alpha \rangle_{N_L,N_H}}{\Delta t} = \frac{\langle \ell_\beta \rangle_{N_H,N_L}}{\Delta t},
\]

\[
r_A(N_A, N_B) = \frac{N_A - \langle \ell_A \rangle_{N_A,N_B}}{N_A \Delta t}, \quad r_B(N_A, N_B) = \frac{N_B - \langle \ell_B \rangle_{N_A,N_B}}{N_B \Delta t},
\]

\[
r \approx \sum_{N_A=0}^{\infty} \sum_{N_B=0}^{\infty} P_{eq}(N_A, N_B)r_A(N_A, N_B) = \sum_{N_A=0}^{\infty} \sum_{N_B=0}^{\infty} P_{eq}(N_A, N_B)r_B(N_A, N_B),
\]

where \( \langle \cdots \rangle_{i,j} \) represents the average of a quantity given that there are \( i \) and \( j \) number of \( A \) and \( B \) proteins initially present respectively, \( N_L \) and \( N_H \) are the most likely number of proteins in the repressed and unrepressed state respectively, and \( P_{eq}(i,j) \) is the probability of having \( N_A = i \) and \( N_B = j \) at relative equilibrium (calculated via discrete-time FSP;\(^{49}\) see Appendix C).

With two proteins present in the system, two complementary feedback metrics can be defined. Auto-feedback, \( F_{A\alpha} \), is defined as

\[
F_{A\alpha} = \frac{\langle \ell_\alpha \ell_A \rangle_{\text{tot}} - \langle \ell_\alpha \rangle_{\text{tot}} \langle \ell_A \rangle_{\text{tot}}}{\sqrt{(\langle \ell_\alpha^2 \rangle_{\text{tot}} - \langle \ell_\alpha \rangle_{\text{tot}}^2)(\langle \ell_A^2 \rangle_{\text{tot}} - \langle \ell_A \rangle_{\text{tot}}^2)}} = \frac{\langle \ell_\beta \ell_B \rangle_{\text{tot}} - \langle \ell_\beta \rangle_{\text{tot}} \langle \ell_B \rangle_{\text{tot}}}{\sqrt{(\langle \ell_\beta^2 \rangle_{\text{tot}} - \langle \ell_\beta \rangle_{\text{tot}}^2)(\langle \ell_B^2 \rangle_{\text{tot}} - \langle \ell_B \rangle_{\text{tot}}^2)}},
\]

(4.5)

where

\[
\langle X \rangle_{\text{tot}} = \sum_{N_A=0}^{\infty} \sum_{N_B=0}^{\infty} \sum_{\ell_\alpha=0}^{M} \sum_{\ell_\beta=0}^{M} \sum_{\ell_A=0}^{N_A} \sum_{\ell_B=0}^{N_B} X P_{eq}(N_A, N_B) P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B}
\]

(4.6)

with \( X \) being any combination of the stochastic variables. Conversely, inter-feedback, \( F_{A\beta} \), is defined as

\[
F_{A\beta} = \frac{\langle \ell_\beta \ell_A \rangle_{\text{tot}} - \langle \ell_\beta \rangle_{\text{tot}} \langle \ell_A \rangle_{\text{tot}}}{\sqrt{(\langle \ell_\beta^2 \rangle_{\text{tot}} - \langle \ell_\beta \rangle_{\text{tot}}^2)(\langle \ell_A^2 \rangle_{\text{tot}} - \langle \ell_A \rangle_{\text{tot}}^2)}} = \frac{\langle \ell_\alpha \ell_B \rangle_{\text{tot}} - \langle \ell_\alpha \rangle_{\text{tot}} \langle \ell_B \rangle_{\text{tot}}}{\sqrt{(\langle \ell_\alpha^2 \rangle_{\text{tot}} - \langle \ell_\alpha \rangle_{\text{tot}}^2)(\langle \ell_B^2 \rangle_{\text{tot}} - \langle \ell_B \rangle_{\text{tot}}^2)}},
\]

(4.7)
For the TS circuit, we expect $0 < F_{A\alpha} < 1$ and $-1 < F_{A\beta} < 0$, but they are not restricted to these ranges.

### 4.1.3 Estimating parameters from stochastic trajectories in M1

To objectively select representative parameter values for a given stochastic TS trajectory (equation 4.1) using ML, we must first describe the likelihood of that trajectory in terms of these parameters. To start, we again discretize the trajectory into $T + 1$ frames with a sampling interval of $\Delta t$. With $i$, $j$ as the number of $A$ proteins at two subsequent frames (say $t$ and $t + 1$) and $k$, $l$ as the number of $B$ proteins at the same two frames, the transition probability $P(j, l, t + 1; i, k, t)$ is then given by

$$P(j, l, t + 1; i, k, t) = \sum_{\ell_\alpha = 0}^{M} \sum_{\ell_A = 0}^{i} \sum_{\ell_\beta = 0}^{M} \sum_{\ell_B = 0}^{k} \delta(\ell_\alpha + \ell_A - j)\delta(\ell_\beta + \ell_B - l)P_{\ell_\alpha, \ell_A, \ell_\beta, \ell_B}. \quad (4.8)$$

Similar to equation 3.8, we must account for rare fluctuations producing more than $M$ proteins in a single frame, so we calculate transition probabilities over $m$ frames as $P(j, l, t + m; i, k, t)$, abbreviated as $P_{(i \to j), (k \to l), m}$, using discrete-time FSP (see Appendix C). This yields the likelihood function as

$$\mathcal{L} = \frac{T}{m} \prod_{n=1}^{T/m} P(N_{A,t+m}, N_{B,t+m}, t + m; N_{A,t}, N_{B,t}, t = m(n-1)) = \prod_{\{i \to j, k \to l\}} P_{\omega_{(i \to j), (k \to l), m}}^{(i \to j), (k \to l), m}, \quad (4.9)$$

where $\lfloor \cdot \rfloor$ is the standard floor division operation, $\omega_{(i \to j), (k \to l), m}$ is the total number of simultaneous $i \to j$ and $k \to l$ transitions over $m$ frames, and $N_{A,t}$ ($N_{B,t}$) denote the number of $A$ ($B$) proteins at frame $t$. The likelihood is then maximized with respect
to \( h_\alpha, h_A, K_{A\alpha}, K_{A\beta} \), and \( M \) to select representative values. As in auto-activation inference, \( m \) is chosen to be the number of frames equivalent to the average residence time between states.

### 4.1.4 Estimating parameters from stochastic trajectories in M3

When considering the likelihood of only one of the stochastic protein expression traces (say protein \( A \)), the two protein joint transition probabilities must be summed over all possible starting values of \( N_B \) and all possible ending values of \( N_B \) producing a new transition probability of

\[
P_2(j, t + 1; i, t) = P_{2,(i\rightarrow j)} = \sum_{N_B=0}^{\infty} \sum_{\ell_A=0}^{M} \sum_{\ell_\beta=0}^{M} P_{\text{eq}}(N_B|i) \delta(\ell_A + \ell_A - j) P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B}
\]

(4.10)

where \( P_{\text{eq}}(N_B|i) \) is the conditional probability of having \( N_B \) number of B proteins given that \( N_A = i \). As before, we must use multi-frame transitions to address rare fluctuations, so we calculate transition probabilities over \( m \) frames as \( P_{2}(j, t+m; i, t) \), abbreviated as \( P_{2,(i\rightarrow j),m} \), using discrete-time FSP (see Appendix C). With this reduction, the likelihood of observing a particular expression trace, given that two protein species are present and interacting, can be calculated as

\[
\mathcal{L} = \prod_{n=1}^{N} P_{2}(N_{A,t+m}, t + m; N_{A,t}, t = m(n - 1)) = \prod_{\{i\rightarrow j\}} P_{2^\omega(i\rightarrow j),m}
\]

(4.11)

where \( \omega_{(i\rightarrow j),m} \) has the same definition as it does in equation 4.9. Like all previous models, \( \mathcal{L} \) is now a function of \( h_\alpha, h_A, K_{A\alpha}, K_{A\beta}, \) and \( M \), and can be maximized with respect to these parameters to determine the representative values.
4.1.5 Creating and analyzing synthetic fluorescence trajectories for M2 and M4

The models described above rely on the input trajectory being presented in discrete protein numbers. However, gene expression is typically measured experimentally by fluorescently labeled reporter proteins. The amount of fluorescence measured per protein, defined as \( f \), is not a fixed number but rather has a distribution of typical values. Consequently, the measured gene expression trajectory is a convolution of the protein number fluctuation inherent to the circuit topology and the fluorescence fluctuation inherent to the experiment. This is a fundamental challenge in data analysis.

In the previous chapter, we established a novel way to address this issue by using a formalism called parallel fluorescence-to-number conversion, or simply PFNC. In PFNC, the likelihood function is constructed by considering the fluorescence per protein distribution along with the protein number transition probabilities. As a result, the likelihood function describes the probability of observing a particular fluorescence trajectory rather than a protein number trajectory and can be directly applied to experimental data. We now extend PFNC to the TS framework in M1 and M3 using fluorescence trajectories of both proteins in the case of M2 and only one of the proteins in M4.

To test MaxCal’s inference capabilities in the case of such limited and convoluted information, we first created synthetic fluorescence trajectories by purposefully “corrupting” the same stochastic protein number trajectories used in M1 and M3. For simplicity, we assume the fluorescence per protein distribution is approximately Gaussian,\(^{139,203,204}\) with an average and standard deviation of \( f_0 \) and \( \sigma \) respectively. Using this assumption, the fluorescence at each time point (having \( N_t \) proteins) is then randomly selected using a convolution of \( N_t \) distributions of fluorescence per
protein, i.e. another Gaussian with an average and standard deviation of $N_t f_0$ and $\sqrt{N_t} \sigma$ respectively. For illustrative purposes, we will demonstrate PFNC on an experimentally relevant noise level (defined as the ratio of $\sigma/f_0$) of 30%.\textsuperscript{139,203,204}

To analyze this synthetic data generated in arbitrary fluorescence, we now construct the corresponding likelihood function by combining protein number path probabilities with the fluorescence per protein distribution. We assume that this fluorescence per protein distribution is known via low-intensity photobleaching experiments.\textsuperscript{207–219} Using similar notation as equation 4.9, the likelihood of observing a given fluorescence trajectory is now calculated as

$$
\mathcal{L} = \prod_{n=1}^{N} \left( \sum_{N_{A,t}=0}^{\infty} \sum_{N_{A,t+m}=0}^{\infty} \sum_{N_{B,t}=0}^{\infty} \sum_{N_{B,t+m}=0}^{\infty} \Phi(N_{A,t}|f_{A,t})\Phi(N_{B,t}|f_{B,t}) \right.
\times P(N_{A,t+m}, N_{B,t+m}, t+m; N_{A,t}, N_{B,t}, t = m(n-1))
\times \Phi(N_{A,t+m}|f_{A,t+m})\Phi(N_{B,t+m}|f_{B,t+m})
$$

where $f_{A,t}$ ($f_{B,t}$) is the amount of fluorescence produced by $A$ ($B$) proteins at frame $t$ and $\Phi(N|f)$ is the conditional probability that $N$ proteins are present given that a fluorescence of $f$ has been observed. These conditional probabilities are calculated via Bayes’ theorem as

$$
\Phi(N|f) = \frac{P_g(f|N) P_{eq}(N)}{P_{tot}(f)}
$$

where $P_g(f|N)$ is the Gaussian fluorescence distribution with an average and standard deviation of $N f_0$ and $\sqrt{N} \sigma$ as mentioned above, $P_{eq}(N)$ is the protein number distribution at relative equilibrium calculated via discrete-time FSP,\textsuperscript{49} and $P_{tot}(f)$ is the fluorescence probability distribution over the entire trace. Using a similar approach, equation 4.11 can be extended to describe the likelihood of observing the fluorescence
trajectory of a single protein as

\[ \mathcal{L} = \prod_{n=1}^{N} \left( \sum_{N_{A,t}=0}^{\infty} \sum_{N_{A,t+m}=0}^{\infty} \Phi(N_{A,t}|f_t) \right) \]

\[ \times P_2(N_{A,t+m}; t + m; N_{A,t}, t = m(n-1))\Phi(N_{A,t+m}|f_{t+m}) \] (4.14)

As in the preceding models, \( \mathcal{L} \) is then maximized with respect to \( h_\alpha, h_A, K_{A\alpha}, K_{A\beta}, \) and \( M \) to determine representative values in both cases.

### 4.2 Results and Discussion

#### 4.2.1 MaxCal accurately infers underlying rates and observables for TS using both protein trajectories (M1)

To further test MaxCal’s ability to model and infer underlying details of more complex circuits, we analyze the TS circuit (modeled using equation 4.1) using the stochastic trajectories of both proteins, \( A \) and \( B \). It is important to note that the present work differs from the previous work\(^{44} \) in four aspects: i) the earlier work did not use synthetic trajectories from a reaction network with mRNA being explicitly present; ii) the sampling time \( (\Delta t) \) was much smaller than typical experimental values; iii) for simplification, it was assumed that \( K_{A\alpha} = 0 \) in equation 4.2; and iv) the inference was based on moments of observables rather than the likelihood of the entire trajectory, causing data-reduction. The caliber is now constructed using the most general set of constraints and the likelihood function (equation 4.9) allows us to directly infer from trajectories, fully utilizing the entire data content. As before, the inference method was applied to ten replicates of 100 trajectories with enough frames to equate to seven days at a sampling rate of five minutes per frame \( (\Delta t = 300 \text{ s}) \).
Using the extracted parameters from each replicate, we computed the average and variance of the effective rates (using equation 4.4) and feedback parameters (using equations 4.5 and 4.7). Comparison of the effective rates and feedbacks to the “true” values from the stochastic input data (table 4.1) demonstrates that MaxCal accurately and consistently infers underlying details even for realistic sampling times. It is also important to note that $K_{A\alpha}$ is non-zero, suggesting a non-trivial role of this constraint in the inference process. Overall, mutual repression should produce net self-promotion. This is reflected in the two feedback terms: $F_{A\beta}$ shows a strong negative correlation between the presence of one protein and the production of the other, while $F_{A\alpha}$ yields a positive value.

<table>
<thead>
<tr>
<th></th>
<th>True Values</th>
<th>M1</th>
<th>M2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{\text{eff}}$ (s$^{-1}$)</td>
<td>$20.0 \times 10^{-3}$</td>
<td>$13.9 \pm 0.9 \times 10^{-3}$</td>
<td>$14.7 \pm 1.4 \times 10^{-3}$</td>
</tr>
<tr>
<td>$p^*_{\text{eff}}$ (s$^{-1}$)</td>
<td>$0.1 \times 10^{-3}$</td>
<td>$0.15 \pm 0.02 \times 10^{-3}$</td>
<td>$0.16 \pm 0.03 \times 10^{-3}$</td>
</tr>
<tr>
<td>$r$ (s$^{-1}$)</td>
<td>$1.0 \times 10^{-3}$</td>
<td>$0.74 \pm 0.05 \times 10^{-3}$</td>
<td>$0.77 \pm 0.06 \times 10^{-3}$</td>
</tr>
<tr>
<td>$\tau$ (s)</td>
<td>$110.4 \times 10^{3}$</td>
<td>$114.8 \pm 96.5 \times 10^{3}$</td>
<td>$109.6 \pm 95.7 \times 10^{3}$</td>
</tr>
<tr>
<td>$S_I$ (bits)</td>
<td>11.1</td>
<td>$11.4 \pm 0.1$</td>
<td>$11.3 \pm 0.2$</td>
</tr>
<tr>
<td>$S_s$ (bits)</td>
<td>10.2</td>
<td>$10.5 \pm 0.2$</td>
<td>$10.4 \pm 0.2$</td>
</tr>
<tr>
<td>$S_{cg}$ (bits)</td>
<td>1.02</td>
<td>$1.02 \pm 0.00$</td>
<td>$1.02 \pm 0.00$</td>
</tr>
<tr>
<td>$F_{A\alpha}$</td>
<td>0.086 $\pm 0.016$</td>
<td>0.105 $\pm 0.011$</td>
<td></td>
</tr>
<tr>
<td>$F_{A\beta}$</td>
<td>$-0.396 \pm 0.017$</td>
<td>$-0.421 \pm 0.014$</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1: Comparison of true and predicted rates and metrics from MaxCal using both trajectories (M1 & M2). The first column reports the true underlying protein production and degradation rates used to generate the synthetic input data ($g = 0.5$ s$^{-1}$, $g^* = 2.5 \times 10^{-3}$ s$^{-1}$, $d = 0.5$ s$^{-1}$, $p = 0.02$ s$^{-1}$, $f = 1.0 \times 10^{-3}$ s$^{-1}$, $f^* = 3.5 \times 10^{-6}$ s$^{-1}$, $b = 2.0 \times 10^{-5}$ s$^{-1}$), the average dwell time in either state, and the corresponding informational entropies. Synthetic data are recorded with $\Delta t = 300$ s. The second and third columns report the average and standard deviation of the same quantities of interest as well as the effective feedback, but extracted using models M1 and M2 on 10 sets of synthetic data, each consisting of 100 trajectories of 7 days.
Comparison of different entropies further highlights MaxCal’s ability to well approximate the inherent fluctuations in the synthetic data. Due to the presence of another protein, the calculation of the overall path entropy is slightly modified from equation 3.11 to include the other species:

\[
S_I = - \sum_{i,j,k,l} P_{i,k} P_{(i\rightarrow j),(k\rightarrow l)} \log_2 (P_{i,k} P_{(i\rightarrow j),(k\rightarrow l)})
\] (4.15)

where \( P_{i,k} \) is the probability of having \( i \) number of A proteins (i.e. \( N_A = i \)) and \( k \) number of B proteins (i.e. \( N_B = k \)) while \( P_{(i\rightarrow j),(k\rightarrow l)} \) is the probability of simultaneously transitioning \( N_A \) from \( i \) to \( j \) and \( N_B \) from \( k \) to \( l \) in a single frame. Since the two stable states in TS are both symmetric, we only consider a single “state entropy”, \( S_s \). For each state, we calculate the path entropy only considering parts of the trajectory that are in that state and \( S_s \) is reported as the average of those values. \( S_{cg} \) is still calculated in the same binary fashion as before. Again, due to symmetry between the states, we compute the average of the dwell time in either state as \( \tau \).

