

# The Molecular Biophysics of Evolutionary and Physiological Adaptation

Thesis by  
Griffin Daniel Chure

In Partial Fulfillment of the Requirements for the  
Degree of  
Doctor of Philosophy in Biochemistry and Molecular Biophysics

**Caltech**

CALIFORNIA INSTITUTE OF TECHNOLOGY  
Pasadena, California

2020  
Defended June 1, 2020

© 2020

Griffin Daniel Chure  
ORCID: 0000-0002-2216-2057

Some rights reserved. This thesis is distributed under a Creative Commons  
Attribution License CC-BY 4.0.

## ACKNOWLEDGEMENTS

Of the many pages in this thesis, this small section is the most difficult to write. This difficulty doesn't come from thinking of *who* to acknowledge. Rather, it comes from the crushing fear of leaving someone out. The past seven years have been some of the most satisfying of my life, both from an intellectual and emotional standpoint. If you're reading this and are hurt you aren't mentioned, please forgive me – your absence comes from the fragility of human memory rather than from a place of malice.

Rob Phillips, my Ph.D. adviser, is a man who escapes description. As such, it feels impossible to properly convey my gratitude for his mentorship and, more importantly, his friendship. His unceasing desire and unique ability to pursue the physical reality of any problem has sculpted my scientific process. Among his many scientific qualities, its his refusal to differentiate between research and teaching that has had the largest impact. I've been lucky enough to travel the world (or, at least four continents) with him teaching principles of physical biology and learning to be a physical biologist along the way. I'll forever be grateful for his willingness to take me as a graduate student and his kindness in treating me as a friend.

Aside from Rob, Justin Bois has perhaps had the strongest influence on my scientific philosophy. Throughout my Ph.D., he has been both a dear friend and statistical shaman. His remarkable ability to teach the complex details of statistical inference, physical biology, and computer programming transformed the way I consider and approach biological questions.

I owe a great deal of my development as a scientist to those with whom I've shared the lab. Heun Jin Lee is a master of all things precise and served as someone I could go to for advice on any topic ranging from the scientific to the deeply personal. His dedication to writing down theory and performing experiments and theory with extreme care has shaped my impression of how science should be per-

formed and, just as importantly, communicated. Alongside Heun Jin, I've had the pleasure to be coauthors with Manuel Razo-Mejia, Soichi Hirokawa, Muir Morrison, Zofii Kaczmarek, Nathan Belliveau, Stephanie Barnes, and Tal Einav. Soichi and Muir have faithfully served as my office-mates for the past six years and have served as a sounding board for many of the ideas presented in this thesis. Muir and Soichi's deep knowledge of physics helped me develop the physical intuition I have today and I will forever owe them for the time they took out of their days help me understand my science. Manuel and Nathan are those who I've spent the most time with at the bench, the whiteboard, and hunched over the keyboard. Alongside Soichi and Muir, they have helped me figure out what makes a good experiment and what makes a bad theory (and vice-versa). Stephanie and Zofii, both stellar scientists in their own right, also served as a reminder of how creative expression is as important as scientific rigor. Tal Einav, a Phillips' lab alumnus who needs no introduction, has taught me much about the nature of collaboration over the years.

Suzy Beeler, though I never had the pleasure of listing as a coauthor, has always someone I could turn to when in need of scientific advice or personal support. Her ability to provide an honest opinion, either scientific or personal has proven to be invaluable over the years. Bill Ireland, Rachel Banks, Rebecca Rousseau, Vahe Galstyan, Niko McCarty, Tom Röschinger, Molly Bassette, Jonathan Gross, Celene Barrera, Matthias Rydenfelt, Kimberly Berry, Gita Mahmoudabadi, and Daniel Jones are others who are or have since left the group to whom I am grateful for their scientific discussions over the years.

Some of my most cherished memories of graduate school stem from my time at the Marine Biological Laboratory, an institution where suspension of disbelief is the rule rather than the exception. While I never took a course, I had the immense privilege of being a teaching assistant for the Physiology summer course from 2015 to 2018 and for Physical Biology of the Cell course in 2018. During this time, I met dozens of graduate students, professors, post docs, research technicians, and more

who have all helped me develop as a scientist and educator. In no particular order, I would like to Celine Aklelmide, George Bell, Ambika Nadkarni, Damien Dudka, Cat Triandifiliou, Mason Kamb, Lizzy Mueller, Chandrima Patra, Miranda Hunter, Kyle Naughton, Cayla Jewett, Emily Meltzer, Roya Huang, Charlotte Strankvist, Simon Alamos, Sean McInally, Karna Gowda, Alina Guna, Nalin Ratinyake, Joe Brzostowski, Carolynn Ott, Steven Wilbert, Ana Gayek, Wallace Marshall, and all of those who have passed through the halls of Loeb Lab with me for forging so many great memories. During these courses, I had the pleasure to teach alongside Rob Brewster, Hernan Garcia, James Boedicker, and Franz Weinert who all passed through Rob's lab in their own right. As I taught alongside them, I learned enough from them to be qualified as yet another one of their students.

While my academics have been focused on the things that are small and alive, I have had a life-long love affair with paleontology where the subjects are often big and always dead. This comes from the opportunities provided to me by Mom and Dad, Lorraine and Dan Chure. Dan and Lorraine Chure. During my father's 38 years as the chief paleontologist at Dinosaur National Monument in Utah, he instilled in me a curiosity for the natural world and its long, tumultuous biological history. My childhood and adolescent years are peppered with fond memories of hiking across the rugged Utah deserts, scanning exposed layers of sandstone for the tell-tale signs of a 200 million year old graveyard. On the weekends, I would work alongside my father at active dig sites, chipping away at the rock hoping to uncover some pieces that help fill in the puzzle of the life history of Earth. The weekdays, however, were spent in the public education system of rural Utah where evolution was a hoax and dinosaur bones were buried by either the devil or the government (or sometimes both). Their commitment to giving me a scientific education is why I'm here 20 years later writing a section of my doctoral thesis. Their unwavering support of my interests and passions (from skateboarding and death metal to coffee and bread baking) have helped pull me through tough parts of grad school and have taught me to savor the good.

The final four years of graduate school have been the most satisfying, and not because my experiments started to work. In July of 2016, I met Barbara de Araujo Soares, a wonderful woman who two years later helped me slice into a wedding cake. Barb's unwavering love and support is why I have made it to where I am today. Without her wit, creativity, cheerfulness and honesty, I don't know how I would have survived the long hours, the cheering successes, and the crushing failures of scientific research. While we have had a blast of the past four years, I think I speak for both of us when I say "até que enfim!"

Stephen Jay Gould, one of my scientific heroes, once wrote "I am, somehow, less interested in the weight and convolutions of Einstein's brain than in the near certainty that people of equal talent have lived and died in cotton fields and sweatshops." This is a quote that has stuck with me since I first read it in middle school. The enormous privilege I have had throughout my life is not lost on me. I write this paragraph as someone with a home with fresh food and running water. As someone with loving family and friends. As someone who is being paid to do what he loves. The physical, emotional, and intellectual security I have been so lucky to have comes at some level at the cost of others.

We stand on the shoulders of giants. These giants are faceless, nameless, and enumerable. For a section of this work dedicated to naming those to whom I am indebted, I want to close it with thanking all those who to me are nameless. To those who lived a harsher life to build a better one for people like me. To those who were more courageous in fighting injustice and cruelty than I. To those who lived and died in dire straits simply because of who they were, where they were born, or what they believed in. This is a debt that can never be repaid, but I hope that I can at least give to the future a fraction of what they gave to me.

## ABSTRACT

Central to any definition of Life is the ability sense changes in ones' environment and respond in kind or, in other words, the ability to adapt. Adaptive phenomena can be found across the biological scales ranging from the nanosecond-scale conformational changes of proteins, to temporary rewiring of metabolic networks, to the 3.5 billion years of evolution that produced the enormous biodiversity we see today. This thesis presents a body of work which attempts to examine the overlap between these three scales of adaptation through the quantitative lens of statistical physics. Namely, we examine how molecular, physiological, and evolutionary adaptation governs a feature common to all life – the regulation of gene expression.

We begin by examining the phenomenon of molecular adaptation in the context of allostery, specifically in the context of allosteric transcriptional repressors. Using simple tools of quasi-equilibrium thermodynamics, we derive and experimentally dissect a quantitative model of how such a repressor adapts to different concentrations of an extracellular inducer molecule, modulating the repressor's activity and thereby gene expression. While the model is relatively simple, it is remarkably powerful in its ability to draw concrete, quantitative predictions about not only the level of gene expression at a given concentration of inducer, but details of how the repressor responds to changes in the inducer concentration. With a few lines of simple mathematics, we are able to use this model to derive a state variable of the simple repression motif which we term the free energy of the regulatory architecture. This permits us to collapse nearly 500 distinct measurements of the level of gene expression onto a master curve defined by this free energy.

We leverage this feature of the model to use data collapse as a method to identify the effects of mutation, a strong evolutionary force responsible for much of the genetic diversity in bacteria. In this chapter, we examine how mutations within the allosteric repressor itself can be mapped to changes in the free energy. The precise value of these free energy shifts and their dependence on the inducer

concentration reveal different classes of mutations with one class affecting only the DNA-repressor interaction and another class governing the allosteric nature of the repressor. We test these pen-and-paper predictions experimentally and illustrate that given sufficient knowledge of how single mutants behave, the complete phenotypic response of pairwise double mutants can be predicted with quantitative accuracy.

With this framework in hand we then turn to exploring how changes in the physiological state of the cell influence the molecular biophysics of the regulatory architecture. We hypothesize that changes in the source of carbon in the growth medium or changes in the growth temperature can be accounted for by the existing the model without any additional parameters. We experimentally show that the parameter values determined in one physiological state are inherited when the available carbon source is verified, but changes in the growth temperature require some additional considerations. This chapter as a whole reveals that, while there remains work to be done both theoretically and experimentally when it comes to temperature variation, thermodynamic models can remain powerful tools to draw predictions of gene expression in different physiological contexts.

Finally, we explore physiological adaptation and cellular decision making where it counts – in the survival of cells to environmental insults. We turn our focus beyond transcriptional regulation and consider the relationship between osmotic shocks, the abundance of mechanosensitive channels, and cellular survival with single cell resolution. Using a combination of quantitative microscopy and tricks of statistical inference, we infer how the probability of a cell surviving an osmotic shock scales as a function of the cell’s number of mechanosensitive channels.

## PUBLISHED CONTENT AND CONTRIBUTIONS

Laxhauber, Kathrin S.; Morrison, M. J; **Chure, G.**; Belliveau, N.M.; Strankvist, C.; Phillips, R. (2020). Theoretical investigation of a genetic switch for metabolic adaptation. PLoS ONE (in press)

Razo-Mejia, M.; Marzen, S.; **Chure, G.**; Taubman, R.; Morrison, M. J.; Phillips, R. (2020). First-principles prediction of the information processing capacity of a simple genetic circuit. BioRxiv (in submission)

**Chure, G.**; Kaczmarek, Z.A.; and Phillips, R. (2019b). Physiological Adaptability and Parametric Versatility in a Simple Genetic Circuit. BioRxiv (under review at Cell Systems) 2019.12.19.878462.

Hirokawa, S.; **Chure, G.**; Belliveau, N.M.; Lovely, G.A.; Anaya, M.; Schatz, D.G.; Baltimore, D.; and Phillips, R. (2019). Sequence-Dependent Dynamics of Synthetic and Endogenous RSSs in V(D)J Recombination. BioRxiv (under review at Nucleic Acids Research) 791954.

**Chure, G.**; Razo-Mejia, M.; Belliveau, N.M.; Einav, T.; Kaczmarek, Z.A.; Barnes, S.L.; and Phillips, R. (2019). Predictive Shifts in Free Energy Couple Mutations to Their Phenotypic Consequences. Proceedings of the National Academy of Sciences 116.

Phillips, R.; Belliveau, N.M.; **Chure, G.**; Garcia, H.G.; Razo-Mejia, M.; and Scholes, C. (2019). Figure 1 Theory Meets Figure 2 Experiments in the Study of Gene Expression. Annu. Rev. Biophys. 48, 121–163.

**Chure, G.** \*; Lee, H.J.\* ; Rasmussen, A.; and Phillips, R. (2018). Connecting the Dots between Mechanosensitive Channel Abundance, Osmotic Shock, and Survival at Single-Cell Resolution. Journal of Bacteriology 200. \* contributed equally

Razo-Mejia, M.\*; Barnes, S.L.\*; Belliveau, N.M.\* ; **Chure, G.**\*; Einav, T.\*; Lewis, M.; and Phillips, R. (2018). Tuning Transcriptional Regulation through Signaling:

x

A Predictive Theory of Allosteric Induction. Cell Systems 6, 456-469.e10. \* contributed equally

## TABLE OF CONTENTS

Acknowledgements . . . . .	iii
Abstract . . . . .	vii
Published Content and Contributions . . . . .	ix
Table of Contents . . . . .	x
List of Illustrations . . . . .	xiv
List of Tables . . . . .	xx
<b>Chapter I: The Phenomenon of Adaptation Across Scales . . . . .</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.2 The Janus Face of Molecules . . . . .	5
1.3 Using Free Energy to Examine Evolutionary Adaptation . . . . .	10
1.4 Topic IV: The Physiological Adaptability of Transient Molecular Interactions . . . . .	14
1.5 On Facing the Elements . . . . .	21
1.6 On Molecular Biophysics and Evolutionary Dynamics . . . . .	24
<b>Chapter II: Through the Intramolecular Grapevine: Signal Processing Via Allosteric Transcription Factors . . . . .</b>	<b>28</b>
2.1 Abstract . . . . .	28
2.2 Introduction . . . . .	29
2.3 Theoretical Model . . . . .	31
2.4 Results . . . . .	37
2.5 Discussion . . . . .	50
2.6 Materials & Methods . . . . .	54
<b>Chapter III: Unknown Knowns, Known Unknowns, and Unforeseen Consequences: Using Free Energy Shifts To Predict Mutant Phenotypes . . . . .</b>	<b>61</b>
3.1 Abstract . . . . .	61
3.2 Introduction . . . . .	62
3.3 Theoretical Model . . . . .	64
3.4 Results . . . . .	69
3.5 Discussion . . . . .	81
3.6 Materials & Methods . . . . .	86
<b>Chapter IV: On The Physiological Adaptability of a Simple Genetic Circuit . . . . .</b>	<b>91</b>
4.1 Abstract . . . . .	91
4.2 Introduction . . . . .	92
4.3 Results . . . . .	95
4.4 Fold-change dependence on carbon quality . . . . .	101
4.5 Discussion . . . . .	106
4.6 Materials & Methods . . . . .	110
<b>Chapter V: 'Water, Water Everywhere, Nor Any Drop to Drink': How Bacteria Adapt To Changes in Osmolarity . . . . .</b>	<b>118</b>

5.1 Abstract . . . . .	118
5.2 Introduction . . . . .	119
5.3 Results . . . . .	121
5.4 Discussion . . . . .	131
5.5 Materials & Methods . . . . .	138
Chapter VI: Supplemental Information For Chapter 2: Signal Processing Via Allosteric Transcription Factors . . . . .	146
6.1 Inferring Allosteric Parameters from Previous Data . . . . .	146
6.2 Variable Repressor Copy Number ( $R$ ) with Multiple Specific Binding Sites ( $N_S > 1$ ) . . . . .	154
6.3 Variable Number of Specific Binding Sites $N_S$ with Fixed Repressor Copy Number ( $R$ ) . . . . .	154
6.4 Competitor Binding Sites . . . . .	155
6.5 Properties of the Induction Response . . . . .	156
6.6 Flow Cytometry . . . . .	159
6.7 Single-Cell Microscopy . . . . .	163
6.8 Fold-Change Sensitivity Analysis . . . . .	168
6.9 Global Fit of All Parameters . . . . .	170
6.10 Applicability of Theory to the Oid Operator Sequence . . . . .	177
6.11 Comparison of Parameter Estimation and Fold-Change Predictions across Strains . . . . .	177
6.12 Properties of Induction Titration Curves . . . . .	179
6.13 Applications to Other Regulatory Architectures . . . . .	185
Chapter VII: Supplemental Information for Chapter III: Predictive Shifts in Free Energy Couple Mutations to Their Phenotypic Consequences . . . . .	192
7.1 Non-Monotonic Behavior of $\Delta F$ Under Changing $K_A$ and $K_I$ . . . . .	192
7.2 Bayesian Parameter Estimation For DNA Binding Mutants . . . . .	196
7.3 Inferring the Free Energy From Fold-Change Measurements . . . . .	210
7.4 Bayesian Parameter Estimation for Inducer Binding Domain Mutants	220
7.5 Additional Characterization of Inducer Binding Domain Mutants .	227
7.6 Comparing Parameter Values To The Literature . . . . .	230
7.7 Parameter Estimation Using All Induction Profiles . . . . .	238
7.8 Generalizability of Data Collapse To Other Regulatory Architectures	241
7.9 Strain and Oligonucleotide Information . . . . .	244
Chapter VIII: Supplemental Information for Chapter IV: The Physiological Adaptability of a Simple Genetic Circuit . . . . .	255
8.1 Non-parametric Inference of Growth Rates . . . . .	255
8.2 Approximating Cell Volume . . . . .	257
8.3 Counting Repressors . . . . .	259
8.4 Parameter Estimation of DNA Binding Energies and Comparison Across Carbon Source . . . . .	275
8.5 Statistical Inference of Entropic Costs . . . . .	279
8.6 Media Recipes and Bacterial Strains . . . . .	285
Chapter IX: Supplemental Information For Chapter 5: How Bacteria Adapt To Changes in Osmolarity . . . . .	289

9.1 Calibration of a Standard Candle . . . . .	292
9.2 Classification of Cell Fates . . . . .	302
9.3 Logistic Regression . . . . .	306
9.4 Classification Of Shock Rate . . . . .	313
9.5 Comparison of Survival Probability with van den Berg et al. 2016 . . .	317
9.6 <i>E. coli</i> Strains . . . . .	317
References . . . . .	322
Chapter X: Questionnaire . . . . .	342
Chapter XI: Consent Form . . . . .	343

## LIST OF ILLUSTRATIONS

<i>Number</i>	<i>Page</i>
1.1 The spatial, temporal, and mechanistic scale of adaptation. . . . .	2
1.2 The phenomenon of enzymatic adaptation revealed in bacterial growth curves. . . . .	4
1.3 A Coarse grained representation of an allosteric molecule. . . . .	7
1.4 Experimental dissection of the inducible simple repression input-output function. . . . .	9
1.5 Collapse of individual induction profiles onto a simple scaling function.	10
1.6 Growth curves of lactose-positive and lactose-negative <i>E. coli</i> strains. .	11
1.7 Mutations lead to predictive shifts in free energy. . . . .	13
1.8 Theoretical prediction and experimental validation of double mutant phenotypes. . . . .	15
1.9 A metabolic hierarchy in a growth medium containing glucose, sorbitol, and glycerol. . . . .	16
1.10 Methods of physiological control used in Chapter 4. . . . .	18
1.11 Performance of a simple thermodynamic model of simple repression in diverse physiological states. . . . .	20
1.12 The connection between mechanosensitive channel number and probability of survival. . . . .	23
1.13 Allocation of cellular resources induces compositional structure in the <i>E. coli</i> proteome. . . . .	26
1.14 The coming interplay between molecular biophysics and evolutionary dynamics. . . . .	27
2.1 Transcriptional regulatory architectures involving an allosteric repressor. . . . .	32
2.2 States and statistical weights for the simple repression motif. . . . .	34

2.3	An experimental pipeline for high-throughput fold-change measurements. . . . .	40
2.4	Predicting induction profiles for different biological control parameters. . . . .	42
2.5	Comparison of predictions against measured and inferred data. . . . .	44
2.6	Predictions and experimental measurements of key properties of induction profiles. . . . .	46
2.7	Coarse graining of promoter occupancy states to a two-state system. . . . .	49
2.8	Collapse of fold-change measurements as a function of the free energy. . . . .	50
3.1	Summary of Chapter 2. A predictive framework for dissection of the simple repression motif . . . . .	65
3.2	Parametric changes due to mutations and their corresponding free-energy shifts. . . . .	70
3.3	Induction profiles and free-energy differences of DNA binding domain mutations in the <i>lac</i> repressor. . . . .	73
3.4	Induction profiles and free-energy differences of inducer binding domain mutants. . . . .	76
3.5	Induction and free-energy profiles of DNA binding and inducer binding double mutants. . . . .	81
3.6	Data collapse of the simple repression regulatory architecture. All data are means of biological replicates. . . . .	85
4.1	Controlling physiological state of <i>E. coli</i> via growth rate modulation by environmental factors. . . . .	98
4.2	Control and quantification of repressor copy number through the binomial partitioning method. . . . .	100
4.3	Scaling of cel size and repressor expression as a function of maximum growth rate . . . . .	102
4.4	Temperature effects on the fold-change in gene expression and free energy. . . . .	107
4.5	A singular theoretical description for the molecular biophysics of physiological and evolutionary adaptation in the simple repression motif. .	111

5.1	Role of mechanosensitive channels during hypo-osmotic shock. . . . .	122
5.2	Control of MscL expression and calculation of channel copy number. .	126
5.3	Experimental approach to measuring survival probability . . . . .	127
5.4	Distributions of survival and death as a function of effective MscL channel number. . . . .	128
5.5	Probability of survival as a function of MscL copy number. . . . .	132
6.1	Multiple sets of parameters yield identical fold-change responses. . .	148
6.2	Induction with variable $R$ and multiple specific binding sites. . . . .	155
6.3	Induction with variable specific sites and fixed $R$ . . . . .	156
6.4	Induction with variable competitor sites, a single specific site, and fixed $R$ . . . . .	157
6.5	Phenotypic properties of induction with multiple specific binding sites.	158
6.6	Phenotypic properties of induction with a single specific site and multiple competitor sites. . . . .	159
6.7	Representative unsupervised gating contours of flow-cytometry data. .	162
6.8	Comparison of experimental methods to determine the fold-change in gene expression. . . . .	163
6.9	Empirical comparison of flow cytometry and single-cell microscopy. .	168
6.10	Sensitivity analysis of the fold-change function to $K_A$ and $K_I$ estimates.	171
6.11	Induction curves using global parameter estimates. . . . .	175
6.12	Key properties of induction profiles as predicted with a global fit us- ing all data. . . . .	176
6.13	Predictions of fold-change for strains with an Oid binding sequence versus experimental measurements with different repressor copy num- bers. . . . .	178
6.14	O1 strain fold-change predictions based on strain-specific parameter estimation of $K_A$ and $K_I$ . . . . .	180
6.15	O2 strain fold-change predictions based on strain-specific parameter estimation of $K_A$ and $K_I$ . . . . .	181

6.16	O3 strain fold-change predictions based on strain-specific parameter estimation of $K_A$ and $K_I$ . . . . .	182
6.17	Dependence of leakiness, saturation, and dynamic range on the operator binding energy and repressor copy number. . . . .	184
6.18	[EC <sub>50</sub> ] and effective Hill coefficient depend strongly on repressor copy number and operator binding energy. . . . .	185
7.1	Non-monotonic behavior of $\Delta F$ with changes in $K_A$ and $K_I$ . . . . .	195
7.2	Prior distributions and prior predictive check for estimation of the DNA binding energy. . . . .	200
7.3	Comparison of averaged posterior and prior distributions for DNA binding energy and homoscedastic error. . . . .	204
7.4	Inferential sensitivity for estimation of the DNA binding energy and homoscedastic error. . . . .	206
7.5	Rank distribution of the posterior samples from simulated data for the DNA binding energy and homoscedastic error. . . . .	207
7.6	Markov chain Monte Carlo samples and posterior predictive check for DNA binding mutant Q18M. . . . .	209
7.7	Prior predictive checks for inference of the mean fold-change . . . . .	213
7.8	Sensitivity measurements and rank statistic distribution of the statistical model estimating $\mu$ and $\sigma$ . . . . .	214
7.9	MCMC sampling output and posterior predictive checks of the statistical model for the mean fold-change and standard deviation. . . . .	215
7.10	Pairwise comparisons of DNA binding mutant estimated induction profiles. . . . .	219
7.11	Prior predictive checks for two hypotheses of inducer binding domain mutants. . . . .	224
7.12	Simulation based calibration of statistical models for inducer binding domain mutants. . . . .	225
7.13	Posterior predictive checks for inducer binding domain mutants where only $K_A$ and $K_I$ are changed. . . . .	227

7.14	Posterior predictive checks for inducer binding domain mutants where all allosteric parameter can change. . . . .	228
7.15	Pairwise comparison of fit strain versus predictions assuming only $K_A$ and $K_I$ are influenced by mutations in the inducer binding domain. . . . .	230
7.16	Pairwise comparison of fit strain versus predictions assuming all allosteric parameters are affected by the mutation in the inducer binding domain. . . . .	231
7.17	Comparison of choice of fit strain on predicted $\Delta F$ profiles for inducer binding domain mutants. . . . .	232
7.18	Degenerate fits of data using parameter values from the literature. . . . .	235
7.19	Induction profiles and predicted change in free energy using parameters estimated from the complete data sets. . . . .	239
7.20	Induction profiles and predicted change in free energy using parameters estimated from the complete data sets for inducer binding domain mutants. . . . .	241
7.21	Various repression-based regulatory architectures and their coarse-grained states. . . . .	244
8.1	Non-parametric estimation of maximum growth rate from bacterial growth curves. . . . .	257
8.2	Growth-rate dependence of cell shape and approximation as a spherocylinder. . . . .	259
8.3	An experimental workflow for determination of a fluorescence calibration factor. . . . .	264
8.4	Experimental sanity checks and inference of a fluorescence calibration factor. . . . .	267
8.5	Representative posterior distributions for inference of the fluorescence calibration factor. . . . .	270
8.6	Origin of a systematic error in repressor counting due to continued growth. . . . .	271
8.7	Cell length distributions of fluorescence snapshots and newborn cells.	273

8.8	Influence of a correction factor on fold-change and the DNA binding energy. . . . .	276
8.9	Pairwise estimation and prediction of DNA binding energies estimated from different carbon sources. . . . .	280
8.10	Sampled posterior probability distributions of entropic penalty parameter inference. . . . .	285
8.11	Pairwise predictions of fold-change in gene expression at different temperatures. . . . .	286
8.12	Fold-change and shift in free energy including a temperature-dependent entropic contribution. . . . .	288
9.1	Characteristic MscL-sfGFP conductance obtained through patch-clamp electrophysiology . . . . .	291
9.2	Measurement of sfGFP maturation as a function of time through flow cytometry. . . . .	293
9.3	Schematic of hierarchical model structure for estimating MscL-sfGFP fluorescence calibration factor. . . . .	299
9.4	Posterior distributions for hyper-parameters and replicate parameters. . . . .	302
9.5	Influence of area correction for Shine-Dalgarno mutants. . . . .	303
9.6	Time lapse of a representative field after osmotic shock and the resulting classifications. . . . .	304
9.7	Representative images of propidium iodide staining after a strong osmotic shock. . . . .	305
9.8	Posterior distributions for logistic regression coefficients evaluated for fast and slow shock rates. . . . .	309
9.9	Survival probability estimation using alternative predictor variables. . . . .	314
9.10	Binning by individual shock rates. . . . .	316
9.11	Coarse graining shock rates into different groups. . . . .	318
9.12	MscL abundance verus survival data reported in van den Berg et al. 2016 with included error. . . . .	319

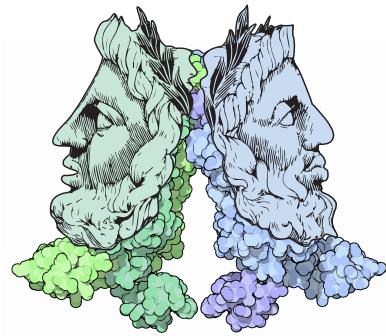
## LIST OF TABLES

<i>Number</i>	<i>Page</i>
3.1 Inferred values of $K_A$ , $K_I$ , and $\Delta\varepsilon_{AI}$ for inducer binding mutants . . . . .	78
5.1 Measured cellular copy numbers of MscL. Asterisk (*) Indicates inferred MscL channel copy number from the total number of detected MscL peptides. . . . .	137
6.1 Instrument settings for data collectuion using the Miltenyi Biotec MAC-SQuant flow cytometer. . . . .	160
6.2 Global parameter estimates and comparison to previously reported values. . . . .	174
6.3 Primers used in this work. . . . .	189
6.4 <i>E. coli</i> strains used in this work. . . . .	190
7.1 Estimated DNA binding energy for DNA binding domain mutants with different repressor copy numbers. Reported values are the median of the posterior distribution with the upper and lower bounds of the 95% credible regions. . . . .	218
7.2 Inferred values of $K_A$ , $K_I$ , and $\Delta\varepsilon_{AI}$ for inducer binding domain mutants. Values reported are the mean of the posterior distribution with the upper and lower bounds of the 95% credible region. . . . .	229
7.3 Thermodynamic parameter values of wild-type LacI from the literature.	236
7.4 Estimated parameters from global fits of data from literature. . . . .	237
7.5 Estimated DNA binding energies for each DNA binding domain mutant using all repressor copy numbers . . . . .	239
7.6 Estimated values for $K_A$ , $K_I$ , and $\Delta\varepsilon_{AI}$ for inducer binding domain mutations using induction profiles of all operator sequences. . . . .	240
7.7 <i>Escherichia coli</i> strains used in this work . . . . .	245
7.8 Oligonucleotides used for mutant generation. . . . .	252

8.1 Summarized parameter estimates of $\epsilon$ and $\sigma$ given a single growth condition. reported as median and upper/lower bounds of 95% credible region. . . . .	279
8.2 M9 minimal medium recipe for each carbon-supplemented medium. .	286
8.3 Bacterial strains used in various physiological states. . . . .	287
9.1 Cell fate classifications and their relative abundances in the complete data set. . . . .	304
9.2 Comparison of morphology-based and dye-based survival classification. . . . .	306
9.3 <i>Escherichia coli</i> strains used in Chapters 5 and 9. . . . .	319
9.4 Oligonucleotide sequences used in Chapters 5 and 9. Bold and italics correspond to Shine-Dalgarno sequence modifications and AT hairpin insertion modifications, respectively. Double bar    indicates a transposon insertion site. . . . .	320

## Chapter 1

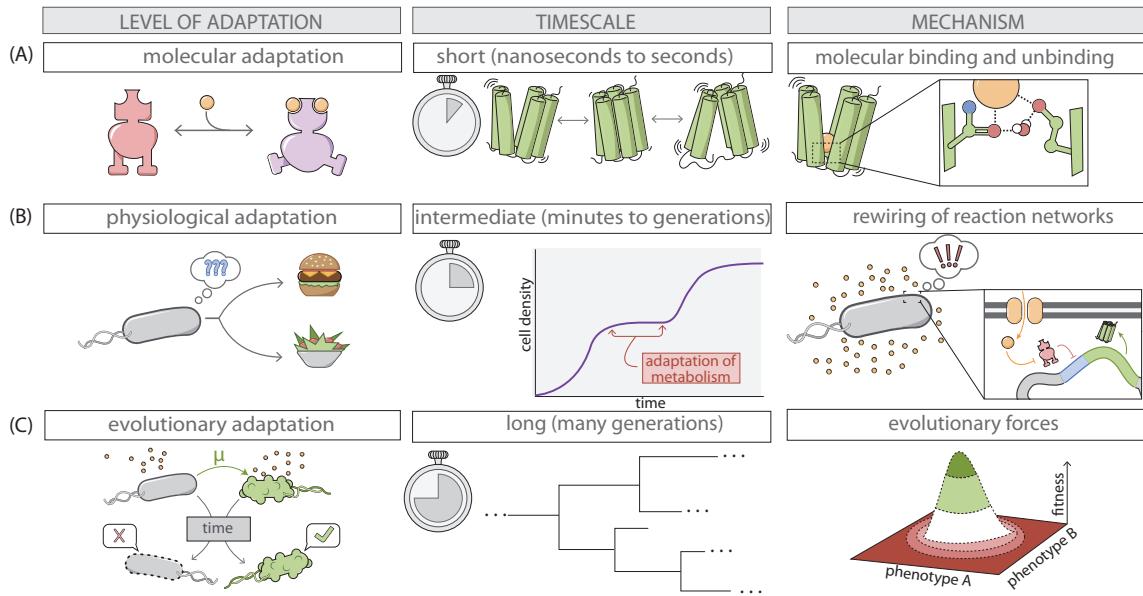
# THE PHENOMENON OF ADAPTATION ACROSS SCALES



### 1.1 Introduction

From archaea thriving in hydrothermal vents on the ocean floor to aspen trees dominating a Coloradan mountainside, all forms of life are unified in their obedience to the whims of their environment. Over the past 3.5 billion years of evolution, life has evolved myriad clever ways to combat (and exploit) environmental fluctuations to amplify reproductive success. The mechanisms behind this adaptation are diverse and traverse the biological scales, ranging from nanosecond-scale conformational switching of proteins (Fig. 1.1(C)), to reconfiguration of metabolic networks to consume different sugars(Fig. 1.1 (B)), to evolutionary trajectories that only become visible over many generations (Fig. 1.1 (C)). While “adaptation” is typically only associated with organisms (at least colloquially), one can use the same language to describe the microscopic operations of molecules.

The idea of molecular adaption is not novel and demands a brief foray into the history of bacterial growth and the dawn of regulatory biology. In the late 1890’s, Emilé Duclaux and his graduate student Frédéric Diénert performed a series of experiments illustrating that the common yeast could only consume galactose after an incubation period with the sugar. This led to a general conclusion that “the production of diastases [enzymes] depends on the manner of nutrition” in which the cultures were grown (Loison, 2013), a phenomenon later coined *enzymatic adap-*



**Figure 1.1: The spatial, temporal, and mechanistic scale of adaptation.** (A) Molecular adaptation in this work is defined through the lens of allostery where the activity of a protein complex is modulated by the reversible binding of a small molecule. These binding and unbinding events lead to rapid changes in protein conformation whose behavior (both energetic and temporal) is comparable to that of thermal motion. (B) Physiological adaptation here is defined as the rewiring of biochemical reaction networks that lead to changes in cellular behavior (such as chemotaxis) or metabolic capacity (such as aerobic to fermentative metabolism). (C) Evolutionary adaptation is recorded in the variation in the genetic sequence of regulatory molecules. Variations in sequence influence the function of the proteins and RNAs they encode which ultimately define the cellular fitness.

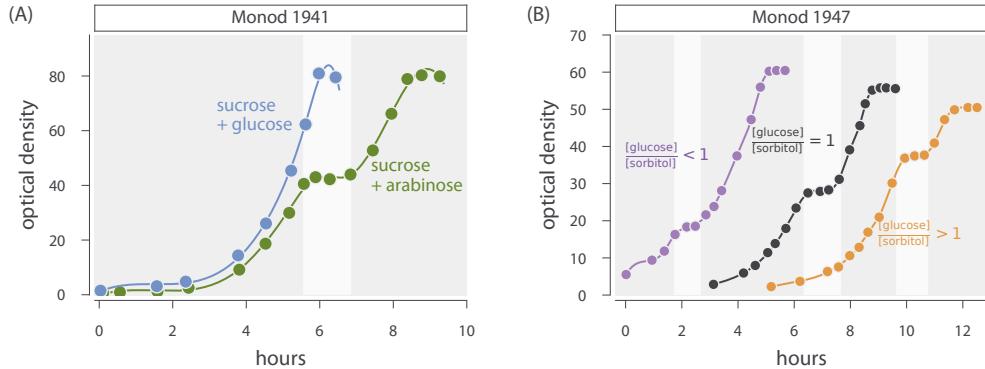
tation. This is one of the first observations of the fact that, while an organism may be able to digest a certain sugar, it may not *always* be able to do so. Rather, there seemed to be certain conditions in which the production or formation of these enzymes could occur. In his doctoral thesis in 1900, Diénert proposed two mechanisms for the origin of enzymatic adaptation observed for galactozymase in *S. cerevisiae* (Loison, 2013). Either (a) the presence of galactose *directly* transformed enzymes already present in the cell into galactozymase or (b) that the galactose activated the production of galactozymase *de novo* (Diénert, 1900).

Nearly half a century later, Jacques Monod would rediscover the phenomenon of enzymatic adaptation, this time in the context of bacterial growth. In his 1941

work, *Sur un phénomène nouveau de croissance complexe dans les cultures bactériennes*, Monod for the first time reported on the phenomenon of diauxic growth, shown in Fig. 1.2 (A). He noted that for some mixtures of carbon sources, the culture grew “kinetically normal” meaning they grew exponentially to saturation (blue points, Fig. 1.2 (A)). However, some mixtures (such as sucrose and arabinose) led to biphasic growth where the culture would grow exponentially, undergo a period where growth had ceased, followed by again by another round of exponential growth (blue points, Fig. 1.2 (A)). Additionally, Monod showed that the onset of this diauxic shift could be tuned by varying the relative concentrations of the carbon sources, revealing a controllable chemical basis for the adaptation (Fig. 1.2 (B)).

Monod immediately made the connection between diauxic growth and enzymatic adaptation (Loison, 2013). Despite his work appearing 40 years after the pioneering work of Ducleaux and Diénert, there had been little progress towards a mechanistic, needless to say quantitative, explanation for the phenomenon. In fact, Monod was particularly disappointed by the teleological explanations where the cells simply changed their behavior to perform only the chemical reactions that were “needed” (Loison, 2013). The teleological approach to much of biology during this time period, especially in the French scientific community, severely bothered Monod. To him, this kind of approach belonged to a pre-scientific era and lacked the “postulate of objectivity” that other fields of science (particularly physics) had adopted (Loison, 2013). Near the middle of the 20th century, Monod published a 60-page treatise on the phenomena of enzymatic adaptation with the level of quantitative rigor he thought it deserved (Monod, 1947). In this work, he set out to progressively deconstruct and invalidate a series of hypotheses for the phenomenon of enzymatic adaptation. In doing so, he laid the groundwork for what would become (in his opinion) his greatest contribution to science, the nature of allosteric transitions (Loison, 2013), a topic that will feature prominently in the remainder of this thesis.

The diauxic growth transitions shown in Fig. 1.2 illustrate adaptive processes



**Figure 1.2: The phenomenon of enzymatic adaptation revealed in bacterial growth curves.** (A) Optical density measurements of *Bacillus subtilis* cultures grown in a mixture of sucrose and either glucose (blue points) or arabinose (green points). Biphasic growth can be observed in the sucrose/arabinose mixture where the pause in growth (white vertical line) corresponds to enzymatic adaptation. Data digitized from Monod (1941). (B) Diauxic growth curves of *Escherichia coli* cells grown on a mixture of glucose and sorbitol in different proportions. Data digitized from Monod (1947). Periods of enzymatic adaptation are highlighted by white vertical lines.

across the biological scales, as were schematized in Fig. 1.1. While it was not known to Monod at the time, we now know that many cases of enzymatic adaptation are driven by the regulation of gene expression. As the bacterial culture approaches the auxic shift, the presence or absence of the substrate is sensed by regulatory molecules that control whether the genes encoding the enzymes for metabolism of the substrate are expressed. This represents the level of **molecular adaptation** where, given binding or unbinding of the substrate molecule, the activity of the regulatory protein is modulated. The amino acid sequence of these proteinaceous regulators are the product of billions of years of **evolutionary adaptation** and define how the regulatory senses and responds to these signals. Finally, the precision with which these genes are regulated are determined by their sensitivity to physiological states, capturing the level of **physiological adaptation**. j

The central aim of this dissertation is to explore the biophysical mechanisms by which these levels of adaptation – molecular, physiological, and evolutionary – are interconnected. Furthermore, in the spirit of Monod, we seek to make our explo-

ration quantitative and leverage the tools of statistical physics to provide precise predictions from pen-and-paper theory that can be rigorously tested through experiment. The remaining sections of this chapter will outline the major topics of this thesis and place them in a historical context alongside the work of Monod. Finally, we will close with a discussion of how these types of models can be used to explore the predictability of evolution.

## 1.2 The Janus Face of Molecules

Monod is perhaps most famous for his discovery of allostery, to which he famously referred to as “the second secret of life” (Monod et al., 1965; Ullmann, 2011). It is fair to say that this “secret” has been now been declassified. Allosteric regulation can be found in all domains of life across varied types of biological processes. Allostery can be found governing the behavior of ion channels (Einav and Phillips, 2017), enzymatic reactions (Einav et al., 2016), chemotaxis (Keymer et al., 2006), G-protein coupled receptors (Canals et al., 2012), quorum sensing (Swem et al., 2008), and transcriptional regulation (Huang et al., 2018), to name a few of many examples. Despite the objective complexity in the molecular structures of all of these allosteric molecules, they can be frequently be reduced to simple cartoons where the details of conformational changes, substrate binding affinities, and more can be massaged into a small set of key details. Fig. 1.3 (A) shows the molecular structures of a variety of allosteric transcriptional repressors (top). While each has their own fascinating structure and continuum of conformational states, all can be coarse grained into a simple cartoon representation (bottom) with an active (red) and inactive (purple) state, both of which possess binding pockets (semicircular notches) for an inducer molecule (orange).

Much as we can reduce the complexity of allosteric molecules schematically, we can enumerate simple mathematical models that describe their behavior. Thermodynamic models built on an assumption of quasi-equilibrium are routinely used to describe complex biological phenomena despite the reality that being in thermodynamic equilibrium is synonymous with being dead. Even with this glaring

assumption, such models have been shown to be exceptionally predictive for a variety of complex systems, especially in modeling molecular binding reactions (Dill and Bromberg, 2010) and allostery writ large (Einav and Phillips, 2017; Einav et al., 2016; Keymer et al., 2006; Phillips, 2015; Swem et al., 2008). As the timescales of binding and unbinding reactions are orders of magnitude smaller than that of many other processes in the cell, it is fair to make the approximation that molecular binding is in equilibrium. Under this assumption, we are granted the powerful mathematical privilege to say that the probability of a given state of the system  $P_{\text{state}}$  follows a Boltzmann distribution,

$$P_{\text{state}} = \frac{e^{-\frac{\epsilon_{\text{state}}}{k_B T}}}{Z}, \quad (1.1)$$

where  $\epsilon_{\text{state}}$  is the energy of that state,  $k_B$  is the Boltzmann constant, and  $T$  is the system temperature. The denominator  $Z$  is the partition function of the system and is the sum

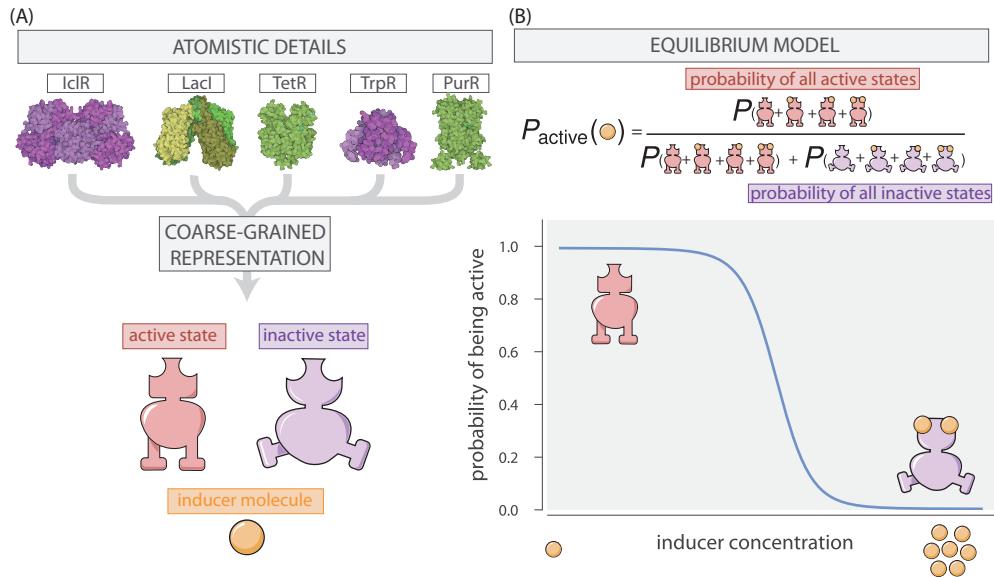
$$Z = \sum_{i \in \text{states}} e^{\frac{\epsilon_i}{k_B T}}, \quad (1.2)$$

ensuring that the distribution is normalized. Therefore, if we are interested in computing the probability of a given allosteric protein being in the active state, we merely have to enumerate all of the Boltzmann weights (given by the numerator in Eq. 1.1) and compute

$$P_{\text{active}} = \frac{\text{sum over all possible active states}}{\text{sum over all possible states}}. \quad (1.3)$$

This probability, defined as a function of the inducer concentration, is shown schematically in Fig. 1.3 (B). While we have passed over some of the more subtle details of this calculation, the plot in Fig. 1.3 (B) presents a *quantitative* prediction of how the activity of an allosteric molecule should scale as a function of the inducer, in this case becoming less active as more inducer is present).

In **Chapter 2** and the associated supplementary **Chapter 6** of this dissertation, we use the Monod-Wyman-Changeux model of allostery (Monod et al., 1965) to build a predictive model of transcriptional regulation where the level of gene expression changes in response to changing activity of an allosteric transcriptional



**Figure 1.3: A Coarse grained representation of an allosteric molecule.** (A) Crystal structures of a variety of allosteric transcription factors are shown at the top. In this thesis, we coarse grain away many of the details to a minimal model (bottom) where the protein can be represent as being either active (red) or inactive (purple), both of which can bind an inducer molecule (orange). (B) By making an assumption of quasi-equilibrium, we can trivially compute a mathematical description of the active probability of an allosteric protein as a function of the inducer concentration (top). In this particular case, the inactive state becomes more probable relative to the active state at higher concentrations of inducer molecule.

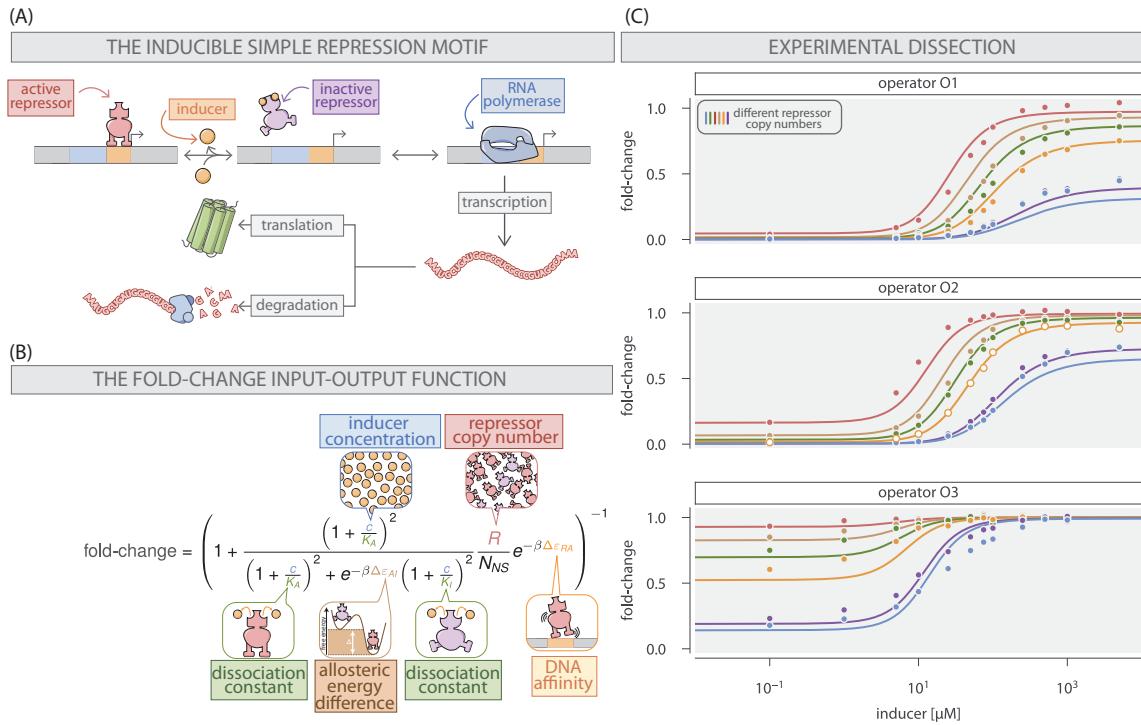
repressor. Using the same tricks given by Eq. 1.1 and Eq. 1.3, we expand upon a previously characterized thermodynamic model of the simple repression motif. This motif, schematized in Fig. 1.4 (A) is not just a convenient abstraction of a regulatory architecture. Rather, this motif is the most ubiquitous regulatory scheme in *E. coli* (???; Gama-Castro et al., 2016) and has been the target of much theoretical and experimental dissection (Bintu et al., 2005a; Brewster et al., 2014; Buchler et al., 2003; Garcia and Phillips, 2011; Phillips et al., 2019; Vilar and Leibler, 2003). However inclusion of allostery in a mathematical sense had yet to be experimentally dissected.

At the beginning of 2016, Manuel Razo-Mejia, Stephanie L. Barnes, Nathan M. Belliveau, Tal Einav, and I joined forces and set out to build a complete theoretical model for allosteric transcriptional regulation coupled with a thorough experimen-

tal dissection. This was no small task and would have likely taken a full Ph.D.’s worth of effort for a single person to do. Yet, within a year of project inception we had submitted a manuscript to preprint servers where all of us were annotated as equal contributors. This experience defined how I view collaboration in scientific research and serves as a shining example of scientific socialism.

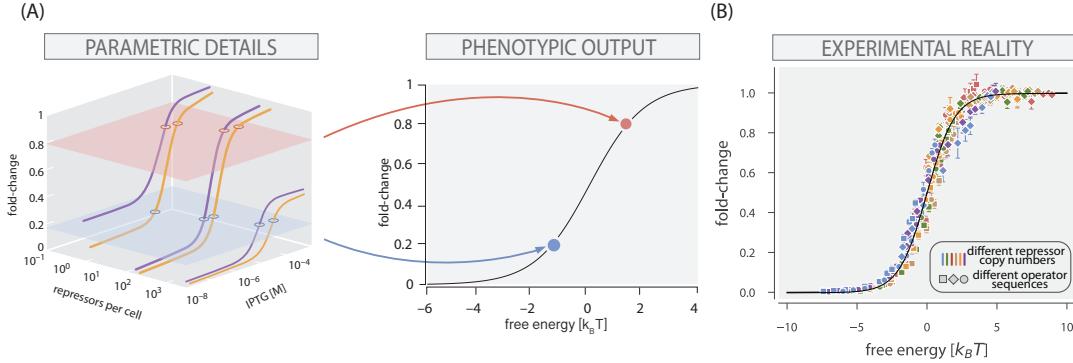
Together, we enumerated a complete thermodynamic model for the inducible simple repression motif and defined a succinct input-output function for the fold-change in gene expression (schematized in Fig. 1.4 (B)). This model, which is explored in depth in Chapter 2, is defined by a minimal set of biophysical parameters, many of which can be directly measured using standard tricks of molecular biology and biochemistry. With a model in hand, we turned to a collection of 17 unique *E. coli* strains, each with different copy numbers of the repressor protein and different regulatory DNA sequences. Using our theoretical model, we inferred the lone two biophysical parameters which we did not know *a priori* from a single experimental strain (white points in middle panel of Fig. 1.4 (C)), and tested our predictions on all other experimental strains. We found the model to be remarkably predictive, suggesting that our “toy” model of an allosteric repressor captured the underlying physics of the system.