As seen in table 4.1, the MaxCal inferred entropy values and dwell time values are in good agreement with the input. Next, we compared the protein number distributions of A and B between the “true” synthetic data and the MaxCal generated model. Figure 4.1 demonstrates that the two distributions are in reasonable agreement. We further tested against additional datasets generated using different parameter values and all metrics are again in good agreement, ensuring the robustness of the proposed methodology.
4.2.2 MaxCal accurately infers underlying rates and observables for TS using fluorescence trajectories of both proteins (M2)

Next, we consider typical conditions encountered experimentally when data is not available in protein number but rather in noisy fluorescence. This is particularly challenging because fluorescence per protein is a random variable itself. Using the synthetically generated fluorescence trajectories of both proteins, the inference method was again applied to ten replicates of 100 trajectories with enough frames to equate to seven days at a sampling rate of five minutes per frame ($\Delta t = 300$ s). Using the likelihood function that incorporates this fluorescence fluctuation (equation 4.12), we find that MaxCal inferred rates, entropies, and dwell times are again in good agreement with their “true” values and variances increase slightly but remain moderate (see table 4.1). Figure 4.2 provides a further comparison by plotting the
protein number and dwell time distributions of both the MaxCal inference and the “true” synthetic data, demonstrating a satisfactory overlap.

![Graph](image)

Figure 4.2: **Predicted distributions agree well with the “true” distributions for M2.** Protein number probability distributions (dark color for high probability) from (A) synthetic input trajectories and (B) predicted MaxCal trajectories. (C) Residence time probability distributions for synthetic input trajectories (blue) and predicted MaxCal trajectories (red). Underlying Gillespie reaction rates and fluorescence values are the same as those used in Table 4.1 (representative MaxCal parameters: $h_\alpha = 0.311$, $h_A = 1.493$, $K_{A\alpha} = -0.037$, $K_{A\beta} = -0.248$, $M = 27$).

### 4.2.3 MaxCal accurately infers underlying rates and observables for TS using only one protein trajectory (M3)

Next, we use the same TS circuit described in equation 4.1 but only provide the trajectory of one protein to MaxCal’s inference machinery. Thus, in M3, MaxCal is provided with information about only one species out of a total of seven species, providing a stronger test for MaxCal’s inferential power. While the same caliber function persists (equation 4.2), a new likelihood function accounting for this data reduction (equation 4.11) was used to determine the Lagrange multipliers. Again using ten replicates of 100 trajectories over seven days with $\Delta t = 300$ s, effective rates and entropy values ($S_I, S_s, S_{cg}$) compare well to their “true” values from synthetic data, as seen in table 4.2. The variances in these quantities increase once again
but remain within an acceptable range. We notice that the magnitude of the auto-feedback $F_{A\alpha}$ and the mutual feedback $F_{A\beta}$ are higher compared to M1. This can be attributed to the lack of knowledge about the second trajectory. This encodes direct information about the cross-talk between $A$ and $B$ by providing the joint distribution of both proteins. In M3, we reduce the data given to our inference machinery and consequently miss the direct cross-correlation captured in M1. The protein number distributions of the MaxCal inferred model also agree reasonably well compared to synthetic data (see figure 4.3). These comparisons show that MaxCal can perform well even when sampling time is sufficiently large and the input data is reduced to one protein.

<table>
<thead>
<tr>
<th></th>
<th>True Values</th>
<th>M3</th>
<th>M4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_{\text{eff}}$ (s$^{-1}$)</td>
<td>20.0 × 10$^{-3}$</td>
<td>22.7 ± 3.1 × 10$^{-3}$</td>
<td>25.8 ± 2.1 × 10$^{-3}$</td>
</tr>
<tr>
<td>$p_{\text{eff}}^*$ (s$^{-1}$)</td>
<td>0.1 × 10$^{-3}$</td>
<td>0.22 ± 0.06 × 10$^{-3}$</td>
<td>0.31 ± 0.07 × 10$^{-3}$</td>
</tr>
<tr>
<td>$r$ (s$^{-1}$)</td>
<td>1.0 × 10$^{-3}$</td>
<td>1.15 ± 0.17 × 10$^{-3}$</td>
<td>1.37 ± 0.12 × 10$^{-3}$</td>
</tr>
<tr>
<td>$\tau$ (s)</td>
<td>110.4 × 10$^{3}$</td>
<td>108.2 ± 97.0 × 10$^{3}$</td>
<td>121.1 ± 106.0 × 10$^{3}$</td>
</tr>
<tr>
<td>$S_I$ (bits)</td>
<td>11.1</td>
<td>10.7 ± 0.2</td>
<td>10.3 ± 0.2</td>
</tr>
<tr>
<td>$S_s$ (bits)</td>
<td>10.2</td>
<td>9.7 ± 0.2</td>
<td>9.5 ± 0.2</td>
</tr>
<tr>
<td>$S_{cg}$ (bits)</td>
<td>1.02</td>
<td>1.02 ± 0.00</td>
<td>1.02 ± 0.00</td>
</tr>
<tr>
<td>$F_{A\alpha}$</td>
<td></td>
<td>0.439 ± 0.137</td>
<td>0.710 ± 0.120</td>
</tr>
<tr>
<td>$F_{A\beta}$</td>
<td>-0.622 ± 0.083</td>
<td>-0.781 ± 0.067</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2: **Comparison of true and predicted rates/metrics from MaxCal using both trajectories (M3 & M4).** The first column reports the true underlying protein production and degradation rates used to generate the synthetic input data (same rates as Table 4.1), the average dwell time in either state, and the corresponding informational entropies. Synthetic data are recorded with $\Delta t = 300$ s. The second and third columns report the average and standard deviation of the same quantities of interest as well as the effective feedback, but extracted using models M3 and M4 on a single trace from 10 sets of synthetic data, each consisting of 100 trajectories of 7 days.
4.2.4 MaxCal accurately infers underlying rates and observables for TS using only one fluorescence trajectory (M4)

As in M2, we consider the experimental situation when data is not available in protein number but rather in noisy fluorescence. This time however, we used the synthetically generated fluorescence trajectories of only one of the protein species (say protein \( A \)) to infer underlying details of the model, specifically ten replicates of 100 trajectories over seven days with \( \Delta t = 300 \text{ s} \). With the likelihood function that incorporates this fluorescence fluctuation (equation 4.14), we find that MaxCal inferred rates, entropies, and dwell times are again in good agreement with their “true” values with reasonable variances (see Table 4.2). The feedback strengths continue to increase higher than M1, most likely for the same data reduction as M3 but not helped at all by data convolution with fluorescence fluctuation. Figure 4.4 provides a further comparison by plotting the protein number and dwell time distributions of both the MaxCal inference and the “true” synthetic data, demonstrating a satis-
factory overlap. This exercise confirms that even for a two-gene mutually repressing
circuit with available data on only one protein given in noisy fluorescence trajectory,
MaxCal is still able to capture fluctuations of the system and infer network details to
reasonable accuracy. However, we notice that M3 and M4 do not capture trajectory
fluctuations as well as M1 and M2 (see entropy metrics in tables 4.1 and 4.2). This
is either due to loss of information (in M3 and M4) or convolution with fluorescence
(in M4). These issues may be augmented when analyzing data for circuits that have
moderate repression (low $F_{A\beta}$) but strong auto-activation (high $F_{A\alpha}$). With these
conditions, M3 and M4 may yield $K_{A\beta} = 0$ ignoring the cross-talk between $A$ and $B$
due to a lack of information. Based on this, it is always advisable to tag both proteins
if their cross-talk is known.

![Figure 4.4: Predicted distributions agree well with the “true” distributions for M4. Protein number probability distributions (dark color for high probability) from (A) synthetic input trajectories and (B) predicted MaxCal trajectories. (C) Residence time probability distributions for synthetic input trajectories (blue) and predicted MaxCal trajectories (red). Underlying Gillespie reaction rates are the same as those used in Table 4.1 (representative MaxCal parameters: $h_\alpha = 0.438$, $h_A = 0.657$, $K_{A\alpha} = -0.020$, $K_{A\beta} = -0.264$, $M = 11.$)](image)

96
4.3 Conclusion

In summary, we have substantially extended the scope of MaxCal to analyze gene networks beyond previous efforts. We applied MaxCal to synthetic data generated from a two-gene mutually-repressing TS circuit with mRNA explicitly modeled. For this circuit, we considered four different scenarios (M1, M2, M3, M4) with increasing levels of difficulty for MaxCal’s inference. First, in M1, we provided information about both proteins to MaxCal. Next, in M2, we provided the same information but in terms of fluorescence instead of protein number to address a typical challenge with data from experiment. To test MaxCal’s inferential power further, we feigned ignorance to the presence of one protein and only provided the fluctuating time trace of the other protein. This model, M3, further extends MaxCal’s applicability to increased numbers of hidden species. Finally, in M4, we provide the same fluctuating time trace of one of the proteins (similar to M3), but in fluorescence instead of protein number. In all four cases, MaxCal is able to describe the underlying fluctuations as quantified by different entropies, average dwell times, and distributions of protein numbers compared against the known model used to generate the synthetic data. Furthermore, MaxCal yields several effective parameters that are not directly visible in experiments, such as the effective protein production and degradation rates in both the basal and repressed states. These inferred rates are in reasonable agreement with the “true” values. Interestingly, MaxCal can also provide effective feedback parameters for promotion or repression between all species from the fluctuations of the data. These parameters add an important set of metrics that further quantify these circuits. This can be particularly useful when analyzing the same circuit topology but varying conditions so as to alter the feedback strength in the system or even eliminate feedback altogether (i.e. bimodality to unimodality). Under stress, genomes of
microbes and cancer cells may evolve circuits with varying degrees of feedback. Another interesting application of MaxCal would be to explore its ability to select between different two different circuit topologies that give rise to same switch-like behavior, e.g. auto-activation (only positive feedback) and TS (mutually repressing two-gene). These examples highlight possible applications of MaxCal to broader problems in synthetic biology beyond the specific model systems investigated here.
Chapter 5

Maximum Caliber Applied to Repressilator Networks

The previous two chapters have demonstrated that information can be extracted from the stochasticity of noisy protein expression trajectories, even when the architecture of the underlying genetic circuit isn’t precisely known. Chapter 3 used a single-gene auto-activation network to demonstrate that it is possible to systematically apply the principle of maximum caliber to extract meaningful metrics and predictions from experimentally realistic data. Chapter 4 went one step further with a two-gene toggle switch circuit to show that even with multiple hidden species (including mRNA), MaxCal’s inferential capabilities can still provide accuracy from limited information. For another step up in complexity, we will now demonstrate the utility of MaxCal on a three-gene repressilator circuit. Not only does this increase the number of species in question, it also analyzes a new oscillatory behavior compared to the switch-like behavior of the previous two circuits, further addressing concerns about the complexity of systems analyzed by MaxCal.
Once again, this “top-down” approach will use the principle of maximum caliber to maximize the path entropy of the system while enforcing certain average values as constraints. In this case, we will again restrict the average protein production and degradation, but also enforce a coupling between the presence of one protein and the production of the next, effectively introducing the circular feedback of the repressilator system. Using simulated trajectories, we will demonstrate that even this simplest model can successfully reproduce the general behaviors of a repressilator circuit. We will also investigate whether further improvement can be gained by introducing further levels of correlation into the caliber function, specifically self-promotion/repression and circular feedback in the opposite direction of oscillation. The inclusion of fluorescence is nontrivial for this system in that the convolution of fluorescence noise from three different signals may compound to make our previous inference methods untenable. However, we will demonstrate that MaxCal accounts for this fluctuation at realistic noise levels, illustrating versatility and applicability to existing experimental systems.

5.1 Materials and Methods

5.1.1 Gillespie Reaction Network

Simulated trajectories are generated by applying a Gillespie algorithm to a three-gene oscillatory circuit known as a repressilator. This circuit is explicitly
defined by the following reaction scheme:

\[
\begin{align*}
\alpha &\rightarrow_{g} \alpha + A; \quad A \rightarrow_{r} \emptyset; \quad \alpha + C \rightarrow_{f} \alpha^* \\
\beta &\rightarrow_{g} \beta + B; \quad B \rightarrow_{r} \emptyset; \quad \beta + A \rightarrow_{f} \beta^* \\
\gamma &\rightarrow_{g} \gamma + C; \quad C \rightarrow_{r} \emptyset; \quad \gamma + B \rightarrow_{f} \gamma^* \\
\alpha^* &\rightarrow_{r} \alpha; \quad \beta^* \rightarrow_{r} \beta; \quad \gamma^* \rightarrow_{r} \gamma
\end{align*}
\]

where generic proteins \(A\), \(B\), and \(C\) are created from their respective genes \(\alpha\), \(\beta\), and \(\gamma\) at a rate of \(g\) and are degraded at a rate of \(r\). To incorporate the circular feedback that is the hallmark of a repressilator circuit, \(A\) can bind and unbind the repressor site of \(\beta\), \(B\) can bind and unbind the repressor site of \(\gamma\), and \(C\) can bind and unbind the repressor site of \(\alpha\) at rates of \(f\) and \(b\) respectively. Binding will send \(\alpha\), \(\beta\), and \(\gamma\) into their deactivated states, \(\alpha^*\), \(\beta^*\), and \(\gamma^*\), where they cannot create their respective proteins. In order to allow oscillation, proteins that are bound are also allowed to degrade at rate \(r\). Rate values are chosen to reproduce oscillatory periods and heights that are representative of experiments while maintaining realistic protein synthesis and degradation rates\(^{22}\) (see Table 5.1). Protein numbers are sampled in increments of \(\Delta t\) over the course of time length \(T\) to mimic experimentally realistic sampling rates.\(^{38}\)

### 5.1.2 MaxCal Model Descriptions

Our baseline Maximum Caliber model accounts for three dynamic behaviors within the system: 1) protein production, 2) protein degradation, and 3) cyclic repression. To account for protein production, each protein species has its own production-path variable, \(\ell_\alpha\), \(\ell_\beta\), and \(\ell_\gamma\), that represents the number of proteins that are created in
<table>
<thead>
<tr>
<th>Rate</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g \text{ (s}^{-1}\text{)}$</td>
<td>$5.0 \times 10^{-2}$</td>
</tr>
<tr>
<td>$r \text{ (s}^{-1}\text{)}$</td>
<td>$3.0 \times 10^{-3}$</td>
</tr>
<tr>
<td>$f \text{ (s}^{-1}\text{)}$</td>
<td>$5.0 \times 10^{-1}$</td>
</tr>
<tr>
<td>$b \text{ (s}^{-1}\text{)}$</td>
<td>$10^{-2}$</td>
</tr>
<tr>
<td>$\Delta t \text{ (s)}$</td>
<td>$60$</td>
</tr>
<tr>
<td>$T \text{ (days)}$</td>
<td>$1$</td>
</tr>
</tbody>
</table>

Table 5.1: Reaction rates used to generate synthetic trajectories.

![Gillespie](image1)
![MaxCal](image2)

Figure 5.1: MaxCal reproduces basic characteristics of repressilator circuit. Direct comparison of synthetic input trajectories (A) and simulated MaxCal trajectories (B). Representative MaxCal parameters: $h_\alpha = -0.460$, $h_A = 1.812$, $K_{A\beta} = -4.635$, $M = 17$.

a discrete time interval of $\Delta t$ and ranges between zero and some predefined maximum, $M$. As for protein degradation, each species also has its own degradation-path variable, $\ell_A$, $\ell_B$, and $\ell_C$, that represents the number of preexisting proteins that remain after a discrete time interval of $\Delta t$ and ranges between zero and the initial number of proteins present, $N_A$, $N_B$, and $N_C$. To constrain the behavior of the system, we introduce Lagrange multipliers to restrict the average values of each path variable, specifically $h_a$, $h_\beta$, $h_\gamma$, $h_A$, $h_B$, and $h_C$. Since all three proteins in equation 5.1 have
symmetric reaction rates, we will assume that all Lagrange multipliers will also be symmetric across protein species (i.e. \( h_\alpha = h_\beta = h_\gamma, h_A = h_B = h_C \)), greatly simplifying our analysis. To enforce cyclic repression, one more Lagrange multiplier, \( K_{A\beta} \), can be introduced to impose a correlation between the presence of one protein and the production of the next, specifically restricting the average values of \( \ell_\beta \ell_A, \ell_\gamma \ell_B, \) and \( \ell_\alpha \ell_C \). With these definitions, the path entropy or “caliber” of our system becomes the following:

\[
C = - \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} \sum_{\ell_\beta=0}^{M} \sum_{\ell_B=0}^{N_B} \sum_{\ell_\gamma=0}^{M} \sum_{\ell_C=0}^{N_C} \left[ P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B,\ell_\gamma,\ell_C} \log P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B,\ell_\gamma,\ell_C} \right. \\
+ h_\alpha(\ell_\alpha + \ell_\beta + \ell_\gamma) P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B,\ell_\gamma,\ell_C} + h_A(\ell_A + \ell_B + \ell_C) P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B,\ell_\gamma,\ell_C} \\
+ K_{A\beta}(\ell_\beta \ell_A + \ell_\gamma \ell_B + \ell_\alpha \ell_C) P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B,\ell_\gamma,\ell_C} \right],
\]

(5.2)

where \( P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B,\ell_\gamma,\ell_C} \) is the probability of observing a path with a particular combination of \( \ell_\alpha, \ell_A, \ell_\beta, \ell_B, \ell_\gamma, \) and \( \ell_C \). Maximizing this caliber function with respect to the individual path probabilities produces

\[
P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B,\ell_\gamma,\ell_C} = Q^{-1} \binom{N_A}{\ell_A} \binom{N_B}{\ell_B} \binom{N_C}{\ell_C} \exp[h_\alpha(\ell_\alpha + \ell_\beta + \ell_\gamma)] \\
+ h_A(\ell_A + \ell_B + \ell_C) + K_{A\beta}(\ell_\beta \ell_A + \ell_\gamma \ell_B + \ell_\alpha \ell_C)]
\]

(5.3)

\[
Q = \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} \sum_{\ell_\beta=0}^{M} \sum_{\ell_B=0}^{N_B} \sum_{\ell_\gamma=0}^{M} \sum_{\ell_C=0}^{N_C} \binom{N_A}{\ell_A} \binom{N_B}{\ell_B} \binom{N_C}{\ell_C} \exp[h_\alpha(\ell_\alpha + \ell_\beta + \ell_\gamma)] \\
+ h_A(\ell_A + \ell_B + \ell_C) + K_{A\beta}(\ell_\beta \ell_A + \ell_\gamma \ell_B + \ell_\alpha \ell_C)].
\]

To produce protein number trajectories to mimic experiment/simulation (Figure 5.1), the path probabilities can be used in a Monte-Carlo fashion by randomly selecting a path according to this distribution, creating and destroying the number of proteins corresponding to that path, and advancing the time of the system by \( \Delta t \). Much
like our previous work\textsuperscript{39,43}, these path probabilities over the discrete time interval of $\Delta t$ can also be propagated through time to calculate the likelihood of stochastic trajectories (from experiment or simulation) using a discretized form of Finite State Projection\textsuperscript{49} (FSP; see Appendix C).

To apply discrete-time FSP to our repressilator system, the number of proteins in a particular state is calculated based on its index $i$ as $N_{A,i} = i \mod N_{\text{max}}$, $N_{B,i} = (i \mod N_{\text{max}}^2)//N_{\text{max}}$, and $N_{C,i} = i//N_{\text{max}}^2$, where $\mod$ is the standard modulo operation, $//$ is the standard floor division operation, and $N_{\text{max}}$ is the maximum number of proteins considered in the finite state space for each protein species. Using these definitions, $i$ ranges between zero and $N_{\text{max}}^3$ with the last state again representing the “sink” state. In the context of MaxCal, the single-frame transition probabilities of the state reaction matrix, $P_{j \rightarrow i}$, can be defined as

$$P_{j \rightarrow i} = \sum_{\ell_{\alpha}=0}^{M} \sum_{\ell_{A}=0}^{N_{A,j}} \sum_{\ell_{\beta}=0}^{N_{B,j}} \sum_{\ell_{\gamma}=0}^{N_{C,j}} \left( \delta(\ell_{\alpha} + \ell_{A} - N_{A,i}) \right)$$

$$\times \delta(\ell_{\beta} + \ell_{B} - N_{B,i}) \delta(\ell_{\gamma} + \ell_{C} - N_{C,i}) P_{\ell_{\alpha},\ell_{A},\ell_{\beta},\ell_{B},\ell_{C}}$$

(5.4)

where $\delta$ is the Dirac delta function. It should be mentioned that in the context of the repressilator circuit, calculating the state reaction matrix, $A$, can be computationally expensive due to the additional degrees of freedom. As such, GPU programming was used to expedite the process (specifically using the python modules of $\text{PyCUDA}$,\textsuperscript{226} $\text{CUDAMat}$,\textsuperscript{227} and $\text{PyOpenCL}$\textsuperscript{226}) and the size of the phase space ($N_{\text{max}}$) should be selected carefully based on the user’s priorities between speed and accuracy.

With this formalism established, we can propagate transition probabilities through multiple frames and calculate the likelihood $\mathcal{L}$ of observing the experimental/simulated
trajectory in increments of \( m \) frames as

\[
\mathcal{L} = \prod_{n=0}^{T/m-1} P(N_A(t + m), N_B(t + m), N_C(t + m), t + m; N_A(t), N_B(t), N_C(t), t = mn)
\]

\[
= \prod_{\{j \rightarrow i\}} P^{\omega_{j \rightarrow i, m}}
\]

(5.5)

where \( N_A(t), N_B(t), \) and \( N_C(t) \) are the number of \( A, B, \) and \( C \) proteins respectively at frame \( t \), \( \omega_{j \rightarrow i, m} \) is the total number of \( j \rightarrow i \) state transitions over \( m \) frames, and the second product is over all possible state transitions. For varying experimental/simulated conditions, representative values of \( h_\alpha, h_A, K_{A\beta}, \) and \( M \) can then be selected by maximizing this likelihood with respect to the parameters of interest. In order to elucidate the proper oscillatory ordering between \( A, B, \) and \( C \), \( m \) should be set to the number of frames closest to one third of the average oscillatory period.

The description above goes through the core execution of MaxCal using model 1, but higher order models can easily be incorporated into this system. Specifically, we only need to modify equations 5.2 and 5.3. First, it is conceivable that an additional level of self-promotion/repression is present in the underlying repressilator circuit being analyzed. To account for this possibility, MaxCal model 2 includes a fourth Lagrange multiplier, \( K_{A\alpha} \), to enforce a correlation between the presence of each protein species (\( \ell_A, \ell_B, \ell_C \)) and its own production (\( \ell_\alpha, \ell_\beta, \ell_\gamma \)). This updates the caliber function in equation 5.2 to

\[
C = - \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} \sum_{\ell_\beta=0}^{N_B} \sum_{\ell_\gamma=0}^{N_C} \left[ P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B,\ell_\gamma,\ell_C} \log P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B,\ell_\gamma,\ell_C} + h_\alpha(\ell_\alpha + \ell_\beta + \ell_\gamma)P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B,\ell_\gamma,\ell_C} + h_A(\ell_A + \ell_B + \ell_C)P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B,\ell_\gamma,\ell_C} + K_{A\alpha}(\ell_\alpha \ell_A + \ell_\beta \ell_B + \ell_\gamma \ell_C)P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B,\ell_\gamma,\ell_C} + K_{A\beta}(\ell_\beta \ell_A + \ell_\gamma \ell_B + \ell_\alpha \ell_C)P_{\ell_\alpha,\ell_A,\ell_\beta,\ell_B,\ell_\gamma,\ell_C} \right].
\]

(5.6)
and the caliber-maximized path probability function in equation 5.3 to

\[ P_{\ell_\alpha, \ell_A, \ell_\beta, \ell_B, \ell_\gamma, \ell_C} = Q^{-1} \left( \begin{array}{c} N_A \\ \ell_A \\ \end{array} \right) \left( \begin{array}{c} N_B \\ \ell_B \\ \end{array} \right) \left( \begin{array}{c} N_C \\ \ell_C \\ \end{array} \right) \exp \left[ h_\alpha (\ell_\alpha + \ell_\beta + \ell_\gamma) \right] (5.7) \]

\[ + h_A(\ell_A + \ell_B + \ell_C) + K_{A_\alpha}(\ell_\alpha \ell_A + \ell_\beta \ell_B + \ell_\gamma \ell_C) + K_{A_\beta}(\ell_\beta \ell_A + \ell_\gamma \ell_B + \ell_\alpha \ell_C); \]

\[ Q = \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} \sum_{\ell_\beta=0}^{M} \sum_{\ell_B=0}^{N_B} \sum_{\ell_\gamma=0}^{M} \sum_{\ell_C=0}^{N_C} \left( \begin{array}{c} N_A \\ \ell_A \\ \end{array} \right) \left( \begin{array}{c} N_B \\ \ell_B \\ \end{array} \right) \left( \begin{array}{c} N_C \\ \ell_C \\ \end{array} \right) \exp \left[ h_\alpha (\ell_\alpha + \ell_\beta + \ell_\gamma) \right] \]

\[ + h_A(\ell_A + \ell_B + \ell_C) + K_{A_\alpha}(\ell_\alpha \ell_A + \ell_\beta \ell_B + \ell_\gamma \ell_C) + K_{A_\beta}(\ell_\beta \ell_A + \ell_\gamma \ell_B + \ell_\alpha \ell_C). \]

FSP is then executed in the same fashion, and as a result, so is ML formalism.

Next, MaxCal model 3 considers the subtle presence of an additional circular promotion/repression in the opposite direction of the preexisting oscillatory order. This is done by introducing a Lagrange multiplier \((K_{A_\gamma})\) to govern the correlation between the presence of one protein \((\ell_A, \ell_B, \ell_C)\) and the production of the previous one in the oscillatory order \((\ell_\gamma, \ell_\alpha, \ell_\beta)\). The caliber function in equation 5.2 then becomes

\[ C = - \sum_{\ell_\alpha=0}^{M} \sum_{\ell_A=0}^{N_A} \sum_{\ell_\beta=0}^{M} \sum_{\ell_B=0}^{N_B} \sum_{\ell_\gamma=0}^{M} \sum_{\ell_C=0}^{N_C} [P_{\ell_\alpha, \ell_A, \ell_\beta, \ell_B, \ell_\gamma, \ell_C} \log P_{\ell_\alpha, \ell_A, \ell_\beta, \ell_B, \ell_\gamma, \ell_C} \]

\[ + h_\alpha(\ell_\alpha + \ell_\beta + \ell_\gamma)P_{\ell_\alpha, \ell_A, \ell_\beta, \ell_B, \ell_\gamma, \ell_C} + h_A(\ell_A + \ell_B + \ell_C)P_{\ell_\alpha, \ell_A, \ell_\beta, \ell_B, \ell_\gamma, \ell_C} \]

\[ + K_{A_\beta}(\ell_\beta \ell_A + \ell_\gamma \ell_B + \ell_\alpha \ell_C)P_{\ell_\alpha, \ell_A, \ell_\beta, \ell_B, \ell_\gamma, \ell_C} \]

\[ + K_{A_\gamma}(\ell_\gamma \ell_A + \ell_\alpha \ell_B + \ell_\beta \ell_C)P_{\ell_\alpha, \ell_A, \ell_\beta, \ell_B, \ell_\gamma, \ell_C} \] (5.8)
producing a caliber-maximized path probability of

\[ P_{\ell, A, \ell, B, \ell, C} = Q^{-1} \left( \frac{N_A}{\ell_A} \right) \left( \frac{N_B}{\ell_B} \right) \left( \frac{N_C}{\ell_C} \right) \exp \left[ h_A(\ell_A + \ell_B + \ell_C) \right] \]

(5.9)

\[ + h_A(\ell_A + \ell_B + \ell_C) + K_{A\beta}(\ell_B + \ell_B + \ell_B + \ell_B) + K_{A\gamma}(\ell_B + \ell_B + \ell_B + \ell_B) \];

\[ Q = \sum_{\ell_A=0}^{M} \sum_{\ell_B=0}^{N_A} \sum_{\ell_B=0}^{N_B} \sum_{\ell_B=0}^{N_C} \left( \frac{N_A}{\ell_A} \right) \left( \frac{N_B}{\ell_B} \right) \left( \frac{N_C}{\ell_C} \right) \exp \left[ h_A(\ell_A + \ell_B + \ell_C) \right] \]

(5.10)

Again, FSP and ML are executed in the same way as before, just with respect to a different set of parameters.

Finally, MaxCal model 4 comes full circle and accounts for both the self-promotion/repression of model 2 and the reverse promotion/repression of model 3. This provides a complete description of all possible crosstalks and is done by incorporating both of their corresponding Lagrange multipliers into its caliber function, producing

\[ C = - \sum_{\ell_A=0}^{M} \sum_{\ell_B=0}^{N_A} \sum_{\ell_B=0}^{N_B} \sum_{\ell_B=0}^{N_C} \left[ P_{\ell, A, \ell, B, \ell, C} \log P_{\ell, A, \ell, B, \ell, C} \right] \]

(5.10)
as well as a caliber-maximized path probability function of

\[
P_{\ell_\alpha, \ell_A, \ell_B, \ell_C, \ell_C} = Q^{-1} \left( \frac{N_A}{\ell_A} \right) \left( \frac{N_B}{\ell_B} \right) \left( \frac{N_C}{\ell_C} \right) \exp[h_\alpha(\ell_\alpha + \ell_\beta + \ell_\gamma)]
\]

\[
+ h_A(\ell_A + \ell_B + \ell_C) + K_{A\alpha}(\ell_\alpha \ell_A + \ell_\beta \ell_B + \ell_\gamma \ell_C)
\]

\[
+ K_{A\beta}(\ell_\beta \ell_A + \ell_\gamma \ell_B + \ell_\alpha \ell_C) + K_{A\gamma}(\ell_\gamma \ell_A + \ell_\alpha \ell_B + \ell_\beta \ell_C) \]

\[
Q = \sum_{\ell_\alpha=0}^M \sum_{\ell_A=0}^{N_A} \sum_{\ell_B=0}^M \sum_{\ell_C=0}^{N_C} \left( \frac{N_A}{\ell_A} \right) \left( \frac{N_B}{\ell_B} \right) \left( \frac{N_C}{\ell_C} \right) \exp[h_\alpha(\ell_\alpha + \ell_\beta + \ell_\gamma)]
\]

\[
+ h_A(\ell_A + \ell_B + \ell_C) + K_{A\alpha}(\ell_\alpha \ell_A + \ell_\beta \ell_B + \ell_\gamma \ell_C)
\]

\[
+ K_{A\beta}(\ell_\beta \ell_A + \ell_\gamma \ell_B + \ell_\alpha \ell_C) + K_{A\gamma}(\ell_\gamma \ell_A + \ell_\alpha \ell_B + \ell_\beta \ell_C) \].
\]

As before, FSP and ML are carried out in the same way, but with respect to the six parameters of \( h_\alpha, h_A, K_{A\alpha}, K_{A\beta}, K_{A\gamma}, \) and \( M \).

### 5.1.3 Predictions and Metrics Used to Assess MaxCal Representations

To provide an accurate and fair comparison of the input trajectory to its corresponding MaxCal representation, multiple different predictions and metrics will be used. First, we can predict effective protein production and degradation rates from our MaxCal parameters that are otherwise unavailable to alternative methods. Given specific protein numbers for each species, the transition probabilities described in equation 5.3 allow us to calculate effective protein production and degradation rates for each species. In the case of production of \( A \), such a rate relates to the average of \( \ell_\alpha \) as

\[
g_A(N_A, N_B, N_C) = \frac{\left( \ell_\alpha \right)_{N_A, N_B, N_C}}{\Delta t}
\]
with symmetric equations for \( B \) and \( C \), where \( \langle \cdots \rangle_{i,j,k} \) represents the average of a quantity given that there are \( i \), \( j \), and \( k \) numbers of \( A \), \( B \), and \( C \) proteins initially present respectively. Accounting for the assumed architecture of our Gillespie simulations, we can predict a total effective protein production rate \( (g_{\text{eff}}) \) to compare with \( g \) in equation 5.1. Specifically, we recognize that a protein species' basal production should be optimally observed when no other protein species are present to influence its production. Thus,

\[
g_{\text{eff}} = \sum_{N_A=0}^{\infty} P_{eq}(N_A|0,0)g_A(N_A,0,0) \tag{5.13}
\]

with symmetric equations for \( B \) and \( C \), where \( P_{eq}(i|0,0) \) is the probability at relative equilibrium (calculated via discrete-time FSP) that there exists \( i \) number of proteins of any one of the three species given that the other two protein species are not present in the system. A similar approach can be taken to calculate an effective degradation rate for \( A \), specifically relating to the average of \( \ell_A \) as

\[
r_A(N_A, N_B, N_C) = \frac{N_A - \langle \ell_A \rangle_{N_A,N_B,N_C}}{N_A \Delta t} \tag{5.14}
\]

with symmetric equations for \( B \) and \( C \). Since proteins should degrade at generally the same rate no matter what protein concentrations are present, we can predict a total effective protein degradation rate \( (r_{\text{eff}}) \) to compare with \( r \) in equation 5.1 by taking a weighted average of these values over the entire phase space of \( N_A \), \( N_B \) and \( N_C \):

\[
r_{\text{eff}} = \sum_{N_A=0}^{\infty} \sum_{N_B=0}^{\infty} \sum_{N_C=0}^{\infty} P_{eq}(N_A, N_B, N_C)r_A(N_A, N_B, N_C) \tag{5.15}
\]