A key feature of this work is a derivation of thermodynamic state variable of this regulatory architecture which we term the *free energy*. This parameter provides an intuition for the effective free energy difference between states of the promoter in which the repressor is bound relative those states in which the repressor is not bound to the promoter. This parameter accounts for all of the ways in which one can tune the parameter values and still achieve the same fold-change in gene expression, as is diagrammed in Fig. 1.5 (A). While we leave the details of this derivation to Chapter 2, we emphasize that this formalism provides a means by which all of the experimental measurements plotted in Fig. 1.4 (C) can be collapsed onto a master curve defined *only* by the free energy, which is illustrated in Fig. 1.5 (B). This scaling, often referred to as “data collapse” in physics, concretely shows that



**Figure 1.4: Experimental dissection of the inducible simple repression input-output function.** (A) Schematic diagram of the inducible simple repression motif. (B) Schematic diagram of the input-output function as is derived in Chapter 2. (C) Experimental measurements of the fold-change in gene expression using the *lac* repressor from *E. coli*. Different rows correspond to different operator sequences and therefore different values for  $\Delta \varepsilon_{RA}$ . Different colors correspond to different values for the average repressor copy number  $R$ . While filled points in the middle panel represent the experimental strain used to infer the values of the inducer dissociation constants. All points correspond to the mean of at least 10 biological replicates.

one has identified the *natural variable* of the system. With this scaling function in hand, we are able to make a measurement of the biophysical parameters, compute the free energy, and make a concrete prediction of what the fold-change in gene expression will be. Or, as we will see in the following section, details of the biophysical parameters can be determined directly from an empirical measurement of the free energy.



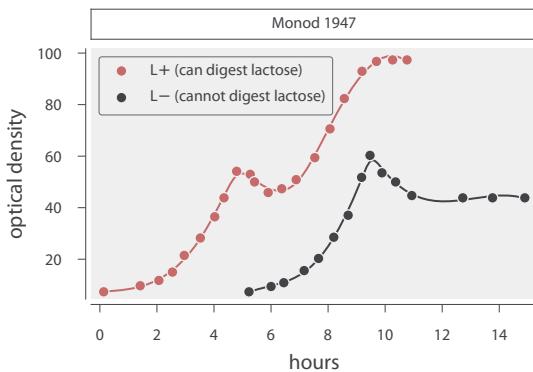
**Figure 1.5: Collapse of individual induction profiles onto a simple scaling function.** (A) Many different combinations of parameter values yield the same value of the fold-change in gene expression, shown as red and blue horizontal planes. Any point on those planes corresponds to a single value of the free energy (middle) and will appear on the master curve. (B) Data presented in Fig. 1.4 (C) collapsed onto the master curve defined by the predicted value of the free energy.

### 1.3 Using Free Energy to Examine Evolutionary Adaptation

Allow us to briefly return to Monod and his biphasic growth curves in the mid 1940's. At this point in scientific history, the French vision of biology had taken a strongly finalistic and vitalistic turn (Loison, 2013). In particular, a neo-Lamarckian view had been employed to explain the phenomenon of enzymatic adaptation where the enzymes appropriate for digesting the substrate could be spontaneously formed out the bacterial cytoplasm and inherited by the cell's descendants, completely independent of genes. In general, this approach to biology deeply frustrated Monod and strongly influenced his desire to "physicalize" the science (Loison, 2013). One tool he knew was critical to this mission was the burgeoning field of genetics. In the mid 1930's Monod undertook a short retreat to Thomas Hunt Morgan's lab at Caltech where he was introduced to genetics which he later remarked to as "biology's first discipline" (Loison, 2013). This visit had a profound impact on Monod, who reflected upon it some three decades later:

*"Upon my return to France, I had again taken up the study of bacterial growth.*

*But my mind remained full of the concepts of genetics and I was confident of its ability to analyze and convinced that one day these ideas would be applied*



**Figure 1.6: Growth curves of lactose-positive and lactose-negative *E. coli* strains on a glucose/lactose mixture.** Black curve shows the growth curve of an *E. coli* strain unable to digest lactose grown on a glucose/lactose mixed medium. Red curve shows a mutant of the same *E. coli* strain which is able to consume lactose. The latter displays a diauxic growth cycle with an adaptive period (highlighted in white), illustrating that enzymatic adaptation is a truly genetic property. Figure adapted from Monod (1947).

*to bacteria.”* (Monod, 1966)

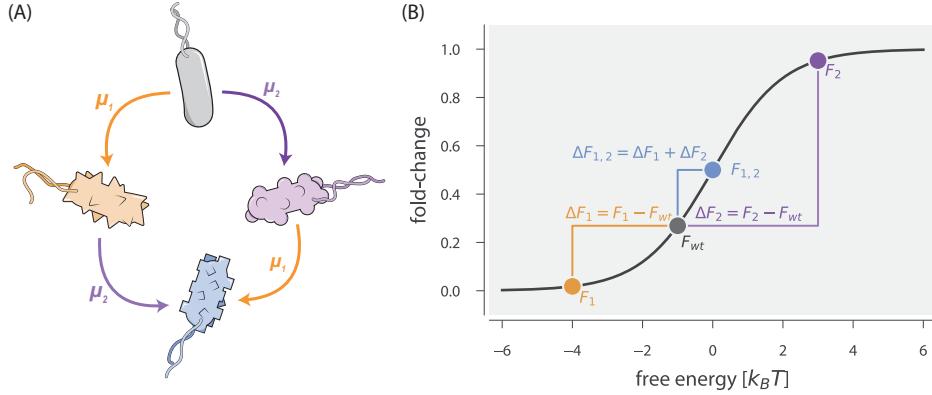
Once he returned to his study of bacterial growth and enzymatic adaptation, he was confronted with incorporating the role of genetic inheritance into his mechanistic explanations. In the mid 1940’s, Monod and his coworkers had begun experimenting with a strain of *E. coli* which was unable to digest lactose ,termed *L-*. When grown on a mixture of glucose and lactose, this strain would not display a diauxic shift and would only be able to consume the glucose in the medium (Fig. 1.6, black). However, Monod and his coworker Alice Audureau discovered a mutation in this strain which *enabled* the digestion of lactose, termed *L+* (Monod, 1947). The growth curve of this strain had the striking feature of diauxic growth. Rather than this mutation merely enabling the digestion of lactose, it did so in a non-constitutive manner and preserved the phenomenon of adaptation. This was an important step forward in Monod’s understanding of enzymatic adaptation (Loison, 2013), revealing that it was a “truly genetic property” (Monod, 1966).

This finding illustrates the level of evolutionary adaptation operating at the level of molecules. While it is difficult to find any literature dissecting this particular

*L+* mutation, it is not difficult to imagine several different mechanisms by how it could be manifest. One such explanation is that this *L+* mutation is within a transcriptional regulator itself where a deficiency in the ability to respond to the presence of lactose (and decreasing glucose concentration) had been restored. Such mutations are the crux of **Chapter 3** and the corresponding supplemental **Chapter 7** of this dissertation.

As summarized in the previous section and discussed in depth in Chapter 2, Chapter 3 and the associated supplemental Chapter 7 of this dissertation focus on the influence of mutation within the allosteric transcription factor. Furthermore, Chapter 3 presents a generic mechanism by which shifts in the free energy can be mapped directly to changes in values of the biophysical parameters. This chapter, much like Chapter 2, was borne out of a wonderful collaboration with Manuel Razo-Mejia, Stephanie L. Barnes, Nathan M. Belliveau, Tal Einav, and Zofii A. Kaczmarek. Being able to launch another collaborative effort afforded us the opportunity to both develop a new theoretical interpretation for how mutations influence the free energy and acquire enough experimental data to thoroughly test it.

The primary conceptual development of Chapter 3 is illustrated in Fig. 1.7. Theoretically, we consider a bacterial strain with an allosteric repressor (which we term the “wild-type” repressor) that has been sufficiently characterized. Given enough parametric knowledge of the system, we can easily compute both its predicted average fold-change in gene expression along with the corresponding free energy. However, once a mutation has been introduced *into the repressor protein* (resulting in a non-synonymous amino acid change), we are once again ignorant *a priori* of what changes, if any, that mutation may have imparted on the system. In Fig. 1.7, we examine two separate hypothetical mutations, shown in purple and orange, which significantly change the character of the system by either increasing or decreasing the fold-change in gene expression, respectively. If we assume that these mutations do not change the underlying physics of the system, we are



**Figure 1.7: Mutations lead to shifts in free energy, permitting prediction of double mutant phenotypes.** Consider a wild-type bacterium which on, on average, exhibits a fold-change of  $\approx 0.3$  and a free energy of  $-1 k_B T$  (grey point in (B)). We can consider that a single mutation (either orange or purple) changes the mean fold-change and therefore the free energy, translating the measurement about the master curve (black line in (B)). Assuming there are no epistatic interactions between the two single mutations, a null hypothesis predicts that for the double mutant (blue bacterium in (A) and point in (B)) $0$ , the net free energy is simply the sum of the individual free energy shifts.

permitted to use the theoretical framework outlined in Chapter 2 and in Fig. 1.4 to characterize each mutation and determine what biophysical parameters have been changed. This permits us to calculate the new free energy of the system ( $F_{\text{mutation } 1}$ ) as well as the shift in free energy from the wild-type value,

$$\Delta F_{\text{mutation } 1} = F_{\text{mutation } 1} - F_{\text{wt}}. \quad (1.4)$$

As will be described in detail in Chapter 3 and the supplemental Chapter 7, the precise value of this free energy shift  $\Delta F$  can be directly computed given sufficient parametric knowledge.

This formalism provides a mathematical hypothesis for how double mutants may behave. Given known values for  $\Delta F$  of each mutation in isolation, can we compute the shift in free energy of the pairwise double mutant  $\Delta F_{\text{mutations } 1 \& 2}$ ? Eq. 1.4 presents a mathematical null hypothesis that the net shift in the free energy is simply the sum of the individual shifts in free energy,

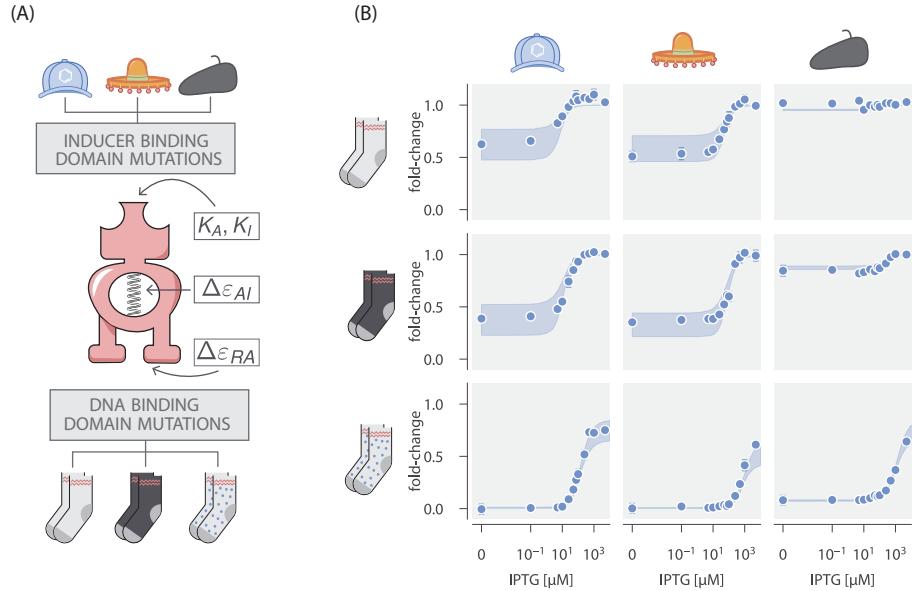
$$\Delta F_{\text{mutations } 1 \& 2} = \Delta F_{\text{mutation } 1} + \Delta F_{\text{mutation } 2}, \quad (1.5)$$

assuming there are no epistatic interactions between the mutations. Given the fact that we can compute the fold-change in gene expression given knowledge of the free energy, we can therefore predict the double mutant phenotype *a priori*, a prediction not possible prior to this work.

Over the course of two years (while this theory was in the works), the experimental cast of characters (Stephanie L. Barnes, Nathan M. Belliveau, Manuel Razo-Mejia, and Zofii A. Kaczmarek) and I made a series of mutations in the LacI repressor that we had characterized in the work presented in Chapter 2. These mutations included three point mutations in the DNA binding domain of the repressor, four mutations in the inducer binding domain, nine double mutants (one inducer binding and one DNA binding each), across four repressor copy numbers and three operator sequences. While this process of strain generation and data collection is not the primary focus of the work, it took  $\approx 80\%$  of the effort. Without them, this work would have remained an untested theoretical novelty. While we leave many of the rich details of this prediction to the reader in Chapter 3, we showcase our experimental success in Fig. 1.8 (B) where the predicted induction profiles of nine double mutants (light blue shaded regions) are overlaid with their experimental measurements (points). The near 100% agreement between theory and experiment illustrates the utility of using free energy shifts as a means to predict new phenotypes.

#### 1.4 Topic IV: The Physiological Adaptability of Transient Molecular Interactions

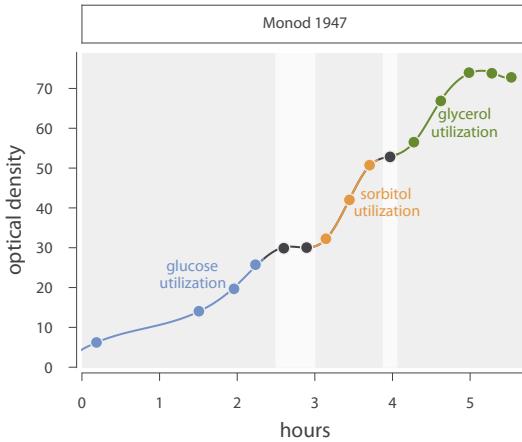
In **Chapters 4 and 5** (and the associated supplementary **Chapters 8 and 9**), we explore the final level of adaptation in Fig. 1.1 – physiological adaptation. We do so in two distinctly different systems. The first (Chapter 4) builds upon our discussion of transcriptional regulation, but now examines how robust the biophysical parameters of the thermodynamic model are to changes in physiology, either by changing the available carbon source or by changing the temperature. Secondly, we examine physiological adaptation in the context of osmoregulation – a true



**Figure 1.8: Theoretical prediction and experimental validation of double mutant phenotypes.** (A) Cartoon representation of the LacI repressor with mutations in the inducer binding domain and DNA binding domain represented by hats and socks, respectively. While the mutations have known chemical features, we characterize each mutation as potentially modifying four biophysical parameters,  $K_A$ ,  $K_I$ ,  $\Delta\epsilon_{AI}$  for inducer binding mutants, or  $\Delta\epsilon_{RA}$  for DNA binding mutants. (B) Predicted induction profiles for pairwise double mutants are shown as blue shaded regions representing the uncertainty in our predictions. Experimental measurements are shown as blue points (means of at least 10 biological replicates). Each row corresponds to a single DNA binding domain mutation and each column to a single inducer binding domain mutation.

matter of life and death in the single-celled world.

Up to this point in our travels through scientific history, we have examined Monod's growth curves in various pairwise combinations of sugars. A feature of note is that the presence of diauxic shifts can be seen in various organisms and for many different types of sugars such as sucrose/arabinose, glucose/sorbitol, and glucose/lactose pairings (Monod, 1947). These combinations reveal that cells are able to juggle dual-input logic systems where the "decision" to digest one carbon source or another relies on monitoring changes in concentrations of either sugar. In his 1947 treatise, Monod showed that this phenomenon was not limited to dual-carbon mixtures and presented a "triauxic" growth curve of *E. coli* grown on a glu-



**Figure 1.9: A metabolic hierarchy in a three-component growth mixture of glucose, sorbitol, and glycerol.** A “triauxic” growth curve illustrating a hierarchy of carbon source metabolism. An *E. coli* culture was grown in a medium with equal parts glucose, sorbitol and glycerol with utilization in that order. Auxic transitions are shown as black points and white vertical lines. Regions of the growth curve where glucose, sorbitol, and glycrol are primarily consumed are colored in blue, orange, and green, respectively. Data digitized from Monod (1947).

cose/sorbitol/glycerol mixture, shown in Fig. 1.9. This result illustrated to Monod that the mechanisms underlying enzymatic adaptation “have the character of *competitive interactions* between different specific enzyme forming systems” (Monod, 1947).

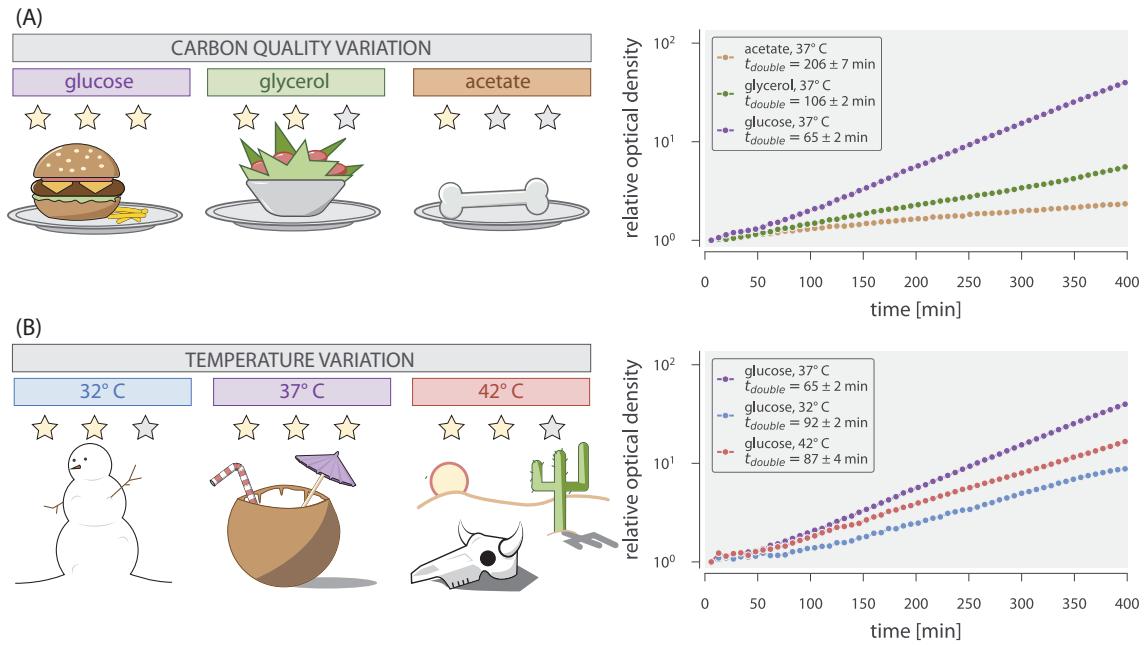
These competing interactions must be resilient to a variety of physiological states. Despite the fact that the carbon atoms in glucose, sorbitol, and glycerol are all ultimately incorporated into the same biomolecules, their pathways to utilization are all distinct and include a variety of different metabolic intermediates. Furthermore, the exponential growth phases in Fig. 1.9 for each carbon source have different growth rates which itself results in large changes in cell volume (Jun et al., 2018; Taheri-Araghi et al., 2015a), genome copy number (Nordström and Dasgupta, 2006), and global gene expression patterns (Hui et al., 2015; Li et al., 2014; Schmidt et al., 2016). Despite these changes in cellular physiology, the regulatory systems underlying enzymatic adaptation still function with binding of transcription factors being ignorant of the majority of possible metabolic states of the cell.

Despite this empirical observation, it has been commonly assumed that the utility of thermodynamic models of gene expression are limited and that the precise values of the biophysical parameters are directly tied to the physiological state in which they were determined. It has even been said that thermodynamic models of gene expression have been a “tactical success, yet strategic failure” in building an understanding of how genomes operate (Phillips et al., 2019).

In **Chapter 4** of this dissertation, we quantitatively assess these assumptions in the context of gene expression by considering the theoretical models built in Chapters 2 and 3 and directly measuring the adaptability of the inducible simple repression regulatory architecture across different physiological states. Namely, we explore how predictive our thermodynamic model can be when modulating either the quality of the carbon source (glucose, glycerol, or acetate, Fig. 1.10 (A)) or by changing the temperature of the growth medium ( $32^\circ\text{ C}$ ,  $37^\circ\text{ C}$ , or  $42^\circ\text{ C}$ , Fig. 1.10 (B)). The culture doubling time varies by nearly a factor of four across the different conditions, illustrating the diversity in physiological states.

How could this variation in cellular physiology be incorporated into our thermodynamic model? Up to this point in this thesis, experiments have been conducted in a growth medium supplemented with glucose held at a balmy  $37^\circ\text{ C}$ . However, nowhere in the thermodynamic model schematized in Fig. 1.4 (B) is it specified *which* carbon source must be present whereas the temperature of the system is explicitly included as a multiplicative factor  $\beta = (k_B T)^{-1}$  in front of the exponentiated terms. These features of the model allow us to make explicit predictions of how these perturbations should influence the observed fold-change in gene expression, if at all.

The parameter that we can say *a priori* is very likely to change is the repressor copy number  $R$ . In Chapters 2 and 3, we knew the total repressor copy number from previous work where the copy numbers were directly measured via quantitative Western blotting in a particular physiological state (Garcia and Phillips, 2011). However, it has been known for nearly three-quarters of a century that the



**Figure 1.10: Control of cellular physiology via carbon source and temperature variation.** (A) Carbon sources used in work presented in Chapter 4 (left) with “star rating” indicating quality of the carbon source. Growth curves for the three carbon sources, all at 37° C are shown on right-hand panel. (B) Growth temperatures explored in Chapter 4 (left) with “star rating” indicating fastest growth rate. Growth curves (right) are shown for the three temperatures, all of which use glucose as the sole carbon source. For right-hand panels in (A) and (B), optical density is computed relative to the initial optical density of the culture.

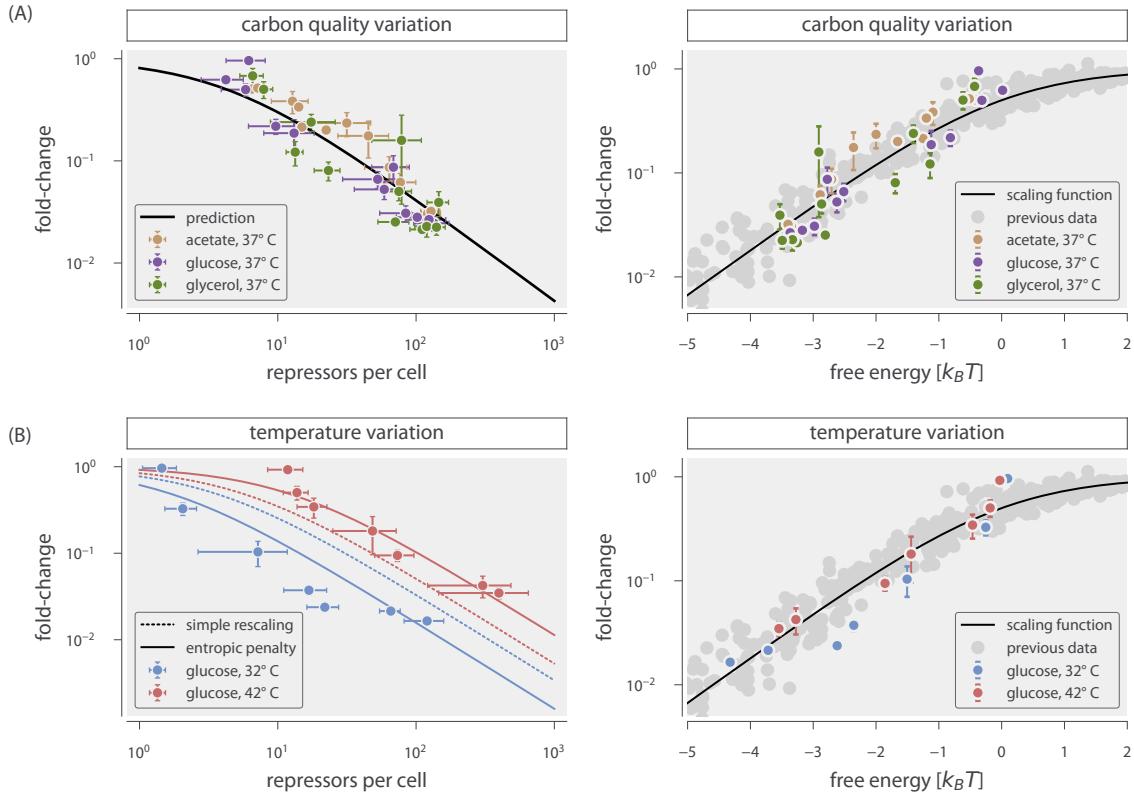
total protein content of the cell scales linearly with the growth rate (Jun et al., 2018; Schaechter et al., 1958), a phenomenon that has recently been queried at the single-protein level through proteomic methods (Li et al., 2014; Peebo et al., 2015; Schmidt et al., 2016; Valgepea et al., 2013). Thus, we cannot assume that the protein copy number of the strains used in Chapters 2 and 3 will not be perturbed. To account for this fact, we used a fluorescence-based method to directly count the number of LacI repressors per cell in each growth condition, a method which is discussed in extensive detail in Chapter 9. This experimental approach, while necessary, is extremely laborious. I am indebted to the work of Zofia A. Kaczmarek for her heroic efforts in conducting a large number of the experiments presented in Chapter 4.

Our work revealed two key features of this thermodynamic model of gene expression. First, we found that the values of the biophysical parameters inferred

from a single physiological state were remarkably predictive when the quality of the carbon source was decreased (Fig. 1.11 (A, left)). This indicates that this genetic circuit is largely insulated from the metabolic state of the cell. This is exemplified in our ability to collapse the measurements across different carbon sources as a function of the free energy, shown in Fig. 1.11 (A, right). For the carbon sources studied in this Chapter, we conclude that this simple thermodynamic model can be considered both tactically and strategically successful.

Yet when it comes to temperature, we find that a simple rescaling of the thermal energy of the system is *not* sufficient to predict the output of this genetic circuit when the temperature is varied (dashed lines in left-hand side of Fig. 1.11, (B)). This is not necessarily a surprising result as binding of transcription factors is not strictly an enthalpic process. Temperature is known to have a strong influence on many material properties of DNA ,such as persistence length and salt release (Goethe et al., 2015), excluded volume effects (Driessens et al., 2014), and repressor-DNA solubility (Elf et al., 2007), to name a few of many effects. To phenomenologically characterize the influence of temperature on the fold-change in gene expression, we considered that there was a constant entropic penalty (though inclusion of a temperature-dependent entropic cost is discussed in Chapter 9). We found that inclusion of this parameter markedly improved the description of the data (solid lines in left-hand side of Fig. 1.11 (B)) and permitted data collapse within experimental noise of data collected at 37° C (right-hand side of Fig. 1.11 (B)).

The inclusion of a phenomenological entropic parameter is not by any means meant to shut the book on temperature effects in this model. Rather, it serves as a representation of what *may* help explain these effects and demands more focused theoretical and experimental work. To say that the current disagreement between theory and experiments embodies the “strategic failure” of thermodynamic models is, in my view, disingenuous. To say so would be to deem the initial failures of elasticity theory to properly predict the influence of impurities and temperature on the elastic constants of materials as a “strategic failure” in material science.



**Figure 1.11: Performance of a simple thermodynamic model of simple repression in diverse physiological states.** (A) Fold-change in gene expression measurements in different carbon sources plotted against the average repressor copy number (left) and free energy (right). Black line in the left-hand panel is the predicted fold-change assuming no parameters are modified. (B) Fold-change measurements at different temperatures plotted as a function of the repressor copy number (left) and free energy (right). Dashed-lines in left-hand plot show the predicted fold-change with a simple rescaling of the thermal energy. Solid lines are predicted fold-change upon inclusion of an entropic penalty. Points on right-hand plot were computed using parameters with an entropic penalty. All measurements and errors displayed are the mean and standard error of three to eight biological replicates. Light-grey points in right hand panels are data from Garcia and Phillips (2011), Brewster et al. (2014), Razo-Mejia et al. (2018), and Chure et al. (2019), all of which were measured in glucose-supplemented media at 37° C.

Initial phenomenological models of the effects of impurities and temperature on elastic properties of solids (Friedel, 1974) led to several decades of focused theoretical and experimental work that resulted in a complete predictive and mechanistic description (Phillips, 2001). Now is the time for a similar approach to biology in the context of temperature and the regulation of gene expression.

### 1.5 On Facing the Elements

The first four chapters of this work encompass myriad perspectives of adaptive processes at the level of transcription regulation. However, just as important as the regulation is the action of the gene that is ultimately expressed. While Monod's work described in the preceding sections was focused on the expression of enzymes, we now turn to yet another level of physiological adaptation in bacteria – the regulation of turgor pressure.

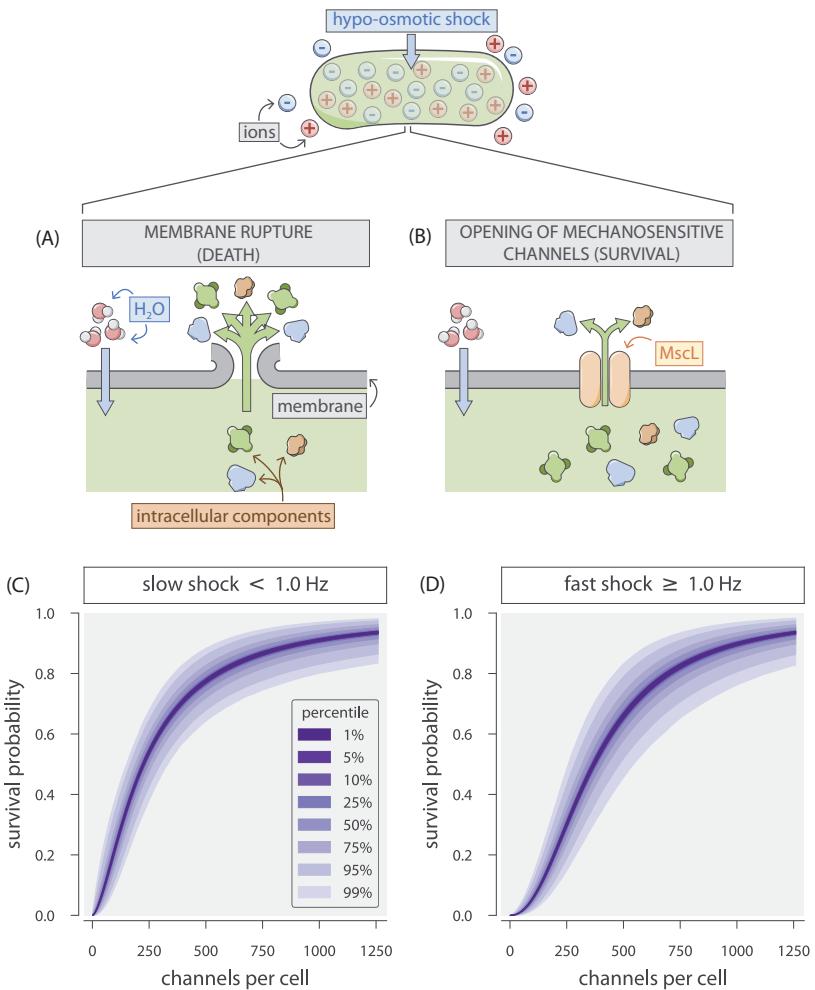
In the wild, microbes are constantly faced with an array of environmental insults ranging from changes in temperature, availability of oxygen for aerobic respiration, and even chemical warfare from neighboring microbial communities (Czaran et al., 2002). One such environmental challenge microbes often face is the variation in the osmolarity of their surroundings. Changes in ion concentrations can result in large volumes of water rushing across the cell membrane, leading to rupture of the membrane and ultimately cell death (Fig. 1.12 (A)). Unsurprisingly, all domains of life have evolved clever mechanisms to combat these osmotic shocks and regulate their internal turgor pressure.

One such mechanism for osmoregulation in *E. coli* is through the action of mechanosensitive ion channels – large, transmembrane structures which sense tension in the cell membrane. Exposure to a hypo-osmotic shock (where water rushes across the cell membrane *into* the cell), a change in membrane tension is sensed by these mechanosensitive channels, triggering a conformational change which opens a pore in the membrane without rupture (Fig. 1.12 (B)). This acts as a pressure release valve, providing a means for turgor pressure to be relieved without a potentially fatal burst. This phenomenon represents yet another system in which

adaptation can be found at the molecular, evolutionary, and physiological levels. In **Chapters 5 and 9** of this dissertation, we explore a fundamental question – how many mechanosensitive channels does a cell need to have an appreciable chance at surviving an osmotic shock?

To approach this question, Heun Jin Lee and I collaborated on the experimental and data analysis components, respectively. This is a project that had been in preparation for several years before I had the privilege of joining the team in the summer of 2017. While the experimental techniques used to probe transcriptional regulation were far from simple, they pale in comparison to those employed by Heun Jin. This project required an enormous amount of molecular biology to generate the necessary strains in which the number of mechanosensitive channels could be tuned across three orders of magnitude and measured with precision. This process involved reworking classic techniques in molecular biology to remove the presence of osmotic shocks which would prove fatal for strains with few or no mechanosensitive channels. On top of the complex biochemistry, Heun Jin developed a clever microfluidic system where osmotic shocks could be imaged in real time at the single cell level. While the majority of Chapter 5 focuses on the analysis and interpretation of the data, none of it would have been possible without Heun Jin's Herculean efforts.

While we leave the details of the inference to Chapter 5 and the supplemental Chapter 9, the survival probability as a function of the total mechanosensitive channel number is given for “slow” and “fast” osmotic shocks in Fig. 1.12 (C) and (D), respectively. The credible regions in this plot illustrate that for an  $\approx 80\%$  chance of surviving either a slow or a fast osmotic shock, at least  $\approx 500$  channels are needed. This number is in agreement with recent proteomic measurements in *E. coli* (Li et al., 2014; Schmidt et al., 2016; Soufi et al., 2015), but are at odds with current theoretical models. While it is difficult to theoretically define a survival/death criterion, current physical models predict only a few mechanosensitive channels (specifically, MscL) are needed to relieve even large increases in membrane ten-



**Figure 1.12: The connection between mechanosensitive channel copy number and probability of survival.** (A) In the absence of mechanosensitive channels, water rushing across the membrane during a hypo-osmotic shock can lead to membrane rupture and large-scale release of intracellular components into the extracellular space, resulting in cell death. (B) In the presence of mechanosensitive channels (specifically, the major *E. coli* mechanosensitive channel MscL as shown in yellow), increased membrane tension results in a conformational change of the channel, resulting in the expulsion of water and some small constituents of the intracellular milieu. The inferred survival probability curves for slow and fast shock exchange rates are shown in (C) and (D), respectively. Different shaded purple regions correspond to different credible regions of the estimates.

sion. These findings illustrate another avenue in which the disagreement between theory and careful, quantitative experiments reveal gaps in our understanding of fundamental biological phenomena.

### 1.6 On Molecular Biophysics and Evolutionary Dynamics

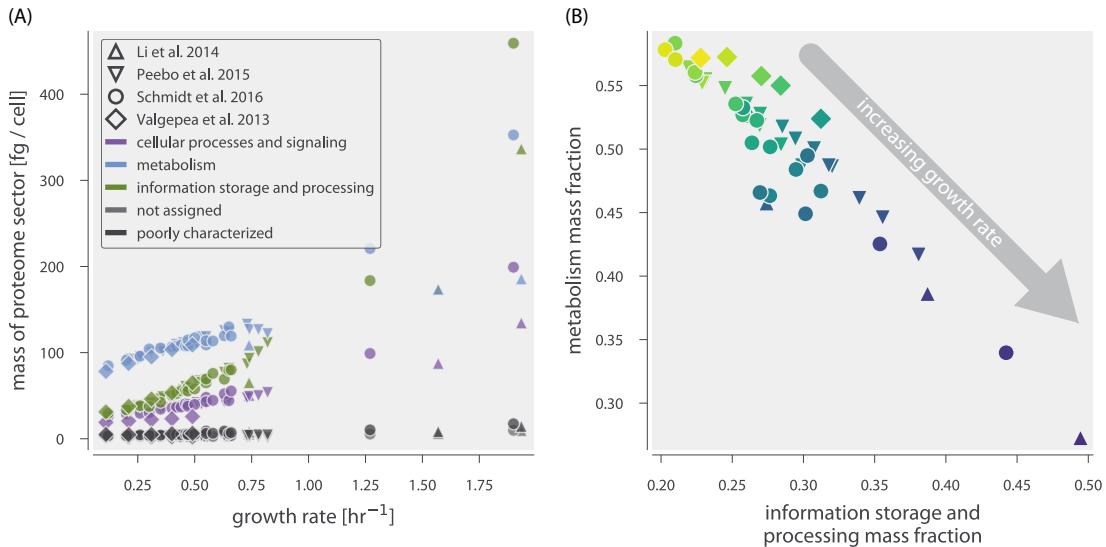
This thesis as a whole presents an attempt to understand how adaptive processes operate in biological systems at a mechanistic level beyond qualitative description. The thermodynamic model derived and explored in Chapter 2 presents a concrete theoretical framework through which we can understand how mutations and environmental perturbations influence the output of a simple genetic circuit with quantitative precision. While the work here specifically explores the *mean* level of gene expression of a population, I've had the privilege to be involved in several projects which explore the complete distribution of gene expression of various regulatory motifs using non-equilibrium models(Laxhuber et al., 2020; Razo-Mejia et al., 2020). Both equilibrium and non-equilibrium approaches, while differing in their fundamental assumptions of the system, can be used to understand how the regulation of gene expression occurs *in vivo* and should be viewed as complementary rather than adversarial approaches.

A combination of these types of approaches will be necessary to attack what I believe is the next great frontier of biological physics – predicting evolution. While this thesis is primarily focused on a single type of regulatory architecture regulating a single promoter via a single species of transcription factor, it is worth remembering that systems-level phenotypes are often complex resulting from the concerted action of an array of biological processes. As was mentioned in our discussion on physiological adaptation, it has been known for nearly a century that the bacterial growth rate is directly correlated to the total protein content of the cell, with recent works illustrating rich phenomenology in the structure of the bacterial proteome as a whole (Hui et al., 2015; Klumpp and Hwa, 2014; Li et al., 2014; Schmidt et al., 2016; Scott et al., 2010).

In collaboration again with Nathan M. Belliveau, we have begun to explore how

the composition of the bacterial proteome is structured at the single-protein level. Fig. 1.13(A) shows a compiled data from a variety of different proteomic data sets (using either quantitative mass spectrometry (Peebo et al., 2015; Schmidt et al., 2016; Valgepea et al., 2013) or ribosomal profiling (Li et al., 2014)) where the abundance of different molecular constituents of the bacterial proteome are plotted as a function of the growth rate. These components, broken down by their functional designation according to their Cluster of Orthologous Groups (COG) annotation (Galperin et al., 2015), reveal varied dependencies on the growth rate. Of note are the COG classes “cellular processes and signaling”, “metabolism”, and “information storage and processing” which all appear to have a correlation between the cellular growth rate and the total mass of that proteome sector. However, when plotted as the total *mass fraction* of the proteome instead of the total mass, a striking result is observed. Fig. 1.13 (B) reveals a very strong, negative correlation between the mass fraction of the proteome dedicated to information storage and processing (including ribosomal and transcriptional machinery) and the proteome fraction dedicated to metabolism. This direct competition for resources between the proteins involved in translation (ribosomes, elongation factors, etc.) and metabolic networks has been shown previously (Hui et al., 2015; Klumpp and Hwa, 2008; Scott et al., 2010) and suggests a strong evolutionary constraint in how resources can be optimally partitioned.

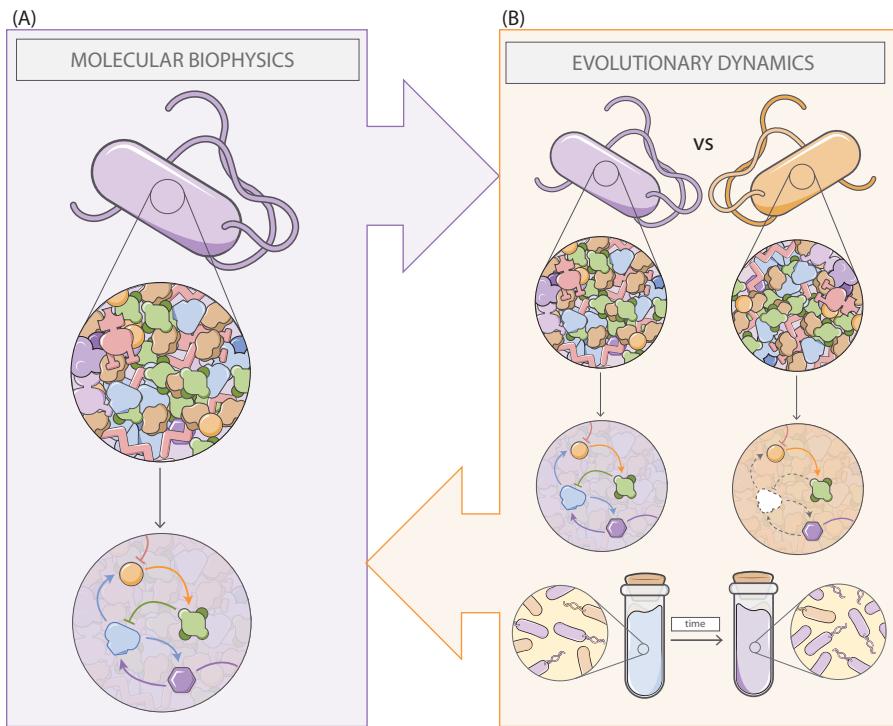
As of this writing, our understanding of the cellular resource allocation visible in Fig. 1.13 remains largely phenomenological (Scott et al., 2014). This is in part due to the tremendously high-dimensional nature of systems-level organization. Our understanding of systems with such huge degrees-of-freedom have classically benefited enormously from the application of statistical mechanics as this thesis shows in the context of transcriptional regulation. The quantitative framework derived and carefully dissected in this thesis, I believe, lays the groundwork to understand how phenomena such as that shown in Fig. 1.13 (B) arise, and perhaps more importantly, evolve. A recent work from Michael Lässig, Ville Mustonen, and Aleksandra Walczak entitled *Predicting evolution* (Lässig et al., 2017) describes



**Figure 1.13: Allocation of cellular resources induces compositional structure in the *E. coli* proteome.** (A) The total proteome mass of the five major annotated COG categories is shown as a function of the experimental growth rate. Different marker shapes represent different data sets. (B) The total fraction of the proteome dedicated to metabolic machinery plotted as a function of the total proteome mass dedicated to the processes of the central dogma. Different shapes correspond to the different data sets shown in (A). Color indicates increasing growth rate from yellow to dark blue. Data shown in this figure come from Peebo et al. (2015) (inverted triangles), Li et al. (2014) (triangles) Schmidt et al. (2016) (circles) and Valgepea et al. (2013) (diamonds)

what the future of evolutionary theory may look like given these types of models. Recent technological advancements in sequencing, microscopy, and computation coupled with theoretical advancements in the biophysics of gene regulation present an opportunity for a rich theoretical dialogue between molecular biophysics and evolutionary dynamics coupled with experimental dissection (Fig. 1.14).

The somewhat recent paradigm shift in our understanding of noise in biological networks illustrates how cross-disciplinary approaches to scientific discovery can solve (and more-often) create new fields of biological inquiry. I can only hope that some of the material described in the coming chapters can help contribute to a systems-biology approach to evolution.



**Figure 1.14: The coming interplay between molecular biophysics and evolutionary dynamics.** (A) Recent progress in our understanding the structure and function of biological networks has resulted in many examples where high-dimensional biological phenomena can be boiled down to effective phenomena. Future work will draw from our understanding of these networks to place them in an evolutionary perspective (B) where the connection between perturbations at the level of nodes in biological networks can be drawn to fitness and evolutionary trajectories can be predicted.

*Chapter 2*

## THROUGH THE INTRAMOLECULAR GRAPEVINE: SIGNAL PROCESSING VIA ALLOSTERIC TRANSCRIPTION FACTORS

A version of this chapter originally appeared as Razo-Mejia, M.\* ; Barnes, S.L.\* ; Belliveau, N.M.\* ; Chure, G.\* ; Einav, T.\* ; Lewis, M.; and Phillips, R. (2018). *Tuning Transcriptional Regulation through Signaling: A Predictive Theory of Allosteric Induction*. Cell Systems 6, 456-469.e10. DOI:<https://doi.org/10.1016/j.cels.2018.02.004>. M.R.M, S.L.B, N.M.B, G.C., and T.E. contributed equally to this work from the theoretical underpinnings to the experimental design and execution. M.R.M, S.L.B, N.M.B, G.C, T.E., and R.P. wrote the paper. M.L. provided extensive guidance and advice.

### 2.1 Abstract

Allosteric regulation is found across all domains of life, yet we still lack simple, predictive theories that directly link the experimentally tunable parameters of a system to its input-output response. To that end, we present a general theory of allosteric transcriptional regulation using the Monod-Wyman-Changeux model. We rigorously test this model using the ubiquitous simple repression motif in bacteria by first predicting the behavior of strains that span a large range of repressor copy numbers and DNA binding strengths and then constructing and measuring their response. Our model not only accurately captures the induction profiles of these strains, but also enables us to derive analytic expressions for key properties such as the dynamic range and  $[EC_{50}]$ . Finally, we derive an expression for the free energy of allosteric repressors that enables us to collapse our experimental data onto a single master curve that captures the diverse phenomenology of the induction profiles.

## 2.2 Introduction

Understanding how organisms sense and respond to changes in their environment has long been a central theme of biological inquiry. At the cellular level, this interaction is mediated by a diverse collection of molecular signaling pathways. A pervasive mechanism of signaling in these pathways is allosteric regulation, in which the binding of a ligand induces a conformational change in some target molecule, triggering a signaling cascade (Lindsley and Rutter, 2006). One of the most important examples of such signaling is offered by transcriptional regulation, where a transcription factors' propensity to bind to DNA will be altered upon binding to an allosteric effector.

Despite the ubiquity of allostery, we largely lack a formal, rigorous, and generalizable framework for studying its effects across the broad variety of contexts in which it appears. A key example of this is transcriptional regulation, in which allosteric transcription factors can be induced or corepressed by binding to a ligand. An allosteric transcription factor can adopt multiple conformational states, each of which has its own affinity for the ligand and for its DNA target site. *In vitro* studies have rigorously quantified the equilibria of different conformational states for allosteric transcription factors and measured the affinities of these states to the ligand (Harman, 2001; Lanfranco et al., 2017). In spite of these experimental observations, the lack of a coherent quantitative model for allosteric transcriptional regulation has made it impossible to predict the behavior of even a simple genetic circuit across a range of regulatory parameters, physiological states of the organism, and evolutionary isoforms of the regulatory sequences.

The ability to predict circuit behavior robustly—that is, across both broad ranges of parameters and regulatory architectures—is important for multiple reasons. First, in the context of a specific gene, accurate prediction demonstrates that all components relevant to the genes' behavior have been identified and characterized to sufficient quantitative precision. Second, in the context of genetic circuits in general, robust prediction validates the model that generated the prediction.

Possessing a validated model also has implications for future work. For example, when we have sufficient confidence in the model, a single data set can be used to accurately extrapolate a system’s behavior in other conditions. Moreover, there is an essential distinction between a predictive model, which is used to predict a system’s behavior given a set of input variables, and a retroactive model, which is used to describe the behavior of data that has already been obtained. We note that even some of the most careful and rigorous analysis of transcriptional regulation often entails only a retroactive reflection on a single experiment. This raises the fear that each regulatory architecture may require a unique analysis that cannot carry over to other systems, a worry that is exacerbated by the prevalent use of phenomenological functions (e.g. Hill functions or ratios of polynomials) that can analyze a single data set but cannot be used to extrapolate a system’s behavior in other conditions (Poelwijk et al., 2011; Rogers et al., 2015; Rohlhill et al., 2017; Setty et al., 2003; Vilar and Saiz, 2013).

This work explores what happens when theory takes center stage, namely, we first write down the equations governing a system and describe its expected behavior across a wide array of experimental conditions, and only then do we set out to experimentally confirm these results. Building upon previous work (Brewster et al., 2014; Garcia and Phillips, 2011; Weinert et al., 2014) and the work of Monod, Wyman, and Changeux (Monod et al., 1965), we present a statistical mechanical rendering of allosteric regulation in the context of induction and corepression (shown schematically in Fig. 2.1 and henceforth referred to as the MWC model) and use it as the basis of parameter-free predictions which we then test experimentally. More specifically, we study the simple repression motif – a widespread bacterial genetic regulatory architecture in which binding of a transcription factor occludes binding of an RNA polymerase, thereby inhibiting transcription initiation. The MWC model stipulates that an allosteric protein fluctuates between two distinct conformations – an active and inactive state – in thermodynamic equilibrium (Monod et al., 1965). During induction, for example, effector binding increases the probability that a repressor will be in the inactive state, weakening its ability to bind to the pro-

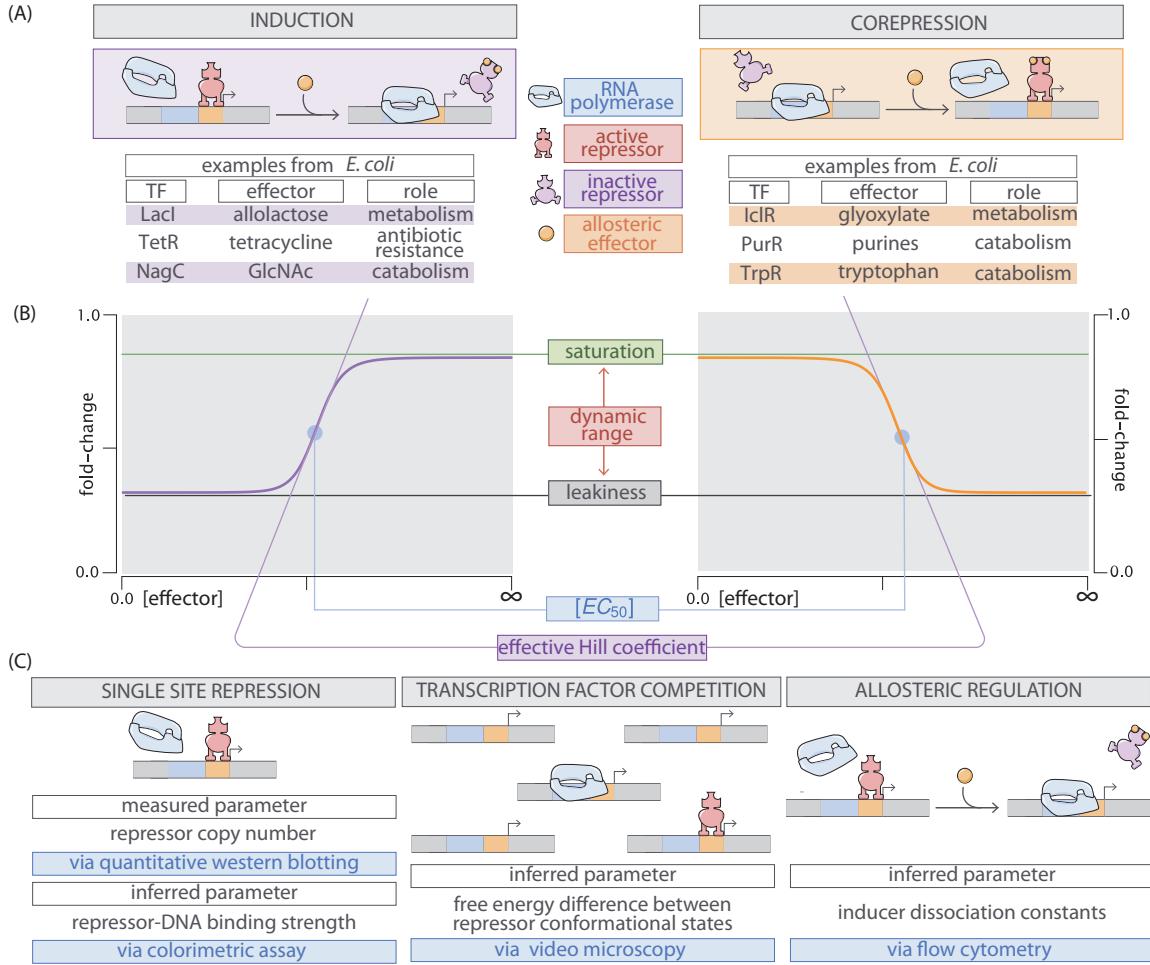
moter and resulting in increased expression. To test the predictions of our model across a wide range of operator binding strengths and repressor copy numbers, we design a genetic construct in *Escherichia coli* in which the binding probability of a repressor regulates gene expression of a fluorescent reporter.