with symmetric equations for \( B \) and \( C \), where \( P_{eq}(i,j,k) \) denotes the probability at relative equilibrium of having \( i \), \( j \), and \( k \) numbers of \( A \), \( B \), and \( C \) proteins respectively.
One novel aspect of the MaxCal modeling system is the ability to calculate effective feedback metrics for different dynamic interactions. For instance, since $K_{A\beta}$ directly enforces cyclic repression, we can quantify the relative amount of cyclic repression in the system by calculating the Pearson correlation coefficient between the presence of one protein ($\ell_A$) and the production of the next ($\ell_\beta$). Specifically, we define this metric as

$$F_{A\beta} = \frac{\langle \ell_\beta \ell_A \rangle_{\text{tot}} - \langle \ell_\beta \rangle_{\text{tot}} \langle \ell_A \rangle_{\text{tot}}}{\sqrt{\langle \ell_\beta^2 \rangle_{\text{tot}} - \langle \ell_\beta \rangle_{\text{tot}}^2 \langle \ell_A^2 \rangle_{\text{tot}} - \langle \ell_A \rangle_{\text{tot}}^2}}$$

$$= \frac{\langle \ell_\gamma \ell_B \rangle_{\text{tot}} - \langle \ell_\gamma \rangle_{\text{tot}} \langle \ell_B \rangle_{\text{tot}}}{\sqrt{\langle \ell_\gamma^2 \rangle_{\text{tot}} - \langle \ell_\gamma \rangle_{\text{tot}}^2 \langle \ell_B^2 \rangle_{\text{tot}} - \langle \ell_B \rangle_{\text{tot}}^2}}$$

$$= \frac{\langle \ell_\alpha \ell_C \rangle_{\text{tot}} - \langle \ell_\alpha \rangle_{\text{tot}} \langle \ell_C \rangle_{\text{tot}}}{\sqrt{\langle \ell_\alpha^2 \rangle_{\text{tot}} - \langle \ell_\alpha \rangle_{\text{tot}}^2 \langle \ell_C^2 \rangle_{\text{tot}} - \langle \ell_C \rangle_{\text{tot}}^2}}$$

where

$$\langle X \rangle_{\text{tot}} = \sum_{N_A=0}^{\infty} \sum_{N_B=0}^{\infty} \sum_{N_C=0}^{\infty} \sum_{N_\alpha=0}^{M} \sum_{N_\beta=0}^{M} \sum_{N_\gamma=0}^{M} \sum_{N_\delta}^{N_C} XP_{eq}(N_A, N_B, N_C) P_{\ell_\alpha, \ell_\beta, \ell_\gamma, \ell_C}$$

with $X$ being any combination of the stochastic variables of interest used to describe protein production and degradation. Furthermore, cyclic interaction could produce indirect forms of feedback within the system. This motivates similar feedback terms representing self promotion/repression,

$$F_{A\alpha} = \frac{\langle \ell_\alpha \ell_A \rangle_{\text{tot}} - \langle \ell_\alpha \rangle_{\text{tot}} \langle \ell_A \rangle_{\text{tot}}}{\sqrt{\langle \ell_\alpha^2 \rangle_{\text{tot}} - \langle \ell_\alpha \rangle_{\text{tot}}^2 \langle \ell_A^2 \rangle_{\text{tot}} - \langle \ell_A \rangle_{\text{tot}}^2}}$$

$$= \frac{\langle \ell_\beta \ell_B \rangle_{\text{tot}} - \langle \ell_\beta \rangle_{\text{tot}} \langle \ell_B \rangle_{\text{tot}}}{\sqrt{\langle \ell_\beta^2 \rangle_{\text{tot}} - \langle \ell_\beta \rangle_{\text{tot}}^2 \langle \ell_B^2 \rangle_{\text{tot}} - \langle \ell_B \rangle_{\text{tot}}^2}}$$

$$= \frac{\langle \ell_\gamma \ell_C \rangle_{\text{tot}} - \langle \ell_\gamma \rangle_{\text{tot}} \langle \ell_C \rangle_{\text{tot}}}{\sqrt{\langle \ell_\gamma^2 \rangle_{\text{tot}} - \langle \ell_\gamma \rangle_{\text{tot}}^2 \langle \ell_C^2 \rangle_{\text{tot}} - \langle \ell_C \rangle_{\text{tot}}^2}}$$

(5.16)
as well as cyclic promotion/repression in the opposite direction,

$$F_{Aγ} = \frac{\langle l γ l A \rangle_{\text{tot}} - \langle l γ \rangle_{\text{tot}} \langle l A \rangle_{\text{tot}}}{\sqrt{\langle (l γ)^2 \rangle_{\text{tot}} - \langle l γ \rangle_{\text{tot}}^2} \langle (l A)^2 \rangle_{\text{tot}} - \langle l A \rangle_{\text{tot}}^2}$$

$$= \frac{\langle l α l B \rangle_{\text{tot}} - \langle l α \rangle_{\text{tot}} \langle l B \rangle_{\text{tot}}}{\sqrt{\langle (l α)^2 \rangle_{\text{tot}} - \langle l α \rangle_{\text{tot}}^2} \langle (l B)^2 \rangle_{\text{tot}} - \langle l B \rangle_{\text{tot}}^2}$$

$$= \frac{\langle l β l C \rangle_{\text{tot}} - \langle l β \rangle_{\text{tot}} \langle l C \rangle_{\text{tot}}}{\sqrt{\langle (l β)^2 \rangle_{\text{tot}} - \langle l β \rangle_{\text{tot}}^2} \langle (l C)^2 \rangle_{\text{tot}} - \langle l C \rangle_{\text{tot}}^2}.$$ (5.19)

To clarify, these feedback metrics work best when comparing the same system, but with different conditions (e.g. experimental environment, simulation rate reactions). Comparing feedback parameters between different systems (e.g. different organisms, reaction networks, MaxCal models) would be an unfair comparison as the entire architecture of interactions has been altered.

For a further quantitative comparison of how well MaxCal encapsulates the behaviors of experiment, we will also consider some observable metrics in addition to the more predictive metrics described above. In a similar fashion as our previous work, 39,43 we will compute three different path entropy metrics. The single-trajectory path entropy, $S_1$, considers the protein number trajectories of $A$, $B$, and $C$ independently of each other and is measured in bits as

$$S_1 = - \sum_{a,b} P_{trj}(a)P_{trj}(a \to b) \log_2(P_{trj}(a)P_{trj}(a \to b))$$ (5.20)

where $P_{trj}(a)$ is the probability of having $a$ proteins at any point in any of the trajectories and $P_{trj}(a \to b)$ is the probability of transitioning from $a$ proteins to $b$ in these trajectories. The total path entropy, $S_{\text{tot}}$, considers all three of the protein number
trajectories together and is calculated as

\[ S_{\text{tot}} = - \sum_{a,b,c,x,y,z} P_{\text{trj}}(a, b, c)P_{\text{trj}}(a \rightarrow x, b \rightarrow y, c \rightarrow z) \]

\[ \log_2(P_{\text{trj}}(a, b, c)P_{\text{trj}}(a \rightarrow x, b \rightarrow y, c \rightarrow z)) \] (5.21)

where \( P_{\text{trj}}(a, b, c) \) is the probability of having \( a \), \( b \), and \( c \) numbers of \( A \), \( B \), and \( C \) proteins respectively during the course of the trajectories and \( P_{\text{trj}}(a \rightarrow x, b \rightarrow y, c \rightarrow z) \) is the probability of transitioning at those timepoints from that state to \( x \), \( y \), and \( z \) numbers of \( A \), \( B \), and \( C \) proteins respectively. Finally, to calculate coarse-grained entropy, \( S_{\text{cg}} \), all three trajectories are first simplified into a single ternary trajectory between three states based on which protein number is at its peak. \( S_{\text{cg}} \) is then calculated in the same manner as equation 5.20.

Due to the oscillatory nature of these trajectories, we will also compare the oscillatory periods and peak heights of both systems. In order to measure a representative period, we put the protein number trajectories through a Fast Fourier Transform (FFT) to break the signal down into its individual frequency components. The frequency with the highest amplitude \( (f_{\text{rep}}) \) is then selected and the representative period is calculated as \( \tau_{\text{rep}} = 1/f_{\text{rep}} \). Peak heights are measured by splitting individual protein number trajectories into “peaks” and “valleys”, by setting a threshold at 25% of the global maximum of protein number. Once a trajectory crosses this threshold, it is considered to be in a “peak” until it returns to zero, at which point it is considered to be in a “valley” until it crosses the threshold again. The maximum value of each peak section is then considered the peak height.
5.1.4 Inclusion of Fluorescence

Models 1 through 4 assume that input trajectories are provided directly in protein number. However, more realistic experimental conditions provide trajectories in terms of fluorescence. The stochastic amount of fluorescence collected from each protein provides another level of fluctuation on top of fluctuation in protein expression and we must be able to account for this throughout our ML parameter estimation protocol. To do this, we include the probability distribution of fluorescence collected per protein into our likelihood function, thus calculating the probability of observing a particular fluorescence trajectory rather than a particular protein number trajectory. Models 5 through 8 will simply be the equivalent of models 1 through 4, but applied to fluorescence trajectories rather than protein number.

In order to test the inference capabilities of our fluorescence-based MaxCal models, we must first create synthetic fluorescence trajectories from our preexisting Gillespie simulations of protein number trajectories by “corrupting” with an experimentally realistic fluorescence per protein distribution. For simplicity’s sake, we assume this distribution to be Gaussian with an average of $f_0$ and a standard deviation of $\sigma$. With this assumption, each timepoint (with protein number $N_t$) can be assigned a random fluorescence value based on convolution of $N_t$ individual fluorescence per protein distributions, i.e. another Gaussian distribution with an average of $N_t f_0$ and a standard deviation of $\sqrt{N_t}\sigma$. As an illustration of realistic experimental conditions, we will perform our analysis on relative noise levels ($\sigma/f_0$) of 30%.

To incorporate fluorescence into our likelihood calculation from equation 5.5, we can combine path probabilities from MaxCal with the probability distribution of fluorescence per protein. We assume that the average and standard deviation of the are known via low-intensity fluorescence experiments. With this information,
we can update equation 5.5 to

$$
\mathcal{L} = \prod_{n=0}^{T/m-1} \sum_{\{i,j\}} \Phi(N_{A,j}|f_A(t))\Phi(N_{B,j}|f_B(t))\Phi(N_{C,j}|f_C(t))
\times P(N_{A,i}, N_{B,i}, N_{C,i}, t + m; N_{A,j}, N_{B,j}, N_{C,j}, t = mn)
\times \Phi(N_{A,i}|f_A(t + m))\Phi(N_{B,i}|f_B(t + m))\Phi(N_{C,j}|f_C(t + m))
$$

(5.22)

where $f_A(t)$, $f_B(t)$, and $f_C(t)$ are the fluorescences measured from $A$, $B$, and $C$ proteins respectively, and $\Phi(N|f)$ is the conditional probability that $N$ proteins are present given that a fluorescence of $f$ was observed. This conditional probability can be calculated using Bayes’ theorem as

$$
\Phi(N|f) = \frac{P_g(f|N)P_{eq}(N)}{P_{tot}(f)}
$$

(5.23)

where $P_g(f|N)$ is the Gaussian fluorescence distribution with an average and standard deviation of $Nf_0$ and $\sqrt{N}\sigma$, $P_{eq}(N)$ is the MaxCal-predicted protein number distribution at relative equilibrium (calculated via discretized FSP), and $P_{tot}(f)$ is the fluorescence probability distribution over the entire trace. Since $\mathcal{L}$ is still a function of the Lagrange multipliers and $M$ no matter which model is being considered, the ML formalism is still applicable and representative values can be selected by maximizing $\mathcal{L}$ with respect to these parameters.

### 5.2 Results and Discussion

#### 5.2.1 MaxCal Models Using Protein Number

Starting with the simplest model 1, we can see that even with just the repression of $K_{A\beta}$, the agreement between input trajectories and MaxCal trajectories is already
Figure 5.2: **Primary behaviors of Model 1 agree well with simulation.** Comparison between synthetic input trajectories (blue) and simulated MaxCal trajectories (green) of (A) oscillatory period, (B) oscillatory height, and (C) one-dimensional protein number distributions. Representative MaxCal parameters: \( h_\alpha = -0.459 \), \( h_A = 1.810 \), \( K_{A\beta} = -4.596 \), \( M = 16 \).

Within reason. Figure 5.2A shows overlapping FFT distributions, suggesting very similar oscillatory frequencies between the two. The distribution of oscillatory heights in Figure 5.2B also shows good alignment. In fact, MaxCal produces similar protein number distributions in both one dimension (Figure 5.2C) and three dimensions (Figure 5.3A, 5.3B).

Looking at numerical comparisons, Table 5.2 demonstrates that while the effective degradation rate \( r_{\text{eff}} \) extracted by MaxCal model 1 is within 10% of the true value, some improvement could be made in the effective production rate as it is in the same order of magnitude, but off by more than 40% of the true value. All three path entropy metrics \( S_1, S_{\text{tot}}, S_{\text{cg}} \) are in very good agreement, all coming within 10% of the “true” value from input trajectories. Finally, the feedback metrics (not readily available in Gillespie algorithms) quantitatively reproduce what we would intuitively expect in that we have strong negative feedback between the presence of one protein and the production of the next \( (F_{A\beta}) \), significant self-promotion \( (F_{A\alpha}) \), and little to no feedback in the reverse direction of oscillation \( (F_{A\gamma}) \).

Moving to higher order models, the addition \( K_{A\alpha} \) and \( K_{A\gamma} \) in models 2 and 3 respectively show marked improvements in some characteristics and only slight
Figure 5.3: **Three-dimensional behaviors of protein number based MaxCal models agree well with simulation.** Comparison of three-dimensional protein number heat maps between (A) synthetic input trajectories and simulated MaxCal trajectories using (B) model 1, (C) model 2, (D) model 3, and (E) model 4.
set-backs in others. Specifically, $g_{\text{eff}}$ improves to within 20% of the “true” value used to create the simulated trajectories, while all three entropy metrics move just slightly further away from the “true” entropies. All feedback metrics change slightly in magnitude, but the trends generally remain the same in that $F_{A\alpha}$ is significantly positive, $F_{A\beta}$ is significantly negative, and $F_{A\gamma}$ is small enough to ignore. In graphical comparisons, oscillatory height distributions (Figure 5.4B, 5.5B) and protein number distributions (Figure 5.4C, 5.5C, 5.3A, 5.3C, 5.3D) do not show any drastic changes from model 1, but slower oscillatory frequencies show up a bit more in Figures 5.4A and 5.5A. However, the peak value dictating the representative frequency and period is still in the same location as input, so this difference does not seem substantial. As would be expected, the likelihood improves in both cases (smaller values of $-\log L$ in the final row of Table 5.2), but model 2 comes out slightly better than model 3. On top of the fact that $F_{A\alpha}$ has a much higher magnitude than $F_{A\gamma}$, this suggests that $K_{A\alpha}$ is a more important addition to the model than $K_{A\gamma}$.

Finally, model 4 includes both $K_{A\alpha}$ and $K_{A\gamma}$ into its caliber and path probability functions (equations 5.10 and 5.11) to provide a complete description of possible feed-

<table>
<thead>
<tr>
<th></th>
<th>Experiment</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_{\text{eff}}$ (10^{-3} s^{-1})</td>
<td>50.0</td>
<td>28.5 ± 0.1</td>
<td>41.9 ± 1.3</td>
<td>46.0 ± 0.2</td>
<td>44.3 ± 0.2</td>
</tr>
<tr>
<td>$r_{\text{eff}}$ (10^{-3} s^{-1})</td>
<td>3.0</td>
<td>2.70 ± 0.01</td>
<td>3.64 ± 0.04</td>
<td>3.29 ± 0.02</td>
<td>3.59 ± 0.02</td>
</tr>
<tr>
<td>$\tau$ (10^{3} s)</td>
<td>3.38 ± 0.18</td>
<td>3.25 ± 0.10</td>
<td>3.38 ± 0.17</td>
<td>3.25 ± 0.13</td>
<td>3.30 ± 0.19</td>
</tr>
<tr>
<td>$\langle H \rangle$</td>
<td>16.9 ± 4.0</td>
<td>16.6 ± 4.3</td>
<td>16.6 ± 5.0</td>
<td>17.3 ± 4.8</td>
<td>16.5 ± 4.9</td>
</tr>
<tr>
<td>$S_1$ (bits)</td>
<td>5.09 ± 0.00</td>
<td>5.39 ± 0.01</td>
<td>5.40 ± 0.03</td>
<td>5.46 ± 0.00</td>
<td>5.41 ± 0.00</td>
</tr>
<tr>
<td>$S_{\text{tot}}$ (bits)</td>
<td>13.3 ± 0.0</td>
<td>13.7 ± 0.0</td>
<td>13.9 ± 0.1</td>
<td>13.9 ± 0.0</td>
<td>13.8 ± 0.0</td>
</tr>
<tr>
<td>$S_{\text{cg}}$ (bits)</td>
<td>1.83 ± 0.00</td>
<td>1.84 ± 0.00</td>
<td>1.77 ± 0.02</td>
<td>1.74 ± 0.00</td>
<td>1.75 ± 0.00</td>
</tr>
<tr>
<td>$F_{A\alpha}$</td>
<td>0.273 ± 0.001</td>
<td>0.546 ± 0.011</td>
<td>0.359 ± 0.001</td>
<td>0.536 ± 0.001</td>
<td></td>
</tr>
<tr>
<td>$F_{A\beta}$</td>
<td>$-0.352 ± 0.000$</td>
<td>$-0.452 ± 0.027$</td>
<td>$-0.359 ± 0.000$</td>
<td>$-0.446 ± 0.001$</td>
<td></td>
</tr>
<tr>
<td>$F_{A\gamma}$</td>
<td>0.057 ± 0.000</td>
<td>0.018 ± 0.003</td>
<td>$-0.041 ± 0.001$</td>
<td>$-0.008 ± 0.002$</td>
<td></td>
</tr>
<tr>
<td>$-\log L$</td>
<td>45,498 ± 77</td>
<td>43,799 ± 52</td>
<td>43,848 ± 63</td>
<td>43,713 ± 55</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.2: Comparison of experimental and MaxCal predicted rates and statistics from protein number based models.
Figure 5.4: **Primary behaviors of Model 2 agree well with simulation.** Comparison between synthetic input trajectories (blue) and simulated MaxCal trajectories (green) of (A) oscillatory period, (B) oscillatory height, and (C) one-dimensional protein number distributions. Representative MaxCal parameters: \( h_\alpha = -0.389, h_A = 1.533, K_{A\alpha} = 0.109, K_{A\beta} = -3.569, M = 3 \).