In total, the work presented here demonstrates that one extremely compact set of parameters can be applied self-consistently and predictively to different regulatory situations including simple repression on the chromosome, cases in which decoy binding sites for repressor are put on plasmids, cases in which multiple genes compete for the same regulatory machinery, cases involving multiple binding sites for repressor leading to DNA looping, and induction by signaling (Boedicker et al., 2013a, 2013b; Brewster et al., 2012, 2014; Garcia and Phillips, 2011; Garcia et al., 2011a). Thus, rather than viewing the behavior of each circuit as giving rise to its own unique input-output response, the MWC model provides a means to characterize these seemingly diverse behaviors using a single unified framework governed by a small set of parameters.

### 2.3 Theoretical Model

#### **Inducible Transcriptional Repression Via The MWC Model of Allostery**

We begin by considering a simple repression genetic architecture in which the binding of an allosteric repressor occludes the binding of RNA polymerase (RNAP) to the DNA (Ackers and Johnson, 1982; Buchler et al., 2003). When an effector molecule (hereafter referred to as an “inducer” for the case of induction) binds to the repressor, it shifts the repressor’s allosteric equilibrium towards the inactive state as specified by the MWC model (Monod et al., 1965). This causes the repressor to bind more weakly to the operator, increasing the probability of RNAP binding the promoter which ultimately leads to gene expression. Simple repression motifs in the absence of inducer have been previously characterized by an equilibrium model where the probability of each state of repressor and RNAP promoter occupancy is dictated by the Boltzmann distribution (Ackers and Johnson, 1982; Bintu et al., 2005b; Brewster et al., 2014; Buchler et al., 2003; Garcia



**Figure 2.1: Transcriptional regulatory motifs involving an allosteric repressor.** (A) We consider a promoter regulated solely by an allosteric repressor in which the active (repressive, red blobs) state of the repressor is energetically favorable in the absence (induction, left panel) or presence (corepression, right panel) of an allosteric effector. Both inducible repression and corepression are ubiquitous regulatory strategies in *E. coli*, several examples of which are given in the tables below each panel. (B) A representative regulatory response (fold-change in gene expression) of the two architectures shown in Panel (A) as a function of the corresponding allosteric effector concentration. Properties of interest to this work are shown schematically upon the regulatory response. (C) Historical progression of thermodynamic modeling of the inducible simple-repression regulatory architecture. Garcia and Phillips (2011) used colorimetric assays and quantitative Western blots to investigate how single-site repression is modified by the repressor copy number and repressor-DNA binding energy. Brewster et al. (2014) used video microscopy to probe how the copy number of the promoter and presence of competing repressor binding sites affect gene expression. Building upon these works, we use flow cytometry to determine the inducer-repressor dissociation constants and demonstrate that with these parameters we can predict *a priori* the behavior of the system for any repressor copy number, DNA binding energy, gene copy number, and inducer concentration.

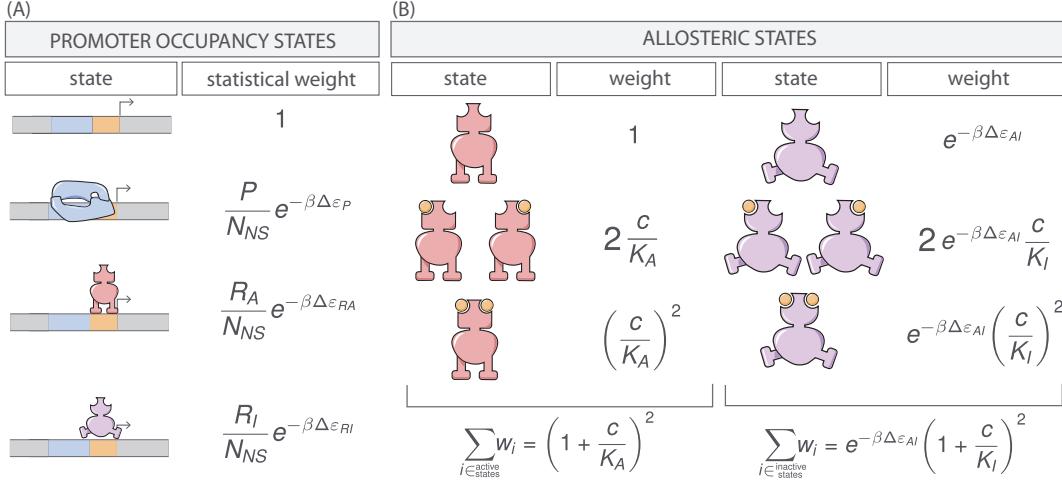
and Phillips, 2011; Vilar and Leibler, 2003) (we note that non-equilibrium models of simple repression have been shown to have the same functional form that we derive below (Phillips, 2015)). We extend these models to consider allostery by accounting for the equilibrium state of the repressor through the MWC model.

Thermodynamic models of gene expression begin by enumerating all possible states of the promoter and their corresponding statistical weights. As shown in Fig. 2.2 (A), the promoter can either be empty, occupied by RNAP, or occupied by either an active or inactive repressor. The probability of binding to the promoter will be affected by the protein copy number, which we denote as  $P$  for RNAP,  $R_A$  for active repressor, and  $R_I$  for inactive repressor. We note that repressors fluctuate between the active and inactive conformation in thermodynamic equilibrium, such that  $R_A$  and  $R_I$  will, on average, remain constant for a given inducer concentration (Monod et al., 1965). We assign the repressor a different DNA binding affinity in the active and inactive state. In addition to the specific binding sites at the promoter, we assume that there are  $N_{NS}$  non-specific binding sites elsewhere (i.e. on parts of the genome outside the simple repression architecture) where the RNAP or the repressor can bind. All specific binding energies are measured relative to the average non-specific binding energy. Thus,  $\Delta\epsilon_P$  represents the energy difference between the specific and non-specific binding for RNAP to the DNA. Likewise,  $\Delta\epsilon_{RA}$  and  $\Delta\epsilon_{RI}$  represent the difference in specific and non-specific binding energies for repressor in the active or inactive state, respectively.

Thermodynamic models of transcription (Ackers and Johnson, 1982; Bintu et al., 2005b, 2005a; Brewster et al., 2014; Buchler et al., 2003; Daber et al., 2011; Garcia and Phillips, 2011; Kuhlman et al., 2007; Vilar and Leibler, 2003; Weinert et al., 2014) posit that gene expression is proportional to the probability that the RNAP is bound to the promoter  $p_{\text{bound}}$ , which is given by

$$p_{\text{bound}} = \frac{\frac{P}{N_{NS}}e^{-\beta\Delta\epsilon_P}}{1 + \frac{R_A}{N_{NS}}e^{-\beta\Delta\epsilon_{RA}} + \frac{R_I}{N_{NS}}e^{-\beta\Delta\epsilon_{RI}} + \frac{P}{N_{NS}}e^{-\beta\Delta\epsilon_P}}, \quad (2.1)$$

with  $\beta = 1/k_B T$  where  $k_B$  is the Boltzmann constant and  $T$  is the temperature



**Figure 2.2: States and weights for the simple repression motif.** (A) Occupancy states of the promoter. RNAP (light blue) and a repressor compete for binding to a promoter of interest. There are  $R_A$  repressors in the active state (red) and  $R_I$  repressors in the inactive state (purple). The difference in energy between a repressor bound to the promoter of interest versus another non-specific site elsewhere on the DNA equals  $\Delta\varepsilon_{RA}$  in the active state and  $\Delta\varepsilon_{RI}$  in the inactive state; the P RNAP have a corresponding energy difference  $\Delta\varepsilon_P$  relative to non-specific binding on the DNA.  $N_{NS}$  represents the number of non-specific binding sites for both RNAP and repressor. (B) Allosteric states of the repressor. A repressor has an active conformation (red, left column) and an inactive conformation (purple, right column), with the energy difference between these two states given by  $\Delta\varepsilon_{AI}$ . The inducer (orange circle) at concentration  $c$  is capable of binding to the repressor with dissociation constants  $K_A$  in the active state and  $K_I$  in the inactive state. The eight states for a dimer with  $n = 2$  inducer binding sites are shown along with the sums of the active and inactive states.

of the system. As  $k_B T$  is the natural unit of energy at the molecular length scale, we treat the products  $\beta\Delta\varepsilon_j$  as single parameters within our model. Measuring  $p_{\text{bound}}$  directly is fraught with experimental difficulties, as determining the exact proportionality between expression and  $p_{\text{bound}}$  is not straightforward. Instead, we measure the fold-change in gene expression due to the presence of the repressor. We define fold-change as the ratio of gene expression in the presence of repressor relative to expression in the absence of repressor (i.e. constitutive expression), namely,

$$\text{fold-change} \equiv \frac{p_{\text{bound}}(R > 0)}{p_{\text{bound}}(R = 0)}. \quad (2.2)$$

We can simplify this expression using two well-justified approximations: (1)

$(P/N_{NS})e^{-\beta\Delta\varepsilon_P} \ll 1$  implying that the RNAP binds weakly to the promoter ( $N_{NS} = 4.6 \times 10^6$ ,  $P \approx 10^3$  (Klumpp and Hwa, 2008),  $\Delta\varepsilon_P \approx -2$  to  $-5 k_B T$  (Brewster et al., 2012), so that  $(P/N_{NS})e^{-\beta\Delta\varepsilon_P} \approx 0.01$ ) and (2)  $(R_I/N_{NS})e^{-\beta\Delta\varepsilon_{RI}} \ll 1 + (R_A/N_{NS})e^{-\beta\Delta\varepsilon_{RA}}$  which reflects our assumption that the inactive repressor binds weakly to the promoter of interest. Using these approximations, the fold-change reduces to the form

$$\text{fold-change} \approx \left(1 + \frac{R_A}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1} \equiv \left(1 + p_{\text{act}}(c)\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1}, \quad (2.3)$$

where in the last step we have introduced the fraction  $p_{\text{act}}(c)$  of repressors in the active state given a concentration  $c$  of inducer, such that  $R_A(c) = p_{\text{act}}(c)R$ . Since inducer binding shifts the repressors from the active to the inactive state,  $p_{\text{act}}(c)$  grows smaller as  $c$  increases.

We use the MWC model to compute the probability  $p_{\text{act}}(c)$  that a repressor with  $n$  inducer binding sites will be active. The value of  $p_{\text{act}}(c)$  is given by the sum of the weights of the active repressor states divided by the sum of the weights of all possible repressor states (see Fig. 2.2 (B)), namely,

$$p_{\text{act}}(c) = \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{-\beta\Delta\varepsilon_{AI}} \left(1 + \frac{c}{K_I}\right)^n}, \quad (2.4)$$

where  $K_A$  and  $K_I$  represent the dissociation constant between the inducer and repressor in the active and inactive states, respectively, and  $\Delta\varepsilon_{AI} = \varepsilon_I - \varepsilon_A$  is the free energy difference between a repressor in the inactive and active state (the quantity  $e^{-\beta\Delta\varepsilon_{AI}}$  is sometimes denoted by  $L$  (Marzen et al., 2013; Monod et al., 1965) or  $K_{RR*}$  (Daber et al., 2011)). In this equation,  $c/K_A$  and  $c/K_I$  represent the change in free energy when an inducer binds to a repressor in the active or inactive state, respectively, while  $e^{-\beta\Delta\varepsilon_{AI}}$  represents the change in free energy when the repressor changes from the active to inactive state in the absence of inducer. Thus, a repressor which favors the active state in the absence of inducer ( $\Delta\varepsilon_{AI} > 0$ ) will be driven towards the inactive state upon inducer binding when  $K_I < K_A$ . The specific case of a repressor dimer with  $n = 2$  inducer binding sites is shown in Fig. 2.2 (B).

Substituting  $p_{\text{act}}(c)$  from Eq. 2.4 into Eq. 2.3 yields the general formula for induction of a simple repression regulatory architecture (Phillips, 2015), namely,

$$\text{fold-change} = \left( 1 + \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{-\beta\Delta\varepsilon_{AI}} \left(1 + \frac{c}{K_I}\right)^n} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}} \right)^{-1}. \quad (2.5)$$

While we have used the specific case of simple repression with induction to craft this model, the same mathematics describe the case of corepression in which binding of an allosteric effector stabilizes the active state of the repressor and decreases gene expression (see Fig. 2.1). Interestingly, we shift from induction (governed by  $K_I < K_A$ ) to corepression ( $K_I > K_A$ ) as the ligand transitions from preferentially binding to the inactive repressor state to stabilizing the active state. Furthermore, this general approach can be used to describe a variety of other motifs such as activation, multiple repressor binding sites, and combinations of activator and repressor binding sites (Bintu et al., 2005b; Brewster et al., 2014; Weinert et al., 2014).

The formula presented in Eq. 2.5 enables us to make precise quantitative statements about induction profiles. Motivated by the broad range of predictions implied by Eq. 2.5, we designed a series of experiments using the *lac* system in *E. coli* to tune the control parameters for a simple repression genetic circuit. As discussed in Fig. 2.1 (C), previous studies have provided well-characterized values for many of the parameters in our experimental system, leaving only the values of the MWC parameters ( $K_A$ ,  $K_I$ , and  $\Delta\varepsilon_{AI}$ ) to be determined. We note that while previous studies have obtained values for  $K_A$ ,  $K_I$ , and  $L = e^{-\beta\Delta\varepsilon_{AI}}$  (Daber et al., 2011; O’Gorman et al., 1980), they were either based upon *in vitro* biochemical experiments or *in vivo* conditions involving poorly characterized transcription factor copy numbers and gene copy numbers. These differences relative to our experimental conditions and fitting techniques led us to believe that it was important to perform our own analysis of these parameters. After inferring these three MWC parameters (see the supplemental Chapter 6 for details regarding the inference of  $\Delta\varepsilon_{AI}$ , which was fitted separately from  $K_A$  and  $K_I$ ), we were able to predict the input/output response of the system under a broad range of experi-

mental conditions. For example, this framework can predict the response of the system at different repressor copy numbers  $R$ , repressor-operator affinities  $\Delta\varepsilon_{RA}$ , inducer concentrations  $c$ , and gene copy numbers.

## 2.4 Results

### Experimental Design

We test our model by predicting the induction profiles for an array of strains that could be made using previously characterized repressor copy numbers and DNA binding energies. Our approach contrasts with previous studies that have parameterized induction curves of simple repression motifs, as these have relied on expression systems where proteins are expressed from plasmids, resulting in highly variable and unconstrained copy numbers (Daber et al., 2009, 2011; Murphy et al., 2007; Sochor, 2014). Instead, our approach relies on a foundation of previous work as depicted in Fig. 2.1 (C). This includes work from our laboratory that used *E. coli* constructs based on components of the *lac* system to demonstrate how the Lac repressor (LacI) copy number  $R$  and operator binding energy  $\Delta\varepsilon_{RA}$  affect gene expression in the absence of inducer (Garcia and Phillips, 2011). Rydenfelt et al. (2014a) extended the theory used in that work to the case of multiple promoters competing for a given transcription factor, which was validated experimentally by Brewster et al. (2014), who modified this system to consider expression from multiple-copy plasmids as well as the presence of competing repressor binding sites.

The present study extends this body of work by introducing three additional biophysical parameters –  $\Delta\varepsilon_{AI}$ ,  $K_A$ , and  $K_I$  – which capture the allosteric nature of the transcription factor and complement the results shown by Garcia and Phillips (2011) and Brewster et al. (2014). Although the current work focuses on systems with a single site of repression, in the Materials & Methods, we utilize data from Brewster et al. (2014), in which multiple sites of repression are explored, to characterize the allosteric free energy difference  $\Delta\varepsilon_{AI}$  between the repressor's active and inactive states. This additional data set is critical because multiple degenerate

sets of parameters can characterize an induction curve equally well, with the  $\Delta\varepsilon_{AI}$  parameter compensated by the inducer dissociation constants  $K_A$  and  $K_I$  (see supplemental Chapter 6). After fixing  $\Delta\varepsilon_{AI}$  as described in the Materials & Methods, we can use data from single-site simple repression systems to determine the values of  $K_A$  and  $K_I$ .

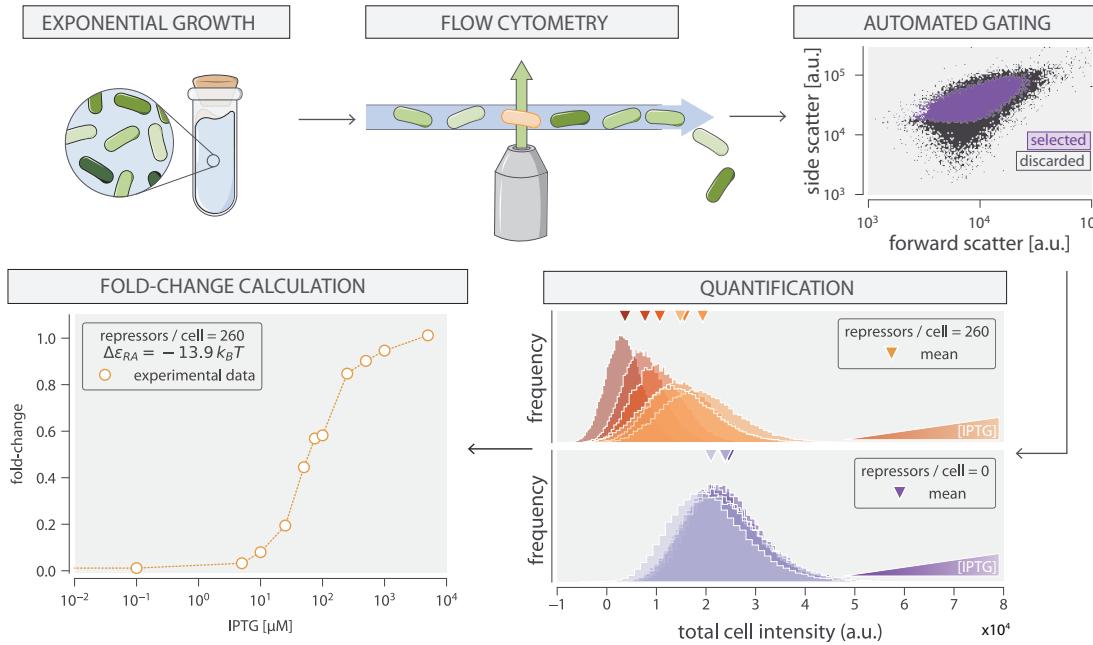
We determine the values of  $K_A$  and  $K_I$  by fitting to a single induction profile using Bayesian inferential methods (Sivia and Skilling, 2006). We then use Eq. 2.5 to predict gene expression for any concentration of inducer, repressor copy number, and DNA binding energy and compare these predictions against experimental measurements. To obtain induction profiles for a set of strains with varying repressor copy numbers, we used modified *lacI* ribosomal binding sites from Garcia and Phillips (2011) to generate strains with mean repressor copy number per cell of  $R = 22 \pm 4$ ,  $60 \pm 20$ ,  $124 \pm 30$ ,  $260 \pm 40$ ,  $1220 \pm 160$ , and  $1740 \pm 340$ , where the error denotes standard deviation of at least three replicates as measured by Garcia and Phillips (2011). We note that  $R$  refers to the number of repressor dimers in the cell, which is twice the number of repressor tetramers reported by Garcia and Phillips (2011); since both heads of the repressor are assumed to always be either specifically or non-specifically bound to the genome, the two repressor dimers in each LacI tetramer can be considered independently. Gene expression was measured using a Yellow Fluorescent Protein (YFP) gene, driven by a *lacUV5* promoter. Each of the six repressor copy number variants were paired with the native O1, O2, or O3 *lac* operator (Oehler et al., 1994) placed at the YFP transcription start site, thereby generating eighteen unique strains. The repressor-operator binding energies ( $O1 \Delta\varepsilon_{RA} = -15.3 \pm 0.2 k_B T$ ,  $O2 \Delta\varepsilon_{RA} = -13.9 k_B T \pm 0.2$ , and  $O3 \Delta\varepsilon_{RA} = -9.7 \pm 0.1 k_B T$ ) were previously inferred by measuring the fold-change of the *lac* system at different repressor copy numbers, where the error arises from model fitting (Garcia and Phillips, 2011). Additionally, we were able to obtain the value  $\Delta\varepsilon_{AI} = 4.5 k_B T$  by fitting to previous data as discussed in the Materials & Methods. We measure fold-change over a range of known IPTG concentrations  $c$ , using  $n = 2$  inducer binding sites per LacI dimer and approximating the num-

ber of non-specific binding sites as the length in base-pairs of the *E. coli* genome,  $N_{NS} = 4.6 \times 10^6$ .

Our experimental pipeline for determining fold-change using flow cytometry is shown in Fig. 2.3. Briefly, cells were grown to exponential phase, in which gene expression reaches steady state (Scott et al., 2010), under concentrations of the inducer IPTG ranging between 0 and 5000  $\mu\text{M}$ . We measure YFP fluorescence using flow cytometry and automatically gate the data to include only single-cell measurements (see Materials & Methods). To validate the use of flow cytometry, we also measured the fold-change of a subset of strains using the established method of single-cell microscopy (see supplemental Chapter 6). We found that the fold-change measurements obtained from microscopy were indistinguishable from that of flow-cytometry and yielded values for the inducer binding constants  $K_A$  and  $K_I$  that were within error.

### Determination of the *in vivo* MWC Parameters

The three parameters that we tune experimentally are shown in Fig. 2.4 (A), leaving the three allosteric parameters ( $\Delta\varepsilon_{AI}$ ,  $K_A$ , and  $K_I$ ) to be determined by fitting. We used previous LacI fold-change data (Brewster et al., 2014) to infer that  $\Delta\varepsilon_{AI} = 4.5 k_B T$  (see Materials & Methods). Rather than fitting  $K_A$  and  $K_I$  to our entire data set of eighteen unique constructs, we performed Bayesian parameter estimation on data from a single strain with  $R = 260$  and an O2 operator ( $\Delta\varepsilon_{RA} = -13.9 k_B T$  (Garcia and Phillips, 2011)) shown in Fig. 2.4(D, white-faced points). Using Markov Chain Monte Carlo, we determine the most likely parameter values to be  $K_A = 139^{+29}_{-22} \mu\text{M}$  and  $K_I = 0.53^{+0.04}_{-0.04} \mu\text{M}$ , which are the modes of their respective distributions, where the superscripts and subscripts represent the upper and lower bounds of the 95<sup>th</sup> percentile of the parameter value distributions (see Fig. 2.4 (B)). Unfortunately, we are not able to make a meaningful value-for-value comparison of our parameters to those of earlier studies (Daber et al., 2009, 2011) because of uncertainties in both gene copy number and transcription factor copy numbers in these studies, as illustrated in supplemental Chapter 6. We then



**Figure 2.3: An experimental pipeline for high-throughput fold-change measurements.** Cells are grown to exponential steady state and their fluorescence is measured using flow cytometry. Automatic gating methods using forward- and side-scattering are used to ensure that all measurements come from single cells (see Materials & Methods). Mean expression is then quantified at different IPTG concentrations (top, blue histograms) and for a strain without repressor (bottom, green histograms), which shows no response to IPTG as expected. Fold-change is computed by dividing the mean fluorescence in the presence of repressor by the mean fluorescence in the absence of repressor. The Python code (`ch2_fig3.py`) used to generate this figure can be found on the thesis GitHub repository.

predicted the fold-change for the remaining seventeen strains with no further fitting (see Fig. 2.4 (C – E)) together with the specific phenotypic properties described in Fig. 2.1(B) and discussed in detail below (see (Fig. 2.4 (F – J))). The shaded regions in Fig. 2.4 (C – E) denote the 95% credible regions. Factors determining the width of the credible regions are explored in the supplemental Chapter 6.

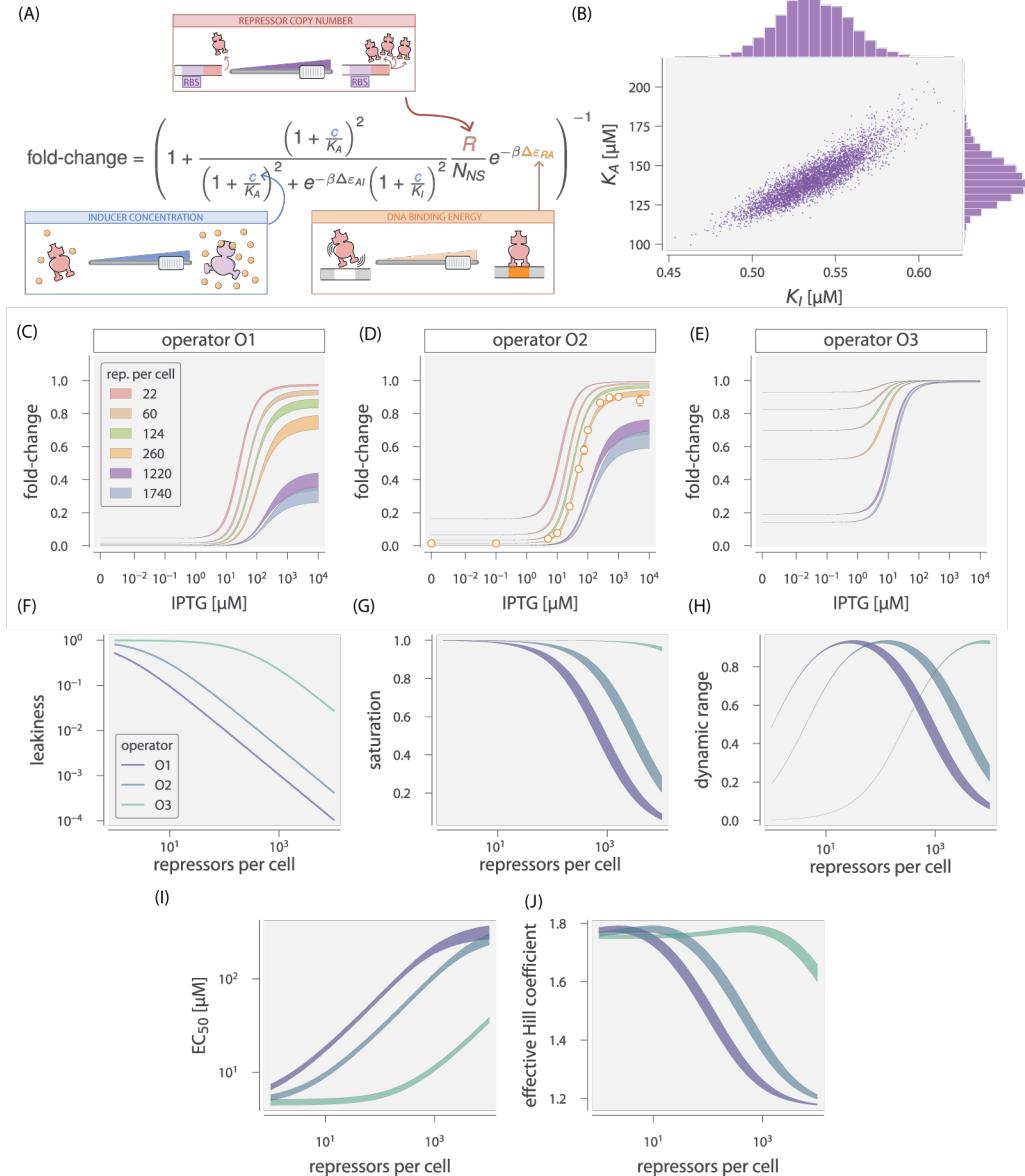
We stress that the entire suite of predictions in Fig. ?? (C – J) is based upon the induction profile of a single strain. Our ability to make such a broad range of predictions stems from the fact that our parameters of interest – such as the repressor copy number and DNA binding energy – appear as distinct physical parameters within our model. While the single data set in Fig. 2.4 could also be fit using a Hill

function, such an analysis would be unable to predict any of the other curves in the figure. Phenomenological expressions such as the Hill function can describe data, but lack predictive power and are thus unable to build our intuition, help us design *de novo* input-output functions, or guide future experiments (Kuhlman et al., 2007; Murphy et al., 2007).

### Comparison of Experimental Measurements with Theoretical Predictions

We tested the predictions shown in Fig. 2.4 by measuring fold-change induction profiles in strains with a broad range of repressor copy numbers and repressor binding energies as characterized in Garcia and Phillips (2011). With a few notable exceptions, the results shown in Fig. 2.5 demonstrate agreement between theory and experiment. We note that there was an apparently systematic shift in the O3  $\Delta\varepsilon_{RA} = -9.7 k_B T$  strains Fig. 2.5 and all of the  $R = 1220$  and  $R = 1740$  strains. This may be partially due to imprecise previous determinations of their  $\Delta\varepsilon_{RA}$  and  $R$  values. By performing a global fit where we infer all parameters including the repressor copy number  $R$  and the binding energy  $\Delta\varepsilon_{RA}$ , we found better agreement for these strains, although a discrepancy in the steepness of the response for all O3 strains remains (see supplemental Chapter 6). We considered a number of hypotheses to explain these discrepancies such as including other states (e.g. non-negligible binding of the inactive repressor), relaxing the weak promoter approximation, and accounting for variations in gene and repressor copy number throughout the cell cycle, but none explained the observed discrepancies. As an additional test of our model, we considered strains using the synthetic Oid operator which exhibits an especially strong binding energy of  $\Delta\varepsilon_{RA} = -17 k_B T$  (Garcia and Phillips, 2011). The global fit agrees well with the Oid microscopy data, though it asserts a stronger Oid binding energy of  $\Delta\varepsilon_{RA} = -17.7 k_B T$  (see supplemental Chapter 6).

To ensure that the agreement between our predictions and data is not an accident of the strain we used to perform our fitting, we also inferred  $K_A$  and  $K_I$  from each of the other strains. As discussed in supplemental Chapter 6 and Fig. 2.4, the



**Figure 2.4: Predicting induction profiles for different biological control parameters.** (A) Schematic representation of experimentally accessible variables. Repressor copy number  $R$  is tuned by changing the sequence of the ribosomal binding site (RBS), DNA binding energy  $\Delta\epsilon_{RA}$  is controlled via the sequence of the operator, and the inducer concentration  $c$  is controlled via a dilution series. (B) Markov Chain Monte Carlo (MCMC) sampling of the posterior distribution of  $K_A$  and  $K_I$ . Each point corresponds to a single MCMC sample. Distribution on top and right represent the marginal posterior probability distribution over  $K_A$  and  $K_I$ , respectively. (C) Predicted induction profiles for strains with various repressor copy numbers and DNA binding energies. White-faced points represent those to which the inducer binding constants  $K_A$  and  $K_I$  were determined. (D) Predicted properties of the induction profiles in (C) using parameter values known *a priori*. The shaded regions denote the 95% credible region. Region between 0 and  $10^{-2}$  μM is scaled linearly with log scaling elsewhere. The Python code (`ch2_fig4.py`) used to generate this figure can be found on the thesis GitHub repository.

inferred values of  $K_A$  and  $K_I$  depend minimally upon which strain is chosen, indicating that these parameter values are highly robust. We also performed a global fit using the data from all eighteen strains in which we fitted for the inducer dissociation constants  $K_A$  and  $K_I$ , the repressor copy number  $R$ , and the repressor DNA binding energy  $\Delta\epsilon_{RA}$  (see supplemental Chapter 6). The resulting parameter values were nearly identical to those fitted from any single strain. For the remainder of the text we continue using parameters inferred from the strain with  $R = 260$  repressors and an O2 operator.

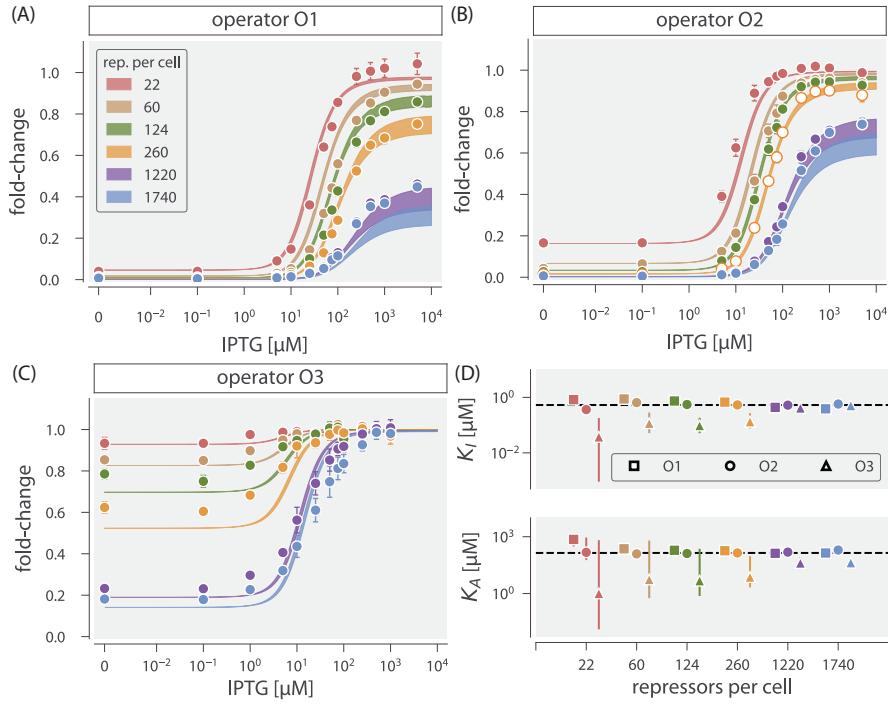
### Predicting the Phenotypic Traits of the Induction Response

The properties shown in Fig. 2.1 (i.e. the leakiness, saturation, dynamic range,  $[EC_{50}]$ , and effective Hill coefficient) are of significant interest to synthetic biology. For example, synthetic biology is often focused on generating large responses (i.e. a large dynamic range) or finding a strong binding partner (i.e. a small  $[EC_{50}]$ ) (Brophy and Voigt, 2014; Shis et al., 2014). While these properties are all individually informative, when taken together they capture the essential features of the induction response. We reiterate that a Hill function approach cannot predict these features *a priori* and furthermore requires fitting each curve individually. The MWC model, on the other hand, enables us to quantify how each trait depends upon a single set of physical parameters as shown by Fig. 2.4 (F – J).

We define these five phenotypic traits using expressions derived from the model presented in Eq. 2.5. These results build upon extensive work by Martins and Swain (2011), who computed many such properties for ligand-receptor binding within the MWC model. We begin by analyzing the leakiness, which is the minimum fold-change observed in the absence of ligand, given by

$$\text{leakiness} = \text{fold-change}(c = 0) = \left( 1 + \frac{1}{1 + e^{-\beta\Delta\epsilon_{AI}}} \frac{R}{N_{NS}} e^{-\beta\Delta\epsilon_{RA}} \right)^{-1}, \quad (2.6)$$

and the saturation, which is the maximum fold change observed in the presence of



**Figure 2.5: Comparison of predictions against measured and inferred data.** Flow cytometry measurements of fold-change over a range of IPTG concentrations for O1, O2, and O3 strains at varying repressor copy numbers, overlaid on the predicted responses. Error bars for the experimental data show the standard error of the mean (eight or more replicates). As discussed in Fig. 2.4, all of the predicted induction curves were generated prior to measurement by inferring the MWC parameters using a single data set (O2  $R = 260$ , shown by white circles in Panel (B)). The predictions may therefore depend upon which strain is used to infer the parameters. The inferred parameter values of the dissociation constants  $K_A$  and  $K_I$  using any of the eighteen strains instead of the O2  $R = 260$  strain. Nearly identical parameter values are inferred from each strain, demonstrating that the same set of induction profiles would have been predicted regardless of which strain was chosen. The points show the mode, and the error bars denote the 95% credible region of the parameter value distribution. Error bars not visible are smaller than the size of the marker. The Python code (`ch2_fig5.py`) used to generate this figure can be found on the thesis GitHub repository.

saturating ligand,

$$\begin{aligned} \text{saturation} &= \text{fold-change}(c \rightarrow \infty) \\ &= \left( 1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}} \left( \frac{K_A}{K_I} \right)^n} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}} \right)^{-1}. \end{aligned} \quad (2.7)$$

Systems that minimize leakiness repress strongly in the absence of effector while systems that maximize saturation have high expression in the presence of effector. Together, these two properties determine the dynamic range of a system's response, which is given by the difference

$$\text{dynamic range} = \text{saturation} - \text{leakiness}. \quad (2.8)$$

These three properties are shown in Fig. 2.4 (F-H). We discuss these properties in greater detail in supplemental Chapter 6. Fig. 2.6 shows that the measurements of these three properties, derived from the fold-change data in the absence of IPTG and the presence of saturating IPTG, closely match the predictions for all three operators.

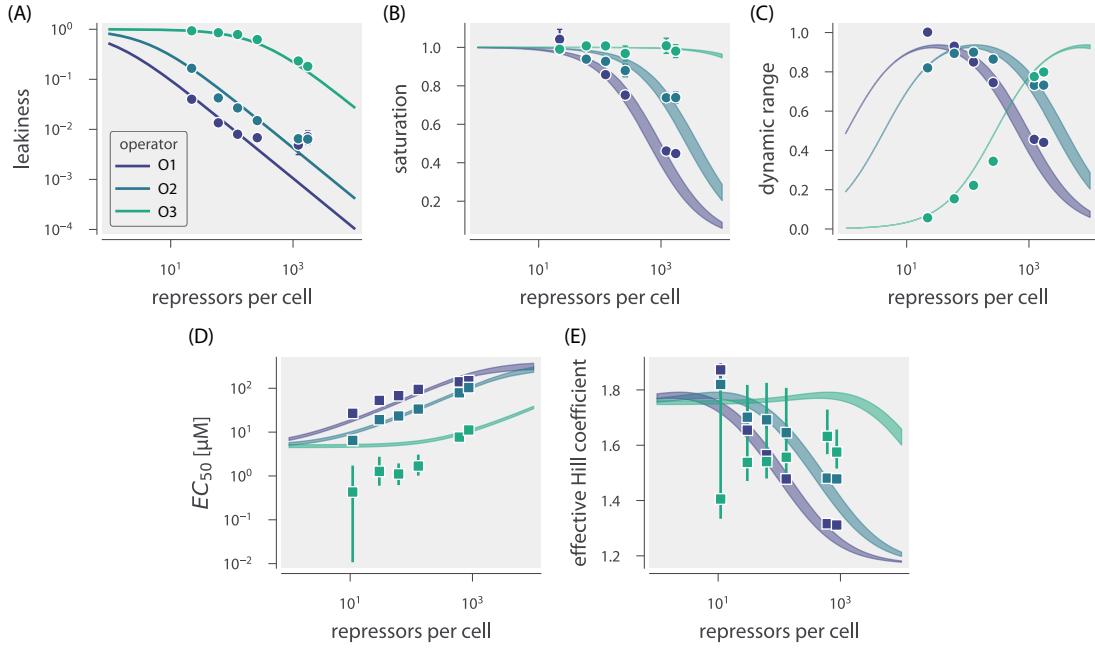
Two additional properties of induction profiles are the  $[EC_{50}]$  and effective Hill coefficient, which determine the range of inducer concentration in which the system's output goes from its minimum to maximum value. The  $[EC_{50}]$  denotes the inducer concentration required to generate a system response halfway between its minimum and maximum value,

$$\text{fold-change}(c = [EC_{50}]) = \frac{\text{leakiness} + \text{saturation}}{2}. \quad (2.9)$$

The effective Hill coefficient  $h$ , which quantifies the steepness of the curve at the  $[EC_{50}]$ , is given by

$$h = \left( 2 \frac{d}{d \log c} \left[ \log \left( \frac{\text{fold-change}(c) - \text{leakiness}}{\text{dynamic range}} \right) \right] \right)_{c=[EC_{50}]} . \quad (2.10)$$

Fig. 2.6 (D-E) shows how the  $[EC_{50}]$  and effective Hill coefficient depend on the repressor copy number. In supplemental Chapter 6, we discuss the analytic forms



**Figure 2.6: Predictions and experimental measurements of key properties of induction profiles.** Data for the leakiness, saturation, and dynamic range are obtained from fold-change measurements in the absence of IPTG and at saturating concentrations of IPTG. The three repressor-operator binding energies in the legend correspond to the O1 operator ( $-15.3 k_B T$ ), O2 operator ( $-13.9 k_B T$ ), and O3 operator ( $-9.7 k_B T$ ). Both the  $[EC_{50}]$  and effective Hill coefficient are inferred by individually fitting each operator-repressor pairing in Fig. 2.5 (C – E) separately to in order to smoothly interpolate between the data points. Error bars for (A – C) represent the standard error of the mean for eight or more replicates; error bars for (D – E) represent the 95% credible region for the parameter found by propagating the credible region of our estimates of  $K_A$  and  $K_I$  into Eq. 2.5. The Python code (ch2\_fig6.py) used to generate this figure can be found on the thesis GitHub repository.

of these two properties as well as their dependence on the repressor-DNA binding energy. Fig. 2.6 (D-E) shows the estimated values of the  $[EC_{50}]$  and the effective Hill coefficient overlaid on the theoretical predictions. Both properties were obtained by fitting to each individual titration curve and computing the  $[EC_{50}]$  and effective Hill coefficient. We find that the predictions made with the single strain closely match those made for each of the strains with O1 and O2 operators, but the predictions for the O3 operator are markedly off. In the supplemental Chapter 6, we show that the large, asymmetric error bars for the O3  $R = 22$  strain arise from

its nearly flat response, where the lack of dynamic range makes it impossible to determine the value of the inducer dissociation constants  $K_A$  and  $K_I$ , as can be seen in the uncertainty of both the  $[EC_{50}]$  and effective Hill coefficient. Discrepancies between theory and data for O3 are improved, but not fully resolved, by performing a global fit or fitting the MWC model individually to each curve (see supplemental Chapter 6). It remains an open question how to account for discrepancies in O3, in particular regarding the significant mismatch between the predicted and fitted effective Hill coefficients.

### Data Collapse of Induction Profiles

Our primary interest heretofore was to determine the system response at a specific inducer concentration, repressor copy number, and repressor-DNA binding energy. However, the cell does not necessarily “care about” the precise number of repressors in the system or the binding energy of an individual operator. The relevant quantity for cellular function is the fold-change enacted by the regulatory system. This raises the question: given a specific value of the fold-change, what combination of parameters will give rise to this desired response? In other words, what trade-offs between the parameters of the system will give rise to the same mean cellular output? These are key questions both for understanding how the system is governed and, as will become evident in the following chapters of this dissertation, can provide insight as to what parameters may be changing in response to a physiological or environmental perturbation. To address these questions, we follow the data collapse strategy used in a number of previous studies (Keymer et al., 2006; Sourjik and Berg, 2002; Swem et al., 2008).

The equilibrium states and statistical weights outlined in Fig. 2.2 (A) can be further coarse grained into two possible states – one state being where the promoter is occupied by the repressor and another being where the promoter is *not* occupied by the repressor (Fig. 2.7 (A)). As the transcriptionally active state and the states in which the repressor is bound are mutually exclusive, we can compute the

probability of the repressor not being bound  $p_{\neg r}$  to the promoter as

$$p_{\neg r} = \frac{\neg r}{r + \neg r}. \quad (2.11)$$

We can now take a similar approach as in Eq. 2.2 and define the fold-change as the probability of the repressor not being bound when repressor is expressed  $p_{\neg r}(R > 0)$  relative to the probability when no repressor is expressed  $p_{\neg r}(R = 0)$ . As the later term is equal to 1, the fold-change in gene expression is directly equivalent to  $p_{\neg r}$  expressed in Eq. 2.11. This form can be algebraically manipulated to the form

$$\text{fold-change} = \frac{1}{1 + \frac{r}{\neg r}} = \frac{1}{1 + e^{-\beta F}} \quad (2.12)$$

where  $F$  can be interpreted as the difference in free energy between the repressor bound and repressor not bound states,

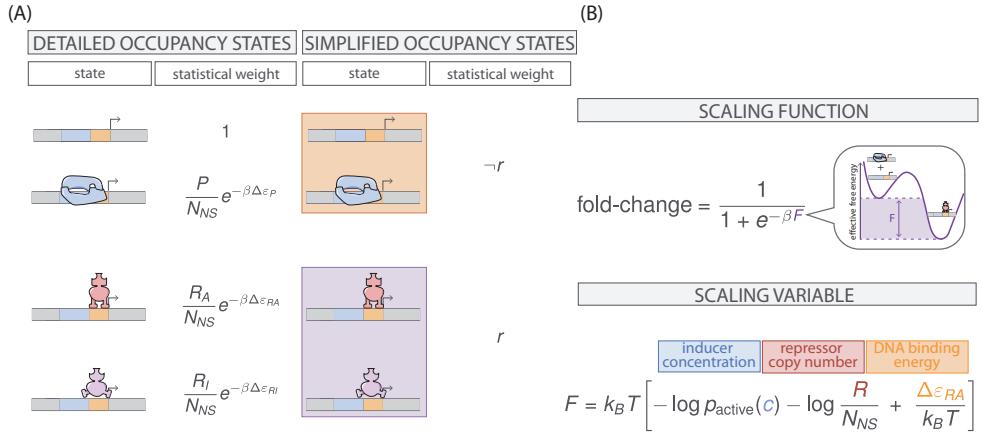
$$F = k_B T [\log \neg r - \log r]. \quad (2.13)$$

As Fig. 2.2 provides mathematical forms for  $r$  and  $\neg r$ ,  $F$  can be directly computed as

$$F = \frac{\Delta \varepsilon_{RA}}{k_B T} - \log \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{-\beta \Delta \varepsilon_{AI}} \left(1 + \frac{c}{K_I}\right)^n} - \log \frac{R}{N_{NS}}. \quad (2.14)$$

The first term in  $F$  denotes the repressor-operator binding energy, the second the contribution from the inducer concentration, and the last the effect of the repressor copy number. We note that elsewhere, this free energy has been dubbed the Bohr parameter since such families of curves are analogous to the shifts in hemoglobin binding curves at different pH known as the Bohr effect (Einav et al., 2016; Mirny, 2010; Phillips, 2015).

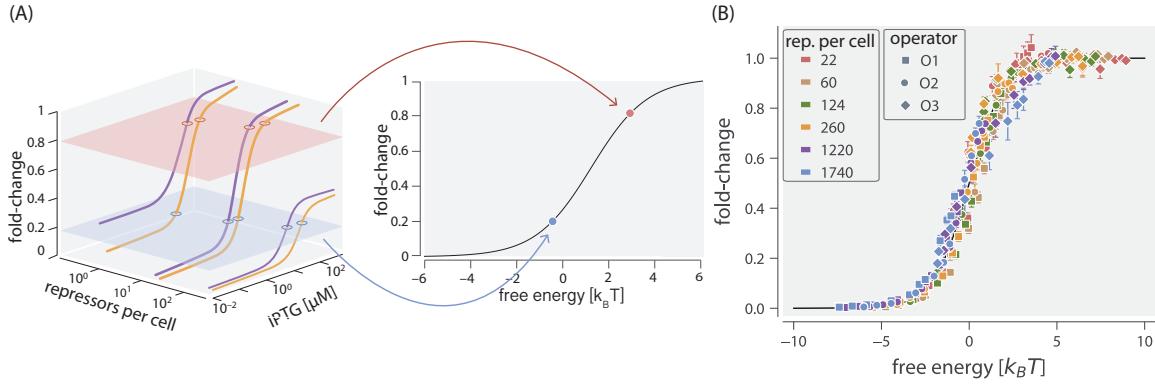
Instead of analyzing each induction curve individually, the free energy provides a natural means to simultaneously characterize the diversity in our eighteen induction profiles. Fig. 2.8 (A) demonstrates how the various induction curves from Fig. 2.4 (C-E) all collapse onto a single master curve, where points from every induction profile that yield the same fold-change are mapped onto the same free



**Figure 2.7: Coarse graining of promoter occupancy states to a two-state system.** (A) The promoter occupancy states shown in Fig. 2.2(A) can be further reduced to a two-state system; one in which the repressor is bound to the promoter ( $r$ ) and one in which it is not ( $\neg r$ ). (B) The fold-change in gene expression can then be evaluated as the probability of the repressor unbound state  $\neg r$  which has the form of a Fermi function (top). The energetic parameter  $F$  denotes the effective free energy difference between the repressor bound and unbound states and can be directly computed (bottom) using the statistical weights in Fig. 2.2.

energy. Fig. 2.8 (B) reveals complete data collapse for the 216 data points in Fig. 2.5 (A – C), demonstrating the close match between the theoretical predictions and experimental measurements across all eighteen strains.

There are many different combinations of parameter values that can result in the same free energy as defined in Eq. 2.14. For example, suppose a system originally has a fold-change of 0.2 at a specific inducer concentration, and then operator mutations increase the  $\Delta \varepsilon_{RA}$  binding energy (Garcia and Phillips, 2011). While this serves to initially increase both the free energy and the fold-change, a subsequent increase in the repressor copy number could bring the cell back to the original fold-change level. Such trade-offs hint that there need not be a single set of parameters that evoke a specific cellular response, but rather that the cell explores a large but degenerate space of parameters with multiple, equally valid paths.



**Figure 2.8: Collapse of fold-change measurements as a function of the free energy.** (A) Any combination of parameters can be mapped to a single physiological response (i.e. fold-change) via the free energy, which encompasses the parametric details of the model. (B) Experimental data from Fig. 2.5 collapse onto a single master curve as a function of the free energy. The free energy for each strain was calculated from Eq. 2.14. using  $n = 2$ ,  $\Delta\epsilon_{AI} = 4.5 k_B T$ ,  $K_A = 139 \mu\text{M}$ ,  $K_I = 0.53 \mu\text{M}$ , and the strain-specific  $R$  and  $\Delta\epsilon_{RA}$ . All data points represent the mean, and error bars are the standard error of the mean for eight or more replicates. The Python code (`ch2_fig8.py`) used to generate this figure can be found on the thesis GitHub repository.

## 2.5 Discussion

Since the early work by Monod, Wyman, and Changeux (Monod et al., 1963, 1965), an array of biological phenomena has been tied to the existence of macromolecules that switch between inactive and active states. Examples can be found in a wide variety of cellular processes, including ligand-gated ion channels (Auerbach, 2012), enzymatic reactions (Einav et al., 2016; Velyvis et al., 2007), chemotaxis (Keymer et al., 2006), quorum sensing (Swem et al., 2008), G-protein coupled receptors (Canals et al., 2012), physiologically important proteins (Levantino et al., 2012; Milo et al., 2007), and beyond. One of the most ubiquitous examples of allosteric regulation is in the context of gene expression, where an array of molecular players bind to transcription factors to influence their ability to regulate gene activity (Li et al., 2014). A number of studies have focused on developing a quantitative understanding of allosteric regulatory systems. The work of Martins and Swain (2011) and Marzen et al. (2013) analytically derives fundamental properties of the MWC model, including the leakiness and dynamic range described in this work, noting

the inherent trade-offs in these properties when tuning the model’s parameters. Work in the Church and Voigt labs, among others, has expanded on the availability of allosteric circuits for synthetic biology (Lutz and Bujard, 1997; Moon et al., 2012; Rogers et al., 2015; Rohlhill et al., 2017). Somewhat recently, Daber et al. (2009) theoretically explored the induction of simple repression within the MWC model and experimentally measured how mutations alter the induction profiles of transcription factors Daber et al. (2011). Vilar and Saiz (2013) analyzed a variety of interactions in inducible *lac*-based systems including the effects of oligomerization and DNA folding on transcription factor induction. Other work has attempted to use the *lac* system to reconcile *in vitro* and *in vivo* measurements (Tungtur et al., 2011).

Although this body of work has done much to improve our understanding of allosteric transcription factors, there have been few attempts to explicitly connect quantitative models to experiments. Here, we generate a predictive model of allosteric transcriptional regulation and then test the model against a thorough set of experiments using well-characterized regulatory components. Specifically, we used the MWC model to build upon a well-established thermodynamic model of transcriptional regulation (Bintu et al., 2005b; Garcia and Phillips, 2011), allowing us to compose the model from a minimal set of biologically meaningful parameters. This model combines both theoretical and experimental insights; for example, rather than considering gene expression directly we analyze the fold-change in expression, where the weak promoter approximation circumvents uncertainty in the RNAP copy number. The resulting model depended upon experimentally accessible parameters, namely, the repressor copy number, the repressor-DNA binding energy, and the concentration of inducer. We tested these predictions on a range of strains whose repressor copy number spanned two orders of magnitude and whose DNA binding affinity spanned  $6 k_B T$ . We argue that one would not be able to generate such a wide array of predictions by using a Hill function, which abstracts away the biophysical meaning of the parameters into phenomenological parameters (Forsén and Linse, 1995). Furthermore, our model reveals system-

atic relationships between behaviors that previously were only determined empirically.

One such property is the dynamic range, which is of considerable interest when designing or characterizing a genetic circuit, is revealed to have an interesting property: although changing the value of  $\Delta\varepsilon_{RA}$  causes the dynamic range curves to shift to the right or left, each curve has the same shape and in particular the same maximum value. This means that strains with strong or weak binding energies can attain the same dynamic range when the value of  $R$  is tuned to compensate for the binding energy. This feature is not immediately apparent from the IPTG induction curves, which show very low dynamic ranges for several of the O1 and O3 strains. Without the benefit of models that can predict such phenotypic traits, efforts to engineer genetic circuits with allosteric transcription factors must rely on trial and error to achieve specific responses (Rogers et al., 2015). Other calculable properties, such as leakiness, saturation,  $[EC_{50}]$  and the effective Hill coefficient, agree well with experimental measurement. One exception is the titration profile of the weakest operator, O3. While performing a global fit for all model parameters marginally improves the prediction of all properties for O3 (see supplemental Chapter 6), a noticeable difference remains when inferring the effective Hill coefficient or the  $[EC_{50}]$ . We further tried including additional states (such as allowing the inactive repressor to bind to the operator), relaxing the weak promoter approximation, accounting for changes in gene and repressor copy number throughout the cell cycle (Jones et al., 2014), and refitting the original binding energies from Garcia and Phillips (2011), but such generalizations were unable to account for the O3 data. It remains an open question as to how the discrepancy between the theory and measurements for O3 can be reconciled.

Despite the diversity observed in the induction profiles of each of our strains, our data are unified by their reliance on fundamental biophysical parameters. In particular, we have shown that our model for fold-change can be rewritten in terms of the free energy, which encompasses all of the physical parameters of the system.

This has proven to be an illuminating technique in a number of studies of allosteric proteins (Keymer et al., 2006; Sourjik and Berg, 2002; Swem et al., 2008). Although it is experimentally straightforward to observe system responses to changes in effector concentration  $c$ , framing the input-output function in terms of  $c$  can give the misleading impression that changes in system parameters lead to fundamentally altered system responses. Alternatively, if one can find the “natural variable” that enables the output to collapse onto a single curve, it becomes clear that the system’s output is not governed by individual system parameters, but rather the contributions of multiple parameters that define the natural variable. When our fold-change data are plotted against the respective free energies for each construct, they collapse cleanly onto a single curve (see Fig. 2.8). This enables us to analyze how parameters can compensate each other. For example, rather than viewing strong repression as a consequence of low IPTG concentration  $c$  or high repressor copy number  $R$ , we can now observe that strong repression is achieved when the free energy  $F(c) \leq -5k_B T$ , a condition which can be reached in a number of ways.

While our experiments validated the theoretical predictions in the case of simple repression, we expect the framework presented here to apply much more generally to different biological instances of allosteric regulation. For example, we can use this model to study more complex systems such as when transcription factors interact with multiple operators (Bintu et al., 2005b). We can further explore different regulatory configurations such as corepression, activation, and coactivation, each of which are found in *E. coli*. This work can also serve as a springboard to characterize not just the mean but the full gene expression distribution and thus quantify the impact of noise on the system (Eldar and Elowitz, 2010). Another extension of this approach would be to theoretically predict and experimentally verify whether the repressor-inducer dissociation constants  $K_A$  and  $K_I$  or the energy difference  $\Delta\epsilon_{AI}$  between the allosteric states can be tuned by making single amino acid substitutions in the transcription factor (Daber et al., 2009; Phillips, 2015). Finally, we expect that the kind of rigorous quantitative description of the allosteric phenomenon provided here will make it possible to construct biophys-

ical models of fitness for allosteric proteins similar to those already invoked to explore the fitness effects of transcription factor binding site strengths and protein stability (Berg et al., 2004; Gerland et al., 2002; Zeldovich and Shakhnovich, 2008). In total, these results show that a thermodynamic formulation of the MWC model supersedes phenomenological fitting functions for understanding transcriptional regulation by allosteric proteins.

## 2.6 Materials & Methods

### Bacterial Strains and DNA Constructs

All strains used in these experiments were derived from *E. coli* K12 MG1655 with the *lac* operon removed, adapted from those created and described in Garcia and Phillips (2011). Briefly, the operator variants and YFP reporter gene were cloned into a pZS25 background which contains a *lacUV5* promoter that drives expression as is shown schematically in Fig. 2.2. These constructs carried a kanamycin resistance gene and were integrated into the *galK* locus of the chromosome using  $\lambda$  Red recombineering Sharan et al. (2009). The *lacI* gene was constitutively expressed via a  $P_{LtetO-1}$  promoter (Lutz and Bujard, 1997), with ribosomal binding site mutations made to vary the LacI copy number as described in Salis et al. (2009) using site-directed mutagenesis (Quickchange II; Stratagene), with further details in Garcia and Phillips (2011). These *lacI* constructs carried a chloramphenicol resistance gene and were integrated into the *ybcN* locus of the chromosome. Final strain construction was achieved by performing repeated P1 transduction (Thomason et al., 2007) of the different operator and *lacI* constructs to generate each combination used in this work. Integration was confirmed by PCR amplification of the replaced chromosomal region and by sequencing. Primers and final strain genotypes are listed in supplemental Chapter 6.

It is important to note that the rest of the *lac* operon (*lacZYA*) was never expressed. The LacY protein is a transmembrane protein which actively transports lactose as well as IPTG into the cell. As LacY was never produced in our strains, we assume that the extracellular and intracellular IPTG concentration was approx-

imately equal due to diffusion across the membrane into the cell as is suggested by previous work (Fernández-Castané et al., 2012).

To make this theory applicable to transcription factors with any number of DNA binding domains, we used a different definition for repressor copy number than has been used previously. We define the LacI copy number as the average number of repressor dimers per cell whereas in Garcia and Phillips (2011), the copy number is defined as the average number of repressor tetramers in each cell. To motivate this decision, we consider the fact that the LacI repressor molecule exists as a tetramer in *E. coli* (Lewis et al., 1996) in which a single DNA binding domain is formed from dimerization of LacI proteins, so that wild-type LacI might be described as dimer of dimers. Since each dimer is allosterically independent (i.e. either dimer can be allosterically active or inactive, independent of the configuration of the other dimer) (Daber et al., 2009), a single LacI tetramer can be treated as two functional repressors. Therefore, we have simply multiplied the number of repressors reported in Garcia and Phillips (2011) by a factor of two.

A subset of strains in these experiments were measured using fluorescence microscopy for validation of the flow cytometry data and results. To aid in the high-fidelity segmentation of individual cells, the strains were modified to constitutively express an mCherry fluorophore. This reporter was cloned into a pZS4\*1 backbone (Lutz and Bujard, 1997) in which mCherry is driven by the *lacUV5* promoter. All microscopy and flow cytometry experiments were performed using these strains.

### Growth Conditions for Flow Cytometry Measurements

All measurements were performed with *E. coli* cells grown to mid-exponential phase in standard M9 minimal media (M9 5X Salts, Sigma-Aldrich M6030; 2 mM magnesium sulfate, Mallinckrodt Chemicals 6066-04; 100  $\mu$ M calcium chloride, Fisher Chemicals C79-500) supplemented with 0.5% (w/v) glucose. Briefly, 500  $\mu$ L cultures of *E. coli* were inoculated into Lysogeny Broth (LB Miller Powder, BD Medical) from a 50% glycerol frozen stock (-80°C) and were grown overnight in

a 2 mL 96-deep-well plate sealed with a breathable nylon cover (Lab Pak - Nitex Nylon, Sefar America Inc. Cat. No. 241205) with rapid agitation for proper aeration. After approximately 12 to 15 hours, the cultures had reached saturation and were diluted 1000-fold into a second 2 mL 96-deep-well plate where each well contained 500  $\mu$ L of M9 minimal media supplemented with 0.5% w/v glucose (anhydrous D-Glucose, Macron Chemicals) and the appropriate concentration of IPTG (Isopropyl  $\beta$ -D-1 thiogalactopyranoside Dioxane Free, Research Products International). These were sealed with a breathable cover and were allowed to grow for approximately eight hours. Cells were then diluted ten-fold into a round-bottom 96-well plate (Corning Cat. No. 3365) containing 90  $\mu$ L of M9 minimal media supplemented with 0.5% w/v glucose along with the corresponding IPTG concentrations. For each IPTG concentration, a stock of 100-fold concentrated IPTG in double distilled water was prepared and partitioned into 100  $\mu$ L aliquots. The same parent stock was used for all experiments described in this work.

### Flow Cytometry

All fold-change measurements were collected on a Miltenyi Biotec MACSquant Analyzer 10 Flow Cytometer graciously provided by the Pamela Björkman lab at Caltech. Detailed information regarding the voltage settings of the photo-multiplier detectors can be found in the supplemental Chapter 6.

Prior to each day's experiments, the analyzer was calibrated using MACSQuant Calibration Beads (Cat. No. 130-093-607) such that day-to-day experiments would be comparable. All YFP fluorescence measurements were collected via 488 nm laser excitation coupled with a 525/50 nm emission filter. Unless otherwise specified, all measurements were taken over the course of two to three hours using automated sampling from a 96-well plate kept at approximately 4° - 10°C on a MACS Chill 96 Rack (Cat. No. 130-094-459). Cells were diluted to a final concentration of approximately  $4 \times 10^4$  cells per  $\mu$ L which corresponded to a flow rate of 2,000-6,000 measurements per second, and acquisition for each well was halted after 100,000 events were detected. Once completed, the data were extracted and immediately

processed using the following methods.

### **Unsupervised Gating of Flow Cytometry Data**

Flow cytometry data will frequently include a number of spurious events or other undesirable data points such as cell doublets and debris. The process of restricting the collected data set to those data determined to be “real” is commonly referred to as gating. These gates are typically drawn manually and restrict the data set to those points which display a high degree of linear correlation between their forward-scatter (FSC) and side-scatter (SSC). The development of unbiased and unsupervised methods of drawing these gates is an active area of research (Aghaeepour et al., 2013; Lo et al., 2008). For our purposes, we assume that the fluorescence level of the population should be log-normally distributed about some mean value. With this assumption in place, we developed a method that allows us to restrict the data used to compute the mean fluorescence intensity of the population to the smallest two-dimensional region of the log(FSC) vs. log(SSC) space in which 40% of the data is found. This was performed by fitting a bivariate Gaussian distribution and restricting the data used for calculation to those that reside within the 40th percentile. This procedure is described in more detail in the supplemental Chapter 6.

### **Experimental Determination of Fold-Change**

For each strain and IPTG concentration, the fold-change in gene expression was calculated by taking the ratio of the population mean YFP expression in the presence of LacI repressor to that of the population mean in the absence of LacI repressor. However, the measured fluorescence intensity of each cell also includes the autofluorescence contributed by the weak excitation of the myriad protein and small molecules within the cell. To correct for this background, we computed the fold change as

$$\text{fold-change} = \frac{\langle I_{R>0} \rangle - \langle I_{\text{auto}} \rangle}{\langle I_{R=0} \rangle - \langle I_{\text{auto}} \rangle}, \quad (2.15)$$

where  $\langle I_{R>0} \rangle$  is the average cell YFP intensity in the presence of repressor,  $\langle I_{R=0} \rangle$  is the average cell YFP intensity in the absence of repressor, and  $\langle I_{\text{auto}} \rangle$  is the average cell autofluorescence intensity, as measured from cells that lack the *lac*-YFP construct.

### Bayesian Parameter Estimation

In this work, we determine the most likely parameter values for the inducer dissociation constants  $K_A$  and  $K_I$  of the active and inactive state, respectively, using Bayesian methods. We compute the probability distribution of the value of each parameter given the data  $D$ , which by Bayes' theorem is given by

$$P(K_A, K_I | D) = \frac{P(D | K_A, K_I)P(K_A, K_I)}{P(D)}, \quad (2.16)$$

where  $D$  is all the data composed of independent variables (repressor copy number  $R$ , repressor-DNA binding energy  $\Delta\varepsilon_{RA}$ , and inducer concentration  $c$ ) and one dependent variable (experimental fold-change).  $P(D | K_A, K_I)$  is the likelihood of having observed the data given the parameter values for the dissociation constants,  $P(K_A, K_I)$  contains all the prior information on these parameters, and  $P(D)$  serves as a normalization constant, which we can ignore in our parameter estimation. Eq. 2.5 assumes a deterministic relationship between the parameters and the data, so in order to construct a probabilistic relationship as required by Eq. 2.16, we assume that the experimental fold-change for the  $i^{\text{th}}$  datum given the parameters is of the form

$$\text{fold-change}_{\text{exp}}^{(i)} = \left( 1 + \frac{\left( 1 + \frac{c^{(i)}}{K_A} \right)^2}{\left( 1 + \frac{c^{(i)}}{K_A} \right)^2 + e^{-\beta\Delta\varepsilon_{AI}} \left( 1 + \frac{c^{(i)}}{K_I} \right)^2} \frac{R^{(i)}}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}^{(i)}} \right)^{-1} + \epsilon^{(i)}, \quad (2.17)$$

where  $\epsilon^{(i)}$  represents the departure from the deterministic theoretical prediction for the  $i^{\text{th}}$  data point. If we assume that these  $\epsilon^{(i)}$  errors are normally distributed with mean zero and standard deviation  $\sigma$ , the likelihood of the data given the

parameters is of the form

$$P(D|K_A, K_I, \sigma) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \prod_{i=1}^n \exp \left[ -\frac{(\text{fold-change}_{\text{exp}}^{(i)} - \text{fold-change}(K_A, K_I, R^{(i)}, \Delta\varepsilon_{RA}^{(i)}, c^{(i)}))^2}{2\sigma^2} \right], \quad (2.18)$$

where  $\text{fold-change}_{\text{exp}}^{(i)}$  is the experimental fold-change and  $\text{fold-change}(\dots)$  is the theoretical prediction. The product  $\prod_{i=1}^n$  captures the assumption that the  $n$  data points are independent. Note that the likelihood and prior terms now include the extra unknown parameter  $\sigma$ . In applying Eq. 2.18, a choice of  $K_A$  and  $K_I$  that provides better agreement between theoretical fold-change predictions and experimental measurements will result in a more probable likelihood.

Both mathematically and numerically, it is convenient to define  $\tilde{k}_A = -\log \frac{K_A}{1\mu\text{M}}$  and  $\tilde{k}_I = -\log \frac{K_I}{1\mu\text{M}}$  and fit for these parameters on a log scale. Dissociation constants are scale invariant, so that a change from  $10\mu\text{M}$  to  $1\mu\text{M}$  leads to an equivalent increase in affinity as a change from  $1\mu\text{M}$  to  $0.1\mu\text{M}$ . With these definitions we assume for the prior  $P(\tilde{k}_A, \tilde{k}_I, \sigma)$  that all three parameters are independent. In addition, we assume a uniform distribution for  $\tilde{k}_A$  and  $\tilde{k}_I$  and a Jeffreys prior for the scale parameter  $\sigma$ . This yields the complete prior

$$P(\tilde{k}_A, \tilde{k}_I, \sigma) \equiv \frac{1}{(\tilde{k}_A^{\max} - \tilde{k}_A^{\min})} \frac{1}{(\tilde{k}_I^{\max} - \tilde{k}_I^{\min})} \frac{1}{\sigma}. \quad (2.19)$$

These priors are maximally uninformative meaning that they imply no prior knowledge of the parameter values. We defined the  $\tilde{k}_A$  and  $\tilde{k}_I$  ranges uniform on the range of  $-7$  to  $7$ , although we note that this particular choice does not affect the outcome provided the chosen range is sufficiently wide.

Putting all these terms together we can now sample from  $P(\tilde{k}_A, \tilde{k}_I, \sigma | D)$  using Markov chain Monte Carlo to compute the most likely parameter as well as the error bars (given by the 95% credible region) for  $K_A$  and  $K_I$ .

## Data Curation

All of the data used in this work as well as all relevant code can be found at the dedicated paper website. Data were collected, stored, and preserved using the Git

version control software in combination with off-site storage and hosting website GitHub. Code used to generate all figures and complete all processing step as and analyses are available on the GitHub repository. Many analysis files are stored as instructive Jupyter Notebooks. The scientific community is invited to fork our repositories and open constructive issues on the GitHub repository.

### Chapter 3

## UNKNOWN KNOWNS, KNOWN UNKNOWNS, AND UNFORSEEN CONSEQUENCES: USING FREE ENERGY SHIFTS TO PREDICT MUTANT PHENOTYPES

A version of this chapter originally appeared as Chure, G; Razo-Mejia, M., Belliveau, N.M.; Kaczmarek, Zofii A.; Einav, T.; Barnes, Stephanie L.; Lewis, M., and Phillips, R. (2019). *Predictive Shifts in Free Energy Couple Mutations to Their Phenotypic Consequences*. PNAS 116(37) DOI: <https://doi.org/10.1073/pnas.1907869116>. G.C., M.R.M, N.M.B., Z.A.K., and S.L.B designed the experiments and collected and analyzed data. G.C. developed theoretical treatment of free energy shifts. G.C., M.R.M, N.M.B., Z.A.K., T.E., S.L.B., and R.P. designed the research project. G.C. and R.P. wrote the paper. M.L. provided guidance and advice.

### 3.1 Abstract

Mutation is a critical mechanism by which evolution explores the functional landscape of proteins. Despite our ability to experimentally inflict mutations at will, it remains difficult to link sequence-level perturbations to systems-level responses. Here, we present a framework centered on measuring changes in the free energy of the system to link individual mutations in an allosteric transcriptional repressor to the parameters which govern its response. We find that the energetic effects of the mutations can be categorized into several classes which have characteristic curves as a function of the inducer concentration. We experimentally test these diagnostic predictions using the well-characterized LacI repressor of *Escherichia coli*, probing several mutations in the DNA binding and inducer binding domains. We find that the change in gene expression due to a point mutation can be captured by modifying only the model parameters that describe the respective domain of the wild-type protein. These parameters appear to be insulated, with mutations in the DNA binding domain altering only the DNA affinity and those in the inducer

binding domain altering only the allosteric parameters. Changing these subsets of parameters tunes the free energy of the system in a way that is concordant with theoretical expectations. Finally, we show that the induction profiles and resulting free energies associated with pairwise double mutants can be predicted with quantitative accuracy given knowledge of the single mutants, providing an avenue for identifying and quantifying epistatic interactions.

### 3.2 Introduction

Thermodynamic treatments of transcriptional regulation have been fruitful in their ability to generate quantitative predictions of gene expression as a function of a minimal set of physically meaningful parameters (Ackers and Johnson, 1982; Bintu et al., 2005b, 2005a; Brewster et al., 2014; Buchler et al., 2003; Daber et al., 2009; Garcia and Phillips, 2011; Kuhlman et al., 2007; Razo-Mejia et al., 2014, 2018; Rydenfelt et al., 2014a; Vilar and Leibler, 2003; Weinert et al., 2014). These models quantitatively describe numerous properties of input-output functions, such as the leakiness, saturation, dynamic range, steepness of response, and the  $[EC_{50}]$  – the concentration of inducer at which the response is half maximal. The mathematical forms of these phenotypic properties are couched in terms of a minimal set of experimentally accessible variables, such as the inducer concentration, transcription factor copy number, and the DNA sequence of the binding site (see Chapter 2 and Razo-Mejia et al. (2018)). While the amino acid sequence of the transcription factor is another controllable variable, it is seldom implemented in quantitative terms considering mutations with subtle changes in chemistry frequently yield unpredictable physiological consequences. In this work, we examine how a series of mutations in either the DNA binding or inducer binding domains of a transcriptional repressor influence the values of the biophysical parameters which govern its regulatory behavior.

We build upon the results presented in Chapter 2 of this thesis and present a theoretical framework for understanding how mutations in the amino acid sequence of the repressor affect different parameters and alter the free energy of the system.

We find that the parameters capturing the allosteric nature of the repressor, the repressor copy number, and the DNA binding specificity contribute independently to the free energy of the system with different degrees of sensitivity. Furthermore, changes restricted to one of these three groups of parameters result in characteristic changes in the free energy relative to the wild-type repressor, providing falsifiable predictions of how different classes of mutations should behave.

Next, we test these descriptions experimentally using the well-characterized transcriptional repressor of the *lac* operon LacI in *E. coli* regulating expression of a fluorescent reporter. We introduce a series of point mutations in either the inducer binding or DNA binding domain. We then measure the full induction profile of each mutant, determine the minimal set of parameters that are affected by the mutation, and predict how each mutation tunes the free energy at different inducer concentrations, repressor copy numbers, and DNA binding strengths. We find in general that mutations in the DNA binding domain only influence DNA binding strength, and that mutations within the inducer binding domain affect only the parameters which dictate the allosteric response. The degree to which these parameters are insulated is notable, as the very nature of allostery suggests that all parameters are intimately connected, thus enabling binding events at one domain to be “sensed” by another.

With knowledge of how a collection of DNA binding and inducer binding single mutants behave, we predict the induction profiles and the free energy changes of pairwise double mutants with quantitative accuracy. We find that the energetic effects of each individual mutation are additive, indicating that epistatic interactions are absent between the mutations examined here. Our model provides a means for identifying and quantifying the extent of epistatic interactions in a more complex set of mutations, and can shed light on how the protein sequence and general regulatory architecture coevolve.

### 3.3 Theoretical Model

This work considers the inducible simple repression regulatory motif depicted in Fig. 3.1 (A) from a thermodynamic perspective which has been thoroughly dissected and tested experimentally (Brewster et al., 2014; Garcia and Phillips, 2011; Razo-Mejia et al., 2018) and is described in depth in Chapter 2. The result of this extensive theory-experiment dialogue is a succinct input-output function schematized in Fig. 3.1 (B) that computes the fold-change in gene expression relative to an unregulated promoter. This function is of the form

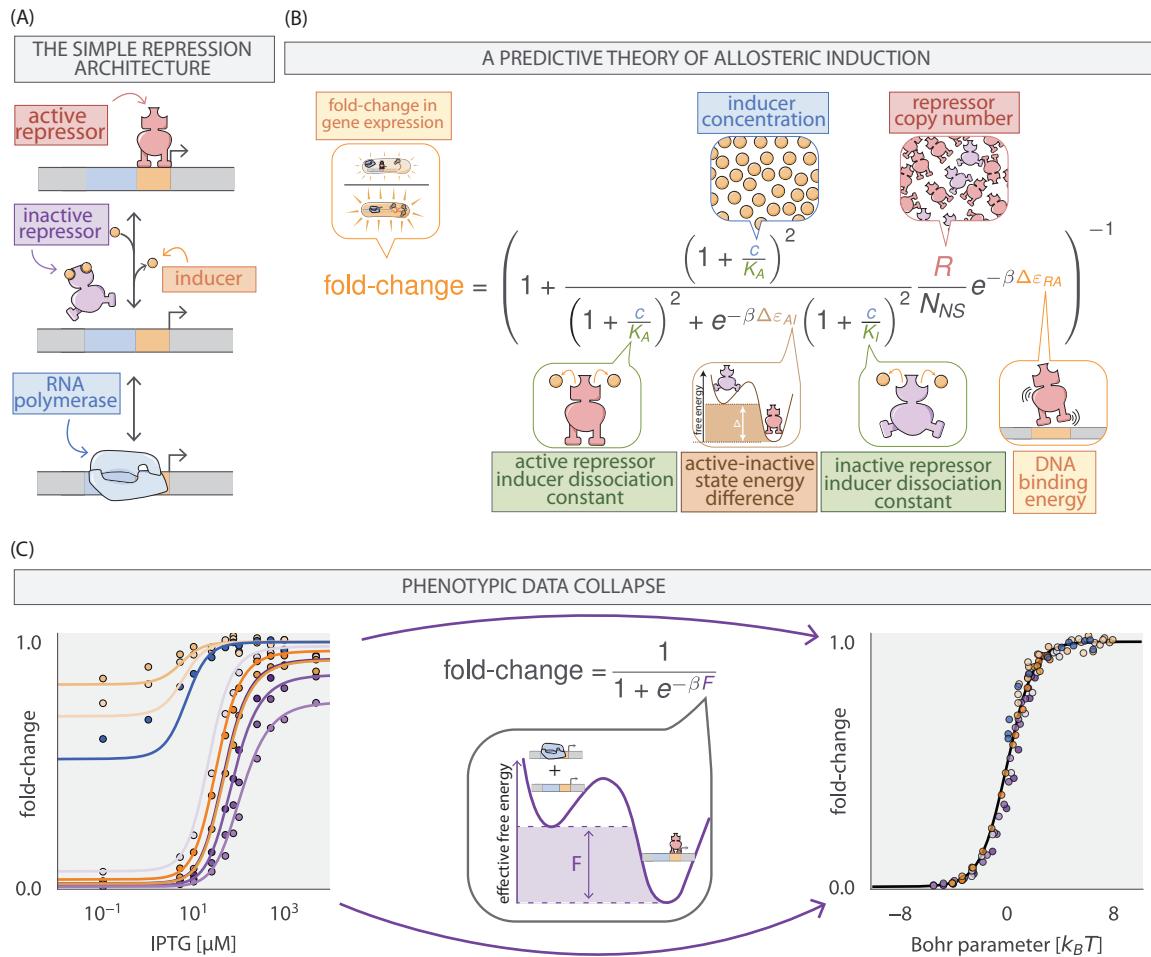
$$\text{fold-change} = \left( 1 + \frac{R_A}{N_{NS}} e^{-\beta \Delta \varepsilon_{RA}} \right)^{-1}, \quad (3.1)$$

where  $R_A$  is the number of active repressors per cell,  $N_{NS}$  is the number of non-specific binding sites for the repressor,  $\Delta \varepsilon_{RA}$  is the binding energy of the repressor to its specific binding site relative to the non-specific background, and  $\beta$  is defined as  $\frac{1}{k_B T}$  where  $k_B$  is the Boltzmann constant and  $T$  is the temperature. While this theory requires knowledge of the number of *active* repressors, we often only know the total number  $R$  which is the sum total of active and inactive repressors. We can define a prefactor  $p_{\text{act}}(c)$  which captures the allosteric nature of the repressor and encodes the probability a repressor is in the active (repressive) state rather than the inactive state for a given inducer concentration  $c$ , namely,

$$p_{\text{act}}(c) = \frac{\left( 1 + \frac{c}{K_A} \right)^n}{\left( 1 + \frac{c}{K_A} \right)^n + e^{-\beta \Delta \varepsilon_{AI}} \left( 1 + \frac{c}{K_I} \right)^n}. \quad (3.2)$$

Here,  $K_A$  and  $K_I$  are the dissociation constants of the inducer to the active and inactive repressor,  $\Delta \varepsilon_{AI}$  is the energetic difference between the repressor active and inactive states, and  $n$  is the number of allosteric binding sites per repressor molecule ( $n = 2$  for LacI). With this in hand, we can define  $R_A$  in Eq. 3.1 as  $R_A = p_{\text{act}}(c)R$ .

A key feature of Eq. 3.1 and Eq. 3.2 is that the diverse phenomenology of the gene expression induction profile can be collapsed onto a single master curve by



**Figure 3.1: A predictive framework for phenotypic and energetic dissection of the simple repression motif.** (A) The inducible simple repression architecture. When in the active state, the repressor (red) binds the cognate operator sequence of the DNA (orange box) with high specificity, preventing transcription by occluding binding of the RNA polymerase (blue rectangle). Upon addition of an inducer molecule, the inactive state (purple) becomes energetically preferable, and the repressor no longer binds the operator sequence with appreciable specificity. Once unbound from the operator, binding of the RNA polymerase (blue) is no longer blocked, and transcription can occur. (B) The simple repression input-output function for an allosteric repressor with two inducer binding sites. The key parameters are identified in speech bubbles. (C) The fold change in gene expression collapses as a function of the free energy. Panel (C, left) shows measurements of the fold change in gene expression as a function of inducer concentration from Razo-Mejia et al. (2018). Points and errors correspond to the mean and SEM of at least 10 biological replicates. The thin lines represent the line of best fit given the model shown in (B). This model can be rewritten as a Fermi function with an energetic parameter  $F$ , which is the energetic difference between the repressor bound and unbound states of the promoter, schematized in C, Middle. The points in (C), Bottom correspond to the data shown in (C, left) collapsed onto a master curve defined by their calculated free energy  $F$ . The solid black line is the master curve defined by the Fermi function shown in (C, Middle). The Python code (ch3\_fig1.py) used to generate this figure can be found on the thesis GitHub repository.

rewriting the input-output function in terms of the free energy  $F$  also called the Bohr parameter (Phillips, 2015),

$$\text{fold-change} = \frac{1}{1 + e^{-\beta F}}, \quad (3.3)$$

where

$$F = -k_B T \log p_{\text{act}}(c) - k_B T \log \left( \frac{R}{N_{NS}} \right) + \Delta \varepsilon_{RA}. \quad (3.4)$$

Hence, if different combinations of parameters yield the same free energy, they will give rise to the same fold-change in gene expression, enabling us to collapse multiple regulatory scenarios onto a single curve. This can be seen in Fig. 3.1 (C) where eighteen unique inducer titration profiles of a LacI simple repression architecture collected and analyzed in Razo-Mejia et al. (2018) collapse onto a single master curve. The tight distribution about this curve reveals that the fold-change across a variety of genetically distinct individuals can be adequately described by a small number of parameters. Beyond predicting the induction profiles of different strains, the method of data collapse inspired by Eq. 3.3 and Eq. 3.4 can be used as a tool to identify mechanistic changes in the regulatory architecture (Swem et al., 2008). Similar data collapse approaches have been used previously in such a manner and have proved vital for distinguishing between changes in parameter values and changes in the fundamental behavior of the system (Keymer et al., 2006; Swem et al., 2008).

Assuming that a given mutation does not result in a non-functional protein, it is reasonable to say that any or all of the parameters in Eq. 3.1 can be affected by the mutation, changing the observed induction profile and therefore the free energy. To examine how the free energy of a mutant  $F^{(\text{mut})}$  differs from that of the wild-type  $F^{(\text{wt})}$ , we define  $\Delta F = F^{(\text{mut})} - F^{(\text{wt})}$ , which has the form

$$\begin{aligned} \Delta F = & -k_B T \log \left( \frac{p_{\text{act}}^{(\text{mut})}(c)}{p_{\text{act}}^{(\text{wt})}(c)} \right) - k_B T \log \left( \frac{R^{(\text{mut})}}{R^{(\text{wt})}} \right) \\ & + (\Delta \varepsilon_{RA}^{(\text{mut})} - \Delta \varepsilon_{RA}^{(\text{wt})}). \end{aligned} \quad (3.5)$$

$\Delta F$  describes how a mutation translates a point across the master curve shown in Fig. 3.1 (C). As we will show in the coming paragraphs (illustrated in Fig. 3.2), this

formulation coarse grains the myriad parameters shown in Eq. 3.1 and Eq. 3.2 into three distinct quantities, each with different sensitivities to parametric changes. By examining how a mutation changes the  $\Delta F$  as a function of the inducer concentration, one can draw conclusions as to which parameters have been modified based solely on the shape of the curve. To help the reader understand how various perturbations to the parameters tune the free energy, we have hosted an interactive figure on the dedicated website for the publication which makes exploration of parameter space a simpler task.

The first term in Eq. 3.5 is the log ratio of the probability of a mutant repressor being active relative to the wild type at a given inducer concentration  $c$ . This quantity defines how changes to any of the allosteric parameters – such as inducer binding constants  $K_A$  and  $K_I$  or active/inactive state energetic difference  $\Delta\varepsilon_{AI}$  – alter the free energy  $F$ , which can be interpreted as the free energy difference between the repressor bound and unbound states of the promoter. Fig. 3.2 (A) illustrates how changes to the inducer binding constants  $K_A$  and  $K_I$  alone alter the induction profiles and resulting free energy as a function of the inducer concentration. In the limit where  $c = 0$ , the values of  $K_A$  and  $K_I$  do not factor into the calculation of  $p_{\text{act}}(c)$  given by Eq. 3.2 meaning that  $\Delta\varepsilon_{AI}$  is the lone parameter setting the residual activity of the repressor. Thus, if only  $K_A$  and  $K_I$  are altered by a mutation, then  $\Delta F$  should be  $0 k_B T$  when  $c = 0$ , illustrated by the overlapping red, purple, and grey curves in the right-hand plot of Fig. 3.2 (A). However, if  $\Delta\varepsilon_{AI}$  is influenced by the mutation (either alone or in conjunction with  $K_A$  and  $K_I$ ), the leakiness will change, resulting in a non-zero  $\Delta F$  when  $c = 0$ . This is illustrated in Fig. 3.2 (B) where  $\Delta\varepsilon_{AI}$  is the only parameter affected by the mutation.

It is important to note that for a mutation which perturbs only the inducer binding constants, the dependence of  $\Delta F$  on the inducer concentration can be non-monotonic. While the precise values of  $K_A$  and  $K_I$  control the sensitivity of the repressor to inducer concentration, it is the ratio  $K_A/K_I$  that defines whether this non-monotonic behavior is observed. This can be seen more clearly when we con-

sider the limit of saturating inducer concentration,

$$\lim_{c \rightarrow \infty} \log \left( \frac{p_{\text{act}}^{(\text{mut})}}{p_{\text{act}}^{(\text{wt})}} \right) \approx \log \left[ \frac{1 + e^{-\beta \Delta \varepsilon_{AI}^{(\text{wt})}} \left( \frac{K_A^{(\text{wt})}}{K_I^{(\text{wt})}} \right)^n}{1 + e^{-\beta \Delta \varepsilon_{AI}^{(\text{wt})}} \left( \frac{K_A^{(\text{mut})}}{K_I^{(\text{mut})}} \right)^n} \right], \quad (3.6)$$

which illustrates that  $\Delta F$  returns to zero at saturating inducer concentration when the ratio  $K_A/K_I$  is the same for both the mutant and wild-type repressors, so long as  $\Delta \varepsilon_{AI}$  is unperturbed. Non-monotonicity can *only* be achieved by changing  $K_A$  and  $K_I$  and therefore serves as a diagnostic for classifying mutational effects reliant solely on measuring the change in free energy. A rigorous proof of this non-monotonic behavior given changing  $K_A$  and  $K_I$  can be found in supplemental Chapter 7.

The second term in Eq. 3.5 captures how changes in the repressor copy number contributes to changes in free energy. It is important to note that this contribution to the free energy change depends on the total number of repressors in the cell, not just those in the active state. This emphasizes that changes in the expression of the repressor are energetically divorced from changes to the allosteric nature of the repressor. As a consequence, the change in free energy is constant for all inducer concentrations, as is schematized in Fig. 3.2 (C). Because the magnitude of the change in free energy scales logarithmically with changing repressor copy number, a mutation which increases expression from 1 to 10 repressors per cell is more impactful from an energetic standpoint ( $k_B T \log(10) \approx 2.3 k_B T$ ) than an increase from 90 to 100 ( $k_B T \log(100/90) \approx 0.1 k_B T$ ). Appreciable changes in the free energy only arise when variations in the repressor copy number are larger than or comparable to an order of magnitude. Changes of this magnitude are certainly possible from a single point mutation, as it has been shown that even synonymous substitutions can drastically change translation efficiency (Frumkin et al., 2018).

The third and final term in Eq. 3.5 is the difference in the DNA binding energy between the mutant and wild-type repressors. All else being equal, if the mutated state binds more tightly to the DNA than the wild type ( $\Delta \varepsilon_{RA}^{(\text{wt})} > \Delta \varepsilon_{RA}^{(\text{mut})}$ ), the net

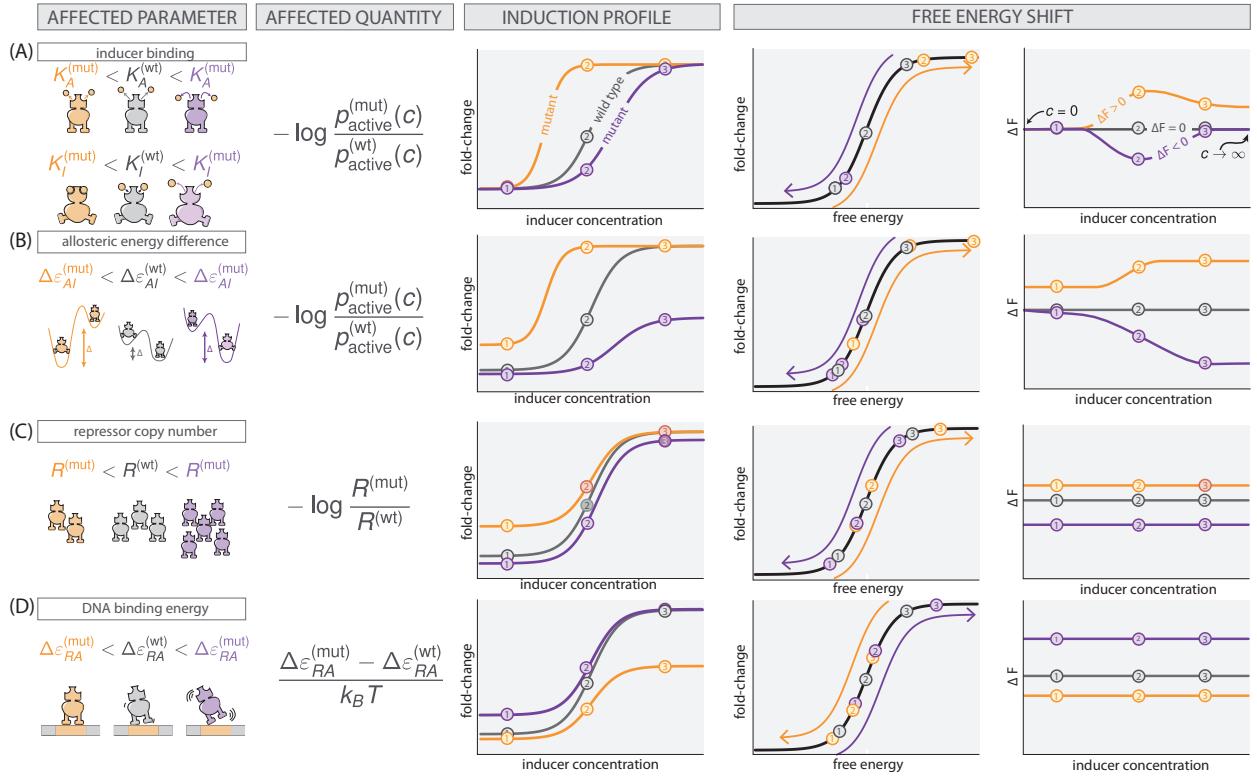
change in the free energy is negative, indicating that the repressor bound states become more energetically favorable due to the mutation. Much like in the case of changing repressor copy number, this quantity is independent of inducer concentration and is therefore also constant Fig. 3.2 (D). However, the magnitude of the change in free energy is linear with DNA binding affinity while it is logarithmic with respect to changes in the repressor copy number. Thus, to change the free energy by  $1 k_B T$ , the repressor copy number must change by a factor of  $\approx 2.3$  whereas the DNA binding energy must change by  $1 k_B T$ .

The unique behavior of each quantity in Eq. 3.5 and its sensitivity with respect to the parameters makes  $\Delta F$  useful as a diagnostic tool to classify mutations. Given a set of fold-change measurements, a simple rearrangement of Eq. 3.3 permits the direct calculation of the free energy, assuming that the underlying physics of the regulatory architecture has not changed. Thus, it becomes possible to experimentally test the general assertions made in Fig. 3.2.

### 3.4 Results

#### DNA Binding Domain Mutants

With this arsenal of analytic diagnostics, we can begin to explore the mutational space of the repressor and map these mutations to the biophysical parameters they control. As one of the most thoroughly studied transcription factors, LacI has been subjected to numerous crystallographic and mutational studies (Daber et al., 2009, 2011; Lewis et al., 1996). One such work generated a set of point mutations in the LacI repressor and examined the diversity of the phenotypic response to different allosteric effectors (Daber et al., 2011). However, several experimental variables were unknown, precluding precise calculation of  $\Delta F$  as presented in the previous section. In Daber et al. (2011), the repressor variants and the fluorescence reporter were expressed from separate plasmids. As the copy numbers of these plasmids fluctuate in the population, both the population average repressor copy number and the number of regulated promoters were unknown. Both of these quantities have been shown previously to significantly alter the measured gene expression



**Figure 3.2: Parametric changes due to mutations and the corresponding free-energy changes for (A) perturbations to  $K_A$  and  $K_I$ , (B) changes to the allosteric energy difference  $\Delta\varepsilon_{AI}$ , (C) changes to repressor copy number, and (D) changes in DNA binding affinity.** The first column schematizes the changed parameters and the second column reflects which quantity in Eq. 3.5 is affected. The third column shows representative induction profiles from mutants which have smaller (purple) and larger (orange) values for the parameters than the wild type (gray). The fourth and fifth columns illustrate how the free energy is changed as a result. Purple and red arrows indicate the direction in which the points are translated about the master curve. Three concentrations (points labeled 1, 2, and 3) are shown to illustrate how each point is moved in free-energy space. An interactive version of this figure can be found on the paper website ([https://www.rpgroup.caltech.edu/mwc\\_mutants](https://www.rpgroup.caltech.edu/mwc_mutants)).

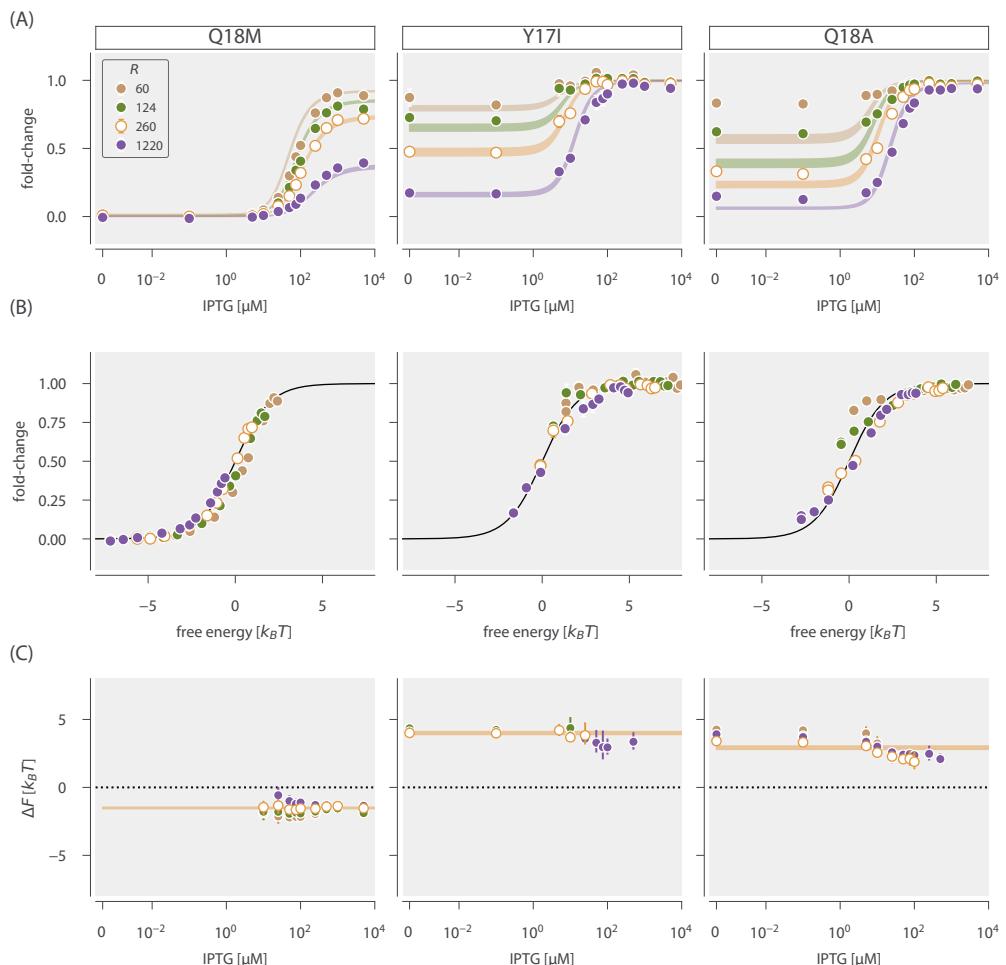
and calculation of  $\Delta F$  is dependent on knowledge of their values. While the approach presented in Daber et al. (2011) considers the Lac repressor as an MWC molecule, the copy numbers of the repressor and the reporter gene were swept into an effective parameter  $R/K_{DNA}$ , hindering our ability to distinguish between changes in repressor copy number or in DNA binding energy. To test our hypothesis of free energy differences resulting from various parameter perturbations, we used the dataset in Daber et al. (2011) as a guide and chose a subset of the mutations to quantitatively dissect. To control copy number variation, the mutant repressors and the reporter gene were integrated into the *E. coli* chromosome where the copy numbers are known and tightly controlled (Garcia and Phillips, 2011; Razo-Mejia et al., 2018). Furthermore, the mutations were paired with ribosomal binding sites where the level of translation of the wild-type repressor had been directly measured previously (Garcia and Phillips, 2011).

We made three amino acid substitutions (Y17I, Q18A, and Q18M) that are critical for the DNA-repressor interaction. These mutations were introduced into the *lacI* sequence used in Garcia and Phillips (2011) with four different ribosomal binding site sequences that were shown (via quantitative Western blotting) to tune the wild-type repressor copy number across three orders of magnitude. These mutant constructs were integrated into the *E. coli* chromosome harboring a Yellow Fluorescent Protein (YFP) reporter. The YFP promoter included the native O2 LacI operator sequence which the wild-type LacI repressor binds with high specificity ( $\Delta\epsilon_{RA} = -13.9 k_B T$ ). The fold-change in gene expression for each mutant across twelve concentrations of IPTG was measured via flow cytometry. As we mutated only a single amino acid with the minimum number of base pair changes to the codons from the wild-type sequence, we find it unlikely that the repressor copy number was drastically altered from those reported in Garcia and Phillips (2011) for the wild-type sequence paired with the same ribosomal binding site sequence. In characterizing the effects of these DNA binding mutations, we take the repressor copy number to be unchanged. Any error introduced by this assumption should be manifest as a larger than predicted systematic shift in the free energy change

when the repressor copy number is varied.

A naïve hypothesis for the effect of a mutation in the DNA binding domain is that *only* the DNA binding energy is affected. This hypothesis appears to contradict the core principle of allostery in that ligand binding in one domain influences binding in another, suggesting that changing parameter modifies them all. The characteristic curves summarized in Fig. 3.2 give a means to discriminate between these two hypotheses by examining the change in the free energy. Using a single induction profile (white-faced points in Fig. 3.3), we estimated the DNA binding energy using Bayesian inferential methods, the details of which are thoroughly discussed in the Materials & Methods as well as in the supplemental Chapter 7. The shaded red region for each mutant in Fig. 3.3 represents the 95% credible region of this fit whereas all other shaded regions are 95% credible regions of the predictions for other repressor copy numbers. We find that redetermining only the DNA binding energy accurately captures the majority of the induction profiles, indicating that other parameters are unaffected. One exception is for the lowest repressor copy numbers ( $R = 60$  and  $R = 124$  per cell) of mutant Q18A at low concentrations of IPTG. However, we note that this disagreement is comparable to that observed for the wild-type repressor binding to the weakest operator in Razo-Mejia et al. (2018), illustrating that our model is imperfect in characterizing weakly repressing architectures. Including other parameters in the fit (such as  $\Delta\varepsilon_{AI}$ ) does not significantly improve the accuracy of the predictions. Furthermore, the magnitude of this disagreement also depends on the choice of the fitting strain (see supplemental Chapter 7).