Figure 5.5: **Primary behaviors of Model 3 agree well with simulation.** Comparison between synthetic input trajectories (blue) and simulated MaxCal trajectories (green) of (A) oscillatory period, (B) oscillatory height, and (C) one-dimensional protein number distributions. Representative MaxCal parameters: \( h_\alpha = -0.230, h_A = 1.655, K_{A\beta} = -4.032, K_{A\gamma} = -0.043, M = 9 \).

back within the system, and by ML formalism, this is the most likely model out of the four protein number based MaxCal models. However, improvement beyond models 2 and 3 in terms of distributions and statistics is fairly minimal. All quantities in the last column of Table 5.2 remain within 20% of their true values with only \( r_{eff} \) being outside 10%. Protein number (Figure 5.6C, 5.3E) and oscillatory height distributions (Figure 5.6B) still overlap quite well with input trajectories, and while lower frequencies are again more present in Figure 5.6A, the representative frequency and period remain identical to input. Taking all of these results into account, the takeaway message is that model 1 will provide a reasonable reproduction of the repressilator circuit,
Figure 5.6: **Primary behaviors of Model 4 agree well with simulation.** Comparison between synthetic input trajectories (blue) and simulated MaxCal trajectories (green) of (A) oscillatory period, (B) oscillatory height, and (C) one-dimensional protein number distributions. Representative MaxCal parameters: $h_\alpha = -0.389$, $h_A = 1.537$, $K_{A\alpha} = 0.063$, $K_{A\beta} = -3.704$, $K_{A\gamma} = -0.014$, $M = 4$.

but more terms will always help (particularly $K_{A\alpha}$). All three higher order models do not show major differences, so if computational resources allow, model 4 should be used to provide a complete description of feedback and the highest likelihood of reproducing experimental/simulated observations.

### 5.2.2 MaxCal Models Using Fluorescence

Every model in the previous section assumes that input trajectories would be given in protein number, an unrealistic condition in experiment. To provide a more realistic assessment of MaxCal’s inferential capabilities, we generated fluorescence trajectories from the same exact protein number trajectories from the previous section by “corrupting” them with Gaussian fluorescence per protein distribution. These fluorescence traces were then given to MaxCal as input and representative parameter values were selected using the ML formalism described in equation 5.22. Relative fluorescence noise levels were set to 30% to best match experimental conditions.

As seen in Figures 5.7-5.11, MaxCal still provides an accurate reproduction of the repressilator circuit even when considering the noisy and convoluted data from fluorescence measurements. As an analog for model 1 in the context of fluorescence...
Figure 5.7: Primary behaviors of Model 5 agree well with simulation. Comparison between synthetic input trajectories (blue) and simulated MaxCal trajectories (green) of (A) oscillatory period, (B) oscillatory height, and (C) one-dimensional protein number distributions. Representative MaxCal parameters: \( h_\alpha = -0.448, h_A = 1.809, K_{A\beta} = -4.646, M = 10 \).

trajectories, model 5 provides a reasonable representation of the underlying genetic repressilator circuit, but could be improved in terms of both likelihood and effective protein production rate. Models 6 through 8, representing the higher order MaxCal models applied to fluorescence trajectories, all show improvement in these areas with \( K_{A\alpha} \) contributing more of an increase in likelihood compared to \( K_{A\gamma} \). If computational resources allow it, model 8 should be used to provide a complete description of possible feedback in the system and the highest likelihood of reproducing input trajectories.

<table>
<thead>
<tr>
<th></th>
<th>Experiment</th>
<th>Model 5</th>
<th>Model 6</th>
<th>Model 7</th>
<th>Model 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>( g_{\text{eff}} ) (10^{-3} \text{ s}^{-1})</td>
<td>50.0</td>
<td>28.1 ± 0.0</td>
<td>39.4 ± 0.1</td>
<td>46.7 ± 0.3</td>
<td>39.2 ± 0.1</td>
</tr>
<tr>
<td>( r_{\text{eff}} ) (10^{-3} \text{ s}^{-1})</td>
<td>3.0</td>
<td>2.70 ± 0.00</td>
<td>3.78 ± 0.02</td>
<td>3.45 ± 0.02</td>
<td>3.79 ± 0.01</td>
</tr>
<tr>
<td>( \tau ) (10^3 \text{ s})</td>
<td>3.38 ± 0.18</td>
<td>3.17 ± 0.09</td>
<td>3.46 ± 0.16</td>
<td>3.08 ± 0.07</td>
<td>3.79 ± 0.37</td>
</tr>
<tr>
<td>( \langle H \rangle )</td>
<td>16.9 ± 4.0</td>
<td>15.5 ± 4.0</td>
<td>15.7 ± 4.5</td>
<td>16.9 ± 4.6</td>
<td>15.6 ± 4.6</td>
</tr>
<tr>
<td>( S_1 ) (bits)</td>
<td>5.09 ± 0.00</td>
<td>5.36 ± 0.01</td>
<td>5.33 ± 0.00</td>
<td>5.44 ± 0.00</td>
<td>5.34 ± 0.01</td>
</tr>
<tr>
<td>( S_{\text{tot}} ) (bits)</td>
<td>13.3 ± 0.00</td>
<td>13.6 ± 0.00</td>
<td>13.6 ± 0.00</td>
<td>13.8 ± 0.00</td>
<td>13.6 ± 0.00</td>
</tr>
<tr>
<td>( S_{\text{eff}} ) (bits)</td>
<td>1.83 ± 0.00</td>
<td>1.84 ± 0.00</td>
<td>1.73 ± 0.00</td>
<td>1.75 ± 0.00</td>
<td>1.74 ± 0.01</td>
</tr>
<tr>
<td>( F_{A\alpha} )</td>
<td>0.285 ± 0.000</td>
<td>0.576 ± 0.001</td>
<td>0.374 ± 0.001</td>
<td>0.575 ± 0.001</td>
<td></td>
</tr>
<tr>
<td>( F_{A\beta} )</td>
<td>-0.369 ± 0.000</td>
<td>-0.498 ± 0.001</td>
<td>-0.370 ± 0.000</td>
<td>-0.497 ± 0.001</td>
<td></td>
</tr>
<tr>
<td>( F_{A\gamma} )</td>
<td>0.061 ± 0.000</td>
<td>0.018 ± 0.001</td>
<td>-0.047 ± 0.001</td>
<td>0.023 ± 0.002</td>
<td></td>
</tr>
<tr>
<td>(- \log L)</td>
<td>93,472 ± 113</td>
<td>88,393 ± 159</td>
<td>89,019 ± 156</td>
<td>88,385 ± 157</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.3: Comparison of experimental and MaxCal predicted rates and statistics from fluorescence based models.
Figure 5.8: Three-dimensional behaviors of fluorescence based MaxCal models agree well with simulation. Comparison of three-dimensional protein number heat maps between (A) synthetic input trajectories and simulated MaxCal trajectories using (B) model 5, (C) model 6, (D) model 7, and (E) model 8.
Figure 5.9: **Primary behaviors of Model 6 agree well with simulation.** Comparison between synthetic input trajectories (blue) and simulated MaxCal trajectories (green) of (A) oscillatory period, (B) oscillatory height, and (C) one-dimensional protein number distributions. Representative MaxCal parameters: \( h_\alpha = -0.420, h_A = 1.521, K_{A\alpha} = 0.113, K_{A\beta} = -3.780, M = 3 \).

Figure 5.10: **Primary behaviors of Model 7 agree well with simulation.** Comparison between synthetic input trajectories (blue) and simulated MaxCal trajectories (green) of (A) oscillatory period, (B) oscillatory height, and (C) one-dimensional protein number distributions. Representative MaxCal parameters: \( h_\alpha = -0.193, h_A = 1.623, K_{A\beta} = -4.317, K_{A\gamma} = -0.048, M = 8 \).

Figure 5.11: **Primary behaviors of Model 8 agree well with simulation.** Comparison between synthetic input trajectories (blue) and simulated MaxCal trajectories (green) of (A) oscillatory period, (B) oscillatory height, and (C) one-dimensional protein number distributions. Representative MaxCal parameters: \( h_\alpha = -0.456, h_A = 1.510, K_{A\alpha} = 0.116, K_{A\beta} = -3.846, K_{A\gamma} = 0.004, M = 3 \).
5.3 Conclusion

We have demonstrated that MaxCal can effectively extract meaningful metrics and predictions from stochastic protein number trajectories generated by the oscillatory repressilator system. Combining MaxCal path probabilities with ML formalism, even our simplest model with only three Lagrange multipliers can quantitatively reproduce multiple behaviors of the underlying genetic circuit, including oscillatory height, period, path entropies, and protein number distributions. Furthermore, MaxCal produces effective feedback metrics to elucidate relationships in a way that is not readily applicable in other simulation methods. Additional improvement can be gained by adding more constraints to the system via Lagrange multipliers, greatly ameliorating predictions of underlying protein production and degradation rates. Finally, using simulated trajectories “corrupted” to mimic fluorescence, we have shown that MaxCal’s inferential capabilities still hold up when applied to experimentally realistic forms of data. This advancement of the scope of maximum caliber demonstrates the practical power of our top-down approach and shows promise for future applications of this formalism to even more complicated genetic circuits. Other synthetic oscillatory systems do not require the assumption of circular symmetry and incorporate both positive and negative feedback to create a tunable and robust circuit.\textsuperscript{228} Future applications to naturally occurring oscillatory circuits like the circadian clock\textsuperscript{47,48} are also possible. In fact, MaxCal is not just limited to the inferential methods presented here. The minimalistic nature of MaxCal can also be useful in circuit design to explore different parameter sets in a time-efficient manner. Traditional “bottom-up” models can lead to overparameterization and as a result can slow down the optimization process.\textsuperscript{29,30} MaxCal enables us to see how circuit fluctuations
depend on various levels of feedback, helping synthetic biologists to control these attributes and ultimately the phenotypes they produce.
Part IV

Noise on a Multicellular Level
Chapter 6

Germband Extension Modeling

The formation of an organism’s head-to-tail (anterior-posterior or AP) body axis is a crucial process in embryonic development. This AP axis elongation is achieved in the early embryo of Drosophila through what is called germ band extension (GBE).\textsuperscript{50,51} Largely in the absence of cell division, the embryonic epithelium more than doubles in length while narrowing in width\textsuperscript{3} using directional neighbor exchange via cell intercalation.\textsuperscript{52,229} In similar fashion to GBE, various tissues utilize intercalation to drive elongation of multiple organ systems\textsuperscript{56–62} as well as the primitive streak of chick embryos\textsuperscript{55} and the ascidian notochord.\textsuperscript{53,54}

During intercalation (Fig. 6.1 A1,A2), the rearrangement of the cell layer goes through a series of clearly defined steps: From the original honeycomb-like hexagonal configuration, the interface between anterior-posterior (AP) cell neighbors (green cells in Fig. 6.1B) contracts into a 4-cell vertex, which is then resolved into a new interface between dorsal-ventral cell neighbors (blue and red cells in Fig. 6.1B). The net effect of this rearrangement is an increase of distance between horizontal (AP) neighbor pairs, and a decrease of the distance between vertical (DV) neighbor pairs. In the nomenclature of developmental biology (which we will use in the following), the initial
vertical interface configuration is called the T1 configuration, the 4-cell vertex is the T2 configuration, and the elongation of a new horizontal interface is the T3 configuration. (It should be noted that in the field of soft matter physics, this entire transition from T1-T2-T3 would be referred to as a T1 type process, while T2 and T3 have entirely different meanings). In Drosophila GBE, this transition is associated with a planar-polarized system of cytoskeletal and junctional proteins, including filamentous actin (F-actin) and the actin-based Myosin II motor protein enriched at AP interfaces, as well as the Bazooka/PAR-3 PDZ-domain protein and the adherens
junction proteins E-Cadherin and Armadillo/β-catenin preferentially localized to DV interfaces.

The fundamental question about the driving forces behind this morphogenetic process has not yet been satisfactorily resolved. There could be several competing mechanisms, including purely local interactions between neighboring cells and/or large-scale global force generation, which integrate the effects of physical actomyosin force generation, asymmetric cytoskeletal architecture, and adhesion properties. How do we test specific mechanisms and select one over the other? Model selection becomes particularly challenging when different mechanisms can broadly explain the same qualitative behavior. A rigorous way to address this question requires a dual approach that combines i) systematic model building based on different competing mechanisms and ii) quantitative testing of these models against live-cell measurements through high-resolution microscopy and computational image analysis. A key aspect of rigorous model building is to extract as many independent observables as possible from data in order to test the performance of a given model. In the case of GBE, we can simultaneously use data from convergence, extension, and tissue morphology. Two competing models can be discriminated based on their prediction of the final state, i.e. the topology of the cellular network at the end of GBE (Fig. 6.1D2 for a typical model and Fig. 6.1E2 for experiment). In cases where two models provide similar final states, further selection can be achieved by following the evolution of the morphology, in particular the nature of extension. In this paper, we adopt this systematic approach of model building that tests predictions against a multitude of experimental data to infer fundamental properties of GBE.

Previous morphogenetic modeling studies of systems other than GBE have reproduced phenomena such as cell sorting in *Xenopus* embryos, cell packing in *Drosophila* wings and eyes, dorsal appendage formation in *Drosophila eggs,*
and wound closure in epithelial tissues of multiple organisms. Physical models of morphogenesis mimic the soap-foam type patterns created by tension energy minimization, and thus often directly lend themselves to soft-matter type approaches. While we borrow tools from these studies to devise an efficient simulation protocol and construct new metrics relevant in the context of GBE, our goal is not only to reproduce qualitative features of the experimental data (i.e. to achieve tissue convergence and extension in the model), but also to provide systematic and quantitative comparison between computational models and experimental data. This would enable us to test whether specific mechanistic ingredients of the model are necessary and sufficient to reproduce experimental data.

The canonical mechanistic model for intercalation in GBE presumes that the driving force for AP junction contraction is anisotropic line tension generated by planar-polarized apical actomyosin contraction. This physical model was motivated by the well-known planar-polarized localization of myosin II, and was supported by findings in proliferating larval wing epithelium that line tension of cell boundaries depends on acto-myosin contractility. Subsequent reports in GBE also found that in laser ablation experiments, AP interfaces have significantly greater recoil speed compared to transverse junctions, suggesting higher line tension. Planar-polarized line tension anisotropy was also observed in vertebrate gastrula mesenchyme cells. The initial first-order computational model of GBE that reproduced intercalation used a vertex-based model for epithelial cell packing and operated under the assumption that anisotropic line tension initiates convergence while extension is driven ‘passively’ by area conservation and perimeter contractility. However, this model has not been tested exhaustively against data during extension, partly due to lack of quantitative data in this phase. Furthermore, a recent paper has proposed that extension proceeds under an ‘active’ mechanism. They argue that an active
mechanism is needed to explain the final orientation of cellular tissue. Nevertheless, the field is still lacking a detailed quantitative comparison with data to systematically rule out different models. For example, how do we know the data comparison can not be improved by altering model parameters instead of adding an ‘active’ mechanism in the energetic function itself? In addition, how do we quantify the performance of the proposed active model in terms of tissue morphology? Are there quantifiable observables that allow us to distinguish whether a ‘passive’ or an ‘active’ model is a better fit to the live-cell data?

Based on these unanswered questions and recent developments, the goals of this study are to (i) create predictive models guided by different mechanisms to compare with experimental data; (ii) identify additional physical observables that allow detailed quantitative comparison between experiment and model; (iii) draw mechanistic conclusions about GBE based on these comparisons. Our analysis reveals the inadequacy of the ‘passive’ model, as it fails to reproduce several important features of the experimental observations during tissue extension. Furthermore, our new quantitative model based on an ‘active’ mechanism during elongation successfully describes several metrics derived from experiment, giving us new fundamental and mechanistic insights into GBE.