Mutations Y17I and Q18A both weaken the affinity of the repressor to the DNA relative to the wild type strain with binding energies of  $-9.9^{+0.1}_{-0.1} k_B T$  and  $-11.0^{+0.1}_{-0.1} k_B T$ , respectively. Here we report the median of the inferred posterior probability distribution with the superscripts and subscripts corresponding to the upper and lower bounds of the 95% credible region. These binding energies are comparable to that of the wild-type repressor affinity to the native LacI opera-



**Figure 3.3: Induction profiles and free-energy differences of DNA binding domain mutations.** Each column corresponds to the highlighted mutant at the top of the figure. Each strain was paired with the native O2 operator sequence. Open points correspond to the strain for each mutant from which the DNA binding energy was estimated. (A) Induction profiles of each mutant at four different repressor copy numbers as a function of the inducer concentration. Points correspond to the mean fold change in gene expression of 6–10 biological replicates. Error bars are the SEM. Shaded regions demarcate the 95% credible region of the induction profile generated by the estimated DNA binding energy. (B) Data collapse of all points for each mutant shown in A using only the DNA binding energy estimated from a single repressor copy number. Points correspond to the average fold change in gene expression of 6–10 biological replicates. Error bars are SEM. Where error bars are not visible, the relative error in measurement is smaller than the size of the marker. (C) The change in the free energy resulting from each mutation as a function of the inducer concentration. Points correspond to the median of the marginal posterior distribution for the free energy. Error bars represent the upper and lower bounds of the 95% credible region. Points in A at the detection limits of the flow cytometer (near fold-change values of 0 and 1) were neglected for calculation of the  $\Delta F$ . The IPTG concentration is shown on a symmetric log scale with linear scaling ranging from 0 to  $10^{-2} \mu\text{M}$  and log scaling elsewhere. The shaded red lines in C correspond to the 95% credible region of our predictions for  $\Delta F$  based solely on estimation of  $\Delta\varepsilon_{RA}$  from the strain with  $R = 260$  repressors per cell. The Python code (ch3\_fig3.py) used to generate this figure can be found on the thesis GitHub

tor sequence O3, with a DNA binding energy of  $-9.7 k_B T$ . The mutation Q18M increases the strength of the DNA-repressor interaction relative to the wild-type repressor with a binding energy of  $-15.43^{+0.07}_{-0.06} k_B T$ , comparable to the affinity of the wild-type repressor to the native O1 operator sequence ( $-15.3 k_B T$ ). It is notable that a single amino acid substitution of the repressor is capable of changing the strength of the DNA binding interaction well beyond that of many single base-pair mutations in the operator sequence (Barnes et al., 2019).

Using the new DNA binding energies, we can collapse all measurements of fold-change as a function of the free energy as shown in Fig. 3.3 (B). This allows us to test the diagnostic power of the decomposition of the free energy described in Fig. 3.2. To compute the  $\Delta F$  for each mutation, we inferred the observed mean free energy of the mutant strain for each inducer concentration and repressor copy number (see Materials & Methods as well as the supplemental Chapter 7 for a detailed explanation of the inference). We note that in the limit of extremely low or high fold-change, the inference of the free energy is either over- or under-estimated, respectively, introducing a systematic error. Thus, points which are close to these limits are omitted in the calculation of  $\Delta F$ . We direct the reader to the supplemental Chapter 7 for a detailed discussion of this systematic error. With a measure of  $F^{(\text{mut})}$  for each mutant at each repressor copy number, we compute the difference in free energy relative to the wild-type strain with the same repressor copy number and operator sequence, restricting all variability in  $\Delta F$  solely to changes in  $\Delta \varepsilon_{RA}$ .

The change in free energy for each mutant is shown in Fig. 3.3 (C). It can be seen that the  $\Delta F$  for each mutant is constant as a function of the inducer concentration and is concordant with the prediction generated from fitting  $\Delta \varepsilon_{RA}$  to a single repressor copy number (orange lines Fig. 3.3 (C)]) This is in line with the predictions outlined in Fig. 3.2 (C) and (D), indicating that the allosteric parameters are “insulated”, meaning they are not affected by the DNA binding domain mutations. As the  $\Delta F$  for all repressor copy numbers collapses onto the prediction, we can say that the expression of the repressor itself is the same or comparable with that of the

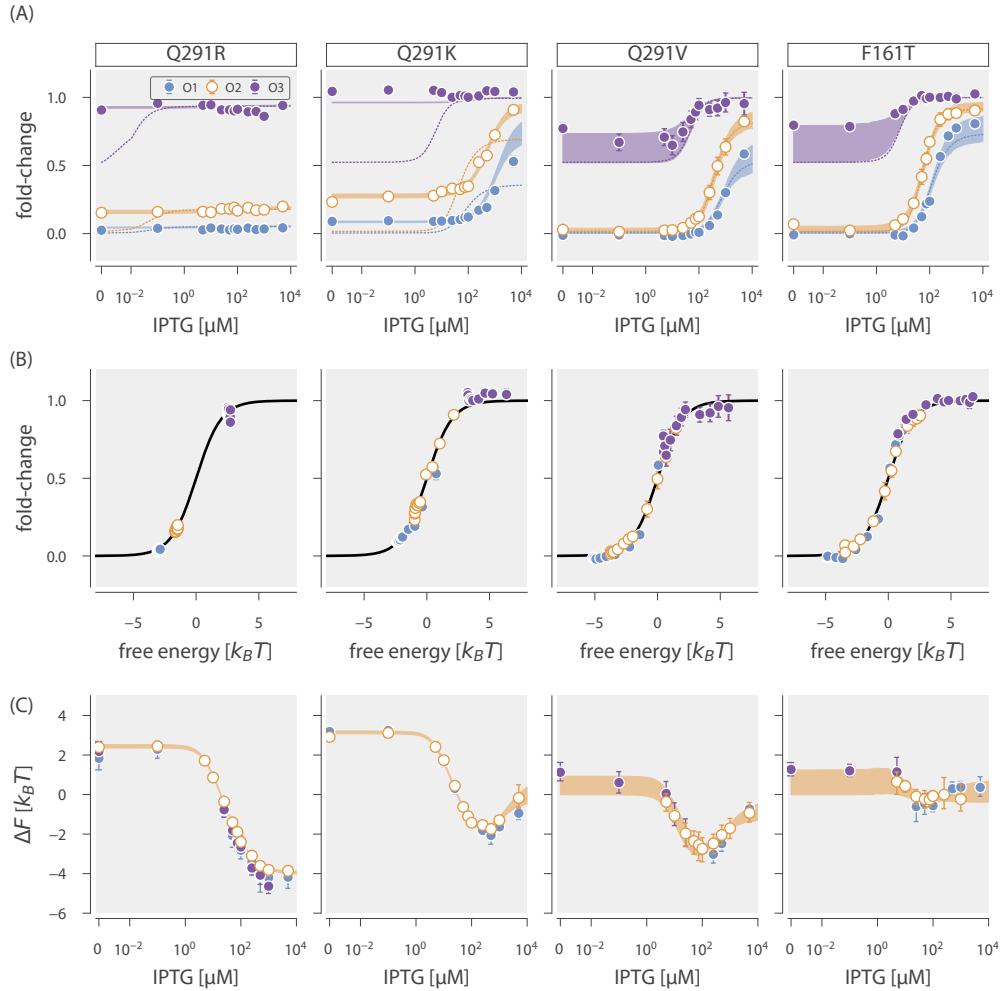
wild type. If the repressor copy number were perturbed in addition to  $\Delta\varepsilon_{RA}$ , one would expect a shift away from the prediction that scales logarithmically with the change in repressor copy number. However, as the  $\Delta F$  is approximately the same for each repressor copy number, it can be surmised that the mutation does not significantly change the expression or folding efficiency of the repressor itself. These results allow us to state that the DNA binding energy  $\Delta\varepsilon_{RA}$  is the only parameter modified by the DNA mutants examined.

### Inducer Binding Domain Mutants

Much as in the case of the DNA binding mutants, we cannot safely assume *a priori* that a given mutation in the inducer binding domain affects only the inducer binding constants  $K_A$  and  $K_I$ . While it is easy to associate the inducer binding constants with the inducer binding domain, the critical parameter in our allosteric model  $\Delta\varepsilon_{AI}$  is harder to restrict to a single spatial region of the protein. As  $K_A$ ,  $K_I$ , and  $\Delta\varepsilon_{AI}$  are all parameters dictating the allosteric response, we consider two hypotheses in which inducer binding mutations alter either all three parameters or only  $K_A$  and  $K_I$ .

We made four point mutations within the inducer binding domain of LacI (F161T, Q291V, Q291R, and Q291K) that have been shown previously to alter binding to multiple allosteric effectors (Daber et al., 2009). In contrast to the DNA binding domain mutants, we paired the inducer binding domain mutations with the three native LacI operator sequences (which have various affinities for the repressor) and a single ribosomal binding site sequence. This ribosomal binding site sequence, as reported in Garcia and Phillips (2011), expresses the wild-type LacI repressor to an average copy number of approximately 260 per cell. As the free energy differences resulting from point mutations in the DNA binding domain can be described solely by changes to  $\Delta\varepsilon_{RA}$ , we continue under the assumption that the inducer binding domain mutations do not significantly alter the repressor copy number.

The induction profiles for these four mutants are shown in Fig. 3.4 (A). Of the



**Figure 3.4: Induction profiles and free-energy differences of inducer binding domain mutants.** Open points represent the strain to which the parameters were fit — namely, the O2 operator sequence. Each column corresponds to the mutant highlighted at the top of the figure. All strains have  $R = 260$  per cell. (A) The fold change in gene expression as a function of the inducer concentration for three operator sequences of varying strength. Dashed lines correspond to the curve of best fit resulting from fitting  $K_A$  and  $K_I$  alone. Shaded curves correspond to the 95% credible region of the induction profile determined from fitting  $K_A$ ,  $K_I$ , and  $\Delta\varepsilon_{AI}$ . Points correspond to the mean measurement of 6–12 biological replicates. Error bars are the SEM. (B) Points in A collapsed as a function of the free energy calculated from redetermining  $K_A$ ,  $K_I$ , and  $\Delta\varepsilon_{AI}$ . (C) Change in free energy resulting from each mutation as a function of the inducer concentration. Points correspond to the median of the posterior distribution for the free energy. Error bars represent the upper and lower bounds of the 95% credible region. Shaded curves are the predictions. IPTG concentration is shown on a symmetric log scaling axis with the linear region spanning from 0 to  $10^{-2} \mu\text{M}$  and log scaling elsewhere. The Python code (ch3\_fig4.py) used to generate this figure can be found on the thesis GitHub repository.

mutations chosen, Q291R and Q291K appear to have the most significant impact, with Q291R abolishing the characteristic sigmoidal titration curve entirely. It is notable that both Q291R and Q291K have elevated expression in the absence of inducer compared to the other two mutants paired with the same operator sequence. Panel (A) in Fig. 3.2 illustrates that if only  $K_A$  and  $K_I$  were being affected by the mutations, the fold-change should be identical for all mutants in the absence of inducer. This discrepancy in the observed leakiness immediately suggests that more than  $K_A$  and  $K_I$  are affected for Q291K and Q291R.

Using a single induction profile for each mutant (shown in Fig. 3.4 as white-faced circles), we inferred the parameter combinations for both hypotheses and drew predictions for the induction profiles with other operator sequences. We find that the simplest hypothesis (in which only  $K_A$  and  $K_I$  are altered) does not permit accurate prediction of most induction profiles. These curves, shown as dotted lines in Fig. 3.4 (A), fail spectacularly in the case of Q291R and Q291K, and undershoot the observed profiles for F161T and Q291V, especially when paired with the weak operator sequence O3. The change in the leakiness for Q291R and Q291K is particularly evident as the expression at  $c = 0$  should be identical to the wild-type repressor under this hypothesis. Altering only  $K_A$  and  $K_I$  is not sufficient to accurately predict the induction profiles for F161T and Q291V, but not to the same degree as Q291K and Q291R. The disagreement is most evident for the weakest operator O3 green lines in 3.4, though we have discussed previously that the induction profiles for weak operators are difficult to accurately describe and can result in comparable disagreement for the wild-type repressor (Razo-Mejia et al., 2018).

Including  $\Delta\varepsilon_{AI}$  as a perturbed parameter in addition to  $K_A$  and  $K_I$  improves the predicted profiles for all four mutants. By fitting these three parameters to a single strain, we are able to accurately predict the induction profiles of other operators as seen by the shaded lines in Fig. 3.4 (A). With these modified parameters, all experimental measurements collapse as a function of their free energy as prescribed by

Eq. 3.3 (Fig. 3.4 (B)). All four mutations significantly diminish the binding affinity of both states of the repressor to the inducer, as seen by the estimated parameter values reported in Table 3.1. As evident in the data alone, Q291R abrogates inducibility outright ( $K_A \approx K_I$ ). For Q291K, the active state of the repressor can no longer bind inducer whereas the inactive state binds with weak affinity. The remaining two mutants, Q291V and F161T, both show diminished binding affinity of the inducer to both the active and inactive states of the repressor relative to the wild-type.

Table 3.1: Inferred values of  $K_A$ ,  $K_I$ , and  $\Delta\epsilon_{AI}$  for inducer binding mutants

Mutant	$K_A$	$K_I$	$\Delta\epsilon_{AI}$ [ $k_B T$ ]	Reference
WT	$139^{+29}_{-22} \mu M$	$0.53^{+0.04}_{-0.04} \mu M$	4.5	Razo-Mejia et al. (2018)
F161T	$165^{+90}_{-65} \mu M$	$3^{+6}_{-3} \mu M$	$1^{+5}_{-2}$	This study
Q291V	$650^{+450}_{-250} \mu M$	$8^{+8}_{-8} \mu M$	$3^{+6}_{-3}$	This study
Q291K	$> 1 \text{ mM}$	$310^{+70}_{-60} \mu M$	$-3.11^{+0.07}_{-0.07}$	This study
Q291R	$9^{+20}_{-9} \mu M$	$8^{+20}_{-8} \mu M$	$-2.35^{+0.01}_{-0.09}$	This study

Given the collection of fold-change measurements, we computed the  $\Delta F$  relative to the wild-type strain with the same operator and repressor copy number. This leaves differences in  $p_{act}(c)$  as the sole contributor to the free energy difference, assuming our hypothesis that  $K_A$ ,  $K_I$ , and  $\Delta\epsilon_{AI}$  are the only perturbed parameters is correct. The change in free energy can be seen in Fig. 3.4 (C). For all mutants, the free energy difference inferred from the observed fold-change measurements falls within error of the predictions generated under the hypothesis that  $K_A$ ,  $K_I$ , and  $\Delta\epsilon_{AI}$  are all affected by the mutation (shaded curves in Fig. 3.4 (C)). The profile of the free energy change exhibits some of the rich phenomenology illustrated in Fig. 3.2 (A) and (B). Q291K, F161T, and Q291V exhibit a non-monotonic dependence on the inducer concentration, a feature that can only appear when  $K_A$  and  $K_I$  are altered. The non-zero  $\Delta F$  at  $c = 0$  for Q291R and Q291K coupled with an

inducer concentration dependence is a telling sign that  $\Delta\varepsilon_{AI}$  must be significantly modified. This shift in  $\Delta F$  is positive in all cases, indicating that  $\Delta\varepsilon_{AI}$  must have decreased, and that the inactive state has become more energetically favorable for these mutants than for the wild-type protein. Indeed the estimates for  $\Delta\varepsilon_{AI}$  (Table 3.1) reveal both mutations Q291R and Q291K make the inactive state more favorable than the active state. Thus, for these two mutations, only  $\approx 10\%$  of the repressors are active in the absence of inducer, whereas the basal active fraction is  $\approx 99\%$  for the wild-type repressor (Razo-Mejia et al., 2018).

We note that the parameter values reported here disagree with those reported in Daber et al. (2011). This disagreement stems from different assumptions regarding the residual activity of the repressor in the absence of inducer and the parametric degeneracy of the MWC model without a concrete independent measure of  $\Delta\varepsilon_{AI}$ . A detailed discussion of the difference in parameter values between our previous work (Garcia and Phillips, 2011; Razo-Mejia et al., 2018), that of Daber et al. (2011), and those of other seminal works can be found in the supplemental Chapter 7.

Taken together, these parametric changes diminish the response of the regulatory architecture as a whole to changing inducer concentrations. They furthermore reveal that the parameters which govern the allosteric response are interdependent and no single parameter is insulated from the others. However, as *only* the allosteric parameters are changed, one can say that the allosteric parameters as a whole are insulated from the other components which define the regulatory response, such as repressor copy number and DNA binding affinity.

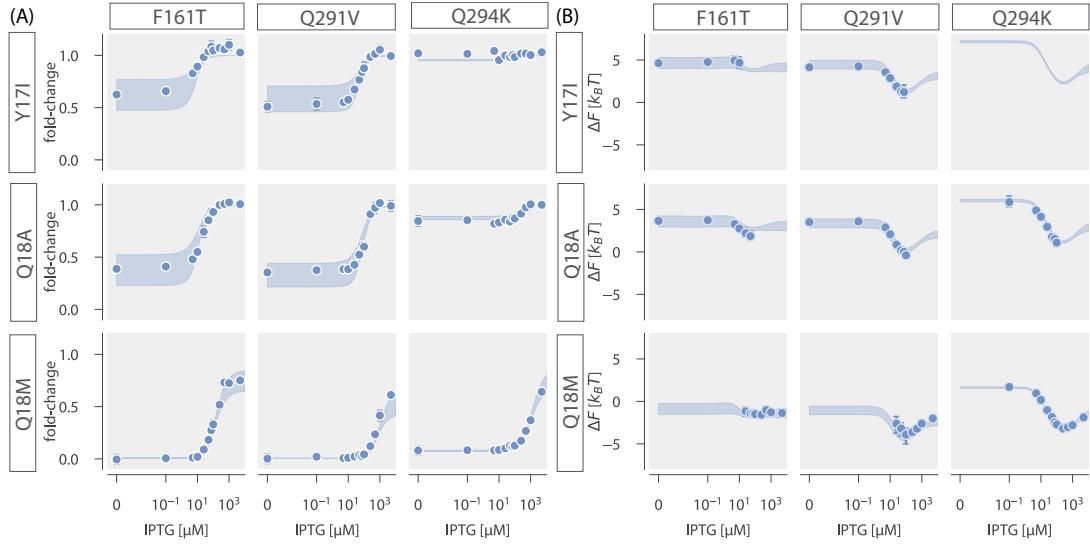
### Predicting Effects of Pairwise Double Mutations

Given full knowledge of each mutation individually, we can draw predictions of the behavior of the pairwise double mutants with no free parameters based on the simplest null hypothesis of no epistasis. The formalism of  $\Delta F$  defined by Eq. 3.5 explicitly states that the contribution to the free energy of the system from the difference in DNA binding energy and the allosteric parameters are strictly additive. Thus, deviations from the predicted change in free energy would suggest epistatic

interactions between the two mutations.

To test this additive model, we constructed nine double mutant strains, each having a unique inducer binding (F161T, Q291V, Q291K) and DNA binding mutation (Y17I, Q18A, Q18M). To make predictions with an appropriate representation of the uncertainty, we computed a large array of induction profiles given random draws from the posterior distribution for the DNA binding energy (determined from the single DNA binding mutants) as well as from the joint posterior for the allosteric parameters (determined from the single inducer binding mutants). These predictions, shown in Fig. 3.5 (A) and (B) as shaded blue curves, capture all experimental measurements of the fold-change (Fig. 3.5 (A)) and the inferred difference in free energy (Fig. 3.5 (B)). The latter indicates that there are no epistatic interactions between the mutations queried in this work, though if there were, systematic deviations from these predictions would shed light on how the epistasis is manifest.

The precise agreement between the predictions and measurements for Q291K paired with either Q18A or Q18M is striking as Q291K drastically changed  $\Delta\varepsilon_{AI}$  in addition to  $K_A$  and  $K_I$ . Our ability to predict the induction profile and free energy change underscores the extent to which the DNA binding energy and the allosteric parameters are insulated from one another. Despite this insulation, the repressor still functions as an allosteric molecule, emphasizing that the mutations we have inserted do not alter the pathway of communication between the two domains of the protein. As the double mutant Y17I-Q291K exhibits fold-change of approximately 1 across all IPTG concentrations (Fig. 3.5 (A)), these mutations in tandem make repression so weak it is beyond the limits which are detectable by our experiments. As a consequence, we are unable to estimate  $\Delta F$  nor experimentally verify the corresponding prediction (grey box in Fig. 3.5 (B)). However, as the predicted fold-change in gene expression is also approximately 1 for all  $c$ , we believe that the prediction shown for  $\Delta F$  is likely accurate. One would be able to infer the  $\Delta F$  to confirm these predictions using a more sensitive method for measuring the



**Figure 3.5: Induction and free-energy profiles of DNA binding and inducer binding double mutants.** (A) Fold change in gene expression for each double mutant as a function of IPTG. Points and errors correspond to the mean and standard error of 6–10 biological replicates. Where not visible, error bars are smaller than the corresponding marker. Shaded regions correspond to the 95% credible region of the prediction given knowledge of the single mutants. These were generated by drawing  $10^4$  samples from the  $\Delta\varepsilon_{RA}$  posterior distribution of the single DNA binding domain mutants and the joint probability distribution of  $K_A$ ,  $K_I$ , and  $\Delta\varepsilon_{AI}$  from the single inducer binding domain mutants. (B) The difference in free energy of each double mutant as a function of the reference free energy. Points and errors correspond to the median and bounds of the 95% credible region of the posterior distribution for the inferred  $\Delta F$ . Shaded lines region are the predicted change in free energy, generated in the same manner as the shaded lines in (A). All measurements were taken from a strain with 260 repressors per cell paired with a reporter with the native O2 LacI operator sequence. In all plots, the IPTG concentration is shown on a symmetric log axis with linear scaling between 0 and  $10^{-2} \mu\text{M}$  and log scaling elsewhere. The Python code (`ch3_fig5.py`) used to generate this figure can be found on the thesis GitHub repository.

fold-change, such as single-cell microscopy or colorimetric assays.

### 3.5 Discussion

Allosteric regulation is often couched as “biological action at a distance”. Despite extensive knowledge of protein structure and function, it remains difficult to translate the coordinates of the atomic constituents of a protein to the precise parameter values which define the functional response, making each mutant its own

intellectual adventure. Bioinformatic approaches to understanding the sequence-structure relationship have permitted us to examine how the residues of allosteric proteins evolve, revealing conserved regions which hint to their function. Co-evolving residues reveal sectors of conserved interactions which traverse the protein that act as the allosteric communication channel between domains (McLaughlin Jr et al., 2012; Reynolds et al., 2011; Suel et al., 2003). Elucidating these sectors has advanced our understanding of how distinct domains “talk” to one another and has permitted direct engineering of allosteric responses into non-allosteric enzymes (Poelwijk et al., 2011; Raman et al., 2016). Even so, we are left without a quantitative understanding of how these admittedly complex networks set the energetic difference between active and inactive states or how a given mutation influences binding affinity. In this context, a biophysical model in which the various parameters are intimately connected to the molecular details can be of use and can lead to quantitative predictions of the interplay between amino-acid identity and system-level response.

By considering how each parameter contributes to the observed change in free energy, we are able to tease out different classes of parameter perturbations which result in stereotyped responses to changing inducer concentration. These characteristic changes to the free energy can be used as a diagnostic tool to classify mutational effects. For example, we show in Fig. 3.2 that modulating the inducer binding constants  $K_A$  and  $K_I$  results in non-monotonic free energy changes that are dependent on the inducer concentration, a feature observed in the inducer binding mutants examined in this work. Simply looking at the inferred  $\Delta F$  as a function of inducer concentration, which requires no fitting of the biophysical parameters, indicates that  $K_A$  and  $K_I$  must be modified considering those are the only parameters which can generate such a response.

Another key observation is that a perturbation to only  $K_A$  and  $K_I$  requires that the  $\Delta F = 0 k_B T$  at  $c = 0$ . Deviations from this condition imply that more than the inducer binding constants must have changed. If this shift in  $\Delta F$  off of  $0 k_B T$

at  $c = 0$  is not constant across all inducer concentrations, we can surmise that the energy difference between the allosteric states  $\Delta\epsilon_{AI}$  must also be modified. We again see this effect for all of our inducer mutants. By examining the inferred  $\Delta F$ , we can immediately say that in addition to  $K_A$  and  $K_I$ ,  $\Delta\epsilon_{AI}$  must decrease relative to the wild-type value as  $\Delta F > 0$  at  $c = 0$ . When the allosteric parameters are fit to the induction profiles, we indeed see that this is the case, with all four mutations decreasing the energy gap between the active and inactive states. Two of these mutations, Q291R and Q291K, make the inactive state of the repressor *more* stable than the active state, which is not the case for the wild-type repressor (Razo-Mejia et al., 2018).

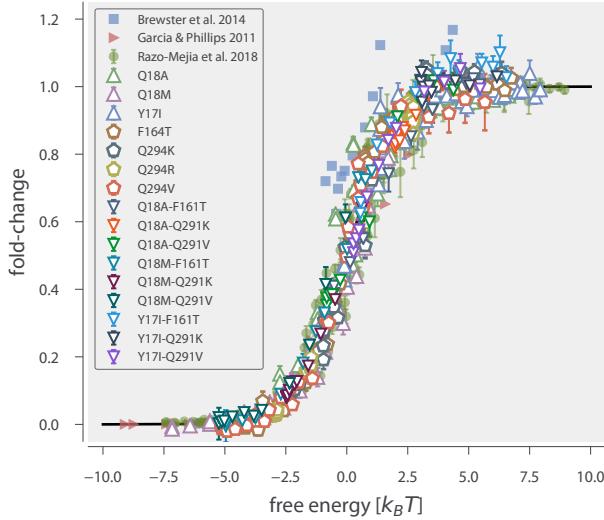
Our formulation of  $\Delta F$  indicates that shifts away from  $0 k_B T$  that are independent of the inducer concentration can only arise from changes to the repressor copy number and/or DNA binding specificity, indicating that the allosteric parameters are untouched. We see that for three mutations in the DNA binding domain,  $\Delta F$  is the same irrespective of the inducer concentration. Measurements of  $\Delta F$  for these mutants with repressor copy numbers across three orders of magnitude yield approximately the same value, revealing that  $\Delta\epsilon_{RA}$  is the sole parameter altered via the mutations.

We note that the conclusions stated above can be qualitatively drawn without resorting to fitting various parameters and measuring the goodness-of-fit. Rather, the distinct behavior of  $\Delta F$  is sufficient to determine which parameters are changing. Here, these conclusions are quantitatively confirmed by fitting these parameters to the induction profile, which results in accurate predictions of the fold-change and  $\Delta F$  for nearly every strain across different mutations, repressor copy numbers, and operator sequence, all at different inducer concentrations. With a collection of evidence as to what parameters are changing for single mutations, we put our model to the test and drew predictions of how double mutants would behave both in terms of the titration curve and free energy profile.

A hypothesis that arises from our formulation of  $\Delta F$  is that a simple summa-

tion of the energetic contribution of each mutation should be sufficient to predict the double mutants (so long as they are in separate domains). We find that such a calculation permits precise and accurate predictions of the double mutant phenotypes, indicating that there are no epistatic interactions between the mutations examined in this work. With an expectation of what the free energy differences should be, epistatic interactions could be understood by looking at how the measurements deviate from the prediction. For example, if epistatic interactions exist which appear as a systematic shift from the predicted  $\Delta F$  independent of inducer concentration, one could conclude that DNA binding energy is not equal to that of the single mutation in the DNA binding domain alone. Similarly, systematic shifts that are dependent on the inducer concentration (i.e. not constant) indicate that the allosteric parameters must be influenced. If the expected difference in free energy is equal to  $0 k_B T$  when  $c = 0$ , one could surmise that the modified parameter must not be  $\Delta\varepsilon_{AI}$  nor  $\Delta\varepsilon_{RA}$  as these would both result in a shift in leakiness, indicating that  $K_A$  and  $K_I$  are further modified.

Ultimately, we present this work as a proof-of-principle for using biophysical models to investigate how mutations influence the response of allosteric systems. We emphasize that such a treatment allows one to boil down the complex phenotypic responses of these systems to a single-parameter description which is easily interpretable as a free energy. The general utility of this approach is illustrated in Fig. 3.6 where gene expression data from previous work along with all of the measurements presented in this work collapse onto the master curve defined by Eq. 3.3. While our model coarse grains many of the intricate details of transcriptional regulation into two states (one in which the repressor is bound to the promoter and one where it is not), it is sufficient to describe a swath of regulatory scenarios. As discussed in the supplemental Chapter 7, any architecture in which the transcription-factor bound and transcriptionally active states of the promoter can be separated into two distinct coarse-grained states can be subjected to such an analysis.



**Figure 3.6: Data collapse of the simple repression regulatory architecture.** All data are means of biological replicates. Where present, error bars correspond to the standard error of the mean of five to fifteen biological replicates. Red triangles indicate data from Garcia and Phillips (2011) obtained by colorimetric assays. Blue squares are data from Brewster et al. (2014) acquired from video microscopy. Green circles are data from Razo-Mejia et al. (2018). obtained via flow cytometry. All other symbols correspond to the work presented here. An interactive version of this figure can be found on the paper website where the different data sets can be viewed in more detail. The Python code (`ch3_fig6.py`) used to generate this figure can be found on the thesis GitHub repository.

Given enough parametric knowledge of the system, it becomes possible to examine how modifications to the parameters move the physiological response along this reduced one-dimensional parameter space. This approach offers a glimpse at how mutational effects can be described in terms of energy rather than Hill coefficients and arbitrary prefactors. While we have explored a very small region of sequence space in this work, coupling of this approach with high-throughput sequencing-based methods to query a library of mutations within the protein will shed light on the phenotypic landscape centered at the wild-type sequence. Furthermore, pairing libraries of protein and operator sequence mutants will provide insight as to how the protein and regulatory sequence coevolve, a topic rich with opportunity for a dialogue between theory and experiment.

### 3.6 Materials & Methods

#### Bacterial Strains and DNA Constructs

All wild-type strains from which the mutants were derived were generated in previous work from the Phillips group (Garcia and Phillips, 2011; Razo-Mejia et al., 2018). Briefly, mutations were first introduced into the *lacI* gene of our pZS3\*1-lacI plasmid (Garcia and Phillips, 2011) using a combination of overhang PCR Gibson assembly as well as QuickChange mutagenesis (Agilent Technologies). The oligonucleotide sequences used to generate each mutant as well as the method are provided in the supplemental Chapter 7.

For mutants generated through overhang PCR and Gibson assembly, oligonucleotide primers were purchased containing an overhang with the desired mutation and used to amplify the entire plasmid. Using the homology of the primer overhang, Gibson assembly was performed to circularize the DNA prior to electroporation into MG1655 *E. coli* cells. Integration of LacI mutants was performed with  $\lambda$  Red recombineering as described in Sharan et al. (2009) and Garcia and Phillips (2011).

The mutants studied in this work were chosen from data reported in Daber et al. (2011). In selecting mutations, we looked for mutants which suggested moderate to strong deviations from the behavior of the wild-type repressor. We note that the variant of LacI used in this work has an additional three amino acids (Met-Val-Asn) added to the N-terminus than the canonical LacI sequence reported in (???). To remain consistent with the field, we have identified the mutations with respect to their positions in the canonical sequence and those in Daber et al. (2011). However, their positions in the raw data files correspond to that of our LacI variant and is noted in the README files associated with the data.

#### Flow Cytometry

All fold-change measurements were performed on a MACSQuant flow cytometer as described in Razo-Mejia et al. (2018). Briefly, saturated overnight cultures 500  $\mu$ L in volume were grown in deep-well 96 well plates covered with a breath-

able nylon cover (Lab Pak - Nitex Nylon, Sefar America, Cat. No. 241205). After approximately 12 to 15 hr, the cultures reached saturation and were diluted 1000-fold into a second 2 mL 96-deep-well plate where each well contained 500  $\mu$ L of M9 minimal media supplemented with 0.5% w/v glucose (anhydrous D-Glucose, Macron Chemicals) and the appropriate concentration of IPTG (Isopropyl  $\beta$ -D-1-thiogalactopyranoside, Dioxane Free, Research Products International). These were sealed with a breathable cover and were allowed to grow for approximately 8 hours until the  $OD_{600nm} \approx 0.3$ . Cells were then diluted ten-fold into a round-bottom 96-well plate (Corning Cat. No. 3365) containing 90  $\mu$ L of M9 minimal media supplemented with 0.5% w/v glucose along with the corresponding IPTG concentrations.

The flow cytometer was calibrated prior to use with MACSQuant Calibration Beads (Cat. No. 130-093-607). During measurement, the cultures were held at approximately 4° C by placing the 96-well plate on a MACSQuant ice block. All fluorescence measurements were made using a 488 nm excitation wavelength with a 525/50 nm emission filter. The photomultiplier tube voltage settings for the instrument are the same as those used in Razo-Mejia et al. (2018) and are listed in supplemental Chapter 6.

The data were processed using an automatic unsupervised gating procedure based on the front and side-scattering values, where we fit a two-dimensional Gaussian function to the  $\log_{10}$  forward-scattering (FSC) and the  $\log_{10}$  side-scattering (SSC) data. Here we assume that the region with highest density of points in these two channels corresponds to single-cell measurements and consider data points that fall within 40% of the highest density region of the two-dimensional Gaussian function. We direct the reader to Razo-Mejia et al. (2018) and supplemental Chapter 6 for further detail and comparison of flow cytometry with single-cell microscopy.

## Bayesian Parameter Estimation

We used a Bayesian definition of probability in the statistical analysis of all mutants in this work. In supplemental Chapter 7, we derive in detail the statistical models used for the various parameters as well as multiple diagnostic tests. Here, we give a generic description of our approach. To be succinct in notation, we consider a generic parameter  $\theta$  which represents  $\Delta\varepsilon_{RA}$ ,  $K_A$ ,  $K_I$ , and/or  $\Delta\varepsilon_{AI}$  depending on the specific LacI mutant.

As prescribed by Bayes' theorem, we are interested in the posterior probability distribution

$$g(\theta | y) \propto f(y | \theta)g(\theta), \quad (3.7)$$

where we use  $g$  and  $f$  to represent probability densities over parameters and data, respectively, and  $y$  to represent a set of fold-change measurements. The likelihood of observing our dataset  $y$  given a value of  $\theta$  is captured by  $f(y | \theta)$ . All prior information we have about the possible values of  $\theta$  are described by  $g(\theta)$ .

In all inferential models used in this work, we assumed that all experimental measurements at a given inducer concentration were normally distributed about a mean value  $\mu$  dictated by Eq. 3.1 with a variance  $\sigma^2$ ,

$$f(y | \theta) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_i^N \exp \left[ -\frac{(y_i - \mu(\theta))^2}{2\sigma^2} \right], \quad (3.8)$$

where  $N$  is the number of measurements in the data set  $y$ .

This choice of likelihood is justified as each individual measurement at a given inducer concentration is a biological replicate and independent of all other experiments. By using a Gaussian likelihood, we introduce another parameter  $\sigma$ . As  $\sigma$  must be positive and greater than zero, we define as a prior distribution a half-normal distribution with a standard deviation  $\phi$ ,

$$g(\sigma) = \frac{1}{\phi} \sqrt{\frac{2}{\pi}} \exp \left[ -\frac{x}{2\phi^2} \right]; x \geq 0, \quad (3.9)$$

where  $x$  is a given range of values for  $\sigma$ . A standard deviation of  $\phi = 0.1$  was chosen given our knowledge of the scale of our measurement error from other experiments. As the absolute measurement of fold-change is restricted between 0

and 1, and given our knowledge of the sensitivity of the experiment, it is reasonable to assume that the error will be closer to 0 than to 1. Further justification of this choice of prior through simulation based methods are given in the supplemental Chapter 7. The prior distribution for  $\theta$  is dependent on the parameter and its associated physical and physiological restrictions. Detailed discussion of our chosen prior distributions for each model can also be found in the supplemental Chapter 7.

All statistical modeling and parameter inference was performed using Markov chain Monte Carlo (MCMC). Specifically, Hamiltonian Monte Carlo sampling was used as is implemented in the Stan probabilistic programming language (Carpenter et al., 2017). All statistical models saved as .stan models and can be accessed at the GitHub repository associated with this work (DOI: 10.5281/zenodo.2721798) or can be downloaded directly from the paper website.

### Inference of Free Energy From Fold-Change Data

While the fold-change in gene expression is restricted to be between 0 and 1, experimental noise can generate fold-change measurements beyond these bounds. To determine the free energy for a given set of fold-change measurements (for one unique strain at a single inducer concentration), we modeled the observed fold-change measurements as being drawn from a normal distribution with a mean  $\mu$  and standard deviation  $\sigma$ . Using Bayes' theorem, we can write the posterior distribution as

$$g(\mu, \sigma | y) \propto g(\mu)g(\sigma) \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_i^N \exp \left[ \frac{-(y_i - \mu)^2}{2\sigma^2} \right] \quad (3.10)$$

where  $y$  is a collection of fold-change measurements. The prior distribution for  $\mu$  was chosen to be uniform between 0 and 1 while the prior on  $\sigma$  was chosen to be half normal, as written in Eq. 3.9. The posterior distribution was sampled independently for each set of fold-change measurements using MCMC. The .stan model for this inference is available on the paper website.

For each MCMC sample of  $\mu$ , the free energy was calculated as

$$F = -\log(\mu^{-1} - 1) \quad (3.11)$$

which is simply the rearrangement of Eq. 3.3. Using simulated data, we determined that when  $\mu < \sigma$  or  $(1 - \mu) < \sigma$ , the mean fold-change in gene expression was over or underestimated for the lower and upper limit, respectively. This means that there are maximum and minimum levels of fold-change that can be detected using flow cytometry which are set by the distribution of fold-change measurements resulting from various sources of day-to-day variation. This results in a systematic error in the calculation of the free energy, making proper inference beyond these limits difficult. This bounds the range in which we can confidently infer this quantity with flow cytometry. We hypothesize that more sensitive methods, such as single cell microscopy, colorimetric assays, or direct counting of mRNA transcripts via Fluorescence *In Situ* Hybridization (FISH) would improve the measurement of  $\Delta F$ . We further discuss details of this limitation in the supplemental Chapter 7.

### Data and Code Availability

All data was collected, stored, and preserved using the Git version control software. Code for data processing, analysis, and figure generation is available on the [GitHub repository] ([https://www.github.com/rpgroup-pboc/mwc\\_mutants](https://www.github.com/rpgroup-pboc/mwc_mutants)) or can be accessed via the paper website. Raw flow cytometry data is stored on the CaltechDATA data repository and can be accessed via DOI 10.22002/D1.1241.

*Chapter 4*

## ON THE PHYSIOLOGICAL ADAPTABILITY OF A SIMPLE GENETIC CIRCUIT

A version of this chapter is currently under review. A preprint is released as Chure, G; Kaczmarek, Z. A.; Phillips, R. *Physiological Adaptability and Parametric Versatility in a Simple Genetic Circuit*. bioRxiv 2019. DOI: <https://doi.org/10.1101/2019.12.19.878462>. G.C. and R.P. designed experiments and developed theoretical models. G.C. and Z.A.K. collected and analyzed data. G.C. and R.P. wrote the paper.

### 4.1 Abstract

The intimate relationship between the environment and cellular growth rate has remained a major topic of inquiry in bacterial physiology for over a century. Now, as it becomes possible to understand how the growth rate dictates the wholesale reorganization of the intracellular molecular composition, we can interrogate the biophysical principles underlying this adaptive response. Regulation of gene expression drives this adaptation, with changes in growth rate tied to the activation or repression of genes covering enormous swaths of the genome. Here, we dissect how physiological perturbations alter the expression of a circuit which has been extensively characterized in a single physiological state. Given a complete thermodynamic model, we map changes in physiology directly to the biophysical parameters which define the expression. Controlling the growth rate via modulating the available carbon source or growth temperature, we measure the level of gene expression from a LacI-regulated promoter where the LacI copy number is directly measured in each condition, permitting parameter-free prediction of the expression level. The transcriptional output of this circuit is remarkably robust, with expression of the repressor being largely insensitive to the growth rate. The predicted gene expression quantitatively captures the observations under different carbon conditions, indicating that the biophysical parameters are indifferent

to the physiology. Interestingly, temperature controls the expression level in ways that are inconsistent with the prediction, revealing temperature-dependent effects that challenge current models. This work exposes the strengths and weaknesses of thermodynamic models in fluctuating environments, posing novel challenges and utility in studying physiological adaptation.

## 4.2 Introduction

Cellular physiology is inextricably tied to the extracellular environment. Fluctuations in nutrient availability and variations in temperature, for example, can drastically modulate the cell's growth rate, which is often used as a measure of the evolutionary fitness (Schaechter et al., 1958). In response to such environmental insults, cells have evolved myriad clever mechanisms by which they can adapt to their changing surroundings, many of which involve restructuring their proteome such that critical processes (i.e. protein translation) are allocated the necessary resources. Recent work exploring this level of adaptation using mass spectrometry, ribosomal profiling, and RNA sequencing have revealed that various classes of genes (termed "sectors") are tuned such that the protein mass fraction of the translational machinery is prioritized over the metabolic and catabolic machinery in nutrient replete environments (Hui et al., 2015; Klumpp and Hwa, 2014; Li et al., 2014; Schmidt et al., 2016; Scott et al., 2014). This drastic reorganization is mediated by the regulation of gene expression, relying on the concerted action of myriad transcription factors. Notably, each gene in isolation is regulated by only one or a few components (Gama-Castro et al., 2016). The most common regulatory architecture in *Escherichia coli* is the simple repression motif in which a transcriptional repressor binds to a single site in the promoter region, occluding binding of an RNA polymerase (Phillips et al., 2019; Rydenfelt et al., 2014b). The simple activation architecture, in which the simultaneous binding of an activator and an RNA polymerase amplifies gene expression, is another common mode of regulation. Combinatorial regulation such as dual repression, dual activation, or combined activation and repression can also be found throughout the genome, albeit with

lower frequency (Phillips et al., 2019). The ubiquity of the simple repression and simple activation motifs illustrate that, for many genes, the complex systems-level response to a physiological perturbation boils down the binding and unbinding of a single regulator to its cognate binding sites.

Despite our knowledge of these modes of regulation, there remains a large disconnect between concrete, physical models of their behavior and experimental validation. The simple repression motif is perhaps the most thoroughly explored theoretically and experimentally (Phillips et al., 2019) where equilibrium thermodynamic (Barnes et al., 2019; Brewster et al., 2014; Garcia and Phillips, 2011; Garcia et al., 2012; Razo-Mejia et al., 2018) and kinetic (Jones et al., 2014; Kepler and Elston, 2001; Ko, 1991; Michel, 2010) models have been shown to accurately predict the level of gene expression in a variety of contexts. While these experiments involved variations of repressor copy number, operator sequence, concentration of an external inducer, and amino acid substitutions, none have explored how the physiological state of the cell as governed by external factors influences gene expression. This is arguably one of the most critical variables one can experimentally tune to understand the role of these regulatory architectures play in cellular physiology writ large.

In this work, we interrogate the adaptability of a simple genetic circuit to various physiological stressors, namely carbon source quality and growth temperature. Following the aforementioned thermodynamic models, we build upon this theory-experiment dialogue by using environmental conditions as an experimentally tunable variable and determine their influence on various biophysical parameters. Specifically, we use physiological stressors to tune the growth rate. One mechanism by which we modulate the growth rate is by exchanging glucose in the growth medium for the poorer carbon sources glycerol and acetate, which decrease the growth rate by a factor of  $\approx 1.5$  and  $\approx 4$  compared to glucose, respectively. We hypothesize that different carbon sources should, if anything, only modulate the repressor copy number seeing as the relationship between growth

rate and total protein content has been rigorously quantified (Jun et al., 2018; Li et al., 2014; Schaechter et al., 1958; Schmidt et al., 2016). Using single-cell time-lapse fluorescence microscopy, we directly measure the copy number of the repressor in each condition. Under a simple hypothesis, all other parameters should be unperturbed, and we can thus rely on previously determined values to make parameter-free predictions of the fold-change in gene expression.

Despite the decrease in growth rate, both the fold-change in gene expression and the repressor copy number remains largely unaffected. We confirm this is the case by examining how the effective free energy of the system changes between carbon sources, a method we have used previously to elucidate parametric changes due to mutations within a transcription factor (Chure et al., 2019) and has been extensively discussed in Chapter 3. This illustrates that the energetic parameters defining the fraction of active repressors and their affinity for the DNA are ignorant of the carbon-dependent physiological states of the cell. Thus, in this context, the values of the biophysical parameters determined in one condition can be used to draw predictions in others.

We then examine how variations in temperature influence the transcriptional output. Unlike in the case of carbon source variation, temperature dependence is explicit in our model: the repressor-DNA binding energy and the energetic difference between the active and inactive states of the repressor are scaled to the thermal energy of the system at 37° C. This is defined via the Boltzmann distribution which states that the probability of a state  $p_{state}$  is related to the energy of that state  $\varepsilon_{state}$  as

$$p_{state} \propto e^{-\varepsilon_{state}/k_B T}, \quad (4.1)$$

where  $k_B$  is the Boltzmann constant and  $T$  is the temperature of the system. Given knowledge of  $T$  for a particular experiment, we can easily draw predictions of the fold-change in gene expression. However, we find the fold-change in gene expression is inconsistent with this simple model, revealing an incomplete description of the energetics. We then examine how entropic effects neglected in the initial esti-

mation of the energetic parameters may play an important role; a hypothesis that is supported when we examine the change in the effective free energy.

The results presented here are, to our knowledge, the first attempts to systematically characterize the growth-dependent effects on biophysical parameters in thermodynamic models of transcription. While some parameters of our model are affected by changing the growth rate, they change in ways that are expected or fall close within our *a priori* predictions, suggesting that such modeling can still be powerful in understanding how adaptive processes influence physiology at the level of molecular interactions.

### 4.3 Results

#### Thermodynamic model

This chapter builds off the theoretical details presented in Chapters 2-3 of this thesis. Here, we once again consider the simple repression motif in which expression of a target gene is regulated by the action of a single allosteric repressor. The key measurable quantity of the following work is the fold-change in expression of the regulated gene. Thermodynamic models described previously (Garcia and Phillips, 2011; Phillips et al., 2019; Razo-Mejia et al., 2018) and in Chapters 2-3 of this work result in a succinct input-output function to quantitatively describe the fold-change in gene expression and is of the form

$$\text{fold-change} = \left(1 + p_{act}(c) \frac{R}{N_{NS}} e^{-\Delta\varepsilon_{RA}/k_B T}\right)^{-1}, \quad (4.2)$$

where  $R$  is the total number of allosteric repressors per cell,  $N_{NS}$  is the number of nonspecific binding sites for the repressor,  $\Delta\varepsilon_{RA}$  is the repressor-DNA binding energy, and  $k_B T$  is the thermal energy of the system. The prefactor  $p_{act}(c)$  defines the probability of the repressor being in the active state at a given concentration of inducer  $c$ . In the absence of inducer,  $p_{act}(c = 0)$  can be written as

$$p_{act}(c = 0) = \left(1 + e^{-\Delta\varepsilon_{AI}/k_B T}\right)^{-1}, \quad (4.3)$$

where  $\Delta\varepsilon_{AI}$  is the energy difference between the active and inactive states. Conditioned on only a handful of experimentally accessible parameters, this model has

been verified using the well-characterized LacI repressor of *Escherichia coli* where parameters such as the repressor copy number and DNA binding affinity (Garcia and Phillips, 2011), copy number of the regulated promoter (Brewster et al., 2014), and the concentration of an extracellular inducer (Razo-Mejia et al., 2018, Chapter 2) can be tuned over orders of magnitude. Chapter 3 and the associated publication (Chure et al., 2019) illustrated that this model permits the mapping of mutations within the repressor protein directly to biophysical parameters in a manner that permits accurate prediction of double mutant phenotypes. All of these applications, however, have been performed in a single physiological state where cells are grown in a glucose-supplemented minimal medium held at 37° C with aeration. In this work, we challenge this model by changing the environmental conditions away from this gold-standard condition, perturbing the physiological state of the cell.

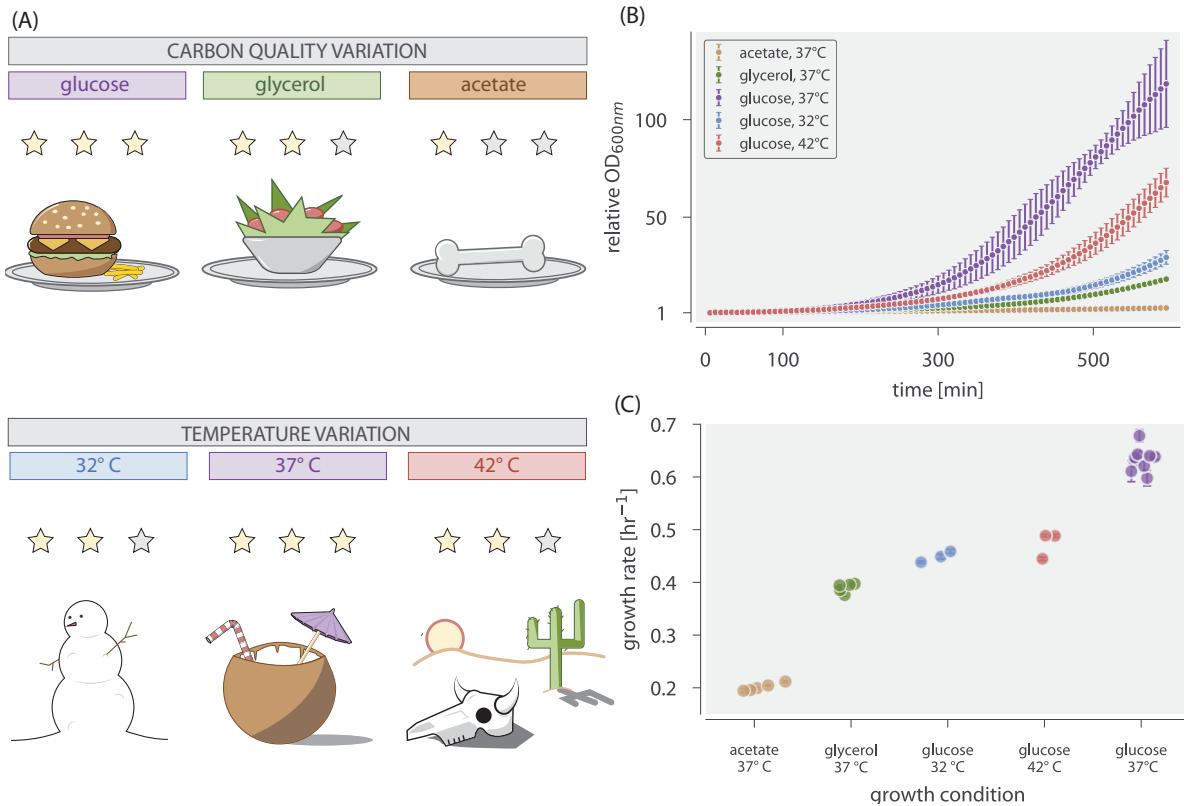
### Experimental Setup

Seminal studies from the burgeoning years of bacterial physiology have demonstrated a strong dependence of the total cellular protein content on the growth rate (Schaechter et al., 1958), a relationship which has been rigorously quantified in recent years using mass spectrometry (Schmidt et al., 2016) and ribosomal profiling (Li et al., 2014). Their combined results illustrate that modulation of the growth rate, either through controlling the available carbon source or the temperature of the growth medium, significantly alters the physiological state of the cell, triggering the reallocation of resources to prioritize expression of ribosome-associated genes. Eq. 4.2 has no explicit dependence on the available carbon source but does depend on the temperature through the energetic parameters  $\Delta\varepsilon_R$  and  $\Delta\varepsilon_{AI}$  which are defined relative to the thermal energy,  $k_B T$ . With this parametric knowledge, we are able to draw quantitative predictions of the fold-change in gene expression in these physiologically distinct states.

We modulated growth of *Escherichia coli* by varying either the quality of the available carbon source (differing ATP yield per C atom) or the temperature of

the growth medium (Fig. 4.1 (A)). All experiments were performed in a defined M9 minimal medium supplemented with one of three carbon sources – glucose, glycerol, or acetate – at concentrations such that the total number of carbon atoms available to the culture remained the same. These carbon sources have been shown to drastically alter growth rate and gene expression profiles, indicating changes in the proteomic composition and distinct physiological states. These carbon sources yield an approximate four-fold modulation of the growth rate with doubling times ranging from  $\approx 220$  minutes to  $\approx 65$  minutes in an acetate or glucose supplemented medium, respectively Fig. 4.1. While the growth temperature was varied over 10° C, both 32° and 42° C result in approximately the same doubling time of  $\approx 90$  min, which is 1.5 times slower than the optimal temperature of 37° C (Fig. 4.1 (B) and (C)).