6.1 Materials and Methods

6.1.1 Vertex-based model

In our model, a group of cells is represented by a two-dimensional cross-section of the germ band tissue perpendicular to the cells’ apical-basal axis, using the assumption that the cross-sectional area is relatively uniform throughout that axis. With this assumption, the three-dimensional cell volume becomes proportional to
the two-dimensional area of an individual cell and the three-dimensional surface area (excluding the apical and basal caps) becomes proportional to its two-dimensional perimeter. In addition, cells are idealized as polygons within the grid. With this simplification, the grid is uniquely characterized through the positions and connectivity of the corners (hereafter referred to as vertices) of each cell. The tissue grid is allowed to evolve over time using a Monte-Carlo-like set of update rules (in our case involving random displacement of individual vertices) with an energy function used for calculating the probability of allowing specific grid updates. The energy function $E$ for a given configuration is defined as

$$E = \alpha \sum_{i}^{\text{cells}} (A_i - A_0)^2 + \Gamma \sum_{i}^{\text{cells}} P_i^2 + \sum_{j}^{\text{ints}} \Lambda_j \ell_j,$$

(6.1)

The first two terms enforce internal constraints, so that increasing a cell’s area above or below a certain target value or increasing a cell’s perimeter to larger values will incur an energetic penalty, representing the cell’s resistance to compression or membrane stretching. The relative contributions of area conservation and perimeter contractility are dictated by $\alpha$ and $\Gamma$, while $A_i$ and $P_i$ are the area and perimeter of the $i^{\text{th}}$ cell for a given configuration, and $A_0$ is the target area of these cells (typically set to the area of an ideal hexagon with a side length of one for ease of calculation). The characteristic energy of membrane bonds formed at the interface between neighboring cells is given by the last term in the energy function. $\Lambda_j$ is the generalized line tension value (incorporating both tensile and adhesive forces$^{260-262}$) of the $j^{\text{th}}$ cell-cell interface having a length $\ell_j$. An interface is defined as the side shared between two cells.
**General anisotropic interfacial interaction**

In order to introduce anisotropy between vertical and non-vertical interfaces (see section 6.1.3 for definition of “vertical” and “non-vertical”), $\Lambda_j$ is assigned different values at the start ($t = 0$) of the simulation,

$$\Lambda_j = \begin{cases} 
\Lambda & \text{for non-vertical interfaces}, \\
\delta & \text{for vertical interfaces}.
\end{cases} \quad (6.2)$$

We impose the condition $\Lambda < \delta$ to ensure that contraction of vertical interfaces is energetically more favorable compared to contraction of non-vertical interfaces. Like reported previously, after full contraction of AP junctions into T2 vertices, the cells’ elastic internal constraints combined with the basic ‘jitter’ of the vertex positions are sufficient to produce elongation of new T3 interfaces (which is necessary for functional tissue extension). At this step of the simulation, we need to assign interfacial tension for these *newly* formed interfaces, giving us two choices: i) In the simplest model, all the new horizontal interfaces are assigned value $\Lambda$, the same value as the transverse interfaces. In this version, the tissue elongates in the horizontal direction simply due to restoring effects from area conservation and perimeter contractility, hence the model is termed ‘passive’. ii) In an alternate model, new interfaces are assigned a value $\epsilon$, with $\epsilon < \Lambda$. This condition simulates a lower net tension along the horizontal compared to the transverse, that could originate either from the activity of medial myosin in surrounding cells or via stabilization through the addition of E-cadherin complexes. In this version, the tissue has an additional energetic advantage to elongating horizontal interfaces beyond just the considerations of area and perimeter conservation. This model is termed ‘active’ due to this direct mechanism.
for elongation. Since the relative contribution of each energetic term is more important than the absolute value of these quantities, we use dimensionless variables in our analysis, \( \tilde{\Gamma} = \Gamma / (\alpha A_0) \), \( \tilde{\Lambda} = \Lambda / (\alpha A_0^{3/2}) \), \( \tilde{\delta} = \delta / (\alpha A_0^{3/2}) \) and \( \tilde{\epsilon} = \epsilon / (\alpha A_0^{3/2}) \).

### 6.1.2 Monte-Carlo-like energy minimization protocol

To allow a configuration of cells to evolve over time and minimize its overall energy, the system goes through a series of iterations where a vertex is randomly selected and randomly displaced based on a symmetric two-dimensional Gaussian probability distribution with a standard deviation of 10% of the side length of an ideal hexagon with area \( A_0 \). Vertices touching the boundary of cellular configurations are allowed to move freely and are included in these selections for perturbation. After each of these perturbations, the simulation algorithm will always accept moves that cause a decrease in energy, while the likelihood of accepting an increase in energy depends on a ratio of Boltzmann factors, a criterion commonly known as the Metropolis criterion:

\[
P(\Delta E) = \begin{cases} 
e^{-\Delta E} & \text{for } \Delta E > 0 \\ 1 & \text{for } \Delta E < 0 \end{cases},
\]

where \( \Delta E \) is the difference between the energies of the configuration before and after the random vertex perturbation. This iterative process is continuously repeated until the total energy of the system has equilibrated (see Appendix E for details). It should be noted that the number of iterations is not directly analogous to experimental time, so comparisons between simulation and experiment in this paper will be represented as functions of the physical parameter of contracting interface length rather than Monte-Carlo steps. More specifically, metrics of cells surrounding all contracting
interfaces of a specific length are averaged and plotted as a function of this length, with T1 interface lengths being negative in magnitude and T3 interface lengths being positive in magnitude (Fig. 6.1B). In order to test the robustness of our result we have also performed simulations using an alternate acceptance criteria (see Appendix E for details).

6.1.3 Imaging protocol and grid initialization

Live-cell imaging data was collected using a spinning-disk laser confocal microscope from Zeiss and Solamere Technologies Group with a 63x/1.4 NA objective. Images begin before stage 7 of Drosophila development, after full cellularization but before any indication of GBE, and continue through the completion of GBE at stage 9. Embryos express both Spider-GFP and Resille-GFP, membrane markers used to delineate cell outlines, and these images are processed using a watershed segmentation algorithm developed in MATLAB. From here, the length scales of these images must be normalized to enable proper comparison with each other as well as with that of simulations. Rather than rescaling the average cell area of a frame to $A_0$, we note that in some conditions, cells will shrink to an equilibrium area below $A_0$ due to constraints other than area conservation. In order to account for this, we rescale cell configurations to make the average cell area of a frame equal to that of the equilibrium area. This rescaling factor is determined by minimizing the energy function for a given set of parameter values (i.e. $\Gamma$, $\Lambda$, $\delta$, $A_0$) and is detailed further in Appendix E.

Upon this rescaling, cellular tissues can be assessed using the same metrics described in section 6.1.4 and directly compared to simulated results. Four different cellular configurations from experiment are used as starting points for simulations. To initialize line tension anisotropy within these grids, we identify all interfaces that
are observed to fully contract throughout the experiment. All interfaces within $\pm 15^\circ$ of the average angle of these interfaces are defined as ‘vertical’ and assigned the anisotropic line tension value of $\bar{\delta}$.

### 6.1.4 Metrics for comparison of model to experiment

Because cell sizes are relative in simulation, direct area and perimeter comparisons to experimental observations are not as meaningful as unitless metrics that depend on relative measurements rather than absolute sizes. As a result, interface lengths will be presented as fractions of equilibrium lengths, i.e. the length of a single interface of a regular hexagon at equilibrium size based on the given conditions (see Appendix E). To differentiate T1 and T3 interfaces in a transition, T1 lengths will be graphically represented as negative lengths and T3 lengths will be positive.

To compare cell shape evolution, we use the shape factor (SF) of each cell, a relative measure of the roundness of cells defined as $SF = P^2/(4\pi A)$, where $A$ and $P$ are the area and perimeter of a given cell. The scaling coefficient of $4\pi$ sets the shape factor of a perfectly round circle as unity, while oblong shapes obtain larger shape factors the more eccentric or ‘spiky’ they become. For comparison, the shape factor of a perfect hexagon is $2\sqrt{3}/\pi \approx 1.10$.

To assess the level of tissue elongation, we use a tissue extension metric defined as the ratio of the separation distance between two initially neighboring columns of cells and the separation between two initially neighboring rows of cells (Fig. 6.1C and Fig. E.1 in Appendix E). As the tissue evolves, cell rows intercalate into each other while cell columns spread apart to make room (Fig. 6.1B), causing the tissue extension metric to increase. The local nature of this metric makes it more accessible for experimental comparison since a global metric involving the entire cell grid would require extended views of entire tissues, a difficult task experimentally. Utilizing
a local directional frame of reference rather than a universal one also adds further
versatility to this metric, since a cell’s directional orientation can rotate throughout
convergent extension, as the tissue reshapes and folds back on itself. For future
comparison, the tissue extension value of a perfectly hexagonal starting grid in T1
configuration is $2/\sqrt{3} \approx 1.15$ while the value for a perfectly hexagonal fully extended
grid in T3 configuration is $2\sqrt{3} \approx 3.46$.

In addition, we use topological metrics to compare the morphology of simulated
and experimental tissues, based in part on previous studies. If a cell grid is largely
composed of hexagons and fairly regular, the distribution of the population percentage
with a certain number of sides will resemble a sharply-peaked function centered at
six sides. On the other hand, if a cell grid is more disorganized it may contain cells
that deviate from hexagons and resemble pentagons, heptagons, octagons, etc. In
this scenario, the grid may contain significant amounts of cells with every possible
number of sides, causing the distribution to be much wider and centered at any
number of sides. To eliminate environmental effects and ensure that we are only
considering inner tissue dynamics and morphology, any cell that touches the outer
environment at any point during a simulation is eliminated from our statistics. All
of these metrics will be used to quantitatively compare experimental and simulated
tissue configurations.
6.2 Results and Discussion

6.2.1 ‘Passive’ line tension model fails to reproduce data during extension.

We first present the results of the simple ‘passive’ line tension model by applying our energy function (equation 6.1) to experimental live-cell movies of in vivo GBE. The values of the model parameters scaling the relative contributions of perimeter contractility and generalized line tension were chosen as $\bar{\Gamma} = 0.04$ and $\bar{\Lambda} = 0.12$ based on previously published data.$^{241}$ The value of the anisotropic line tension for vertical interfaces was set as $\bar{\delta} = 0.30$ in order to best fit the progression of the shape factor during the converging interface lengths. To account for any misalignment between different movies in regards to timing within GBE, the live cell movies were aligned to an effective ‘convergence midpoint’ at $t = 0$. This time point was determined by noting the first instance where the percentage of pentagons in the tissue exceeds the percentage of hexagons, resembling the onset of T2 or transition between T1 and T3 (see Fig. 6.1B). While the contracted state of the tissue is in principle geometrically less favorable than the initial hexagonal configuration, the incorporated line tension anisotropy causes an energy decrease during the contraction phase (see negative time values of Fig. 6.2B), as expected. However, during the second half of this process when the initial line tension anisotropy has disappeared, there is a curious lack of energetic reduction despite the fact that the tissue continues to extend (positive time values of Fig. 6.2B). This finding contradicts the expectation that tissue reorganization must be associated with lowering of energy. As a result, we must conclude that the ‘passive’ model provides an incorrect explanation of experimental convergent extension.

Furthermore, if the same initial experimental configurations are allowed to evolve independently via simulation within this ‘passive’ model using the parameter values
Figure 6.2: ‘Passive’ model produces accurate convergence, but misrepresents extension. (A1,A2,A3) Experimental cell grids throughout the progression of GBE, as labeled in subfigure (B). (B) Energetic evolution of experimental cell grids over time with no tension anisotropy (purple) and the simplest tension anisotropy model (red). Comparison of (C) cellular shape factor and (D) tissue extension between ‘passive’ simulation (blue) and experiment (red) as a function of contracting interface length. (E) Comparison of final cell grid topologies between ‘passive’ simulation (blue) and experiment (red). Both simulation ($\bar{\Gamma} = 0.04$, $\bar{\Lambda} = 0.12$, $\bar{\delta} = 0.30$, $N = 4$ starting configurations from experiment, 10 simulations each, 2268 cells total) and experiment ($N = 4$ embryos, 268 cells total) begin with the same initial cellular configurations (orange).

mentioned above, several metrics of tissue morphology during the extension phase of GBE are inconsistent with the predictions from our simulations (Fig. 6.2 C,D, and E). For example, consider the dynamics of the shape factor, a relative measure of how oblong a cell is (Fig. 6.2C, see section 6.1.4 for definition). The initial contraction
of an interface during T1 reduces the number of sides of the two participating cells (e.g. conversion from hexagons to pentagons). This reduction of the number of sides is necessarily accompanied by an increase of the perimeter-to-area ratio, captured by the increase in the shape factor (see negative lengths of Fig. 6.2C). After full contraction and formation of a higher-order vertex in T2, the creation of a new interface should have the inverse effect, in that it allows the participating cells to add a side (e.g. changing from a pentagon to a hexagon shape) and therefore create a ‘rounder’ and energetically more favorable shape, reflected in the model by a sharp drop in the shape factor (positive lengths of Fig. 6.2C, blue). Surprisingly, we did not observe a similar drop in the shape factor associated with new interface elongation in the experimental data (positive lengths of Fig. 6.2C, red). While cells increase their shape factor as expected during contraction, they fail to reduce it significantly during the elongation phase and retain high values into full extension, indicating that in experiment, cells do not return to the more ‘round’ shape. The energetics of perimeter and area elasticity also dictate a specific pace of tissue extension as a function of interface length, which is determined by the energetic advantage of new interface elongation compared to that of shape factor minimization. Similar to the shape factor metric, the gradual tissue extension produced during the convergence phase of simulation agrees well with experiment, but deviates significantly during the extension phase due to a sudden increase in pace (Fig. 6.2D, blue). The basic line tension model also predicts that simulations starting with ordered hexagonal grids taken directly from experiment (Fig. 6.1D1 and 6.1E1) should return to ordered hexagonal final configurations (see Fig. 6.1D2 and 6.2E, blue). However, we observe in experiment that the final configuration after GBE is significantly more disordered than the initial configuration with a large percentage of non-hexagonal cells (Fig. 6.1E2 and 6.2E, red).
6.2.2 Softer conditions during GBE do not resolve discrepancies between model and experiment.

In the preceding section, we have shown that there are major discrepancies between the predictions of the ‘passive’ line tension model and experiment by comparing several straightforward metrics of tissue morphology. What modifications to the model are needed to resolve these discrepancies? One immediate hypothesis is that perhaps the energy function is adequate, but the magnitudes of $\tilde{\Gamma}$ and $\tilde{\Lambda}$ are incorrect. In the simulations above, we used parameter values based on a previously published model of wing epithelium,\textsuperscript{241} and while these values can also been used in the context of GBE, to our knowledge, they have not been independently and rigorously validated in germ band epithelium. Thus, can we improve the agreement between simulation and experiment by simply altering the parameter values without changing the basic premise of the model?

While exploring different parameter values, we draw attention to two major regions of parameter space: i) stiff conditions and ii) soft conditions, as characterized by Farhadifar et al.\textsuperscript{241} Stiff network conditions, shown in the gray region in Fig. 6.3A, are dominated by elastic considerations and have one lowest energy cellular configuration, hence making them resistant to external forces or perturbations. Soft network conditions on the other hand, represented by the cyan region in Fig. 6.3A, are dominated by adhesive considerations and have a degenerate lowest energy state. Multiple minima make the system more likely to absorb external forces or perturbations introduced to the system and respond differently, hence the term soft. There is a third region (see dashed region of Fig. 6.3A) where tension and contractility are so dominant that no lowest energy state exists and the entire cellular grid shrinks down into a point. The study of Farhadifar et al.\textsuperscript{241} showed that the experimental model system of wing
Figure 6.3: **Softer conditions are not enough to resolve discrepancies.** (A) Phase space of the normalized parameters for perimeter contractility and generalized line tension, where gray represents stiff conditions, cyan represents soft conditions, and the individual points correspond to the simulation parameters used in (B)-(D). Comparison of (B) final cell grid topologies as well as (C) cellular shape factor and (D) tissue extension as a function of contracting interface length between simulations using parameters from literature ($\bar{\Gamma} = 0.04$, $\bar{\Lambda} = 0.12$, $\bar{\delta} = 0.30$, $N = 4$ starting configurations from experiment, 10 simulations each, 2268 cells total, blue), simulations using softer parameters ($\bar{\Gamma} = 0.04$, $\bar{\Lambda} = -0.06$, $\bar{\delta} = 0.05$, $N = 4$ starting configurations from experiment, 10 simulations each, 2425 cells total, magenta; $\bar{\Gamma} = 0.04$, $\bar{\Lambda} = -0.16$, $\bar{\delta} = -0.10$, $N = 4$ starting configurations from experiment, 10 simulations each, 2437 cells total, green; $\bar{\Gamma} = 0.04$, $\bar{\Lambda} = -0.24$, $\bar{\delta} = -0.21$, $N = 4$ starting configurations from experiment, 10 simulations each, 2485 cells total, black), and experimental observations ($N = 4$ embryos, 268 cells total, red).

Epithelium is a **stiff** network and they further established that cells in stiff networks have a strong area dependency on their number of sides while soft networks have no such dependency. Tissues throughout GBE retain this area dependency (Fig. 6.4, red) implying that they are likely to reside within stiff conditions, but to the best of our knowledge, ideal simulation conditions (i.e. $\bar{\Gamma}$, $\bar{\Lambda}$, $\bar{\delta}$) have not been determined for GBE. It is conceivable that the more disordered morphology experimentally observed in GBE is caused by operating in a **softer** network than originally thought.
This is an intuitively appealing idea that could potentially explain the failure of the experimental grids to relax back to hexagonal configurations during the extension phase.

To test this hypothesis, we approached the soft condition by focusing on a particular region of the phase space, specifically by maintaining a constant amount of contractility while gradually decreasing the level of generalized tension/adhesion in our simulations (Fig. 6.3A, blue, magenta, green, and black points). Softer conditions would suggest that GBE occurs in an environment either less dominated by tension or aided by a stronger contribution from adhesive mechanisms throughout the entire tissue. Biologically, this could happen through stabilization of adhesion sites, or through reorganization of cellular actomyosin networks after cellularization. As $\Lambda$ decreases and the system approaches softer conditions, we do not see any significant improvement in the agreement of the progression of shape factor and tissue extension as newly formed interfaces extend (see positive lengths in Fig. 6.3C, 6.3D). Furthermore, even at $\Lambda = -0.16$, the final topologies of softer conditions are not significantly more disordered than the final topologies of our original stiff conditions and are still in disagreement with the heterogeneous final topologies of the experiment (Fig. 6.3B, 6.5A and 6.5B).
While soft conditions restrict the choice of $\Gamma$ and $\Lambda$, $\delta$ is left as a free parameter to explore and we use the value of $\delta = -0.10$ (when $\Lambda = -0.16$) to best match the contracting lengths of the shape factor graph with experiment (negative lengths of Fig. 6.3C). However, it is important to note that $\delta$ primarily dictates behaviors during the contraction phase and has no role during the extension phase. Therefore, an exhaustive search in $\delta$ can not improve the agreement with data during the extension phase. For completeness, a similar parameter search was conducted by maintaining a constant value of $\Lambda$ and decreasing $\Gamma$ and the same general outcomes were produced. Considering these results, we conclude that a shift towards softer conditions is not sufficient to accurately describe the experimental system in full.

6.2.3 Presence of ‘active’ extension mechanism is a plausible explanation based on experimental observations.

The analysis above suggests that the detailed exploration of the parameter space within the ‘passive’ model is not sufficient to reproduce experimental observations.
Specifically, it is not enough to resolve the discrepancies between the ‘passive’ model and the experimental morphology during the elongation phase. This motivates us to explore an alternative hypothesis: an ‘active’ extension mechanism for newly formed interfaces could make newly elongating horizontal T3 interfaces energetically more favorable than the existing transverse interfaces. The biophysical basis for this mechanism could be local interface anisotropies (such as the targeted recruitment of E-cadherin protein or actin to actively stabilize new cell-cell contacts) or possibly global ‘external’ force generation mechanisms. In our computational model, we implement this ‘active’ model by assigning newly formed interfaces a different line tension value than existing transverse interfaces (see section 6.1.1 for details). The generalized tension value of these interfaces is initialized as $\bar{\epsilon}$, with the constraint that $\bar{\epsilon} < \bar{\Lambda}$ to ensure that these newly formed interfaces have an energetic incentive to extend over time. For consistency, the values of $\bar{\Gamma}$, $\bar{\Lambda}$, and $\bar{\delta}$ were returned to the original values proposed in the literature and the value of $\bar{\epsilon} = -0.12$ was chosen to best fit final ex-
perimental topologies (Fig. 6.6A). Using these parameter values, we predicted the progressions of shape factor and tissue extension and compared them against experimental data.

Figure 6.7: **Active extension mechanism resolves energetic discrepancies from experiment.** Energetic evolution of experimental cell grids over time with no tension anisotropy (purple), only first-order tension anisotropy (red), and both first-order and second-order anisotropy (cyan) \( (\overline{\Gamma} = 0.04, \overline{\Lambda} = 0.12, \overline{\delta} = 0.30, \overline{\epsilon} = -0.12, N = 4 \) embryos, 268 cells total).

The simulation results show that an active elongation mechanism in a stiff network successfully recapitulates the unique behaviors of the shape factor and tissue extension metrics during the extension phase of GBE. In the case of the shape factor (Fig. 6.6B), the energetic advantage from elongating a new interface is so large that elongation can take place even when that elongation makes the cells more oblong. As for the extension metric, in passive extension conditions (Fig. 6.2D), the degree of local tissue extension is dependent on the extension of the tissue’s neighboring cells since they must be able to make room for this reconfiguration. With active extension conditions (Fig. 6.6C), the energetic gain of interface extension is high enough that this dependency is no longer dominant and new interfaces can elongate even when neighboring cells do not expand into the AP direction. This added mechanism can also reconcile the larger degree of disorder in final tissue topologies as seen in the population percentage histogram in Fig. 6.6A. Figure 6.5C also shows a sample final cell grid from simulation using the ‘active’ extension model that is in qualitative agreement with experimental tissues (Fig. 6.1E2). As expected, application of the energy function to the experimental data in Fig. 6.2B using the ‘active’ mechanism also produces the desired continuously decreasing energy
function (see Fig. 6.7, cyan). In summary, the simulated predictions from our ‘active’
extension model provide consistently better agreement with experimentally measured
morphology compared to the ‘passive’ line tension model, supporting an alternative
tissue elongation mechanism - local or global - that has not yet been explored in
depth.

6.3 Conclusion

Our results indicate that while the ‘passive’ line tension model is reasonably suc-
cessful at describing contraction and generating higher-order vertices as nucleations
for elongation of new interfaces, the mechanism for convergent extension must be
more intricate, since major discrepancies arise when comparing such models to quanti-
tative data. The ‘passive’ model for extension predicts that tissue grids should return
to a relatively ordered hexagonal morphology during the extension phase, while dif-
ferent metrics of tissue morphology show that experimentally observed tissues retain
the more disordered topology of the contracted state, even after reaching full tis-
sue extension. Two possible explanations for this discrepancy were explored in this
study. First, we explored whether the discrepancy could be due to an incorrect choice
of parameters, since parameter values were motivated by a wing epithelium study that
may be inappropriate to describe germ band epithelium. To investigate this, we
analyzed sets of parameters associated with soft conditions. In biological settings,
this could be realized through a stabilization of cell-cell adhesion sites over time or
a lower contribution of actomyosin forces than previously predicted. While softer
grid conditions with lower generalized tension values produce a slight improvement
in the agreement of cellular shape factor and tissue extension, the level of topological
disagreement between simulation and experiment is too large to establish these softer conditions as a legitimate rationale to justify the ‘passive’ line tension model.

Next, we changed the ‘passive’ model to an ‘active’ model. In this novel ‘active’ model, we introduce an additional level of anisotropy during the elongation phase for the newly formed interfaces. Biologically, the additional anisotropy at a later stage can be envisioned as a direct targeting of adhesive E-cadherin molecules or structural F-actin networks to stabilize these new contacts, but it would also be consistent with external force generating mechanisms, or a combination of these mechanisms as recently proposed.\textsuperscript{10} This version of the GBE model is novel and it produces a much closer agreement with experiment in three observables: i) it retains high cellular shape factors throughout interface elongation, ii) it produces more gradual tissue extension throughout interface elongation, and iii) it generates heterogeneous topologies matching those observed in experimental configurations. This explanation shifts the attention from the previously all-important nucleation step of interface contraction to the relatively unexplored interface extension step of GBE. Our systematic model development is able to test these different mechanisms of GBE through a direct combination of modeling and quantitative analysis of experimental data at a level not achieved before. This exhaustive study highlights the need for maximal usage of qualitative and quantitative data to carefully select competing models in morphogenesis, in line with other areas of physical biology.
Part V

Concluding Remarks
Across multiple length scales, this work highlights the prevalence of stochasticity throughout biology and how noisy information can be utilized to advance design principles in synthetic biology. At a molecular scale, disorder presents itself in the form of IDPs. Despite these proteins not having a defined folded state, our analytical model can quantitatively predict the conformation size of any given sequence and even identify specific “hot spots” within the sequence that produce significant changes in conformation upon post-translational modification. This high-throughput nature leads to a better understanding of IDPs at a proteomic level and could drive conformational design principles for future studies. At the scale of gene networks, stochasticity is perfectly exemplified by stochastic gene expression. Using multiple motifs ubiquitous throughout biology, we have shown that protein number fluctuations are inherent to \textit{in vivo} conditions. Through the versatile modeling system of MaxCal, this stochasticity can be exploited by researchers to make meaningful predictions from the full information contained in noisy experimental trajectories. This added knowledge provides quantitative insight into microbial evolution under stress and ultimately allows us to produce specific genetic phenotypes through circuit design. Finally, at the multicellular organismal scale, epithelial tissues take disordered geometries to achieve sufficient extension via single-cell extension mechanisms. Randomness and fluctuation do not always imply disarray and intractibility, but instead can convey adaptability and possibility. As such, these characteristics should be embraced by the field of biophysics.
Bibliography


159


[206] McLean, P.; Smolke, C.; Salit, M. bioRxiv 2016,


164


Appendix A

Derivation of Renormalized Kuhn Length

Here, we outline the derivation of equation 1.5. A variational approach, originally introduced by Edwards-Singh\textsuperscript{120} to compute the average size of homopolymers and later extended by others to polyelectrolytes\textsuperscript{121–123} and heteropolymers,\textsuperscript{16} is used to give an alternate formulation to the free energy method presented in the main text. Within this variational scheme, we map the total Hamiltonian ($H_t$ in equation 1.1), with all interactions, to a renormalized Hamiltonian ($H_r$) such that

$$\beta H_r = \left( \frac{3}{2l_r} \int_0^L ds \left( \frac{dR(s)}{ds} \right)^2 \right)$$

(A.1)

where $l_r$ is the renormalized Kuhn length and is a function of the excluded volume, three-body and electrostatic interaction parameters. The ensemble average of some physical observable $A$, with respect to the total Hamiltonian $H_t$ up to the first order in $(H_r - H_t)$, can be expressed as

$$\langle A \rangle = \langle A \rangle_r + \langle A \rangle_r \langle (H_r - H_t) \rangle_r - \langle A(H_r - H_t) \rangle_r + O[(H_r - H_t)^2]$$

(A.2)
where \( \langle \ldots \rangle_r \) denotes averages with respect to the renormalized Hamiltonian \( H_r \), and \( \langle \ldots \rangle \) denotes average over the original Hamiltonian \( H_t \). Averages are computed over all possible configurations, and thus involve functional integrals. The effective Kuhn length \( l_r \) can be determined by demanding \( \langle A \rangle \approx \langle A \rangle_r \). Therefore, the equation for the renormalized Kuhn length is given by

\[
\langle A \rangle_r \langle (H_r - H_t) \rangle_r = \langle A(H_r - H_t) \rangle_r. \tag{A.3}
\]

The choice of \( A \) depends on the quantity of interest. For the current problem, we are interested in determining the average end-to-end distance, \( R_{ee} \), hence we set \( A = [R_{ee}]^2 = (R(L) - R(0))^2 \). This yields the following equation:

\[
\left\langle \left( \frac{1}{l_r} - \frac{1}{l} \right) \int_0^L ds \left( \frac{dR(s)}{ds} \right)^2 \right\rangle_r - \left( \langle R_{ee}^2 \rangle \right)_r \left\langle \frac{3}{2} \left( \frac{1}{l_r} - \frac{1}{l} \right) \int_0^L ds \left( \frac{dR(s)}{ds} \right)^2 \right\rangle_r
\]

\[
= \left( \langle R_{ee}^2 \rangle \right)_r \left\langle l \int_0^L ds \int_0^s ds' \omega(s, s') \delta[R(s) - R(s')] \right\rangle_r
\]

\[
- \left( \langle R_{ee}^2 \rangle \right)_r \left\langle l \int_0^L ds \int_0^s ds' \omega(s, s') \delta[R(s) - R(s')] \right\rangle_r
\]

\[
+ \left( \langle R_{ee}^2 \rangle \right)_r \left\langle l_b \left( \frac{l}{2} \right) \int_0^L ds \int_0^s ds' q(s) q(s') \exp(-\kappa[R(s) - R(s')]) \right\rangle_r
\]

\[
- \left( \langle R_{ee}^2 \rangle \right)_r \left\langle l_b \left( \frac{l}{2} \right) \int_0^L ds \int_0^s ds' q(s) q(s') \exp(-\kappa[R(s) - R(s')]) \right\rangle_r
\]

\[
+ \left( \langle R_{ee}^2 \rangle \right)_r \left\langle \frac{l}{2} \int_0^L ds \int_0^s ds' \int_0^{s'} ds'' \delta[R(s) - R(s')] \delta[R(s') - R(s'')] \right\rangle_r
\]

\[
- \left( \langle R_{ee}^2 \rangle \right)_r \left\langle \frac{l}{2} \int_0^L ds \int_0^s ds' \int_0^{s'} ds'' \delta[R(s) - R(s')] \delta[R(s') - R(s'')] \right\rangle_r
\]

Using properties of Gaussian chains and calculations presented elsewhere,\textsuperscript{16,111,120} the term in the left hand side of equation A.4 simplifies to

\[
\left( \frac{1}{l_r} - \frac{1}{l} \right) \int_0^L ds \left( \frac{dR(s)}{ds} \right)^2 \right\rangle_r - \left( \langle R_{ee}^2 \rangle \right)_r \left\langle \frac{3}{2} \left( \frac{1}{l_r} - \frac{1}{l} \right) \int_0^L ds \left( \frac{dR(s)}{ds} \right)^2 \right\rangle_r
\]

\[
= l_r^2 \left( \frac{1}{l_r} - \frac{1}{l} \right). \tag{A.5}
\]
Following our earlier work (see Appendix of Reference\textsuperscript{16}), the excluded volume interaction terms in the right hand side of equation A.4 can be rewritten as,

\[
\langle (R_{ee})^2 \int_0^L ds \int_0^s ds' \omega(s, s') \delta[R(s) - R(s')] \rangle_r - \langle (R_{ee})^2 \rangle_r \langle \int_0^L ds \int_0^s ds' \omega(s, s') \delta[R(s) - R(s')] \rangle_r \tag{A.6}
\]

\[
= -\frac{l_r^2}{9} \left[ \int_0^L ds \int_0^s ds' \omega(s, s')(s - s')^2 \int \frac{d^3kk^2}{(2\pi)^3} \exp \left( -\frac{k^2l_r|s - s'|}{6} \right) \right]
\]

where the last equality follows using the properties of a Gaussian chain propagator.\textsuperscript{111,120} Following our previous work and using the relation below,

\[
\int \frac{d^3k}{(2\pi)^3} \frac{k^2}{l_r^2} \exp \left( -\frac{k^2l_r|s - s'|}{6} \right) = \left( \frac{1}{2\pi} \right)^{5/2} \left( \frac{3}{l_r|s - s'|} \right)^{5/2} \tag{A.7}
\]

we simplify equation A.6 as,

\[
\langle (R_{ee})^2 \int_0^L ds \int_0^s ds' \omega(s, s') \delta[R(s) - R(s')] \rangle_r - \langle (R_{ee})^2 \rangle_r \langle \int_0^L ds \int_0^s ds' \omega(s, s') \delta[R(s) - R(s')] \rangle_r \tag{A.8}
\]

\[
= -\frac{1}{l_r^{1/2}} \left( \frac{3}{2\pi} \right)^{3/2} \left[ \int_0^L ds \int_0^s ds' \omega(s, s') \frac{1}{(s - s')^{1/2}} \right]
\]

Next, following our earlier work (see Appendix of reference\textsuperscript{16}), the electrostatic contributions in equation A.4 can be simplified as,

\[
\langle (R_{ee})^2 \frac{l_b}{l^2} \int_0^L ds \int_0^s ds' q(s) q(s') \exp\left( -\kappa|R(s) - R(s')| \right) \frac{\exp\left( -\kappa|R(s) - R(s')| \right)}{|R(s) - R(s')|} \rangle_r - \langle (R_{ee})^2 \rangle_r \langle \frac{l_b}{l^2} \int_0^L ds \int_0^s ds' q(s) q(s') \exp\left( -\kappa|R(s) - R(s')| \right) \frac{\exp\left( -\kappa|R(s) - R(s')| \right)}{|R(s) - R(s')|} \rangle_r \tag{A.9}
\]

\[
= -\frac{4\pi l_b}{l^2} \frac{l_r^2}{9} \left[ \int_0^L ds \int_0^s ds' q(s) q(s')(s - s')^2 \int \frac{d^3kk^2}{(2\pi)^3(k^2 + \kappa^2)} \exp \left( -\frac{k^2l_r|s - s'|}{6} \right) \right]
\]
We now simplify the above integral by using the identity

$$
\int \frac{d^3k k^2}{(k^2 + \kappa^2)} \exp \left( -\kappa^2 l_v |s - s'| \right) = 4\pi \left[ \frac{\pi^{1/2}}{4} \left( \frac{6}{l_v (s - s')} \right)^{3/2} - \frac{\pi^{1/2}}{2\kappa^2} \left( \frac{6}{l_v (s - s')} \right)^{1/2} + \frac{\pi}{2\kappa^3} \exp \left( \frac{\kappa^2 l_v (s - s')}{6} \right) \right]
$$

(A.10)

Taking the zero salt limit ($\kappa = 0$), we can further simplify equation A.9 as

$$
\left\langle \frac{(R_{ee})^2}{l^2} \int_0^L ds \int_0^s ds' q(s)q(s') \frac{1}{|R(s) - R(s')|} \right\rangle_r
- \left\langle \frac{(R_{ee})^2}{l^2} \int_0^L ds \int_0^s ds' q(s)q(s') \frac{1}{|R(s) - R(s')|} \right\rangle_r
= \frac{1}{l^2} \left[ \int_0^L ds \int_0^s ds' q(s)q(s')(s - s')^{1/2} \right]
$$

(A.11)