The growth rate dependence of the proteome composition suggests that changing physiological conditions could change the total repressor copy number of the cell. As emphasized in Chapter 2, it can be difficult to differentiate between a change in repressor copy number  $R$  and the allosteric energy difference  $\Delta\varepsilon_{AI}$  as there are many combinations of parameter values that yield the same fold-change. To combat this degeneracy, we used a genetically engineered strain of *E. coli* in which the expression of the repressor copy number and its regulated gene product (YFP) can be simultaneously measured. This strain, used previously to interrogate the transcription factor titration effect (Brewster et al., 2014), is diagrammed in Fig. 4.2 (A). A dimeric form of the LacI repressor N-terminally tagged with an mCherry fluorophore is itself regulated through the action of the TetR repressor whose level of activity can be modulated through the addition of the allosteric effector anhydrous tetracycline (ATC). This dual repression genetic circuit allows for the expression of the LacI repressor to be tuned over several orders of magnitude. This is demonstrated in Fig. 4.2 (B) where a titration of ATC in the growth medium results in a steady increase in the expression of the LacI-mCherry gene product (red lines and points) which in turn represses expression of the YFP reporter (yellow lines and points).



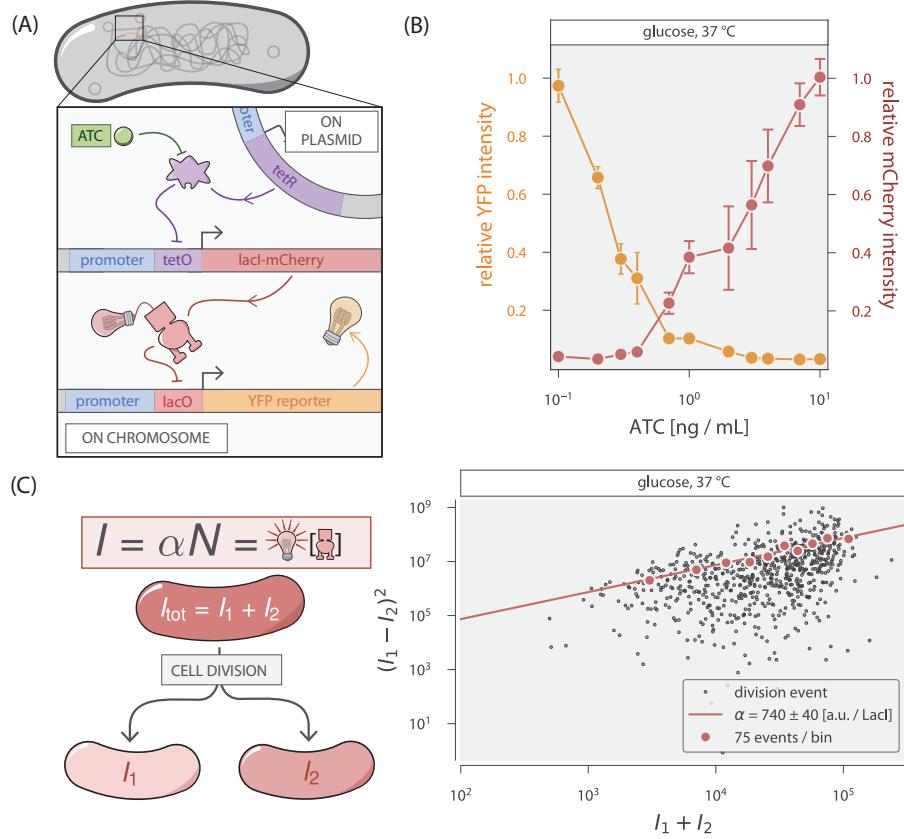
**Figure 4.1: Control of growth rate via environmental factors.** (A) Bacterial growth can be controlled by varying the available carbon source (top panel) or temperature (bottom panel). (B) Bulk bacterial growth curves under all conditions illustrated in (A). The  $y$ -axis is the optical density measurements at 600 nm relative to the initial value. Interval between points is  $\approx 6$  min. Points and errors represent the mean and standard deviation of three to eight biological replicates. (C) Inferred maximum growth rate of each biological replicate for each condition. Points represent the doubling time computed from the maximum growth rate. Error bars correspond to the standard deviation of the inferred growth rate. Where not visible, error bars are smaller than the marker. The Python code (`ch4_fig1.py`) used to generate this figure can be found on the thesis GitHub repository.

While the mCherry fluorescence is proportional to the repressor copy number, it is not a direct measurement as the fluorescence of a single LacI-mCherry dimer is unknown *a priori*. Using video microscopy, we measure the partitioning statistics of the fluorescence intensity into two sibling cells after division (Fig. 4.2 (C)). This method, described in detail in the Materials & Methods as well as in Rosenfeld et al. (2002), Rosenfeld et al. (2005), and Brewster et al. (2014), reveals a linear relationship between the variance in intensity between two sibling cells and the intensity of the parent cell, the slope of which is equal to the brightness of a single LacI repressor. Since this measurement is performed simultaneously with measurement of the expression of the YFP reporter, this calibration factor was determined for each unique experimental replicate. We direct the reader to the supplemental Chapter 8 for a more thorough discussion of this inference.

### Scaling of Gene Expression With Growth Rate

Given the single-cell resolution of our experimental method, we examined how the cell volume and repressor copy number scaled across the different growth conditions at different levels of ATC induction. In agreement with the literature (Jun et al., 2018; Schaechter et al., 1958; Shehata and Marr, 1975) our measurement reveals a strong linear dependence of the cell volume on the choice of carbon source, but no significant dependence on temperature (Fig. 4.3 (A) and (B)). Additionally, these findings are consistent across different ATC induction regimes. Together, these observations confirm that the particular details of our experimental system does not introduce unintended physiological consequences.

Using a fluorescence calibration factor determined for each experimental replicate (see Fig. 4.2 (C) and the Materials & Methods), we estimated the number of repressors per cell from snapshots of the mCherry signal intensity of each induction condition. Fig. 4.3 (C) reveals a remarkable insensitivity of the repressor copy number on the growth rate under different carbon sources. Despite the change in cellular volume, the mean number of repressors expressed at a given induction condition is within error between all carbon sources. Previous work using mass



**Figure 4.2: Control and quantification of repressor copy number.** (A) The dual repression expression system. The inducible repressor TetR (purple blob) is expressed from a low-copy-number plasmid in the cell and represses expression of the LacI-mCherry construct by binding to its cognate operator (tetO). In the presence of anhydrous tetracycline (ATC, green sphere), the inactive state of TetR becomes energetically favorable, permitting expression of the LacI-mCherry construct (red). This in turn binds to the lacO operator sequence repressing the expression of the reporter Yellow Fluorescent Protein (YFP, yellow lightbulb). (B) An ATC titration curve showing anticorrelated YFP (yellow) and mCherry (red) intensities. Reported values are scaled to the maximum mean fluorescence for each channel. Points and errors correspond to the mean and standard error of eight biological replicates. (C) Determination of a fluorescence calibration factor. After cessation of LacI-mCherry expression, cells are allowed to divide, partitioning the fluorescently tagged LacI repressors into the two daughter cells (left panel). The total intensity of the parent cell is equivalent to the summed intensities of the daughters. The squared fluctuations in intensity of the two sibling cells is linearly related to the parent cell with a slope  $\alpha$ , which is the fluorescence signal measured per partitioned repressor (right panel). Black points represent single divisions and red points are the means of 50 division events. Line corresponds to linear fit to the black points with a slope of  $\alpha = 740 \pm 40$  a.u. per LacI. The Python codes (ch4\_fig2b.py) and (ch4\_fig2c.py) used to generate this figure can be found on the thesis GitHub repository.

spectrometry, a higher resolution method, has shown that there is a slight dependence of LacI copy number on growth rate expressed from its native promoter (Schmidt et al., 2016). It is possible that such a dependence exists in our experimental setup, but is not detectable with our lower resolution method. We also observe an insensitivity of copy number to growth rate when the temperature of the system is tuned (Fig. 4.3 (D)), though two aberrant points with large error obfuscates the presence of a growth rate dependence at high concentrations of ATC. For concentrations below 7 ng / mL, however, the repressor copy number remains constant across conditions. With no significant change in the repressor copy number and thus no dependence on the carbon source in our theoretical model, we can immediately draw predictions of the fold-change in gene expression in different growth media.

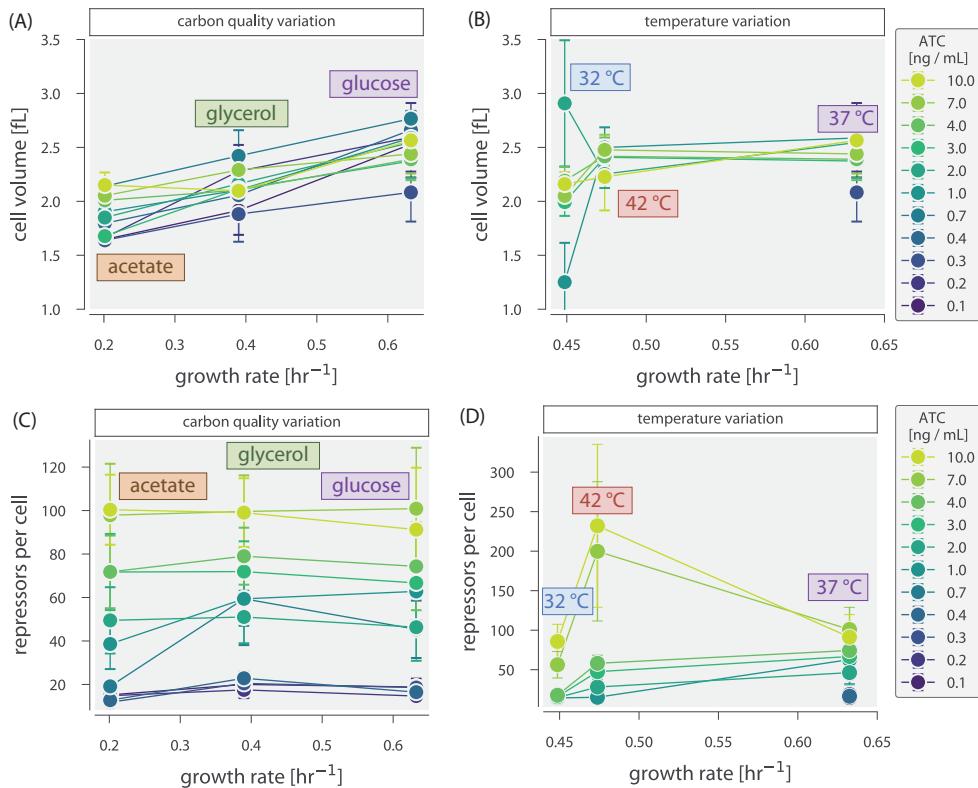
#### 4.4 Fold-change dependence on carbon quality

Given *a priori* knowledge of the biophysical parameter values (Garcia and Phillips, 2011; Razo-Mejia et al., 2018) present in Eq. 4.2 and Eq. 4.3 and direct measurement of the repressor copy number, we made measurements of the fold-change in gene expression for each growth medium to test the prediction (Fig. ?? (A)). We find that the measurements fall upon the predicted theoretical curve within error, suggesting that the values of the energetic terms in the model are insensitive to changing carbon sources. This is notable as glucose, glycerol, and acetate are metabolized via different pathways, changing the metabolite and protein composition of the cytosol (Kim et al., 2007; Martínez-Gómez et al., 2012). This result underscores the utility of these thermodynamic parameters as quantitative traits in the study of growth-condition dependent gene expression.

We have shown in the preceding chapters that Eq. 4.2 can be rewritten into a Fermi function of the form

$$\text{fold-change} = \frac{1}{1 + e^{-F/k_B T}}, \quad (4.4)$$

where  $F$  is the effective free energy difference between the repressor bound and



**Figure 4.3: Scaling of cell size and repressor expression as a function of maximum growth rate.** Dependence of cell volume on maximum growth rate under varying (A) carbon sources and (B) temperatures. Points and errors correspond to the mean and standard errors of five to eight biological replicates. The cell volume was calculated by approximating each cell as a spherocylinder and using measurements of the short-and long axis lengths of each segmentation mask. Measured volumes are from snapshots of a nonsynchronously growing culture. The measured repressor copy number for each ATC induction condition (colored lines) as a function of the growth rate for various (C) carbon sources and (D) temperatures. Points and errors represent the mean and standard error for five to eight biological replicates. Colors correspond to the ATC induction concentration ranging from 10 ng/mL (yellow) to 0.1 ng/mL (black).

unbound states of the promoter as described in Chapters 2 and 3. For the case of an allosteric simple repression architecture, and given knowledge of the values of the biophysical parameters,  $F$  can be directly calculated as

$$F = k_B T \left[ -\log \left( 1 + e^{-\Delta\varepsilon_{AI}/k_B T} \right)^{-1} - \log \frac{R}{N_{NS}} + \frac{\Delta\varepsilon_{RA}}{k_B T} \right]. \quad (4.5)$$

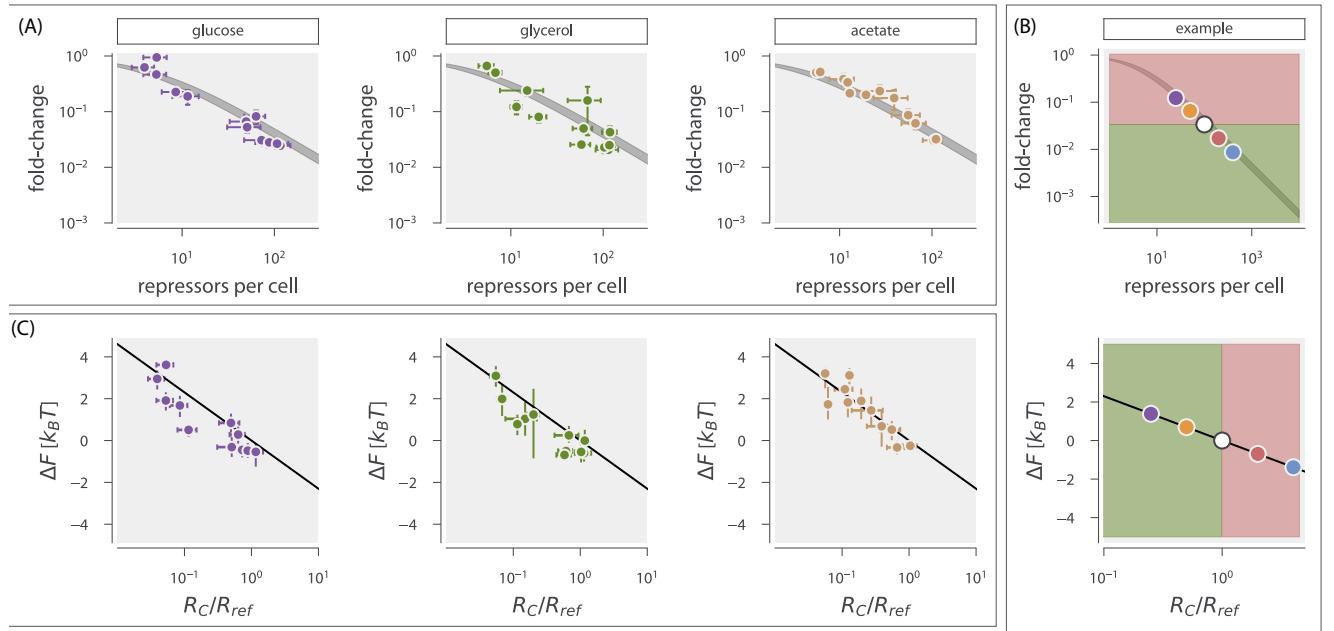
We have recently interrogated how this formalism can be used to map mutations within the repressor to biophysical parameters by examining the difference in free energy between a mutant and reference (wild-type) strain,  $\Delta F = F_{mut} - F_{ref}$  (Chure et al., 2019 and Chapter 3). This approach revealed that different parametric changes yield characteristic response functions to changing inducer concentrations. Rather than using wild-type and mutant variants of the repressor, we can choose a reference condition and compare how the free energy changes between different growth media. Here, we choose the reference condition to be a sample grown at 37° C with glucose as the available carbon source and a repressor copy number  $R = 100$  per cell. Under the hypothesis that the only variable parameter in these growth conditions is the repressor copy number  $R$ , the shift in free energy  $\Delta F$  becomes

$$\Delta F = F_C - F_{ref} = -\log \frac{R_C}{R_{ref}}, \quad (4.6)$$

where  $F_C$  and  $R_C$  correspond to the free energy and repressor copy number of the different growth conditions. This concise prediction serves as a quantitative measure of how robust the energetic parameters  $\Delta\varepsilon_{RA}$  and  $\Delta\varepsilon_{AI}$  are in units of  $k_B T$  (Fig. ?? (B)). In using free energy shifts as a diagnostic, one can immediately determine the effect of the perturbation on the parameter values by quantifying the disagreement between the observed and predicted  $\Delta F$  as the parameters and the free energy are both in the same natural units.

We inferred the observed free energy for the fold-change measurements shown in Fig. ?? (A) (as described previously (Chure et al., 2019) and in Chapter 7) and compared it to the theoretical prediction of Eq. 4.6, shown in Fig. ?? (C). Again, we see the observed change in free energy is in strong agreement with our theoretical

predictions. This agreement indicates that the free energy shift  $\Delta F$  can be used in multiple contexts to capture the energetic consequences of physiological and evolutionary perturbations between different states of the system. The insensitivity of the biophysical parameters to these distinctly different physiological states demonstrates that  $\Delta\varepsilon_{AI}$  and  $\Delta\varepsilon_R$  are material properties of the repressor defined by the intricate hydrogen bonding networks of its constituent amino acids rather than by the chemical constituency of its surroundings. Tuning temperature, however, can change these material properties.



### ## Fold-change dependence on temperature variation

Unlike the identity of the carbon source, the temperature of the system is explicitly stated in Eq. 4.2 and Eq. 4.3 where  $\Delta\varepsilon_{RA}$  and  $\Delta\varepsilon_{AI}$  are defined relative to the thermal energy of the system in which they were determined. This scaling is mathematically quantified as  $k_B T$  dividing the exponentiated terms in Eq. 4.2 and Eq. 4.3. As all biophysical parameters were determined at a reference temperature of 37° C, any change in the growth temperature must be included as a correction factor. The simplest approach is to rescale the energy by the relative change in temperature. This is a simple multiplicative factor of  $\phi_T = k_B T_{ref} / k_B T_{exp}$  where  $T_{ref}$  is the reference temperature of 37° C and  $T_{exp}$  is the experimental temperature. This is

an intuitive result since an increase in temperature relative to the reference results in  $\phi_T < 1$ , weakening the binding. Similarly, decreasing the temperature scales  $\phi_T > 1$ , strengthening the binding relative to that of the reference temperature.

Fig. 4.4 (A) shows the measured fold-change in gene expression (points) plotted against the theoretical prediction with this correction factor (orange line). It is immediately evident that a simple rescaling of the energetic parameters is not sufficient for the 32° C condition and slightly underestimates the fold-change in the 42° C condition. To identify the source of this disagreement, we can again examine the free energy shift  $\Delta F$ . As both  $\Delta\epsilon_{AI}$  and  $\Delta\epsilon_R$  are scaled to the thermal energy,  $\Delta F$  defined as  $F_T - F_{ref}$  can be directly calculated as

$$\Delta F = k_B T_{exp} \left[ -\log \frac{1 + e^{-\frac{\Delta\epsilon_{AI}}{k_B T_{ref}}}}{1 + e^{-\frac{\phi_T \Delta\epsilon_{AI}}{k_B T_{exp}}}} - \log \frac{R_T}{R_{ref}} \right] + \Delta\epsilon_R (1 - \phi_T). \quad (4.7)$$

This prediction along with the empirically determined  $\Delta F$  is shown in Fig. 4.4 (B). Again, we see that this simple correction factor significantly undershoots or overshoots the observed  $\Delta F$  for 32° C and 42° C, respectively, indicating that there are temperature dependent effects that are not accounted for in the simplest null model of temperature dependence.

The model described by Eq. 4.2 and Eq. 4.3 and subsumes the myriad rich dynamical processes underlying protein binding and conformational changes into two effective energies,  $\Delta\epsilon_R$  and  $\Delta\epsilon_{AI}$ . By no means is this done to undercut the importance of these details in transcriptional regulation. Rather, it reduces the degrees of freedom in this objectively complex system to the set of the details critical to particular conditions in which we want to draw predictions. All prior dissections of this thermodynamic model have been performed at a single temperature, abrogating the need to consider temperature dependent effects. As we now vary temperature, we must consider details that are swept into the effective energies.

The model presented here only considers entropy by enumerating the multiplicity of states in which the repressor can bind to the DNA nonspecifically, resulting in

terms of the form  $R / N_{NS}$ . However, there are many other temperature-dependent entropic contributions to the effective energies such as the fraction of repressors bound to DNA versus in solution (Elf et al., 2007; Kao-Huang et al., 1977), the vibrational entropy of the repressor (Goethe et al., 2015), or conformational entropy of the genome (Driessens et al., 2014; Mondal et al., 2011). We can consider the effective energies  $\Delta\varepsilon_R$  and  $\Delta\varepsilon_{AI}$  as having generic temperature dependent-entropic components  $\Delta S_R$  and  $\Delta S_{AI}$ ,

$$\Delta\varepsilon_R = \Delta H_R - T\Delta S_R, \quad (4.8)$$

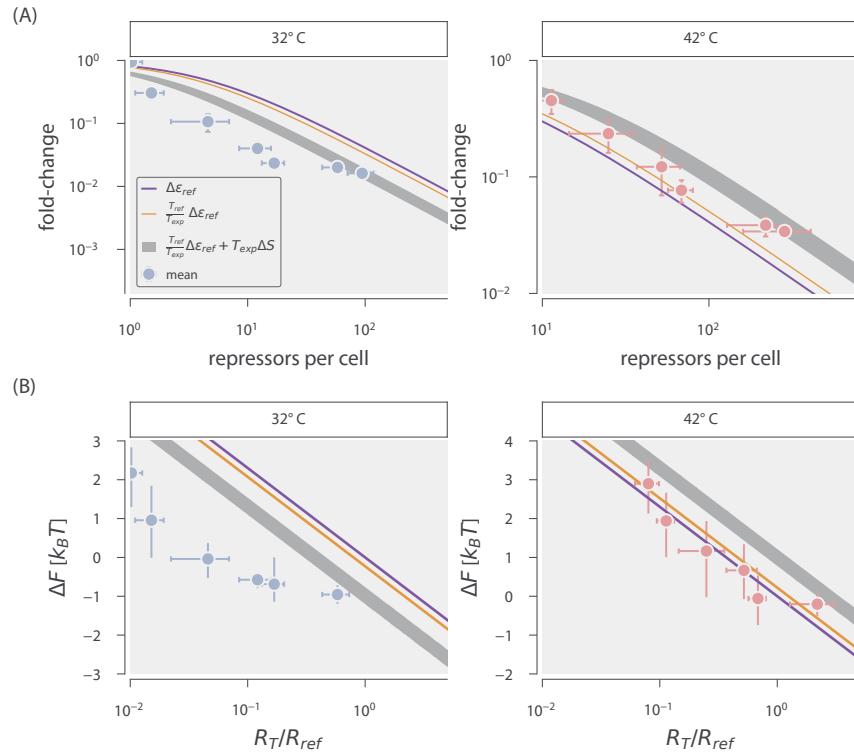
and

$$\Delta\varepsilon_{AI} = \Delta H_{AI} - T\Delta S_{AI}, \quad (4.9)$$

where  $\Delta H_R$  and  $\Delta H_{AI}$  is the enthalpic contribution to the energies  $\Delta\varepsilon_R$  and  $\Delta\varepsilon_{AI}$ , respectively. Given the fold-change measurements at 32° C and 42° C, we estimated the entropic parameters  $\Delta S_R$  and  $\Delta S_{AI}$  under the constraints that at 37° C,  $\Delta\varepsilon_R = -13.9 k_B T$  and  $\Delta\varepsilon_{AI} = 4.5 k_B T$  (see supplemental Chapter 8 for more discussion on this parameter estimation). The grey shaded lines in Fig. 4.4 show the result of this fit where the width represents the 95% credible region of the prediction given the estimated values of  $\Delta S_R$  and  $\Delta S_{AI}$ . Including this phenomenological treatment of the entropy improves the prediction of the fold-change in gene expression (Fig. 4.4 (A)) as well as shift in free energy (Fig. 4.4 (B)). This phenomenological description suggests that even small shifts in temperature can drastically alter the expression of a genetic circuit simply by tuning hidden entropic effects rather than scaling the difference in affinity between specific and nonspecific binding.

## 4.5 Discussion

The past century of work in bacterial physiology has revealed a rich molecular complexity that drives cellular growth rate (Jun et al., 2018). A key finding of this body of work is that the composition of the proteome is highly dependent on the specific growth condition, with entire classes of genes being up- or down-regulated to ensure that enough resources are allocated towards maintaining a pool of active ribosomes (Hui et al., 2015; Scott et al., 2014). These studies have led to a



**Figure 4.4: Temperature effects on the fold-change in gene expression and free energy.** (A) The fold-change in gene expression for growth in glucose supplemented medium at 32° C (left) and 42° C (right). Points and errors correspond to the mean and standard error of five biological replicates. Predictions of the fold-change are shown without correcting for temperature (purple), with multiplicative scaling (orange), and with an entropic penalty (grey). The width of the prediction of the entropic penalty is the 95% credible region. (B) Predicted and observed shifts in the free energy for growth in glucose medium at 32° C (left) and 42° C (right). Points correspond to the median of the inferred shift in free energy. Vertical error bars indicate the bounds of the 95% credible region. Horizontal position and error corresponds to the mean and standard error for the repressor count over five to eight biological replicates. The Python code (`ch4_fig5.py`) used to generate this figure can be found on the thesis GitHub repository.

coarse-grained view of global gene expression where physiological perturbations substantially change the molecular composition of the cell, obfuscating the utility of using thermodynamic models of individual regulatory elements across physiological states. In this work, we rigorously examine how robust the values of the various biophysical parameters are to changes in cellular physiology.

We first examined how nutrient fluctuations dictate the output of this architecture. We took three carbon sources with distinct metabolic pathways and varying quality and measured the level of gene expression, hypothesizing that the values of the biophysical parameters to be independent of the growth medium. We found that even when the growth rate is varied across a wide range (220 minute doubling time in acetate to 60 minute doubling time in glucose supported medium), there is no significant change to the fold-change in gene expression or in the expression of the transcription factor itself, within the resolution of our experiments. Given numerous quantitative studies of the proteomic composition reveal a dependence on protein content with growth rate (Hui et al., 2015; Li et al., 2014; Schmidt et al., 2016), we find this robustness to be striking.

Schmidt et al. (2016) found that the native expression of LacI has a weak positive correlation with the growth rate. The native LacI promoter region is solely regulated by activation via the cAMP Receptor Protein (CRP), a broadly acting dual regulator in *E. coli* (Gama-Castro et al., 2016). This is in contrast to the LacI expression system used in the present work where the promoter is negatively regulated by the TetR repressor, itself expressed from a low-copy number plasmid. Furthermore, the expression of LacI in this work is tuned by the addition of the allosteric effector of TetR, ATC, adding yet another layer of allosteric regulation on LacI expression. The significant difference in the regulatory mechanisms between the native and synthetic circuit used in this work makes the two findings difficult to directly compare. Regardless, our finding that the fold-change in gene expression is unaltered from one carbon source to another illustrates that the values of the biophysical parameters  $\Delta\epsilon_R$  and  $\Delta\epsilon_{AI}$  remain unperturbed, permitting

quantitative prediction of gene expression across numerous physiological states.

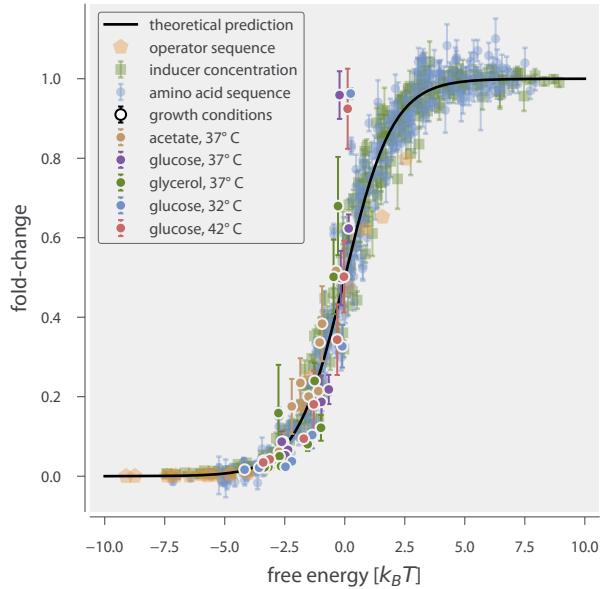
However, in varying the temperature, we find that the predictive utility of the biophysical parameters values determined at 37° C is diminished, indicating that there are hidden effects not explicitly accounted for in our thermodynamic model. The measurements of the fold-change in gene expression are under- or over-estimated when the temperature is increased or decreased, respectively, when one simply rescales the energetic terms by the relative change in temperature. There are many features of transcriptional regulation that are not explicitly considered in our coarse-graining of the architecture into a two-state model. Recently, it has been suggested that the phenomenon of allostery writ large should be framed in the language of an ensemble of states rather than a simple active/inactive distinction (Motlagh et al., 2014). While our recent work illustrates that a two-state rendering of an allosteric repressor is highly predictive in a variety of situations (Chure et al., 2019; Razo-Mejia et al., 2018), we must now consider details which are dependent on the temperature of the system. In Fig. 4.4, we demonstrate that incorporating a temperature-dependent entropic cost to the energetic terms significantly improves the description of the experimental data. This is not to say, however, that this is now an open-and-closed case for what precisely defines this entropic cost. Rather, we conclude that the phenomenology of the temperature dependence can be better described by the inclusion of a correction factor that is linearly dependent on the system temperature. Biology is replete with phenomena which can introduce such an effect, including changes to the material properties of the repressor and DNA (Goethe et al., 2015; Mondal et al., 2011), excluded volume effects (Driessens et al., 2014), and solubilities (Elf et al., 2007; Kao-Huang et al., 1977; Yakovchuk et al., 2006). Understanding the mechanistic underpinnings of temperature dependence in elasticity theory was borne out of similar phenomenological characterization (Friedel, 1974) and required a significant dialogue between theory and experiment (Phillips, 2001). Further work is now needed to develop a theory of temperature effects in the regulation of gene expression.

The effective free energy  $F$ , as defined in Eq. ??, is a state variable of the simple repression regulatory architecture. This is illustrated in Fig. 4.5 where fold-change measurements from a wide array of conditions (and measurement techniques) can be collapsed onto the same theoretical description. Evolutionary perturbations (such as mutations in the operator or repressor sequence), physiological changes (such as modulations of the growth rate), or changes in the level of activity of the repressor (due to changes in inducer concentration) do not change the fundamental physics of the system and can all be described by changes in the free energy relative to one another. While such a statement is not “surprising”, we can now say it with quantitative confidence and use this principle to probe the degree to which physiological perturbations influence the biophysical parameters writ large. With such a framework in hand, we are in the auspicious position to take a predictive approach towards understanding how this regulatory architecture evolves in experimental settings, shedding light on the interplay between biophysical parameters, organismal fitness, and the fundamental forces of evolution.

## 4.6 Materials & Methods

### Bacterial Strains and Growth Media

Three genotypes were primarily used in this work, all in the genetic background of *Escherichia coli* MG1655-K12 and all derived from those used in Brewster et al. (2014). For each experiment, an autofluorescence control was used which contained no fluorescent reporters (except for a CFP volume marker used for segmentation in Brewster et al., 2014) which had the *lacI* and *lacZYA* genes deleted from the chromosome. The constitutive expression strain ( $\Delta lacI; \Delta lacZYA$ ) included a YFP reporter gene integrated into the *galK* locus of the chromosome along with a kanamycin resistance cassette. The experimental strains in which LacI expression was controlled contained a *lacI-mCherry* fluorescent fusion integrated into the *ybcN* locus of the chromosome along with a chloramphenicol resistance cassette. This cassette was later deleted from the chromosome using FLP/FRT recombination (Schlake and Bode, 1994; Zhu and Sadowski, 1995). The strain was then trans-



**Figure 4.5: A singular theoretical description for the molecular biophysics of physiological and evolutionary adaptation in the simple repression regulatory architecture.** Measurements of the fold-change in gene expression varying the sequence of the operator site (orange pentagon, Garcia and Phillips, 2011), concentration of the extracellular inducer (green squares, Razo-Mejia et al., 2018 and Chapter 2), amino-acid sequence of the repressor (blue points, Chure et al., 2019 and Chapter 3), and the various growth conditions queried in this work can be collapsed as a function of the effective free energy. Error bars correspond to the standard error of five to 10 biological replicates. The Python code (`ch4_fig6.py`) used to generate this figure can be found on the thesis GitHub repository.

formed with plasmid (pZS3-pN25-tetR following notation described in Lutz and Bujard, 1997) constitutively expressing the TetR repressor along with a chloramphenicol resistance cassette. All bacterial strains and plasmids used in this work are reported in the supplemental Chapter 8.

### Bacterial Growth Curves

Bacterial growth curves were measured in a multi-well plate reader (BioTek Cytation5) generously provided by the David Van Valen lab at Caltech. Cells constitutively expressing YFP were grown overnight to saturation in LB broth (BD Medical) at 37° C with aeration. Once saturated, cells were diluted 1000 fold into 50 mL of the desired growth medium and were allowed to grow at the appropriate exper-

imental temperature with aeration for several hours until an  $OD_{600nm} \approx 0.3$  was reached. Cells were then diluted 1:10 into the desired growth media at the desired temperature. After being thoroughly mixed, 500  $\mu\text{L}$  aliquots were transferred to a black-walled 96-well plate (Brooks Automation Incorporated, Cat No. MGB096-1-2-LG-L), leaving two rows and two columns of wells on each side of the plate filled with sterile growth medium to serve as blanks and buffer against temperature variation. The plate was then transferred to the pre-warmed plate reader.  $OD_{600nm}$  measurements were made every five minutes for 12 to 24 hours until cultures had saturated. In between measurements, the plate incubated at the appropriate temperature with a linear shaking mode. We found that double-orbital shaking modes led to the formation of cell aggregates which gave inconsistent measurements.

### **Estimation of Bacterial Growth Rate**

Non-parametric estimation of the maximum growth rate was performed using the FitDeriv Python package as described in Swain et al. (2016). Using this approach, the bacterial growth curve is modeled as a Gaussian process in which the measured growth at a given time point is modeled as a Gaussian distribution whose mean is dependent on the mean of the neighboring time points. This allows for smooth interpolation between adjacent measurements and calculation of second derivatives without an underlying parametric model. The reported growth rates are the maximum value inferred from the exponential phase of the experimental growth curve.

### **Growth Conditions**

Parent strains (autofluorescence control,  $\Delta lacI$  constitutive control, and the ATC-inducible LacI-mCherry strain) were grown in LB Miller broth (B.D. Medical, Cat. No. 244620 ) at 37° C with vigorous aeration until saturated. Cells were then diluted between 1000 and 5000 fold into 3 mL of M9 minimal medium (B.D. Medical, Cat. No. 248510). The bacterial strain expressing the tetracycline-inducible LacI-mCherry was diluted into six 3 mL cultures with differing concentrations of ATC

(Chemodex, Cat. No. CDX-A0197-T78) ranging from 0.1 to 10 ng / mL to induce expression of the transcription factor. These concentrations were reached by dilution from 1  $\mu$ g / mL stock in 50% ethanol. All cultures were shielded from ambient light using either aluminum foil or via an enclosure and were grown at the appropriate experimental temperature with aeration until an OD<sub>600nm</sub> of approximately 0.25 – 0.35. Due to pipetting errors, cultures reached OD<sub>600nm</sub>  $\approx$  0.3 at slightly different points in time. To ensure that strains could be directly compared, all strains were back diluted by several fold until equivalent. When all samples reached the appropriate OD<sub>600nm</sub>, the cells were harvested for imaging.

### **Imaging Sample Preparation**

Prior to the preparation of cell cultures for imaging, a 2% (w/v) agarose substrate (UltraPure, Thermo Scientific) was prepared and allowed to reach room temperature. For experiments conducted at 42°C, 4% (w/v) agarose substrates were prepared. Briefly, the agarose was mixed with the appropriate growth medium in a 50 mL conical polystyrene tube and then microwaved until molten. A 300 to 500  $\mu$ L aliquot was then sandwiched between two glass coverslips to ensure a flat imaging surface. Once solidified, the agarose pads were cut into squares approximately 0.5 cm per side.

To determine the calibration factor between fluorescence and protein copy number, production the fluorophore in question must be halted such that all differences in intensity between two daughter cells result from binomial partitioning of the fluorophores and not from continuing expression. This was achieved by removing the anhydrous tetracycline inducer from the growth medium through several washing steps as outlined in Brewster et al. (2014). Aliquots of 100  $\mu$ L from each ATC-induced culture were combined and pelleted at 13000 $\times g$  for 2 minutes. The supernatant (containing ATC) was aspirated and replaced with 1 mL of M9 growth medium without ATC. The pellet was resuspended and pelleted at 13000 $\times g$ . This wash step was repeated three times to ensure residual ATC had been removed and LacI-mCherry production was halted. After the final wash, the cell pellet was

resuspended in 1 mL of M9 medium and diluted ten-fold. A 1  $\mu$ L aliquot of the diluted mixture was then spotted onto an agarose substrate containing the appropriate carbon source. The remaining bacterial cultures (autofluorescence control, constitutive expression control, and the ATC-induced samples) were diluted ten-fold into sterile M9 medium. This dilution was thoroughly mixed and 1  $\mu$ L aliquots were spotted onto agarose substrates lacking the carbon source.

Once the spotted cells had dried onto the agarose substrates (about 5 to 10 minutes after deposition), the agarose pads were inverted and pressed onto a glass bottom dish (Electron Microscopy Sciences, Cat. No. 70674-52) and sealed with parafilm. Strips of double stick tape were added to the edge of the dish to help immobilize the sample on the microscope stage and minimize drift.

### Microscopy

All imaging was performed on a Nikon Ti-Eclipse inverted microscope outfitted with a SOLA LED fluorescence illumination system. All images were acquired on a Andor Zyla 5.5 sCMOS camera (Oxford Instruments Group). The microscope body and stage was enclosed in a plexiglass incubation chamber (Haison, approximately 1° C regulation control) connected to an external heater. Temperature of the stage was measured via a thermometer which controlled heating of the system.

All static images (i.e. images from which fold-change and repressor counts were calculated) were measured in an identical manner. Ten to fifteen fields of view containing on average 25 cells were imaged using phase contrast and fluorescence excitation. Fluorescence exposures were each 5 seconds while the phase contrast exposure time was between 75 ms and 150 ms. This procedure was repeated for each unique strain and ATC induction concentration.

To compute the calibration factor for that day of imaging, time-lapse images were taken of a separate agarose pad covered in cells containing various levels of LacI-mCherry. Fifteen to twenty positions were marked, choosing fields of view containing 20 to 50 cells. Cells were allowed to grow for a period of 90 to 120

minutes (depending on the medium-dependent growth rate) with phase contrast images taken every 5 to 10 minutes. At the final time-point, both phase contrast and fluorescence images were acquired using the same settings for the snapshots. Once the experiment was completed, images were exported to .tif format and transferred to cold storage and a computational cluster for analysis.

### **Lineage Tracking**

Cells were segmented and lineages reconstructed using the deep-learning-based bacterial segmentation software SuperSegger v1.1 (Cass et al., 2017; Stylianidou et al., 2016) operated through MATLAB (Mathworks, v2017b). We found that the pre-trained network constants 100XEcM9 (packaged with the SuperSegger software) worked well for all growth conditions tested in this work. The generated files (`clist.mat`) associated with each sample and position were parsed using bespoke Python scripts to reconstruct lineages and apply appropriate filtering steps before calculating the fluorescence calibration factor. We direct the reader to the SI text for details of our data validation procedure to ensure proper lineage tracking.

### **Calculation of the Calibration Factor**

To estimate the calibration factor  $\alpha$ , we used a Bayesian definition of probability to define a posterior distribution of  $\alpha$  conditioned on intensity measurements of sibling cells after division. We direct the reader to supplemental Chapter 8 for a detailed discussion of the inferential procedures and estimation of systematic error. We give a brief description of the inference below.

We are interested in determining the fluorescence of a single LacI-mCherry repressor dimer given a set of intensity measurements of sibling cells,  $[I_1, I_2]$ . The intensity of a given cell  $I$  is related to the number of LacI-mCherry dimers it is expressing by a multiplicative factor  $\alpha$  which can be enumerated mathematically as

$$I = \alpha N, \quad (4.10)$$

where  $N$  is the total number of LacI-mCherry dimers. We can define the poste-

prior probability distribution of  $\alpha$  conditioned on the intensity measurements using Bayes' theorem as

$$g(\alpha | [I_1, I_2]) = \frac{f([I_1, I_2] | \alpha)g(\alpha)}{f([I_1, I_2])}. \quad (4.11)$$

where we have used  $g$  and  $f$  as probability densities over parameters and data, respectively. The denominator of this expression (the evidence) is equivalent to the first term of the numerator (the likelihood) marginalized over  $\alpha$ . In this work, this term serves as normalization factor and can be neglected.

Assuming that no more LacI-mCherry dimers are produced during cell division, the number of repressors that each sibling cell receives after division of the parent cell is binomially distributed with a probability  $p$ . We can make the approximation that partitioning of the repressors is even such that  $p = 1/2$ . The validity of this approximation is discussed in detail in Chapter 8. Using Eq. 4.10 and the change-of-variables formula, we can define the likelihood  $g([I_1, I_2] | \alpha)$  as

$$g([I_1, I_2] | \alpha) = \frac{1}{\alpha^k} \prod_i^k \frac{\Gamma\left(\frac{I_{1i}+I_{2i}}{\alpha} + 1\right)}{\Gamma\left(\frac{I_{1i}}{\alpha} + 1\right)\Gamma\left(\frac{I_{2i}}{\alpha} + 1\right)} 2^{-\frac{I_{1i}+I_{2i}}{\alpha}}, \quad (4.12)$$

where  $k$  is the total number of sibling pairs measured.

With a likelihood defined, we must now define a functional form for  $g(\alpha)$  which describes all prior information known about the calibration factor knowing nothing about the actual measurements. Knowing that we design the experiments such that only  $\approx 2/3$  of the dynamic range of the camera is used and  $\alpha$  cannot be less than or equal to zero, we can define a half-normal distribution with a standard deviation of  $\sigma$  as

$$g(\alpha) = \sqrt{\frac{2}{\pi\sigma^2}} \exp\left[\frac{-\alpha^2}{2\sigma^2}\right]; \forall \alpha > 0. \quad (4.13)$$

where the standard deviation is large, for example,  $\sigma = 500$  a.u. / fluorophore. We evaluated the posterior distribution using Markov chain Monte Carlo (MCMC) as is implemented in the Stan probabilistic programming language (Carpenter et al., 2017). The .stan file associated with this model along with the Python code used to execute it can be accessed on the paper website.

## Counting Repressors

Given an estimation for  $\alpha$  for each experiment, we calculate the total number of repressors per cell from

$$R = \frac{I_{mCherry}}{\alpha}. \quad (4.14)$$

However, as discussed in detail in Chapter 8, a systematic error in the repressor count is introduced due to division in the asynchronous culture between the cessation of LacI-mCherry production and the actual imaging. The entire sample preparation procedure is  $\approx 30 - 60$  min, during which time some cells complete a cell division, thereby diluting the total repressor count. To ensure that the measured number of repressors corresponded to the measured YFP intensity, we restricted our dataset for all experiments to cells that had a pole-to-pole length  $\ell \geq 3.5 \mu\text{m}$ , indicating that they had likely not undergone a division during the sample preparation.

## Code and Data Availability

All code used in this work is available on the paper website and associated GitHub repository(DOI: 0.5281/zenodo.3560369). This work also used the open-source software tools SuperSegger v.1.1(Cass et al., 2017; Stylianidou et al., 2016) for lineage tracking and FitDeriv v.1.03 (Swain et al., 2016) for the nonparametric estimation of growth rates. Raw image files are large (1.8 Tb) and are therefore available upon request. The `clist.mat` files used to calculate fold-change and to assign sibling cells can be accessed via the associated GitHub repository via (DOI: 0.5281/zenodo.3560369) or through Caltech DATA under the DOI: 10.22002/D1.1315.

*Chapter 5*

## 'WATER, WATER EVERYWHERE, NOR ANY DROP TO DRINK': HOW BACTERIA ADAPT TO CHANGES IN OSMOLARITY

A version of this chapter was published as Chure, G.\* ; Lee, H.J.\* ; Rasmussen, A.; and Phillips, R. (2018). *Connecting the Dots between Mechanosensitive Channel Abundance, Osmotic Shock, and Survival at Single-Cell Resolution*. Journal of Bacteriology 200. DOI: 10.1128/JB.00460-18 (\* contributed equally). G.C., H.J.L, and R.P. designed and planned experiments. G.C. and H.J.L performed experiments. H.J.L constructed bacterial strains. A.R. performed electrophysiology experiments. G.C. performed data analysis and figure generation. G.C. and R.P. wrote the manuscript

### 5.1 Abstract

Rapid changes in extracellular osmolarity are one of many insults microbial cells face on a daily basis. To protect against such shocks, *Escherichia coli* and other microbes express several types of transmembrane channels that open and close in response to changes in membrane tension. In *E. coli*, one of the most abundant channels is the mechanosensitive channel of large conductance (MscL). While this channel has been heavily characterized through structural methods, electrophysiology, and theoretical modeling, our understanding of its physiological role in preventing cell death by alleviating high membrane tension remains tenuous. In this work, we examine the contribution of MscL alone to cell survival after osmotic shock at single-cell resolution using quantitative fluorescence microscopy. We conducted these experiments in an *E. coli* strain which is lacking all mechanosensitive channel genes save for MscL, whose expression was tuned across 3 orders of magnitude through modifications of the Shine-Dalgarno sequence. While theoretical models suggest that only a few MscL channels would be needed to alleviate even large changes in osmotic pressure, we find that between 500 and 700 channels per cell are needed to convey upwards of 80% survival. This number agrees with the

average MscL copy number measured in wild-type *E. coli* cells through proteomic studies and quantitative Western blotting. Furthermore, we observed zero survival events in cells with fewer than  $\approx 100$  channels per cell. This work opens new questions concerning the contribution of other mechanosensitive channels to survival, as well as regulation of their activity.

## 5.2 Introduction

Changes in the extracellular osmolarity can be a fatal event for the bacterial cell. Upon a hypo-osmotic shock, water rushes into the cell across the membrane, leaving the cell with no choice but to equalize the pressure. This equalization occurs either through damage to the cell membrane (resulting in death) or through the regulated flux of water molecules through transmembrane protein channels (Fig 1A). Such proteinaceous pressure release valves have been found across all domains of life, with the first bacterial channel being described in 1987 (Martinac et al., 1987). Over the past thirty years, several more channels have been discovered, described, and (in many cases) biophysically characterized. *E. coli*, for example, has seven of these channels (one MscL and six MscS homologs) which have varied conductance, gating mechanisms, and expression levels. While they have been the subject of much experimental and theoretical dissection, much remains a mystery with regard to the roles their abundance and interaction with other cellular processes play in the greater context of physiology (Bavi et al., 2016; Bialecka-Fornal et al., 2012, 2015; Edwards et al., 2012; Naismith and Booth, 2012; Ursell et al., 2008; van den Berg et al., 2016).

Of the seven channels in *E. coli*, the mechanosensitive channel of large conductance (MscL) is one of the most abundant and the best characterized. This channel has a large conductance (3 nS) and mediates the flux of water molecules across the membrane via a  $\sim 3$  nm wide pore in the open state (Cruickshank et al., 1997; Haswell et al., 2011). Molecular dynamics simulations indicate that a single open MscL channel permits the flux of  $4 \times 10^9$  water molecules per second, which is an order of magnitude larger than a single aquaporin channel (BNID 100479)

(Louhivuori et al., 2010; Milo et al., 2010). This suggests that having only a few channels per cell could be sufficient to relieve even large changes in membrane tension. Electrophysiological experiments have suggested a small number of channels per cell (Booth et al., 2005; Hase et al., 1997), however, more recent approaches using quantitative Western blotting, fluorescence microscopy, and proteomics have measured several hundred MscL per cell (Bialecka-Fornal et al., 2012; Schmidt et al., 2016; Soufi et al., 2015). To further complicate matters, the expression profile of MscL appears to depend on growth phase, available carbon source, and other environmental challenges (Bialecka-Fornal et al., 2012; Soufi et al., 2015; Stokes et al., 2003). While there are likely more than just a few channels per cell, why cells seem to need so many and the biological rationale behind their condition-dependent expression both remain a mystery.

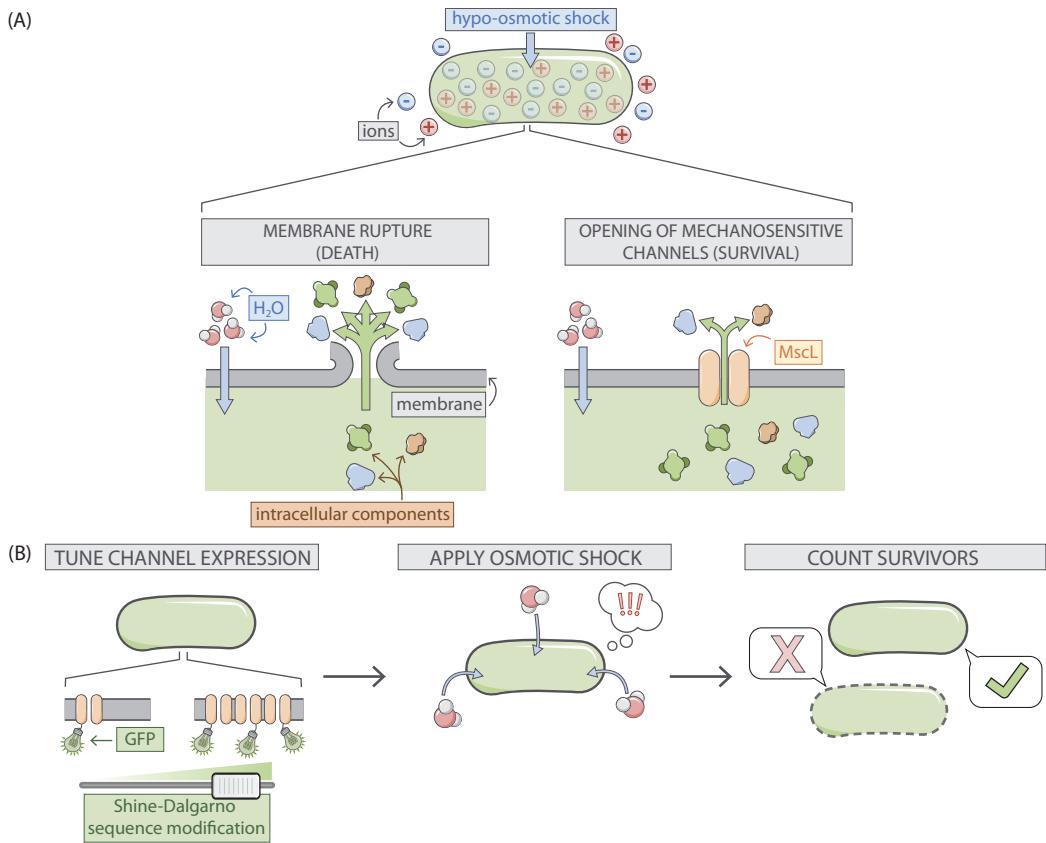
While their biochemical and biophysical characteristics have received much attention, their connection to cell survival is understudied. Drawing such a direct connection between channel copy number and survival requires quantitative *in vivo* experiments. To our knowledge, the work presented in van den Berg et al. 2016 (van den Berg et al., 2016) is the first attempt to simultaneously measure channel abundance and survivability for a single species of mechanosensitive channel. While the measurement of channel copy number was performed at the level of single cells using super-resolution microscopy, survivability after a hypoosmotic shock was assessed in bulk plating assays which rely on serial dilutions of a shocked culture followed by counting the number of resulting colonies after incubation. Such bulk assays have long been the standard for querying cell viability after an osmotic challenge. While they have been highly informative, they reflect only the mean survival rate of the population, obfuscating the variability in survival of the population. The stochastic nature of gene expression results in a noisy distribution of MscL channels rather than a single value, meaning those found in the long tails of the distribution have quite different survival rates than the mean but are lost in the final calculation of survival probability.