We now focus on the contribution of the three body term in equation A.4. Using the transformation

$$
\delta^3(r) = \int \frac{d^3k}{(2\pi)^3} \exp(i k \cdot r),
$$

(A.12)

we can rewrite

$$
\left\langle \frac{(R_{ee})^2}{l^2} \omega_3 l^3 \int_0^L ds \int_0^s ds' \int_0^{s''} ds'' \delta[R(s) - R(s')][\delta[R(s') - R(s'')]] \right\rangle_r
- \left\langle \frac{(R_{ee})^2}{l^2} \omega_3 l^3 \int_0^L ds \int_0^s ds' \int_0^{s''} ds'' \delta[R(s) - R(s')][\delta[R(s') - R(s'')]] \right\rangle_r
= \left\langle \frac{(R_{ee})^2}{l^2} \frac{d^3k_1}{(2\pi)^3} \int_0^L ds \int_0^s ds' \int_0^{s''} ds'' \exp(i k_1 \cdot [R(s) - R(s')]) \exp(i k_2 \cdot [R(s') - R(s'')]) \right\rangle_r
- \left\langle \frac{(R_{ee})^2}{l^2} \frac{d^3k_1}{(2\pi)^3} \int_0^L ds \int_0^s ds' \int_0^{s''} ds'' \exp(i k_1 \cdot [R(s) - R(s')]) \exp(i k_2 \cdot [R(s') - R(s'')]) \right\rangle_r
$$

(A.13)

Combining equations A.8, A.11, and A.13, and switching to discrete notation by setting $s = pl$, $s' = ml$, $s'' = nl$, $x = l_v/l$ and $L = Nl$, we get the final equation 1.5 where $\Omega$, $Q$, and $B$ are defined by equation 1.3.
Appendix B

Simple Approximation for $\langle m_1 \rangle$ and $\langle (\Delta m_1)^2 \rangle$

We take the logarithm of equation 2.3 and use Stirling’s approximation to obtain

$$
\ln P(m_1, m_2) = -(M - m_1 - m_2) \ln(M - m_1 - m_2) - m_1 - m_2 - (N_1 - m_1) \ln(N_1 - m_1) - (N_2 - m_2) \ln(N_2 - m_2) - m_1 \ln m_1 - m_2 \ln m_2 + m_1 \ln x_1 + m_2 \ln x_2 + A
$$

(B.1)

where all the terms independent of $m_1$ and $m_2$ have been included in the term $A$ as normalization. Expanding $\ln P(m_1, m_2)$ about its maximum at $m_1 = m_1^*$, $m_2 = m_2^*$ and keeping terms up to the second order, we get

$$
\ln P(m_1, m_2) \approx \ln P(m_1^*, m_2^*) + \frac{d^2 \ln P(m_1, m_2)}{dm_1^2} \bigg|_{m_1 = m_1^*} \frac{(m_1 - m_1^*)^2}{2} \\
+ \frac{d^2 \ln P(m_1, m_2)}{dm_2^2} \bigg|_{m_2 = m_2^*} \frac{(m_2 - m_2^*)^2}{2} \\
+ \frac{d^2 \ln P(m_1, m_2)}{dm_1 dm_2} \bigg|_{m_1 = m_1^*, m_2 = m_2^*} (m_1 - m_1^*)(m_2 - m_2^*)
$$

(B.2)
where,

$$\frac{d\ln P(m_1, m_2)}{dm_1} \bigg|_{m_1 = m_1^*} = 0 \text{ yields } x_1 = \frac{m_1^*}{(M - m_1^* - m_2^*)(N_1 - m_1^*)} \quad (B.3)$$

and

$$\frac{d\ln P(m_1, m_2)}{dm_2} \bigg|_{m_2 = m_2^*} = 0 \text{ yields } x_2 = \frac{m_2^*}{(M - m_1^* - m_2^*)(N_2 - m_2^*)} \quad (B.4)$$

Combining equation B.1 and equation B.2, we simplify further to get

$$P(m_1, m_2) = \mathcal{N} \exp \left( -\frac{P_1(m_1 - m_1^*)^2}{2} - \frac{P_2(m_2 - m_2^*)^2}{2} - Q(m_1 - m_1^*)(m_2 - m_2^*) \right) \quad (B.5)$$

where $\mathcal{N}$ is the normalization constant and

$$Q = \frac{1}{M - m_1^* - m_2^*}; \quad P_1 = Q + \frac{1}{N_1 - m_1^*} + \frac{1}{m_1^*}; \quad P_2 = Q + \frac{1}{N_2 - m_2^*} + \frac{1}{m_2^*} \quad (B.6)$$

Using properties of Gaussian distribution, we identify the average ($\bar{m}_1$) and variance ($\sigma_1^2$) of $m_1$ as

$$\bar{m}_1 = m_1^* \text{ and } \sigma_1^2 = \frac{P_2}{P_1 P_2 - Q^2} \quad (B.7)$$
Appendix C

Application of Finite State Projection to Maximum Caliber

Within the context of MaxCal, time progresses in discrete increments of $\Delta t$ while the original formalism of Finite State Projection (FSP) works within the context of continuous time. This requires us to make some slight modifications. For the discrete-time formulation of FSP, we use the same probability density column vector, $\mathbf{P}(t)$, of length $N+1$ where $N$ is the number of states being considered in our finite state space. Each element in this vector, $P_i(t)$, is the probability at timepoint $t$ of being in a particular state, $i$, with state indexing ranging from zero to $N$. The last state ($i = N$) is considered the “sink” state where probability that enters states outside of our defined finite state space can never escape. Where our discretized FSP differs from the original formulation of FSP is in the calculation of the state reaction matrix, $\mathbf{A}$, of size $N + 1 \times N + 1$. Each element of $\mathbf{A}$ represents a transition probability from some initial state (depending on the column index) to some final state (depending on the row index) using the same state indexing as in $\mathbf{P}(t)$. As mentioned above, any probability transitioning outside of our finite state space must enter the sink state.
and not be allowed to transition out of it. The individual elements of $A$ are therefore given as

\[
A_{ij} = \begin{cases} 
P_{i \to j} & i, j < N \\
0 & i < N, j = N \\
1 - \sum_{k=0}^{i-1} A_{kj} & i = N
\end{cases}
\] (C.1)

where $j$ is the index of the initial state, $i$ is the index of the final state, and $P_{i \to j}$ is the probability of transitioning from state $j$ to state $i$ in a single increment of $\Delta t$.

With this definition of $A$, we can propagate our probability density vector $P$ through a discrete number ($m$) of $\Delta t$ increments via matrix multiplication with $A$ that many times:

\[
P(t + m) = A^m P(t).
\] (C.2)

As in the original formulation of FSP, the last element of $P(t + m)$ is the amount of probability in the sink state and represents an effective measure of error between $P(t + m)$ and the true probability density function at $t + m$. If this error is too large, $N$ can be increased, expanding our finite state space and decreasing the amount of probability absorbed by the sink state. The steady state protein number distribution can also be calculated via FSP by propagating the probability distribution sufficiently far in time to reach relative equilibrium (e.g. 10 times the average dwell time or period).

The utility of FSP is particularly advantageous during value selection for the Lagrange multipliers of MaxCal via maximum likelihood. Given a stochastic trajectory of protein expression (via experiment or simulation), how does one select values for these parameters that adequately represent that trajectory, and more importantly, the underlying genetic circuit that generated it? Considering trajectories frame-by-
frame, we only need equation 5.3 to calculate the likelihood \( L \) of observing that exact trajectory for a particular set of values for these parameters. The logical choice of parameter values would then be the values that produce the highest likelihood. However, experiment and Gillespie algorithms do not have upper limits on protein production analogous to \( M \) in MaxCal. Rare jumps in protein number over a single time step of \( \Delta t \) (i.e. greater than \( M \)) disproportionately punish the likelihood of otherwise suitable parameters. To account for this, we can use FSP to calculate transition probabilities over multiple time increments \( (m \text{ frames}) \) as

\[
P_{j \rightarrow i, m} = (A^m)_{ij}, \quad (C.3)
\]

thus making these larger jumps theoretically possible over multiple frames and avoiding the erroneous penalty of outlying events.
Appendix D

Application of MaxCal to Alternate Auto-Activation Circuits

To further test the accuracy of MaxCal, the inference method described in the main text was applied to an alternate model of self-promotion that excludes the effects of RNA and has monomers binding to the promoter site rather than dimers:

\[
\alpha \xrightarrow{g} \alpha + A ; \ A \xrightarrow{r} \varnothing ; \ \alpha + A \xrightarrow{f_p/b_p} \alpha^* ; \ \alpha^* \xrightarrow{g^*} \alpha^* + A \quad (D.1)
\]

where some generic protein \( A \) is created from its corresponding gene \( \alpha \) at a rate of \( g \), degrades at a rate of \( r \), and binds to the promoter site, \( \alpha \), with forward and backward rates of \( f_p \) and \( b_p \) respectively. This sends \( \alpha \) into or out of its activated state \( \alpha^* \), which creates protein \( A \) at a much faster rate \( g^* \). This again captures the essentials of a positive feedback mechanism, but represents a different level of non-linearity and cooperativity in Hill-type models. This circuit is motivated by the earlier work of Lipshtat et al.\(^{42}\) demonstrating that bimodality in toggle switch circuits can be obtained without cooperative binding. Using reaction rates similar to those utilized
for the model in the main text, the inferred rates and distributions are displayed in Table D.1 and Figure D.1 respectively. These results demonstrate that an acceptable level of accuracy can be generated using MaxCal, regardless of the exact molecular underpinnings of the circuit being considered.

<table>
<thead>
<tr>
<th></th>
<th>True Values</th>
<th>Predicted Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g$ (s$^{-1}$)</td>
<td>$5.0 \times 10^{-3}$</td>
<td>$6.2 \pm 0.1 \times 10^{-3}$</td>
</tr>
<tr>
<td>$g^*$ (s$^{-1}$)</td>
<td>$50.0 \times 10^{-3}$</td>
<td>$45.8 \pm 1.4 \times 10^{-3}$</td>
</tr>
<tr>
<td>$r$ (s$^{-1}$)</td>
<td>$1.0 \times 10^{-3}$</td>
<td>$1.01 \pm 0.03 \times 10^{-3}$</td>
</tr>
<tr>
<td>$\tau_{L \rightarrow H}$ (s)</td>
<td>$59.0 \times 10^3$</td>
<td>$85.2 \pm 3.0 \times 10^3$</td>
</tr>
<tr>
<td>$\tau_{H \rightarrow L}$ (s)</td>
<td>$78.7 \times 10^3$</td>
<td>$105.5 \pm 5.5 \times 10^3$</td>
</tr>
<tr>
<td>$S_I$ (bits)</td>
<td>8.86</td>
<td>9.23 $\pm$ 0.03</td>
</tr>
<tr>
<td>$S_h$ (bits)</td>
<td>9.38</td>
<td>9.02 $\pm$ 0.02</td>
</tr>
<tr>
<td>$S_I$ (bits)</td>
<td>6.25</td>
<td>7.66 $\pm$ 0.02</td>
</tr>
<tr>
<td>$S_{cg}$ (bits)</td>
<td>1.02</td>
<td>1.01 $\pm$ 0.01</td>
</tr>
</tbody>
</table>

Table D.1: Comparison of true rates and predicted rates using MaxCal on alternate self-promotion model. The first column reports “true” underlying protein synthesis and degradation rates used to create synthetic input data ($f_p = 3.56 \times 10^{-6}$ s$^{-1}$, $b_p = 1.65 \times 10^{-5}$ s$^{-1}$), average residence times in the high and low states, and corresponding path informational entropies. Synthetic input data was recorded at $\Delta t = 300$s. The second column reports the average and standard deviation of the same quantities of interest, but extracted using the MaxCal model on ten sets of synthetic data, each consisting of 100 trajectories of 7 days.

Figure D.1: Predicted distributions for alternate model agree well with the “true” distributions. (A) Protein number probability distributions from synthetic input trajectories (blue) and predicted MaxCal trajectories (red). (B) Low state and (C) high state residence time probability distributions for synthetic input trajectories (blue) and predicted MaxCal trajectories (red). Underlying Gillespie reaction rates are the same as those used in Table D.1.
Appendix E

Detailed Protocols for Cell Grid Simulations

Energy minimization and equilibration for cell grid simulations

Simulations of intercalating cellular grids are allowed to continue until the total energy of the system has ‘equilibrated.’ This equilibrium is achieved when the relative fluctuation (defined as the standard deviation divided by the average) of the energy of the entire grid over the last half of the total number of iterations of the simulation is less than 0.005. To ensure that the outcomes of simulations were not dependent on the choice of perturbation acceptance criterion, we also created simulations using the Glauber acceptance criterion [Glauber R. J. Math. Phys. 4, 294 (1963)] in our algorithm:

\[ P(\Delta E) = \frac{e^{\Delta E/2}}{e^{\Delta E/2} + e^{-\Delta E/2}}, \]  

(E.1)

where \( \Delta E \) is the difference between the energies of the configuration before and after the random vertex perturbation. While the dynamics of the system as a function of iterations were slightly faster, the dynamics as a function of contracting interface length were largely identical.
**Edge swapping protocol for cell grid simulations**

When converging interfaces shorten to within a specified “joining” radius (see Table E.1), the two nodes at either end of the interface are automatically conjoined into one higher order node at the average position of the two nodes. From that point on, every time this higher node is selected for perturbation, it has a defined probability of spontaneously splitting into two separate nodes (see Table E.1) with each possible orientation (i.e. back into the vertical direction and newly elongating into the horizontal direction) being equally likely. If the node happens to split into two, these nodes are placed the same distance apart as the previously mentioned joining radius to create a new interface and this interface is assigned a generalized line tension based on the rules defined in section 6.1.1 of the main text.

**Equilibrium length calculation for a given parameter set**

As mentioned in section 6.1.3 of the main text, in stiff conditions, cells will shrink to equilibrium areas below the target area \( A_0 \) described in equation 6.1 of the main text due to the two constraints other than area conservation. To calculate this for any given set of parameters, let’s reconsider the energy function \( E \) from equation 6.1 of the main text. If we consider an infinite cell grid without generalized tension/adhesion anisotropy, each individual cell is contributing an energy of \( E_i \) defined as

\[
E_i = \alpha (A_i - A_0)^2 + \Gamma P_i^2 + \frac{\Lambda}{2} P_i. \tag{E.2}
\]

If we then assume every cell to be an ideal hexagon of the same size, this equation becomes

\[
E_i = \alpha \left( \frac{3\sqrt{3}}{2} \ell_i^2 - \frac{3\sqrt{3}}{2} \ell_0^2 \right)^2 + 36\Gamma \ell_i^2 + 3\Lambda \ell_i. \tag{E.3}
\]
where \( \ell_i \) is the side length of each hexagonal cell and \( \ell_0 \) is the side length of a hexagonal cell with area \( A_0 \) (typically set to one for ease of calculation). The equilibrium length can then easily be calculated by setting the derivative of equation E.3 equal to zero and solving for \( \ell_i \).

### Geometric Parameters

<table>
<thead>
<tr>
<th></th>
<th>( 3\sqrt{3}\ell_0^2/2 = 2.598 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energetic target area, ( A_0 )</td>
<td></td>
</tr>
<tr>
<td>Hexagon side length at target area, ( \ell_0 )</td>
<td>1</td>
</tr>
</tbody>
</table>

### Perturbation Parameters

<table>
<thead>
<tr>
<th></th>
<th>( 0.1\ell_0 = 0.1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard deviation of perturbation Gaussian</td>
<td></td>
</tr>
<tr>
<td>Node joining/splitting radius</td>
<td>( 0.05\ell_0 = 0.05 )</td>
</tr>
<tr>
<td>Node splitting probability</td>
<td>20%</td>
</tr>
</tbody>
</table>

### Model Specific Parameters

<table>
<thead>
<tr>
<th>Condition Description</th>
<th>Passive</th>
<th>Soft</th>
<th>Active</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perimeter contractility, ( \bar{\Gamma} )</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>Generalized line tension, ( \bar{\Lambda} )</td>
<td>0.12</td>
<td>-0.16</td>
<td>0.12</td>
</tr>
<tr>
<td>Converging anisotropic line tension, ( \bar{\delta} )</td>
<td>0.30</td>
<td>-0.10</td>
<td>0.30</td>
</tr>
<tr>
<td>Elongating anisotropic line tension, ( \bar{\epsilon} )</td>
<td>0.12</td>
<td>-0.16</td>
<td>-0.12</td>
</tr>
<tr>
<td>Hexagon side length at equilibrium area, ( \ell_{eq} )</td>
<td>( 0.7587\ell_0 )</td>
<td>( 0.9405\ell_0 )</td>
<td>( 0.7587\ell_0 )</td>
</tr>
</tbody>
</table>

Table E.1: Parameter values used in cell grid simulations.
Figure E.1: Sample calculation of the tissue extension metric using experimental images. Since we would like to associate this local metric with a singular contracting interface length, only the surrounding four cells are used to calculate the tissue extension. The length of the red line, $x$, is the distance between the centroids of two cells in initially neighboring columns. The lengths of the green and blue lines, $y_1$ and $y_2$, are the distances from the centroids of the upper and lower cells to the neighboring row between them, which are averaged to keep the number of contributing rows and columns equal. The ratio of the distance between neighboring columns and the distance between neighboring rows can then be used as a relative metric of how far the local tissue has been elongated, as can be seen by the value of 1.25 before neighbor exchange and 2.42 after neighbor exchange.