In this work, we present an experimental system to quantitatively probe the interplay between MscL copy number and survival at single-cell resolution, as is seen in Fig. 5.1(B). We generated an *E. coli* strain in which all seven mechanosensitive channels had been deleted from the chromosome followed by a chromosomal integration of a single gene encoding an MscL-super-folder GFP (sfGFP) fusion protein. To explore copy number regimes beyond those of the wild-type expression level, we modified the Shine-Dalgarno sequence of this integrated construct, allowing us to cover nearly three decades of MscL copy number. To probe survivability, we exposed cells to a large hypo-osmotic shock at controlled rates in a flow cell under a microscope, allowing the observation of the single-cell channel copy number and the resulting survivability of single cells. With this large set of single cell measurements, we approach the calculation of survival probability in a manner that is free of binning bias which allows the reasonable extrapolation of survival probability to copy numbers outside of the observed range. In addition, we show that several hundred channels are needed to convey high rates of survival and observe a minimum number of channels needed to permit any degree of survival.

### 5.3 Results

#### Quantifying the single-cell MscL copy number

The principal goal of this work is to examine the contribution of a single mechanosensitive channel species to cell survival under a hypo-osmotic shock. While this procedure could be performed for any species of channel, we chose MscL as it is the most well characterized and one of the most abundant species in *E. coli*. To probe the contribution of MscL alone, we integrated an *mscL* gene encoding an MscL super-folder GFP (sfGFP) fusion into a strain in which all seven known mechanosensitive channel genes were deleted from the chromosome (Edwards et al., 2012). Chromosomal integration imposes strict control on the gene copy number compared to plasmid borne expression systems, which is important to minimize variation in channel expression across the population and provide condi-



**Figure 5.1: Role of mechanosensitive channels during hypo-osmotic shock.** (A) A hypo-osmotic shock results in a large difference in the osmotic strength between the intracellular and extracellular spaces. As a result, water rushes into the cell to equalize this gradient increasing the turgor pressure and tension in the cell membrane. If no mechanosensitive channels are present and membrane tension is high (left panel), the membrane ruptures releasing intracellular content into the environment resulting in cell death. If mechanosensitive channels are present (right panel) and membrane tension is beyond the gating tension, the mechanosensitive channel MscL opens, releasing water and small intracellular molecules into the environment thus relieving pressure and membrane tension. (B) The experimental approach undertaken in this work. The number of mechanosensitive channels tagged with a fluorescent reporter is tuned through modification of the Shine-Dalgarno sequence of the *mscL* gene. The cells are then subjected to a hypo-osmotic shock and the number of surviving cells are counted, allowing the calculation of a survival probability.

tions more representative of native cell physiology. Abrogation of activity, mislocalization, or cytotoxicity are all inherent risks associated with creating chimeric reporter constructs. In Chapter 9, we carefully dissect the functionality of this protein through electrophysiology (Fig. S1), measure the rate of fluorophore maturation (Fig. S2), and quantify potential aggregates (Figs. S3 and S4). To the best of our knowledge, the MscL-sfGFP fusion protein functions identically to the wild-type, allowing us to confidently draw conclusions about the physiological role this channel plays in wild-type cells.

To modulate the number of MscL channels per cell, we developed a series of mutants which were designed to decrease the expression relative to wild-type. These changes involved direct alterations of the Shine-Dalgarno sequence as well as the inclusion of AT hairpins of varying length directly upstream of the start codon which influences the translation rate and hence the number of MscL proteins produced Fig. 5.2. The six Shine-Dalgarno sequences used in this work were chosen using the RBS binding site strength calculator from the Salis Laboratory at the Pennsylvania State University (Espah Borujeni et al., 2014; Salis et al., 2009). While the designed Shine-Dalgarno sequence mutations decreased the expression relative to wild-type as intended, the distribution of expression is remarkably wide spanning an order of magnitude.

To measure the number of MscL channels per cell, we determined a fluorescence calibration factor to translate arbitrary fluorescence units per cell to protein copy number. While there have been numerous techniques developed over the past decade to directly measure this calibration factor, such as quantifying single-molecule photobleaching constants or measuring the binomial partitioning of fluorescent proteins upon cell division (Bialecka-Fornal et al., 2012; Elowitz et al., 2002), we used *a priori* knowledge of the mean MscL-sfGFP expression level of a particular *E. coli* strain to estimate the average fluorescence of a single channel. In Bialecka-Fornal et al. 2012 (Bialecka-Fornal et al., 2012), the authors used single-molecule photobleaching and quantitative Western blotting to probe the expres-

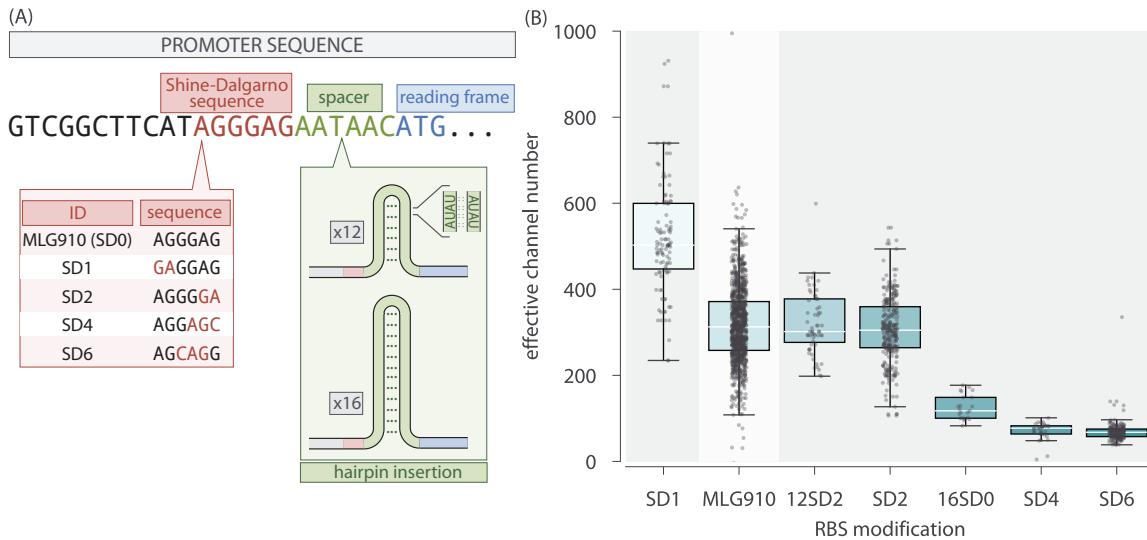
sion of MscL-sfGFP under a wide range of growth conditions. To compute a calibration factor, we used the strain MLG910 (*E. coli* K12 MG1655  $\phi$ (mscL-sfGFP)) as a “standard candle”, highlighted in white in Fig. 5.2 (B). This standard candle strain was grown and imaged in identical conditions in which the MscL count was determined through fluorescence microscopy. The calibration factor was computed by dividing the mean total cell fluorescence by the known MscL copy number, resulting in a measure of arbitrary fluorescence units per MscL channel. Details regarding this calculation and appropriate propagation of error as well as its sensitivity to varying growth media can be found in the Materials & Methods as well as supplemental Chapter 9.

While it is seemingly straightforward to use this calibration factor to determine the total number of channels per cell for wild-type or highly expressing strains, the calculation for the lowest expressing strains is complicated by distorted cell morphology. We observed that as the channel copy number decreases, cellular morphology becomes increasingly aberrant with filamentous, bulging, and branched cells becoming more abundant. This morphological defect has been observed when altering the abundance of several species of mechanosensitive channels, suggesting that they play an important role in general architectural stability (Bialecka-Fornal et al., 2012, 2015). As these aberrant morphologies can vary widely in size and shape, calculating the number of channels per cell becomes a more nuanced endeavor. For example, taking the total MscL copy number for these cells could skew the final calculation of survival probability as a large but severely distorted cell would be interpreted as having more channels than a smaller, wild-type shaped cell (Fig. S7B). To correct for this pathology, we computed the average expression level per unit area for each cell and multiplied this by the average cellular area of our standard candle strain which is morphologically indistinguishable from wild-type *E. coli*, allowing for the calculation of an effective channel copy number. The effect of this correction can be seen in Chapter 9, which illustrate that there is no other correlation between cell area and channel expression.

Our calculation of the effective channel copy number for our suite of Shine-Dalgarno mutants is shown in Fig. 5.2(B). The expression of these strains cover nearly three orders of magnitude with the extremes ranging from approximately four channels per cell to nearly one thousand. While the means of each strain are somewhat distinct, the distributions show a large degree of overlap, making one strain nearly indistinguishable from another. This variance is a quantity that is lost in the context of bulk scale experiments but can be accounted for via single-cell methods.

### **Performing a single-cell hypo-osmotic challenge assay**

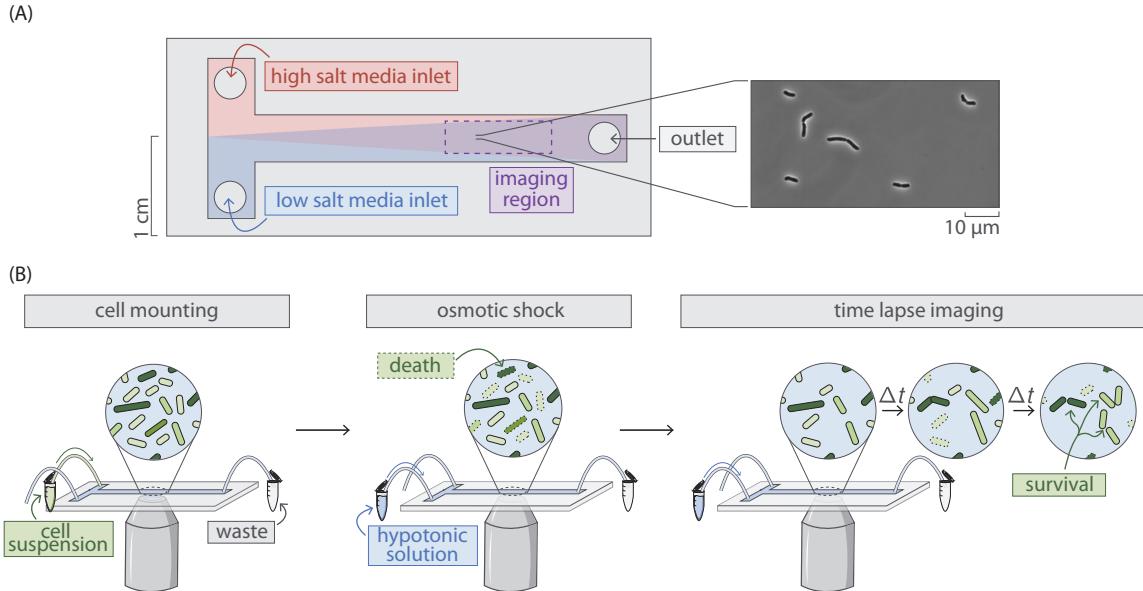
To measure the channel copy number of a single cell and query its survival after a hypo-osmotic shock, we used a custom-made flow cell in which osmotic shock and growth can be monitored in real time using video microscopy (Fig. fig. 5.3(A)). The design and characterization of this device has been described in depth previously and is briefly described in the Materials & Methods (Bialecka-Fornal et al., 2015). Using this device, cells were exposed to a large hypo-osmotic shock by switching between LB Lennox medium supplemented with 500 mM NaCl and LB Lennox media alone. All six Shine-Dalgarno modifications shown in Fig. 5.2(B) (excluding MLG910) were subjected to a hypo-osmotic shock at controlled rates while under observation. After the application of the osmotic shock, the cells were imaged every sixty seconds for four to six hours. Each cell was monitored over the outgrowth period and was manually scored as either a survivor, fatality, or inconclusive observation. The criteria used for scoring death were the same as those previously described in Bialecka-Fornal et al. 2015 (Bialecka-Fornal et al., 2015). Survivors were defined as cells that underwent multiple divisions post-shock. To qualify as survivors, cells must undergo at least two divisions, although more typically, four to eight divisions are observed without any signs of slowing down. Imaging is stopped when the survivors cells begin to go out of focus or overlap each other. Survivors do not show any sign of ceasing division. More information regarding this classification can be found in the Materials & Methods as well as the supple-



**Figure 5.2: Control of MscL expression and calculation of channel copy number.**

(A) Schematic view of the expression modifications performed in this work. The beginning portion of the native *mscL* sequence is shown with the Shine-Dalgarno sequence, spacer region, and start codon shaded in red, green, and blue, respectively. The Shine-Dalgarno sequence was modified through the Salis lab Ribosomal Binding Strength calculator (Espah Borujeni et al., 2014; Salis et al., 2009). The wild-type sequence (MLG910) is shown in black with mutations for the other four Shine-Dalgarno mutants highlighted in red. Expression was further modified by the insertion of repetitive AT bases into the spacer region, generating hairpins of varying length which acted as a thermodynamic barrier for translation initiation.

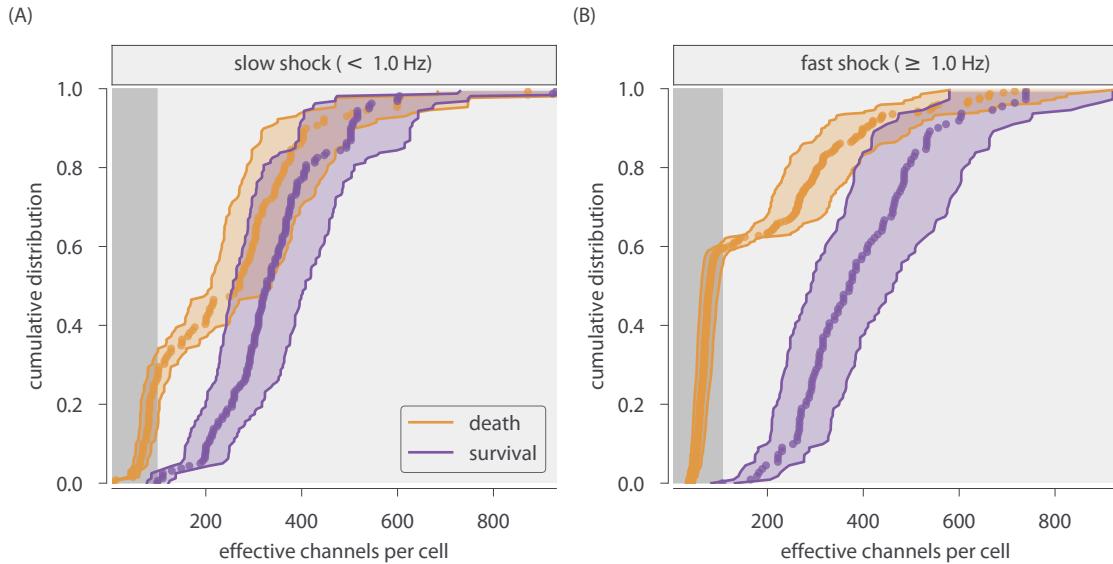
(B) Variability in effective channel copy number is computed using the standard candle. The boxes represent the interquartile region of the distribution, the center line displays the median, and the whiskers represent 1.5 times the maximum and minimum of the interquartile region. Individual measurements are denoted as black points. The strain used for calibration of channel copy number (MLG910) is highlighted in white. The Python code (`ch5_fig2.py`) used to generate this figure can be found on the thesis GitHub repository.



**Figure 5.3: Experimental approach to measuring survival probability.** (A) Layout of a home-made flow cell for subjecting cells to osmotic shock. Cells are attached to a polyethylenimine functionalized surface of a glass coverslip within the flow chamber by loading a dilute cell suspension through one of the inlets. (B) The typical experimental procedure. Cells are loaded into a flow chamber as shown in (A) and mounted to the glass coverslip surface. Cells are subjected to a hypo-osmotic shock by flowing hypotonic medium into the flow cell. After shock, the cells are monitored for several hours and surviving cells are identified.

mental Chapter 9. The brief experimental protocol can be seen in Fig. 5.3(B).

Due to the extensive overlap in expression between the different Shine-Dalgarno mutants (see Fig. fig. 5.2(B)), computing the survival probability by treating each mutant as an individual bin obfuscates the relationship between channel abundance and survival. To more thoroughly examine this relationship, all measurements were pooled together with each cell being treated as an individual experiment. The hypo-osmotic shock applied in these experiments was varied across a range of 0.02 Hz (complete exchange in 50 s) to 2.2 Hz (complete exchange in 0.45 s). Rather than pooling this wide range of shock rates into a single data set, we chose to separate the data into “slow shock” ( $< 1.0$  Hz) and “fast shock” ( $\geq 1.0$  Hz) classes. Other groupings of shock rate were explored and are discussed in Chapter 9. The cumulative distributions of channel copy number separated by



**Figure 5.4: Distributions of survival and death as a function of effective channel number.** (A) Empirical cumulative distributions of channel copy number separated by survival (purple) or death (orange) after a slow ( $< 1.0 \text{ Hz}$ ) osmotic shock. (B) The empirical cumulative distribution for a fast ( $\geq 1.0 \text{ Hz}$ ) osmotic shock. Shaded purple and orange regions represent the 95% credible region of the effective channel number calculation for each cell. Shaded grey stripe signifies the range of channels in which no survival was observed. The Python code (`ch5_fig4.py`) used to generate this figure can be found on the thesis GitHub repository.

survival are shown in Fig. 5.4. In these experiments, survival was never observed for a cell containing less than approximately 100 channels per cell, indicated by the grey shaded region in Fig. 5.4. This suggests that there is a minimum number of channels needed for survival on the order of 100 per cell. We also observe a slight shift in the surviving fraction of the cells towards higher effective copy number, which matches our intuition that including more mechanosensitive channels increases the survival probability.

### Prediction of survival probability as a function of channel copy number

There are several ways by which the survival probability can be calculated. The most obvious approach would be to group each individual Shine-Dalgarno mutant as a single bin and compute the average MscL copy number and the survival

probability. Binning by strain is the most frequently used approach for such measurements and has provided valuable insight into the qualitative relationship of survival on other physiological factors (Bialecka-Fornal et al., 2015; van den Berg et al., 2016). However the copy number distribution for each Shine-Dalgarno mutant (Fig. 5.2(B)) is remarkably wide and overlaps with the other strains. We argue that this coarse-grained binning negates the benefits of performing single-cell measurements as two strains with different means but overlapping quartiles would be treated as distinctly different distributions.

Another approach would be to pool all data together, irrespective of the Shine-Dalgarno mutation, and bin by a defined range of channels. Depending on the width of the bin, this could allow for finer resolution of the quantitative trend, but the choice of the bin width is arbitrary with the *a priori* knowledge that is available. Drawing a narrow bin width can easily restrict the number of observed events to small numbers where the statistical precision of the survival probability is lost. On the other hand, drawing wide bins increases the precision of the estimate, but becomes further removed from a true single-cell measurement and represents a population mean, even though it may be a smaller population than binning by the Shine-Dalgarno sequence alone. In both of these approaches, it is difficult to extrapolate the quantitative trend outside of the experimentally observed region of channel copy number. Here, we present a method to estimate the probability of survival for any channel copy number, even those that lie outside of the experimentally queried range.

To quantify the survival probability while maintaining single-cell resolution, we chose to use a logistic regression model which does not require grouping data into arbitrary bins and treats each cell measurement as an independent experiment. Logistic regression is an inferential method to model the probability of a Boolean or categorical event (such as survival or death) given one or several predictor variables and is commonly used in medical statistics to compute survival rates and dose response curves (Anderson et al., 2003; Mishra et al., 2016). The primary

assumption of logistic regression is that the log-odds probability of survival  $p_s$  is linearly dependent on the predictor variable, in our case the log channels per cell  $N_c$  with a dimensionless intercept  $\beta_0$  and slope  $\beta_1$ ,

$$\log \frac{p_s}{1 - p_s} = \beta_0 + \beta_1 \log N_c. \quad (5.1)$$

Under this assumption of linearity,  $\beta_0$  is the log-odds probability of survival with no MscL channels. The slope  $\beta_1$  represents the change in the log-odds probability of survival conveyed by a single channel. As the calculated number of channels in this work spans nearly three orders of magnitude, it is better to perform this regression on  $\log N_c$  as regressing on  $N_c$  directly would give undue weight for lower channel copy numbers due to the sparse sampling of high-copy number cells. The functional form shown in Eq. eq. 5.1 can be derived directly from Bayes' theorem and is shown in Chapter 9. If one knows the values of  $\beta_0$  and  $\beta_1$ , the survival probability can be expressed as

$$p_s = \frac{1}{1 + N_c^{-\beta_1} e^{-\beta_0}}. \quad (5.2)$$

In this analysis, we used Bayesian inferential methods to determine the most likely values of the coefficients and is described in detail in the supplemental Chapter 9.

The results of the logistic regression are shown in Fig. 5.5. We see a slight rightward shift the survival probability curve under fast shock relative to the slow shock case, reaffirming the conclusion that survival is also dependent on the rate of osmotic shock (Bialecka-Fornal et al., 2015). This rate dependence has been observed for cells expressing MscL alongside other species of mechanosensitive channels, but not for MscL alone. This suggests that MscL responds differently to different rates of shock, highlighting the need for further study of rate dependence and the coordination between different species of mechanosensitive channels. Fig. 5.5 also shows that several hundred channels are required to provide appreciable protection from osmotic shock. For a survival probability of 80%, a cell must have approximately 500 to 700 channels per cell for a fast and slow shock, respectively. The results from the logistic regression are showed as continuous colored curves. The individual cell measurements separated by survival and death

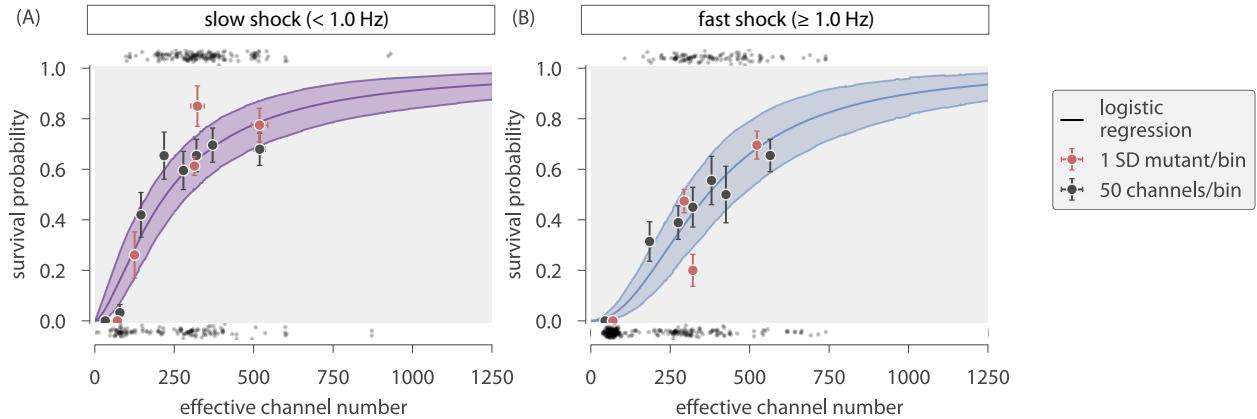
are shown at the top and bottom of each plot, respectively, and are included to provide a sense of sampling density.

Over the explored range of MscL copy number, we observed a maximum of 80% survival for any binning method. The remaining 20% survival may be attained when the other species of mechanosensitive channels are expressed alongside MscL. However, it is possible that the flow cell method performed in this work lowers the maximal survival fraction as the cells are exposed to several, albeit minor, mechanical stresses such as loading into the flow cell and chemical adherence to the glass surface. To ensure that the results from logistic regression accurately describe the data, we can compare the survival probabilities to those using the binning methods described earlier (red and black points, Fig. 5.5). Nearly all binned data fall within error of the prediction (see Materials & Methods for definition of error bar on probability), suggesting that this approach accurately reflects the survival probability and gives license to extrapolate the estimation of survival probability to regions of outside of our experimentally explored copy number regime.

Thus far, we've dictated that for a given rate of osmotic shock (i.e. "fast" or "slow"), the survival probability is dependent only on the number of channels. In Chapter 9, we show the result of including other predictor variables, such as area and shock rate alone. In such cases, including other predictors resulted in pathological curves showing that channel copy number is the most informative out of the available predictor variables.

#### 5.4 Discussion

One of the most challenging endeavors in the biological sciences is linking the microscopic details of cellular components to the macro-scale physiology of the organism. This formidable task has been met repeatedly in the recent history of biology, especially in the era of DNA sequencing and single molecule biochemistry. For example, the scientific community has been able to connect sickle-cell anemia to a single amino acid substitution in Hemoglobin which promotes precipitation under a change in O<sub>2</sub> partial pressure (Feeling-Taylor et al., 2004; Finch et



**Figure 5.5: Probability of survival as a function of MscL copy number.** (A) Estimated survival probability for survival under slow shock as a function of channel copy number. (B) The estimated survival probability of survival under a fast shock as a function of channel copy number. Solid curves correspond to the most probable survival probability from a one-dimensional logistic regression. Shaded regions represent the 95% credible regions. Points at the top and bottom of plots represent individual cell measurements which survived and perished, respectively. The red and black points correspond to the survival probability estimated via binning by Shine-Dalgarno sequence and binning by groups of 50 channels per cell, respectively. Horizontal error bars represent the standard error of the mean from at least 25 measurements. Vertical error bars represent the certainty of the probability estimate given  $n$  survival events from  $N$  total observations. The Python code (`ch5_fig5.py`) used to generate this figure can be found on the thesis GitHub repository.

al., 1973; Perutz and Mitchison, 1950). Others have assembled a physical model that quantitatively describes chemosensation in bacteria (Berg and Purcell, 1977) in which the arbiter of sensory adaptation is the repeated methylation of chemoreceptors (Colin and Sourjik, 2017; Krembel et al., 2015a, 2015b; Sourjik and Berg, 2002). In the past ~50 years alone, numerous biological and physical models of the many facets of the central dogma have been assembled that give us a sense of the interplay between the genome and physiology. For example, the combination of biochemical experimentation and biophysical models have given us a picture of how gene dosage affects furrow positioning in *Drosophila* (Liu et al., 2013), how recombination of V(D)J gene segments generates an extraordinarily diverse antibody repertoire (Lovely et al., 2015; Schatz and Baltimore, 2004; Schatz and Ji, 2011), and how telomere shortening through DNA replication is intrinsically tied

to cell senescence (Herbig et al., 2004; Victorelli and Passos, 2017), to name just a few of many such examples.

By no means are we “finished” with any of these topics. Rather, it’s quite the opposite in the sense that having a handle on the biophysical knobs that tune the behavior opens the door to a litany of new scientific questions. In the case of mechanosenstaion and osmoregulation, we have only recently been able to determine some of the basic facts that allow us to approach this fascinating biological phenomenon biophysically. The dependence of survival on mechanosensitive channel abundance is a key quantity that is missing from our collection of critical facts. To our knowledge, this work represents the first attempt to quantitatively control the abundance of a single species of mechanosensitive channel and examine the physiological consequences in terms of survival probability at single-cell resolution. Our results reveal two notable quantities. First, out of the several hundred single-cell measurements, we never observed a cell which had less than approximately 100 channels per cell and survived an osmotic shock, irrespective of the shock rate. The second is that between 500 and 700 channels per cell are needed to provide  $\geq 80\%$  survival, depending on the shock rate.

Only recently has the relationship between the MscL copy number and the probability of survival been approached experimentally. In van den Berg et al. (2016), the authors examined the contribution of MscL to survival in a genetic background where all other known mechanosensitive channels had been deleted from the chromosome and plasmid-borne expression of an MscL-mEos3.2 fusion was tuned through an IPTG inducible promoter (van den Berg et al., 2016). In this work, they measured the single-cell channel abundance through super-resolution microscopy and queried survival through bulk assays. They report a nearly linear relationship between survival and copy number, with approximately 100 channels per cell conveying 100% survival. This number is significantly smaller than our observation of approximately 100 channels as the *minimum* number needed to convey any observable degree of survival.

The disagreement between the numbers reported in this work and in van den Berg et al. (2016) may partially arise from subtle differences in the experimental approach. The primary practical difference is the magnitude of the osmotic shock. van den Berg et al. applied an approximately 600 mOsm downshock in bulk whereas we applied a 1 Osm downshock, which would lead to lower survival (Levina et al., 1999). In their work, the uncertainty in both the MscL channel count and survival probability is roughly 30%. Given this uncertainty, it is reasonable to interpret that the number of channels needed for complete protection from osmotic downshock is between 100 and 250 per cell. The uncertainty in determining the number of channels per cell is consistent with the observed width of the channel number distribution of the Shine-Dalgarno sequence mutants used in this work (Fig. fig. 5.2(B)). A unique property of the single-cell measurements performed in this work allow is the direct observation of survival or death of individual cells. We find that morphological classification and classification through a propidium iodide staining agree within 1% (Chapter 9). The bulk plating assays, as are used in van den Berg et al. (2016), rely on colony formation and outgrowth to determine survival probability. As is reported in their supplemental information, the precision in this measurement is around 30% (Fig. S14). Accounting for this uncertainty brings both measurements within a few fold where we still consistently observe lower survival for a given channel number. This remaining disagreement may be accounted for by systematic uncertainty in both experimental methods.

For example, variation in the length of outgrowth, variable shock rate, and counting statistics could bias towards higher observed survival rates in ensemble plating assays. During the outgrowth phase, the control sample not exposed to an osmotic shock is allowed to grow for approximately 30 minutes in a high-salt medium before plating. The shocked cells, however, are allowed to grow in a low-salt medium. We have found that the difference between the growth rates in these two conditions can be appreciable (approximately 35 minutes versus 20 minutes, respectively). Cells that survived an osmotic shock may have a growth advantage relative to the control sample if the shock-induced lag phase is less than

the outgrowth, leading to higher observed survival rates (Levina et al., 1999). This is one possible explanation for the survival rates which are reported in excess of 100%. Cells that survived an osmotic shock may have a growth advantage relative to the normalization sample if the shock-induced lag phase is less than the outgrowth, leading to higher observed survival rates, even surpassing 100%. We have performed these assays ourselves and have observed survival rates above of 100% (ranging from 110% to 125%) with an approximate 30% error (see Fig. S3 in Bialecka-Fornal et al. (2012)) which we concluded to arise from differences in growth rate. We also note that survival rates greater than 100% are observed in van den Berg et al. (2016). For strains that have survival rates between 80% and 100% the uncertainty is typically large, making it difficult to make precise statements regarding when full survival is achieved.

It has been shown that there is a strong inverse relationship between the rate of osmotic shock and survival probability (Bialecka-Fornal et al., 2015). Any experiment in which the shock was applied more slowly or quickly than another would bias toward higher or lower survivability, respectively. The shocks applied in bulk assays are often performed manually which can be highly variable. We note that in our experiments, we frequently observe cells which do not separate and form chains of two or more cells. In plating assays, it is assumed that colonies arise from a single founding cell however a colony formed by a cluster of living and dead cells would be interpreted as a single surviving cell, effectively masking the death of the others in the colony forming unit. This too could bias the measurement toward higher survival rates. Single-cell shock experiments can also have systematic errors which can bias the results towards lower survival rates. Such errors are associated with handling of the cells such as shear damage from loading into the flow cell, adhering the cells to the coverslip, and any chemical perturbations introduced by the dye used to measure the shock rate.

Despite these experimental differences, the results of this work and van den Berg et al. (2016) are in agreement that MscL must be present at the level of 100 or more

channels per cell in wild-type cells to convey appreciable survival. As both of these works were performed in a strain in which the only mechanosensitive channel was MscL, it remains unknown how the presence of the other channel species would alter the number of MscL needed for complete survival. In our experiments, we observed a maximum survival probability of approximately 80% even with close to 1000 MscL channels per cell. It is possible that the combined effort of the six other mechanosensitive channels would make up for some if not all of the remaining 20%. To explore the contribution of another channel to survival, van den Berg et al. also queried the contribution of MscS, another mechanosensitive channel, to survival in the absence of any other species of mechanosensitive channel. It was found that over the explored range of MscS channel copy numbers, the maximum survival rate was approximately 50%, suggesting that different mechanosensitive channels have an upper limit to how much protection they can confer. Both van den Berg et al. and our work show that there is still much to be learned with respect to the interplay between the various species of mechanosensitive channel as well as their regulation.

Recent work has shown that both magnitude and the rate of osmotic down shock are important factors in determining cell survival (Bialecka-Fornal et al., 2015). In this work, we show that this finding holds true for a single species of mechanosensitive channel, even at high levels of expression. One might naïvely expect that this rate-dependent effect would disappear once a certain threshold of channels had been met. Our experiments, however, show that even at nearly 1000 channels per cell the predicted survival curves for a slow ( $< 1.0$  Hz) and fast ( $\geq 1.0$  Hz) are shifted relative to each other with the fast shock predicting lower rates of survival. This suggests either we have not reached this threshold in our experiments or there is more to understand about the relationship between abundance, channel species, and the shock rate.

Some experimental and theoretical treatments suggest that only a few copies of MscL or MscS should be necessary for 100% protection given our knowledge

of the conductance and the maximal water flux through the channel in its open state (Booth, 2014; Louhivuori et al., 2010). However, recent proteomic studies have revealed average MscL copy numbers to be in the range of several hundred per cell, depending on the condition, as can be seen in Table 5.1 (Li et al., 2014; Schmidt et al., 2016; Soufi et al., 2015). Studies focusing solely on MscL have shown similar counts through quantitative Western blotting and fluorescence microscopy (Bialecka-Fornal et al., 2012). Electrophysiology studies have told another story with copy number estimates ranging between 4 and 100 channels per cell (Blount et al., 1999; Booth et al., 2005; Stokes et al., 2003). These measurements, however, measure the active number of channels. The factors regulating channel activity in these experiments could be due to perturbations during the sample preparation or reflect some unknown mechanism of regulation, such as the presence or absence of interacting cofactors (Schumann et al., 2010). The work described here, on the other hand, measures the *maximum* number of channels that could be active and may be able to explain why the channel abundance is higher than estimated by theoretical means. There remains much more to be learned about the regulation of activity in these systems. As the *in vivo* measurement of protein copy number becomes accessible through novel single-cell and single-molecule methods, we will continue to collect more facts about this fascinating system and hopefully connect the molecular details of mechanosensation with perhaps the most important physiological response – life or death.

Table 5.1: Measured cellular copy numbers of MscL.

Asterisk (\*) Indicates inferred MscL channel copy number from the total number of detected MscL peptides.

Reported channels per cell	Method	Reference
480 ± 103	Western blotting	Bialecka-Fornal et al. (2012)
560*	Ribosomal profiling	Li et al. (2014)
331*	Mass spectrometry	Schmidt et al. (2016)
583*	Mass spectrometry	Soufi et al. (2015)

---

Reported channels per cell	Method	Reference
4 - 5	Electrophysiology	Stokes et al. (2003)
10 - 100	Electrophysiology	Booth et al. (2005)
10 - 15	Electrophysiology	Blount et al. (1999)

---

## 5.5 Materials & Methods

### Bacterial strains and growth conditions

The bacterial strains are described in Table 9.1. The parent strain for the mutants used in this study was MJF641 (Edwards et al., 2012), a strain which had all seven mechanosensitive channels deleted. The MscL-sfGFP coding region from MLG910 (Bialecka-Fornal et al., 2012) was integrated into MJF641 by P1 transduction, creating the strain D6LG-Tn10. Selection pressure for MscL integration was created by incorporating an osmotic shock into the transduction protocol, which favored the survival of MscL-expressing stains relative to MJF641 by ~100-fold. Screening for integration candidates was based on fluorescence expression of plated colonies. Successful integration was verified by sequencing. Attempts to transduce RBS-modified MscL-sfGFP coding regions became increasingly inefficient as the targeted expression level of MscL was reduced. This was due to the decreasing fluorescence levels and survival rates of the integration candidates. Consequently, Shine-Dalgarno sequence modifications were made by inserting DNA oligos with lambda Red-mediated homologous recombination, i.e., recombineering (Sharan et al., 2009). The oligos had a designed mutation (Fig. fig. 5.2) flanked by ~25 base pairs that matched the targeted MscL region (Table S2). A two-step recombineering process of selection followed by counter selection using a *tetA-sacB* gene fusion cassette (Li et al., 2013) was chosen because of its capabilities to integrate with efficiencies comparable to P1 transduction and not leave antibiotic resistance markers or scar sequences in the final strain.

To prepare the strain D6LG-Tn10 for this scheme, the Tn10 transposon containing the *tetA* gene needed to be removed to avoid interference with the *tetA-sacB*

cassette. Tn10 was removed from the middle of the *ycjM* gene with the primer Tn10delR (Table S2) by recombineering, creating the strain D6LG (SD0). Counter selection against the *tetA* gene was promoted by using agar media with fusaric acid (Bochner et al., 1980; Li et al., 2013). The *tetA-sacB* cassette was PCR amplified out of the strain XTL298 using primers MscLSPSac and MscLSPSacR (Table S2). The cassette was integrated in place of the spacer region in front of the MscL start codon of D6LG (SD0) by recombineering, creating the intermediate strain D6LTetSac. Positive selection for cassette integration was provided by agar media with tetracycline. Finally, the RBS modifying oligos were integrated into place by replacing the *tetA-sacB* cassette by recombineering. Counter selection against both *tetA* and *sacB* was ensured by using agar media with fusaric acid and sucrose (Li et al., 2013), creating the Shine-Dalgarno mutant strains used in this work.

Strain cultures were grown in 5 mL of LB-Lennox media with antibiotic (apramycin) overnight at 37°C. The next day, 50  $\mu$ L of overnight culture was inoculated into 5 mL of LB-Lenox with antibiotic and the culture was grown to OD<sub>600nm</sub> ~ .25. Subsequently, 500  $\mu$ L of that culture was inoculated into 5 mL of LB-Lennox supplemented with 500mM of NaCl and the culture was regrown to OD<sub>600nm</sub> ~0.25. A 1 mL aliquot was taken and used to load the flow cell.

### Flow cell

All experiments were conducted in a home-made flow cell as is shown in Fig. 5.3(A). This flow cell has two inlets which allow media of different osmolarity to be exchanged over the course of the experiment. The imaging region is approximately 10 mm wide and 100  $\mu$ m in depth. All imaging took place within 1 – 2 cm of the outlet to avoid imaging cells within a non-uniform gradient of osmolarity. The interior of the flow cell was functionalized with a 1:400 dilution of polyethylenimine prior to addition of cells with the excess washed away with water. A dilute cell suspension in LB Lennox with 500 mM NaCl was loaded into one inlet while the other was connected to a vial of LB medium with no NaCl. This hypotonic medium was clamped during the loading of the cells.

Once the cells had adhered to the polyethylenimine coated surface, the excess cells were washed away with the 500 mM NaCl growth medium followed by a small (~20  $\mu$ L) air bubble. This air bubble forced the cells to lay flat against the imaging surface, improving the time-lapse imaging. Over the observation period, cells not exposed to an osmotic shock were able to grow for 4–6 divisions, showing that the flow cell does not directly impede cell growth.

### Imaging conditions

All imaging was performed in a flow cell held at 30°C on a Nikon Ti-Eclipse microscope outfitted with a Perfect Focus system enclosed in a Haison environmental chamber (approximately 1°C regulation efficiency). The microscope was equipped with a 488 nm laser excitation source (CrystaLaser) and a 520/35 laser optimized filter set (Semrock). The images were collected on an Andor iXon EM+ 897 EM-CCD camera and all microscope and acquisition operations were controlled via the open source  $\mu$ Manager microscope control software (Edelstein et al., 2014). Once cells were securely mounted onto the surface of the glass coverslip, between 15 and 20 positions containing 5 to 10 cells were marked and the coordinates recorded. At each position, a phase contrast and GFP fluorescence image was acquired for segmentation and subsequent measurement of channel copy number. To perform the osmotic shock, LB media containing no NaCl was pulled into the flow cell through a syringe pump. To monitor the media exchange, both the high salt and no salt LB media were supplemented with a low-affinity version of the calcium-sensitive dye Rhod-2 (250 nM; TEF Labs) which fluoresces when bound to Ca<sup>2+</sup>. The no salt medium was also supplemented with 1 $\mu$ M CaCl<sub>2</sub> to make the media mildly fluorescent and the exchange rate was calculated by measuring the fluorescence increase across an illuminated section of one of the positions. These images were collected in real time for the duration of the shock. The difference in measured fluorescence between the pre-shock images and those at the end of the shock set the scale of a 500 mM NaCl down shock. The rate was calculated by fitting a line to the middle region of this trace. Further details regarding this procedure can be

found in Bialecka-Fornal et al. (2015).

### Image Processing

Images were processed using a combination of automated and manual methods. First, expression of MscL was measured via segmenting individual cells or small clusters of cells in phase contrast and computing the mean pixel value of the fluorescence image for each segmented object. The fluorescence images were passed through several filtering operations which reduced high-frequency noise as well as corrected for uneven illumination of the excitation wavelength.

Survival or death classification was performed manually using the CellProfiler plugin for ImageJ software (NIH). A survivor was defined as a cell which was able to undergo at least two division events after the osmotic down shock. Cell death was recognized by stark changes in cell morphology including loss of phase contrast through ejection of cytoplasmic material, structural decomposition of the cell wall and membrane, and the inability to divide. To confirm that these morphological cues corresponded with cell death, we probed cell viability on a subset of our strains after osmotic shock through staining with propidium iodide, a DNA intercalating dye commonly used to identifying dead cells (LIVE/DEAD BacLight Bacterial Cell Viability Assay, Thermo Fisher). We found that our classification based on morphology agreed with that based off of staining within 1%. More information regarding these experiments can be found in Chapter 9. Cells which detached from the surface during the post-shock growth phase or those which became indistinguishable from other cells due to clustering were not counted as survival or death and were removed from the dataset completely. A region of the cell was manually marked with 1.0 (survival) or 0.0 (death) by clicking on the image. The xy coordinates of the click as well as the assigned value were saved as an .xml file for that position.

The connection between the segmented cells and their corresponding manual markers was automated. As the manual markings were made on the first phase contrast image after the osmotic shock, small shifts in the positions of the cell made

one-to-one mapping with the segmentation mask non-trivial. The linkages between segmented cell and manual marker were made by computing all pairwise distances between the manual marker and the segmented cell centroid, taking the shortest distance as the true pairing. The linkages were then inspected manually and incorrect mappings were corrected as necessary.

All relevant statistics about the segmented objects as well as the sample identity, date of acquisition, osmotic shock rate, and camera exposure time were saved as .csv files for each individual experiment. A more in-depth description of the segmentation procedure as well as the relevant code can be accessed as a Jupyter Notebook at ([http://rpgroup.caltech.edu/mscl\\_survival](http://rpgroup.caltech.edu/mscl_survival)).

### **Calculation of effective channel copy number**

To compute the MscL channel copy number, we relied on measuring the fluorescence level of a bacterial strain in which the mean MscL channel copy number was known via fluorescence microscopy (Bialecka-Fornal et al., 2012). *E. coli* strain MLG910, which expresses the MscL-sfGFP fusion protein from the wild-type SD sequence, was grown under identical conditions to those described in Bialecka-Fornal et al. 2015 in LB Miller medium (BD Medical Sciences) to an OD<sub>600nm</sub> of ~0.3. The cells were then diluted ten fold and immobilized on a rigid 2% agarose substrate and placed onto a glass bottom petri dish and imaged in the same conditions as described previously.

Images were taken of six biological replicates of MLG910 and were processed identically to those in the osmotic shock experiments. A calibration factor between the average cell fluorescence level and mean MscL copy number was then computed. We assumed that all measured fluorescence (after filtering and background subtraction) was derived from the MscL-sfGFP fusion,

$$\langle I_{\text{tot}} \rangle = \alpha \langle N \rangle, \quad (5.3)$$

in which  $\alpha$  is the calibration factor and  $\langle N \rangle$  is the mean cellular MscL-sfGFP copy number as reported in Bialecka-Fornal et al. (2012). To correct for errors in segmen-

tation, the intensity was computed as an areal density  $\langle I_A \rangle$  and was multiplied by the average cell area  $\langle A \rangle$  of the population. The calibration factor was therefore computed as

$$\alpha = \frac{\langle I_A \rangle \langle A \rangle}{\langle N \rangle}. \quad (5.4)$$

We used Bayesian inferential methods to compute this calibration factor taking measurement error and replicate-to-replicate variation into account. The resulting average cell area and calibration factor was used to convert the measured cell intensities from the osmotic shock experiments to cell copy number. The details of this inference are described in depth in the supplemental Chapter 9.

### Logistic regression

We used Bayesian inferential methods to find the most probable values of the coefficients  $\beta_0$  and  $\beta_1$  and the appropriate credible regions and is described in detail in the supplemental information (*Logistic Regression*). Briefly, we used Markov chain Monte Carlo (MCMC) to sample from the log posterior distribution and took the most probable value as the mean of the samples for each parameter. The MCMC was performed using the Stan probabilistic programming language (Carpenter et al., 2017) and all models can be found on the GitHub repository ([http://github.com/rpgroup-pboc/mscl\\_survival](http://github.com/rpgroup-pboc/mscl_survival)).

### Calculation of survival probability error

The vertical error bars for the points shown in Fig. 5.5 represent our uncertainty in the survival probability given our measurement of  $n$  survivors out of a total  $N$  single-cell measurements. The probability distribution of the survival probability  $p_s$  given these measurements can be written using Bayes' theorem as

$$g(p_s | n, N) = \frac{f(n | p_s, N)g(p_s)}{f(n | N)}, \quad (5.5)$$

where  $g$  and  $f$  represent probability density functions over parameters and data, respectively. The likelihood  $f(n | p_s, N)$  represents the probability of measuring  $n$  survival events, given a total of  $N$  measurements each with a probability of sur-

vival  $p_s$ . This matches the story for the Binomial distribution and can be written as

$$f(n | p_s, N) = \frac{N!}{n!(N-n)!} p_s^n (1-p_s)^{N-n}. \quad (5.6)$$

To maintain maximal ignorance we can assume that any value for  $p_s$  is valid, such that is in the range  $[0, 1]$ . This prior knowledge, represented by  $g(p_s)$ , can be written as

$$g(p_s) = \begin{cases} 1 & 0 \leq p_s \leq 1 \\ 0 & \text{otherwise} \end{cases}. \quad (5.7)$$

We can also assume maximal ignorance for the total number of survival events we could measure given  $N$  observations,  $f(n | N)$ . Assuming all observations are equally likely, this can be written as

$$f(n | N) = \frac{1}{N+1} \quad (5.8)$$

where the addition of one comes from the possibility of observing zero survival events. Combining Eq. 5.6, Eq. 5.7, and Eq. 5.8, the posterior distribution  $g(p_s | n, N)$  is

$$g(p_s | n, N) = \frac{(N+1)!}{n!(N-n)!} p_s^n (1-p_s)^{N-n}. \quad (5.9)$$

The most probable value of  $p_s$ , where the posterior probability distribution given by Eq. 5.9 is maximized, can be found by computing the point at which derivative of the log posterior with respect to  $p_s$  goes to zero,

$$\frac{d \log g(p_s | n, N)}{dp_s} = \frac{n}{p_s} - \frac{N-n}{1-p_s} = 0. \quad (5.10)$$

Solving Eq. 5.10 for  $p_s$  gives the most likely value for the probability,

$$p_s^* = \frac{n}{N}. \quad (5.11)$$

So long as  $N \gg np_s^*$ , Eq. 5.9 can be approximated as a Gaussian distribution with a mean  $p_s^*$  and a variance  $\sigma_{p_s}^2$ . By definition, the variance of a Gaussian distribution

is computed as the negative reciprocal of the second derivative of the log posterior evaluated at  $p_s = p_s^*$ ,

$$\sigma_{p_s}^2 = - \left( \frac{d^2 \log g(p_s | n, N)}{dp_s^2} \Bigg|_{p_s=p_s^*} \right)^{-1}. \quad (5.12)$$

Evaluating Eq. 5.12 yields

$$\sigma_{p_s}^2 = \frac{n(N-n)}{N^3}. \quad (5.13)$$

Given Eq. 5.11 and Eq. 5.13, the most-likely survival probability and estimate of the uncertainty can be expressed as

$$p_s = p_s^* \pm \sigma_{p_s}. \quad (5.14)$$

### **Data and software availability**

All raw image data is freely available and is stored on the CaltechDATA Research Data Repository. The raw Markov chain Monte Carlo samples are stored as .csv files on CaltechDATA. All processed experimental data, Python, and Stan code used in this work are freely available through the paper GitHub repository accessible through DOI: 10.5281/zenodo.1252524. The scientific community is invited to fork our repository and open constructive issues.

## Chapter 6

### SUPPLEMENTAL INFORMATION FOR CHAPTER 2: SIGNAL PROCESSING VIA ALLOSTERIC TRANSCRIPTION FACTORS

A version of this chapter originally appeared as Razo-Mejia, M.\* ; Barnes, S.L.\* ; Belliveau, N.M.\* ; Chure, G.\* ; Einav, T.\* ; Lewis, M.; and Phillips, R. (2018). *Tuning Transcriptional Regulation through Signaling: A Predictive Theory of Allosteric Induction*. Cell Systems 6, 456-469.e10. DOI:<https://doi.org/10.1016/j.cels.2018.02.004>. M.R.M, S.L.B, N.M.B, G.C., and T.E. contributed equally to this work from the theoretical underpinnings to the experimental design and execution. M.R.M, S.L.B, N.M.B, G.C, T.E., and R.P. wrote the paper. M.L. provided extensive guidance and advice.

#### **6.1 Inferring Allosteric Parameters from Previous Data**

The fold-change profile described by features three unknown parameters  $K_A$ ,  $K_I$ , and  $\Delta\varepsilon_{AI}$ . In this section, we explore different conceptual approaches to determining these parameters. We first discuss how the induction titration profile of the simple repression constructs used in this paper are not sufficient to determine all three MWC parameters simultaneously, since multiple degenerate sets of parameters can produce the same fold-change response. We then utilize an additional data set from Brewster et al. (2014) to determine the parameter  $\Delta\varepsilon_{AI} = 4.5 k_B T$ , after which the remaining parameters  $K_A$  and  $K_I$  can be extracted from any induction profile with no further degeneracy.

#### **Degenerate Parameter Values**

In this section, we discuss how multiple sets of parameters may yield identical fold-change profiles. More precisely, we show that if we try to fit the data in to the fold-change and extract the three unknown parameters ( $K_A$ ,  $K_I$ , and  $\Delta\varepsilon_{AI}$ ), then multiple degenerate parameter sets would yield equally good fits. In other

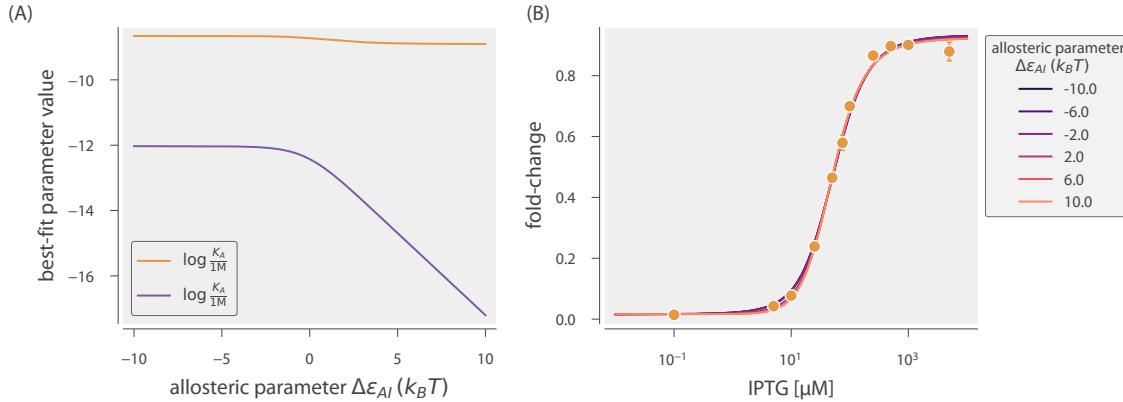
words, this data set alone is insufficient to uniquely determine the actual physical parameter values of the system. This problem persists even when fitting multiple data sets simultaneously as illustrated later in this chapter.

In Fig. 6.1 we fit the  $R = 260$  data by fixing  $\Delta\varepsilon_{AI}$  to the value shown on the  $x$ -axis and determine the parameters  $K_A$  and  $K_I$  given this constraint. We use the fold-change function but with  $\beta\Delta\varepsilon_{RA}$  modified to the form  $\beta\Delta\tilde{\varepsilon}_{RA}$  in to account for the underlying assumptions used when fitting previous data (as is defined in the following section).

The best-fit curves for several different values of  $\Delta\varepsilon_{AI}$  are shown in Fig. 6.1 (B). Note that these fold-change curves are nearly overlapping, demonstrating that different sets of parameters can yield nearly equivalent responses. Without more data, the relationships between the parameter values shown in represent the maximum information about the parameter values that can be extracted from the data. Additional experiments which independently measure any of these unknown parameters could resolve this degeneracy. For example, NMR measurements could be used to directly measure the fraction  $(1 + e^{-\beta\Delta\varepsilon_{AI}})^{-1}$  of active repressors in the absence of IPTG (Boulton and Melacini, 2016; Gardino et al., 2003).

### Computing $\Delta\varepsilon_{AI}$

As shown in the previous section, the fold-change response of a single strain is not sufficient to determine the three MWC parameters ( $K_A$ ,  $K_I$ , and  $\Delta\varepsilon_{AI}$ ), since degenerate sets of parameters yield nearly identical fold-change responses. To circumvent this degeneracy, we now turn to some previous data from the *lac* system in order to determine the value of  $\Delta\varepsilon_{AI}$  in for the induction of the Lac repressor. Specifically, we consider two previous sets of work from: (i) Garcia and Phillips (2011) and (ii) Brewster et al. (2014), both of which measured fold-change with the same simple repression system in the absence of inducer ( $c = 0$ ) but at various repressor copy numbers  $R$ . The original analysis for both data sets assumed that in the absence of inducer all of the Lac repressors were in the active state. As a result, the effective binding energies they extracted were a convolution of the



**Figure 6.1: Multiple sets of parameters yield identical fold-change responses.** (A) The data for the O2 strain ( $\Delta\epsilon_{RA} = -13.9 k_B T$ ) with  $R = 260$  Fig. 2.5(C) was fit using Eq. 2.5 with  $n = 2$ . The allosteric energy difference  $\Delta\epsilon_{AI}$  is forced to take on the value shown on the  $x$ -axis, while  $K_A$  and  $K_I$  are fit freely. (B) The resulting best-fit functions for several values of  $\Delta\epsilon_{AI}$  yield nearly identical fold-change responses. The Python code (ch6\_figS1.py) used to generate this figure can be found on the thesis GitHub repository.

DNA binding energy  $\Delta\epsilon_{RA}$  and the allosteric energy difference  $\Delta\epsilon_{AI}$  between the Lac repressor's active and inactive states. We refer to this convoluted energy value as  $\Delta\tilde{\epsilon}_{RA}$ . We first disentangle the relationship between these parameters in Garcia and Phillips and then use this relationship to extract the value of  $\Delta\epsilon_{AI}$  from Brewster et al. (2014).

Garcia and Phillips determined the total repressor copy numbers  $R$  of different strains using quantitative Western blots. Then they measured the fold-change at these repressor copy numbers for simple repression constructs carrying the O1, O2, O3, and Oid *lac* operators integrated into the chromosome. These data were then fit to the following thermodynamic model to determine the repressor-DNA binding energies  $\Delta\tilde{\epsilon}_{RA}$  for each operator,

$$\text{fold-change}(c = 0) = \left( 1 + \frac{R}{N_{NS}} e^{-\beta \Delta\tilde{\epsilon}_{RA}} \right)^{-1}. \quad (6.1)$$

Note that this functional form does not exactly match our fold-change in the limit

$c = 0$ ,

$$\text{fold-change}(c = 0) = \left( 1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}} \right)^{-1}, \quad (6.2)$$

since it is missing the factor  $\frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}}$  which specifies what fraction of repressors are in the active state in the absence of inducer,

$$\frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} = p_A(0). \quad (6.3)$$

In other words, Garcia and Phillips (2011) assumed that in the absence of inducer, all repressors were active. In terms of our notation, the convoluted energy values  $\Delta\tilde{\varepsilon}_{RA}$  extracted by Garcia and Phillips (namely,  $\Delta\tilde{\varepsilon}_{RA} = -15.3 k_B T$  for O1 and  $\Delta\tilde{\varepsilon}_{RA} = -17.0 k_B T$  for Oid) represent

$$\beta\Delta\tilde{\varepsilon}_{RA} = \beta\Delta\varepsilon_{RA} - \log \left( \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} \right). \quad (6.4)$$

Note that if  $e^{-\beta\Delta\varepsilon_{AI}} \ll 1$ , then nearly all of the repressors are active in the absence of inducer so that  $\Delta\tilde{\varepsilon}_{RA} \approx \Delta\varepsilon_{RA}$ . In simple repression systems where we definitively know the value of  $\Delta\varepsilon_{RA}$  and  $R$ , we can use Eq. 6.4 to determine the value of  $\Delta\varepsilon_{AI}$  by comparing with experimentally determined fold-change values. However, the binding energy values that we use from Garcia and Phillips (2011) are effective parameters  $\Delta\tilde{\varepsilon}_{RA}$ . In this case, we are faced with an undetermined system in which we have more variables than equations, and we are thus unable to determine the value of  $\Delta\varepsilon_{AI}$ . In order to obtain this parameter, we must turn to a more complex regulatory scenario which provides additional constraints that allow us to fit for  $\Delta\varepsilon_{AI}$ .

A variation on simple repression in which multiple copies of the promoter are available for repressor binding (for instance, when the simple repression construct is on plasmid) can be used to circumvent the problems that arise when using  $\Delta\tilde{\varepsilon}_{RA}$ . This is because the behavior of the system is distinctly different when the number of active repressors  $p_A(0)R$  is less than or greater than the number of available promoters  $N$ . Repression data for plasmids with known copy number  $N$  allows us to perform a fit for the value of  $\Delta\varepsilon_{AI}$ .

To obtain an expression for a system with multiple promoters  $N$ , we follow Weinert et al. (2014), writing the fold-change in terms of the the grand canonical ensemble as

$$\text{fold-change} = \frac{1}{1 + \lambda_r e^{-\beta \Delta \varepsilon_{RA}}}, \quad (6.5)$$

where  $\lambda_r = e^{\beta \mu}$  is the fugacity and  $\mu$  is the chemical potential of the repressor. The fugacity will enable us to easily enumerate the possible states available to the repressor.

To determine the value of  $\lambda_r$ , we first consider that the total number of repressors in the system,  $R_{\text{tot}}$ , is fixed and given by

$$R_{\text{tot}} = R_S + R_{NS}, \quad (6.6)$$

where  $R_S$  represents the number of repressors specifically bound to the promoter and  $R_{NS}$  represents the number of repressors nonspecifically bound throughout the genome. The value of  $R_S$  is given by

$$R_S = N \frac{\lambda_r e^{-\beta \Delta \varepsilon_{RA}}}{1 + \lambda_r e^{-\beta \Delta \varepsilon_{RA}}}, \quad (6.7)$$

where  $N$  is the number of available promoters in the cell. Note that in counting  $N$ , we do not distinguish between promoters that are on plasmid or chromosomally integrated provided that they both have the same repressor-operator binding energy (Weinert et al., 2014). The value of  $R_{NS}$  is similarly given by

$$R_{NS} = N_{NS} \frac{\lambda_r}{1 + \lambda_r}, \quad (6.8)$$

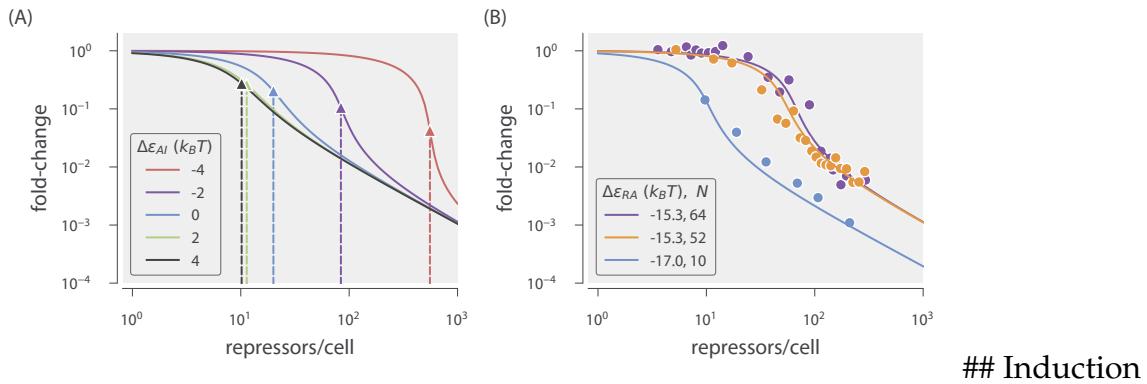
where  $N_{NS}$  is the number of non-specific sites in the cell (recall that we use  $N_{NS} = 4.6 \times 10^6$  for *E. coli*). Substituting in into the modified yields the form

$$p_A(0)R_{\text{tot}} = \frac{1}{1 + e^{-\beta \Delta \varepsilon_{AI}}} \left( N \frac{\lambda_r e^{-\beta \Delta \varepsilon_{RA}}}{1 + \lambda_r e^{-\beta \Delta \varepsilon_{RA}}} + N_{NS} \frac{\lambda_r}{1 + \lambda_r} \right), \quad (6.9)$$

where we recall from Eq. 6.4 that  $\beta \Delta \varepsilon_{RA} = \beta \Delta \tilde{\varepsilon}_{RA} + \log \left( \frac{1}{1 + e^{-\beta \Delta \varepsilon_{AI}}} \right)$ . Numerically solving for  $\lambda_r$  and plugging the value back into yields a fold-change function in which the only unknown parameter is  $\Delta \varepsilon_{AI}$ .

With these calculations in hand, we can now determine the value of the  $\Delta\epsilon_{AI}$  parameter. shows how different values of  $\Delta\epsilon_{AI}$  lead to significantly different fold-change response curves. Thus, analyzing the specific fold-change response of any strain with a known plasmid copy number  $N$  will fix  $\Delta\epsilon_{AI}$ . Interestingly, the inflection point of occurs near  $p_A(0)R_{\text{tot}} = N$  (as shown by the triangles in Fig. ??), so that merely knowing where the fold-change response transitions from concave down to concave up is sufficient to obtain a rough value for  $\Delta\epsilon_{AI}$ . We note, however, that for  $\Delta\epsilon_{AI} \geq 5 k_B T$ , increasing  $\Delta\epsilon_{AI}$  further does not affect the fold-change because essentially every repressors will be in the active state in this regime. Thus, if the  $\Delta\epsilon_{AI}$  is in this regime, we can only bound it from below.

We now analyze experimental induction data for different strains with known plasmid copy numbers to determine  $\Delta\epsilon_{AI}$ . shows experimental measurements of fold-change for two O1 promoters with  $N = 64$  and  $N = 52$  copy numbers and one Oid promoter with  $N = 10$  from Brewster et al. (2014). By fitting these data to Eq. 6.5, we extracted the parameter value  $\Delta\epsilon_{AI} = 4.5 k_B T$ . Substituting this value into shows that 99% of the repressors are in the active state in the absence of inducer and  $\Delta\tilde{\epsilon}_{RA} \approx \Delta\epsilon_{RA}$ , so that all of the previous energies and calculations made by Garcia and Phillips (2011) and Brewster et al. (2014) were accurate.



**of Simple Repression with Multiple Promoters or Competitor Sites** We made the choice to perform all of our experiments using strains in which a single copy of our simple repression construct had been integrated into the chromosome. This stands in contrast to the methods used by a number of other studies (Daber et al.,

2009, 2011; Oehler et al., 1994; Setty et al., 2003; Shis et al., 2014; Sochor, 2014; Vilal and Saiz, 2013), in which reporter constructs are placed on plasmid, meaning that the number of constructs in the cell is not precisely known. It is also common to express repressor on plasmid to boost its copy number, which results in an uncertain value for repressor copy number. Here we show that our treatment of the MWC model has broad predictive power beyond the single-promoter scenario we explore experimentally, and indeed can account for systems in which multiple promoters compete for the repressor of interest. Additionally, we demonstrate the importance of having precise control over these parameters, as they can have a significant effect on the induction profile.

### Chemical Potential Formulation to Calculate Fold-Change

In this section, we discuss a simple repression construct which we generalize in two ways from the scenario discussed in Chapter 2. First, we will allow the repressor to bind to  $N_S$  identical specific promoters whose fold-change we are interested in measuring, with each promoter containing a single repressor binding site ( $N_S = 1$  in Chapter 2). Second, we consider  $N_C$  identical competitor sites which do not regulate the promoter of interest, but whose binding energies are substantially stronger than non-specific binding ( $N_C = 0$  in Chapter 2). As in Chapter 2, we assume that the rest of the genome contains  $N_{NS}$  non-specific binding sites for the repressor. Using the formalism described in the previous section, we can write the fold-change in the grand canonical ensemble as

$$\text{fold-change} = \frac{1}{1 + \lambda_r e^{-\beta \Delta \varepsilon_{RA}}}, \quad (6.10)$$

where  $\lambda_r$  is the fugacity of the repressor and  $\Delta \varepsilon_{RA}$  represents the energy difference between the repressor's binding affinity to the specific operator of interest relative to the repressor's non-specific binding affinity to the rest of the genome.

We now expand our definition of the total number of repressors in the system,  $R_{\text{tot}}$ , so that it is given by

$$R_{\text{tot}} = R_S + R_{NS} + R_C, \quad (6.11)$$

where  $R_S$ ,  $R_{NS}$ , and  $R_C$  represent the number of repressors bound to the specific promoter, a non-specific binding site, or to a competitor binding site, respectively. The value of  $R_S$  is given by

$$R_S = N_S \frac{\lambda_r e^{-\beta \Delta \varepsilon_{RA}}}{1 + \lambda_r e^{-\beta \Delta \varepsilon_{RA}}}, \quad (6.12)$$

where  $N_S$  is the number of specific binding sites in the cell. The value of  $R_{NS}$  is similarly given by

$$R_{NS} = N_{NS} \frac{\lambda_r}{1 + \lambda_r}, \quad (6.13)$$

where  $N_{NS}$  is the number of non-specific sites in the cell (recall that we use  $N_{NS} = 4.6 \times 10^6$  for *E. coli*), and  $R_C$  is given by

$$R_C = N_C \frac{\lambda_r e^{-\beta \Delta \varepsilon_C}}{1 + \lambda_r e^{-\beta \Delta \varepsilon_C}}, \quad (6.14)$$

where  $N_C$  is the number of competitor sites in the cell and  $\Delta \varepsilon_C$  is the binding energy of the repressor to the competitor site relative to its non-specific binding energy to the rest of the genome.

To account for the induction of the repressor, we replace the total number of repressors  $R_{\text{tot}}$  in Eq. 6.11 by the number of active repressors in the cell,  $p_{\text{act}}(c)R_{\text{tot}}$ . Here,  $p_{\text{act}}$  denotes the probability that the repressor is in the active state (Eq. 2.4),

$$p_{\text{act}}(c) = \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{-\beta \Delta \varepsilon_{AI}} \left(1 + \frac{c}{K_I}\right)^n}. \quad (6.15)$$

Substituting Eq. 6.12 into the modified Eq. 6.11 yields

$$p_{\text{active}}(c)R_{\text{tot}} = N_S \frac{\lambda_r e^{-\beta \Delta \varepsilon_{RA}}}{1 + \lambda_r e^{-\beta \Delta \varepsilon_{RA}}} + N_{NS} \frac{\lambda_r}{1 + \lambda_r} + N_C \frac{\lambda_r e^{-\beta \Delta \varepsilon_C}}{1 + \lambda_r e^{-\beta \Delta \varepsilon_C}}. \quad (6.16)$$

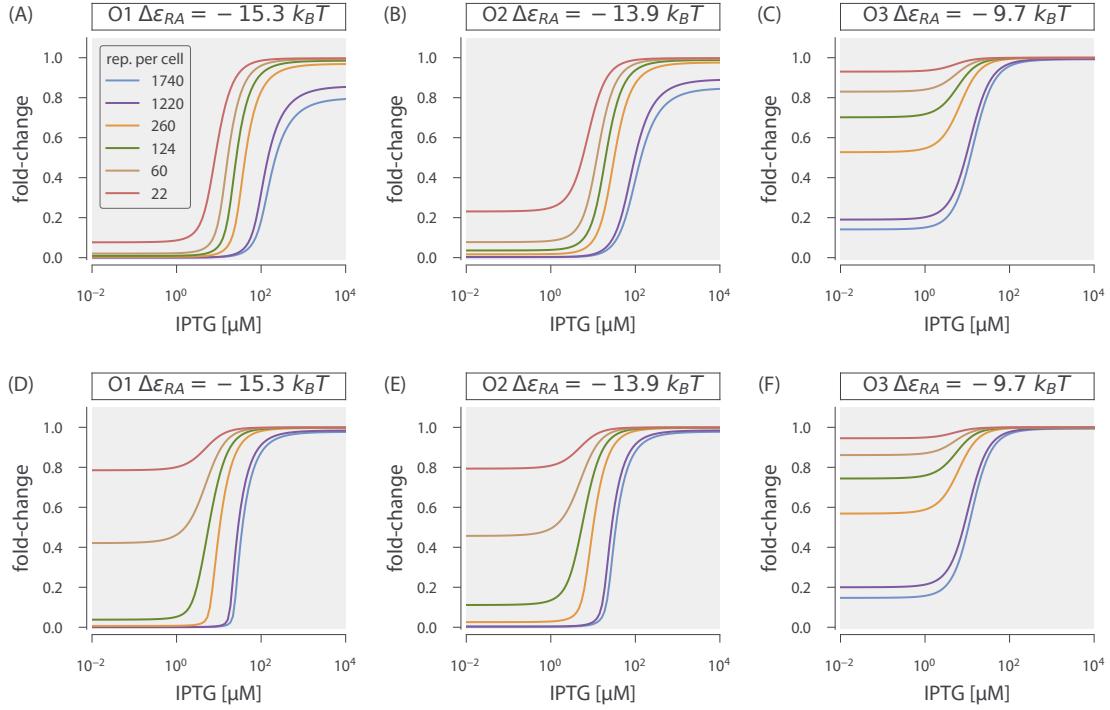
For systems where the number of binding sites  $N_S$ ,  $N_{NS}$ , and  $N_C$  are known, together with the binding affinities  $\Delta \varepsilon_{RA}$  and  $\Delta \varepsilon_C$ , we can solve numerically for  $\lambda_r$  and then substitute it into to obtain a fold-change at any concentration of inducer  $c$ . In the following sections, we will theoretically explore the induction curves given by Eq. 6.16 for a number of different combinations of simple repression binding sites, thereby predicting how the system would behave if additional specific or competitor binding sites were introduced.

## 6.2 Variable Repressor Copy Number ( $R$ ) with Multiple Specific Binding Sites ( $N_S > 1$ )

In Chapter 2, we consider the induction profiles of strains with varying  $R$  but a single, specific binding site  $N_S = 1$ . Here we predict the induction profiles for similar strains in which  $R$  is varied, but  $N_S > 1$ , as shown in Fig. 6.2. The top row shows induction profiles in which  $N_S = 10$  and the bottom row shows profiles in which  $N_S = 100$ , assuming three different choices for the specific operator binding sites given by the O1, O2, and O3 operators. These values of  $N_S$  were chosen to mimic the common scenario in which a promoter construct is placed on either a low or high copy number plasmid. A few features stand out in these profiles. First, as the magnitude of  $N_S$  surpasses the number of repressors  $R$ , the leakiness begins to increase significantly, since there are no longer enough repressors to regulate all copies of the promoter of interest. Second, in the cases where  $\Delta\varepsilon_{RA} = -15.3 k_B T$  for the O1 operator or  $\Delta\varepsilon_{RA} = -13.9 k_B T$  for the O2 operator, the profiles where  $N_S = 100$  are notably sharper than the profiles where  $N_S = 10$ , and it is possible to achieve dynamic ranges approaching 1. Finally, it is interesting to note that the profiles for the O3 operator where  $\Delta\varepsilon_{RA} = -9.7 k_B T$  are nearly indifferent to the value of  $N_S$ .

## 6.3 Variable Number of Specific Binding Sites $N_S$ with Fixed Repressor Copy Number ( $R$ )

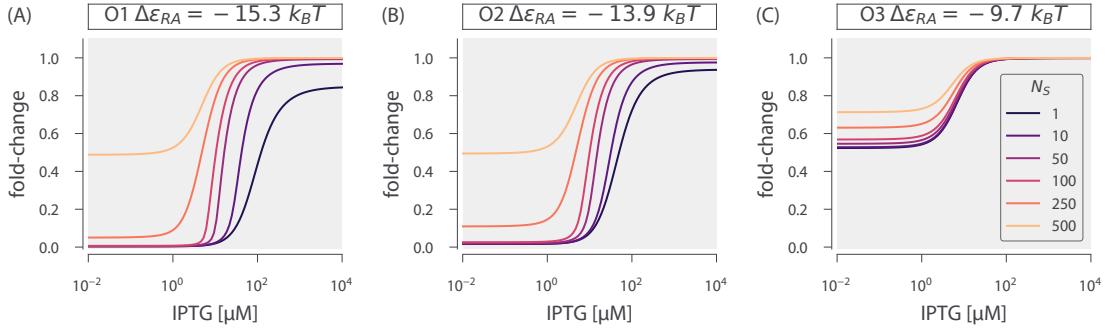
The second set of scenarios we consider is the case in which the repressor copy number  $R = 260$  is held constant while the number of specific promoters  $N_S$  is varied (see Fig. 6.3). Again, we see that leakiness is increased significantly when  $N_S > R$ , though all profiles for  $\Delta\varepsilon_{RA} = -9.7 k_B T$  exhibit high leakiness, making the effect less dramatic for this operator. Additionally, we find again that adjusting the number of specific sites can produce induction profiles with maximal dynamic ranges. In particular, the O1 and O2 profiles with  $\Delta\varepsilon_{RA} = -15.3$  and  $-13.9 k_B T$ , respectively, have dynamic ranges approaching 1 for  $N_S = 50$  and 100.



**Figure 6.2: Induction with variable  $R$  and multiple specific binding sites.** Induction profiles are shown for strains with variable  $R$  and  $\Delta\epsilon_{RA} = -15.3, -13.9$ , or  $-9.7 \text{ } k_B T$ . The number of specific sites,  $N_S$ , is held constant at 10 as  $R$  and  $\Delta\epsilon_{RA}$  are varied.  $N_S$  is held constant at 100 as  $R$  and  $\Delta\epsilon_{RA}$  are varied. These situations mimic the common scenario in which a promoter construct is placed on either a low or high copy number plasmid. The Python code (`ch6_figS3.py`) used to generate this figure can be found on the thesis GitHub repository.

#### 6.4 Competitor Binding Sites

An intriguing scenario is presented by the possibility of competitor sites elsewhere in the genome. This serves as a model for situations in which a promoter of interest is regulated by a transcription factor that has multiple targets. This is highly relevant, as the majority of transcription factors in *E. coli* have at least two known binding sites, with approximately 50 transcription factors having more than ten known binding sites (Rydenfelt et al., 2014b; Schmidt et al., 2016). If the number of competitor sites and their average binding energy is known, however, they can be accounted for in the model. Here, we predict the induction profiles for strains in which  $R = 260$  and  $N_S = 1$ , but there is a variable number of competitor sites  $N_C$



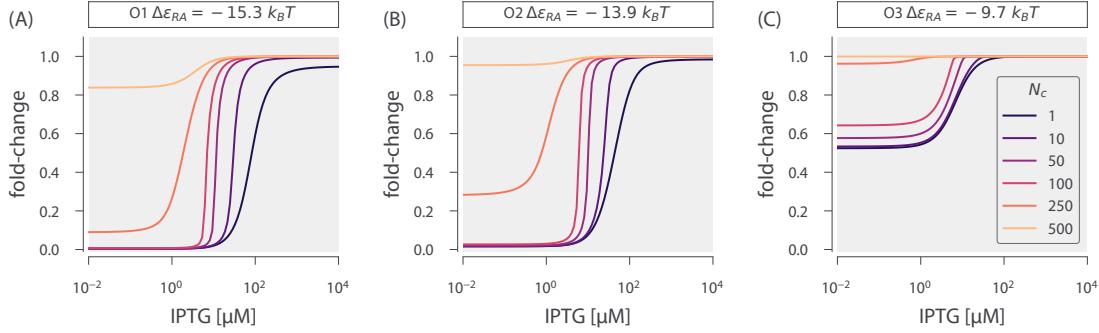
**Figure 6.3: Induction with variable specific sites and fixed  $R$ .** Induction profiles are shown for strains with  $R = 260$  and  $\Delta\epsilon_{RA} = -15.3 \text{ } k_B T$ ,  $\Delta\epsilon_{RA} = -13.9 \text{ } k_B T$ , or  $\Delta\epsilon_{RA} = -9.7 \text{ } k_B T$ . The number of specific sites  $N_S$  is varied from 1 to 500. The Python code (`ch6_figS4.py`) used to generate this figure can be found on the thesis GitHub repository.

with a strong binding energy  $\Delta\epsilon_C = -17.0 \text{ } k_B T$ . In the presence of such a strong competitor, when  $N_C > R$  the leakiness is greatly increased, as many repressors are siphoned into the pool of competitor sites. This is most dramatic for the case where  $\Delta\epsilon_{RA} = -9.7 \text{ } k_B T$ , in which it appears that no repression occurs at all when  $N_C = 500$ . Interestingly, when  $N_C < R$  the effects of the competitor are not especially notable.

## 6.5 Properties of the Induction Response

As discussed in the main body of the paper, our treatment of the MWC model allows us to predict key properties of induction responses. Here, we consider the leakiness, saturation, and dynamic range (diagrammed in Fig. 2.1) by numerically solving Eq. 6.16 in the absence of inducer,  $c = 0$ , and in the presence of saturating inducer  $c \rightarrow \infty$ . Using Eq. 6.15, the former case is given by

$$R_{\text{tot}} \frac{1}{1 + e^{-\beta\Delta\epsilon_{AI}}} = N_S \frac{\lambda_r e^{-\beta\Delta\epsilon_{RA}}}{1 + \lambda_r e^{-\beta\Delta\epsilon_{RA}}} + N_{NS} \frac{\lambda_r}{1 + \lambda_r} + N_C \frac{\lambda_r e^{-\beta\Delta\epsilon_C}}{1 + \lambda_r e^{-\beta\Delta\epsilon_C}}, \quad (6.17)$$



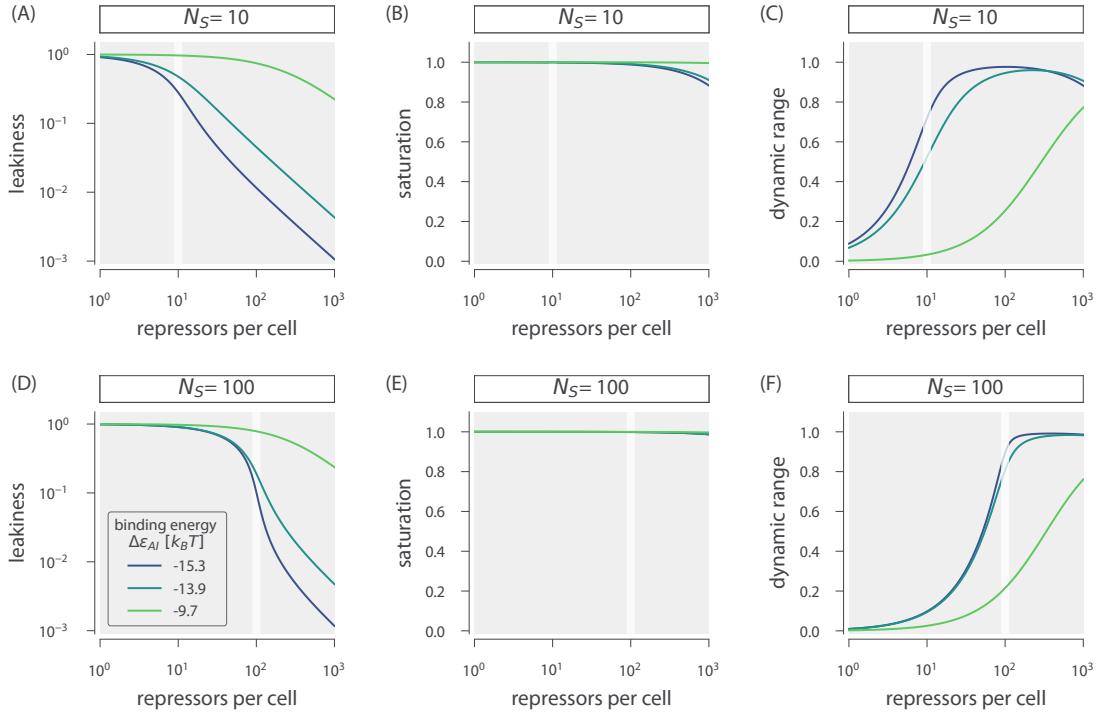
**Figure 6.4: Induction with variable competitor sites, a single specific site, and fixed  $R$ .** Induction profiles are shown for strains with  $R = 260$ ,  $N_s = 1$ , and  $\Delta\epsilon_{RA} = -15.3 \text{ } k_B T$  for the O1 operator,  $\Delta\epsilon_{RA} = -13.9 \text{ } k_B T$  for the O2 operator, or  $\Delta\epsilon_{RA} = -9.7 \text{ } k_B T$  for the O3 operator. The number of specific sites,  $N_C$ , is varied from 1 to 500. This mimics the common scenario in which a transcription factor has multiple binding sites in the genome. The Python code (ch6\_figS5.py) used to generate this figure can be found on the thesis GitHub repository.

whereupon substituting in the value of  $\lambda_r$  into Eq. ?? will yield the leakiness. Similarly, the limit of saturating inducer is found by determining  $\lambda_r$  from the form

$$R_{\text{tot}} \frac{1}{1 + e^{-\beta\Delta\epsilon_{AI}} \left( \frac{K_A}{K_I} \right)^2} = N_S \frac{\lambda_r e^{-\beta\Delta\epsilon_{RA}}}{1 + \lambda_r e^{-\beta\Delta\epsilon_{RA}}} + N_{NS} \frac{\lambda_r}{1 + \lambda_r} + N_C \frac{\lambda_r e^{-\beta\Delta\epsilon_C}}{1 + \lambda_r e^{-\beta\Delta\epsilon_C}}. \quad (6.18)$$

In Fig. 6.5, we show how the leakiness, saturation, and dynamic range vary with  $R$  and  $\Delta\epsilon_{RA}$  in systems with  $N_S = 10$  or  $N_S = 100$ . An inflection point occurs where  $N_S = R$ , with leakiness and dynamic range behaving differently when  $R < N_S$  than when  $R > N_S$ . This transition is more dramatic for  $N_S = 100$  than for  $N_S = 10$ . Interestingly, the saturation values consistently approach 1, indicating that full induction is easier to achieve when multiple specific sites are present. Moreover, dynamic range values for O1 and O2 strains with  $\Delta\epsilon_{RA} = -15.3$  and  $-13.9 \text{ } k_B T$  approach 1 when  $R > N_S$ , although when  $N_S = 10$  there is a slight downward dip owing to saturation values of less than 1 at high repressor copy numbers.

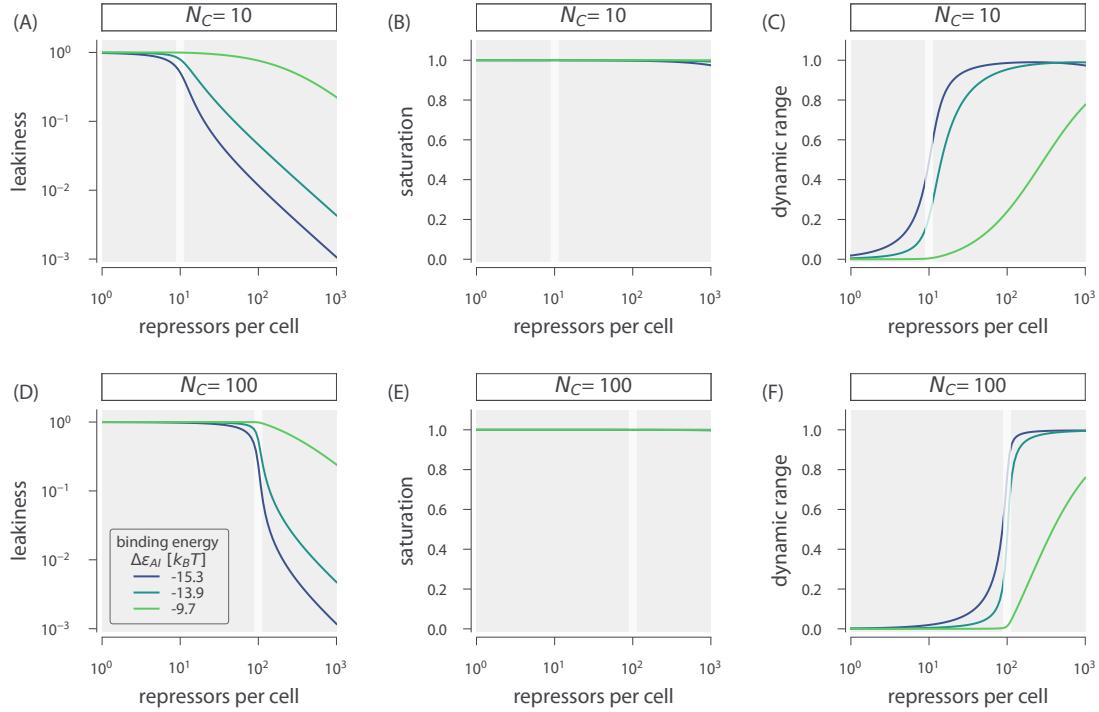
In Fig. 6.6, we similarly show how the leakiness, saturation, and dynamic range vary with  $R$  and  $\Delta\epsilon_{RA}$  in systems with  $N_S = 1$  and multiple competitor



**Figure 6.5: Phenotypic properties of induction with multiple specific binding sites.** The leakiness, saturation, and dynamic range are shown for systems with number of specific binding sites  $N_S = 10$  or  $N_S = 100$ . The vertical white line indicates the point at which  $N_S = R$ . The Python code (ch6\_figS6.py) used to generate this figure can be found on the thesis GitHub repository.

sites  $N_C = 10$  or  $N_C = 100$ . Each of the competitor sites has a binding energy of  $\Delta\epsilon_C = -17.0 \text{ } k_B T$ . The phenotypic profiles are very similar to those for multiple specific sites shown in , with sharper transitions at  $R = N_C$  due to the greater binding strength of the competitor site. This indicates that introducing competitors has much the same effect on the induction phenotypes as introducing additional specific sites, as in either case the influence of the repressors is damped when there are insufficient repressors to interact with all of the specific binding sites.

This section gives a quantitative analysis of the nuances imposed on induction response in the case of systems involving multiple gene copies as are found in the vast majority of studies on induction. In these cases, the intrinsic parameters of the MWC model get entangled with the parameters describing gene copy number.



**Figure 6.6: Phenotypic properties of induction with a single specific site and multiple competitor sites.** The leakiness, saturation, and dynamic range are shown for systems with a single specific binding site  $N_S = 1$  and a number of competitor sites  $N_C = 10$  or  $N_C = 100$ . All competitor sites have a binding energy of  $\Delta\epsilon_C = -17.0 \text{ } k_B T$ . The vertical white line indicates the point at which  $N_C = R$ . The Python code (`ch6_figS7.py`) used to generate this figure can be found on the thesis GitHub repository.

## 6.6 Flow Cytometry

In this section, we provide information regarding the equipment used to make experimental measurements of the fold-change in gene expression in the interests of transparency and reproducibility. We also provide a summary of our unsupervised method of gating the flow cytometry measurements for consistency between experimental runs.

### Equipment

Due to past experience using the Miltenyi Biotec MACSQuant flow cytometer during the Physiology summer course at the Marine Biological Laboratory, we used the same flow cytometer for the formal measurements in this work graciously pro-

vided by the Pamela Björkman lab at Caltech. All measurements were made using an excitation wavelength of 488 nm with an emission filter set of 525/50 nm. This excitation wavelength provides approximately 40% of the maximum YFP absorbance , and this was found to be sufficient for the purposes of these experiments. A useful feature of modern flow cytometry is the high-sensitivity signal detection through the use of photomultiplier tubes (PMT) whose response can be tuned by adjusting the voltage. Thus, the voltage for the forward-scatter (FSC), side-scatter (SSC), and gene expression measurements were tuned manually to maximize the dynamic range between autofluorescence signal and maximal expression without losing the details of the population distribution. Once these voltages were determined, they were used for all subsequent measurements. Extremely low signal producing particles were discarded before data storage by setting a basal voltage threshold, thus removing the majority of spurious events. The various instrument settings for data collection are given in Table 6.1.

Table 6.1: Instrument settings for data collectuion us-  
ing the Miltenyi Biotec MACSQuant flow cytometer.

<b>Laser</b>	<b>Channel</b>	<b>Sensor Voltage</b>
488 nm	Forward-Scatter (FSC)	423 V
488 nm	Side-Scatter (SSC)	537 V
488 nm	Intensity (B1 Filter, 525/50 nm)	790 V
488 nm	Trigger (debris threshold)	24.5V

### Unsupervised Gating

Flow cytometry data will frequently include a number of spurious events or other undesirable data points such as cell doublets and debris. The process of restricting the collected data set to those data determined to be “real” is commonly referred to as gating. These gates are typically drawn manually and restrict the data set to those points which display a high degree of linear correlation between their forward-scatter (FSC) and side-scatter (SSC). The development of unbiased

and unsupervised methods of drawing these gates is an active area of research (Aghaeepour et al., 2013).

For this study, we used an automatic unsupervised gating procedure to filter the flow cytometry data based on the front and side-scattering values returned by the MACSQuant flow cytometer. We assume that the region with highest density of points in these two channels corresponds to single-cell measurements. Everything extending outside of this region was discarded in order to exclude sources of error such as cell clustering, particulates, or other spurious events.

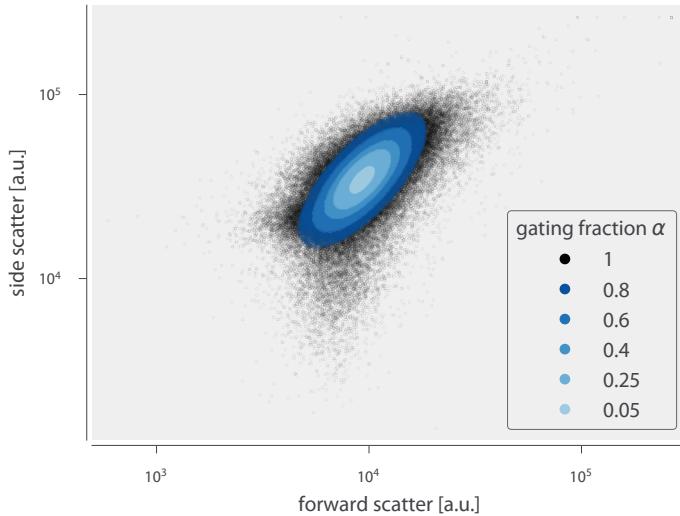
In order to define the gated region we fit a two-dimensional Gaussian function to the  $\log_{10}$  forward-scattering (FSC) and the  $\log_{10}$  side-scattering (SSC) data. We then kept a fraction  $\alpha \in [0, 1]$  of the data by defining an elliptical region given by

$$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \leq \chi_\alpha^2(p), \quad (6.19)$$

where  $\mathbf{x}$  is the  $2 \times 1$  vector containing the  $\log(\text{FSC})$  and  $\log(\text{SSC})$ ,  $\boldsymbol{\mu}$  is the  $2 \times 1$  vector representing the mean values of  $\log(\text{FSC})$  and  $\log(\text{SSC})$  as obtained from fitting a two-dimensional Gaussian to the data, and  $\boldsymbol{\Sigma}$  is the  $2 \times 2$  covariance matrix also obtained from the Gaussian fit.  $\chi_\alpha^2(p)$  is the quantile function for probability  $p$  of the chi-squared distribution with two degrees of freedom. Fig. 6.7 shows an example of different gating contours that would arise from different values of  $\alpha$  in Eq. 6.19. In this work, we chose  $\alpha = 0.4$  which we deemed was a sufficient constraint to minimize the noise in the data. The specific code where this gating is implemented can be found in GitHub repository.

### Comparison of Flow Cytometry with Other Methods

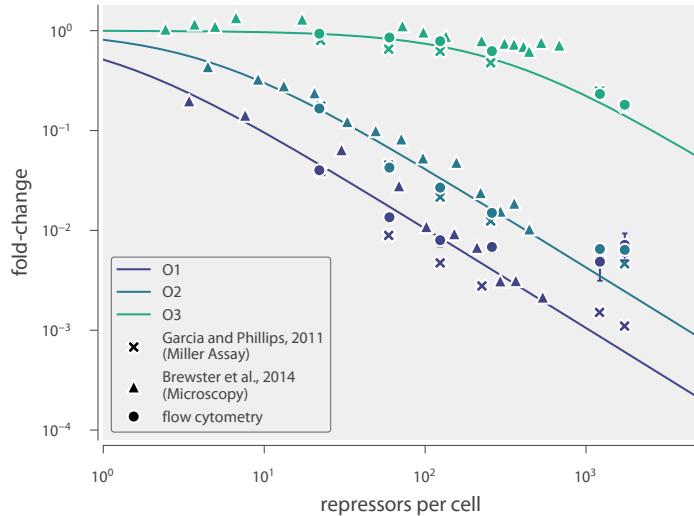
Previous work from the Phillips' lab experimentally determined fold-change for similar simple repression constructs using a variety of different measurement methods (Garcia et al., 2011a). Garcia and Phillips used the same background strains as the ones used in this work, but gene expression was measured with Miller assays based on colorimetric enzymatic reactions with the LacZ protein (Garcia and Phillips, 2011). The experiments in Brewster et al. (2014) (as well as in Chapter



**Figure 6.7: Representative unsupervised gating contours of flow-cytometry data.** Points indicate individual flow cytometry measurements of forward scatter and side scatter. Colored contours indicate arbitrary gating contours ranging from 100% ( $\alpha = 1$ ) to 5% ( $\alpha = 0.05$ ). All measurements shown in Chapters 2 and 3 in this work were made by computing the mean fluorescence from the 40<sup>th</sup> percentile ( $\alpha = 0.4$ ). The Python code (ch6\_figS8.py) used to generate this figure can be found on the thesis GitHub repository.

4 of this dissertation) used a LacI dimer with the tetramerization region replaced with an mCherry tag, where the fold-change was measured as the ratio of the gene expression rate rather than a single snapshot of the gene output.

Fig. 6.8 shows the comparison of these methods along with the flow cytometry method used in this work. The consistency of these three readouts validates the quantitative use of flow cytometry and unsupervised gating to determine the fold-change in gene expression. However, one important caveat revealed by this figure is that the sensitivity of flow cytometer measurements is not sufficient to accurately determine the fold-change for the high repressor copy number strains in O1 without induction. Instead, a method with a large dynamic range such as the Miller assay is needed to accurately resolve the fold-change at such low levels of expression.



**Figure 6.8: Comparison of experimental methods to determine the fold-change.** The fold-change in gene expression of equivalent simple-repression constructs has been determined using three independent methods: flow cytometry (Chapter 2), colorimetric Miller assays (Garcia and Phillips (2011)), and video microscopy (Brewster et al. (2014)). All three methods give consistent results, although flow cytometry measurements lose accuracy for fold-change less than 0.01. Note that the repressor-DNA binding energies  $\Delta\epsilon_{RA}$  used for the theoretical predictions were determined in Garcia and Phillips (2011). The Python code (ch6\_figS9.py) used to generate this figure can be found on the thesis GitHub repository.

## 6.7 Single-Cell Microscopy

In this section, we detail the procedures and results from single-cell microscopy verification of our flow cytometry measurements. Our previous measurements of fold-change in gene expression have been measured using bulk-scale Miller assays (Garcia and Phillips, 2011) or through single-cell microscopy (Brewster et al., 2014). In this work, flow cytometry was an attractive method due to the ability to screen through many different strains at different concentrations of inducer in a short amount of time. To verify our results from flow cytometry, we examined two bacterial strains with different repressor-DNA binding energies ( $\Delta\epsilon_{RA}$ ) of  $-13.9 k_B T$  and  $-15.3 k_B T$  with  $R = 260$  repressors per cell using fluorescence microscopy and estimated the values of the parameters  $K_A$  and  $K_I$  for direct comparison between the two methods. For a detailed explanation of the Python code implementation

of the processing steps described below, please see this paper's GitHub repository.

### Strains and Growth Conditions

Cells were grown in an identical manner to those used for measurement via flow cytometry (see Materials & Methods of Chapter 2). Briefly, cells were grown overnight (between 10 and 13 hours) to saturation in rich media broth (LB) with  $100 \mu\text{g} \cdot \text{mL}^{-1}$  spectinomycin in a deep-well 96 well plate at  $37^\circ\text{C}$ . These cultures were then diluted 1000-fold into  $500 \mu\text{L}$  of M9 minimal medium supplemented with 0.5% glucose and the appropriate concentration of the inducer IPTG. Strains were allowed to grow at  $37^\circ\text{C}$  with vigorous aeration for approximately 8 hours. Prior to mounting for microscopy, the cultures were diluted 10-fold into M9 glucose minimal medium in the absence of IPTG. Each construct was measured using the same range of inducer concentration values as was performed in the flow cytometry measurements (between 100 nM and 5 mM IPTG). Each condition was measured in triplicate in microscopy whereas approximately ten measurements were made using flow cytometry.

### Imaging Procedure

During the last hour of cell growth, an agarose mounting substrate was prepared containing the appropriate concentration of the IPTG inducer. This mounting substrate was composed of M9 minimal medium supplemented with 0.5% glucose and 2% agarose (Life Technologies UltraPure Agarose, Cat. No. 16500100). This solution was heated in a microwave until molten followed by addition of the IPTG to the appropriate final concentration. This solution was then thoroughly mixed and a  $500 \mu\text{L}$  aliquot was sandwiched between two glass coverslips and was allowed to solidify.

Once solid, the agarose substrates were cut into approximately  $10 \text{ mm} \times 10 \text{ mm}$  squares. An aliquot of one to two microliters of the diluted cell suspension was then added to each pad. For each concentration of inducer, a sample of the autofluorescence control, the  $\Delta lacI$  constitutive expression control, and the experimental

strain was prepared yielding a total of thirty-six agarose mounts per experiment. These samples were then mounted onto two glass-bottom dishes (Ted Pella Wilco Dish, Cat. No. 14027-20) and sealed with parafilm.

All imaging was performed on a Nikon Ti-Eclipse inverted fluorescent microscope outfitted with a custom-built laser illumination system and operated by the open-source MicroManager control software (Edelstein et al., 2014). The YFP fluorescence was imaged using a CrystaLaser 514 nm excitation laser coupled with a laser-optimized (Semrock Cat. No. LF514-C-000) emission filter.

For each sample, between fifteen and twenty positions were imaged allowing for measurement of several hundred cells. At each position, a phase contrast image, an mCherry image, and a YFP image were collected in that order with exposures on a time scale of ten to twenty milliseconds. For each channel, the same exposure time was used across all samples in a given experiment. All images were collected and stored in `ome.tiff` format. All microscopy images are available on the CaltechDATA online repository under DOI: 10.22002/D1.229.

## Image Processing

### Correcting Uneven Illumination

The excitation laser has a two-dimensional gaussian profile. To minimize non-uniform illumination of a single field of view, the excitation beam was expanded to illuminate an area larger than that of the camera sensor. While this allowed for an entire field of view to be illuminated, there was still approximately a 10% difference in illumination across both dimensions. This nonuniformity was corrected for in post-processing by capturing twenty images of a homogeneously fluorescent plastic slide (Autofluorescent Plastic Slides, Chroma Cat. No. 920001) and averaging to generate a map of illumination intensity at any pixel  $I_{YFP}$ . To correct for shot noise in the camera (Andor iXon+ 897 EMCCD), twenty images were captured in the absence of illumination using the exposure time used for the experimental data. Averaging over these images produced a map of background noise at any

pixel  $I_{\text{dark}}$ . To perform the correction, each fluorescent image in the experimental acquisition was renormalized with respect to these average maps as

$$I_{\text{flat}} = \frac{I - I_{\text{dark}}}{I_{\text{YFP}} - I_{\text{dark}}} \langle I_{\text{YFP}} - I_{\text{dark}} \rangle, \quad (6.20)$$

where  $I_{\text{flat}}$  is the renormalized image and  $I$  is the original fluorescence image.

### Cell Segmentation

Each bacterial strain constitutively expressed an mCherry fluorophore from a low copy-number plasmid. This served as a volume marker of cell mass allowing us to segment individual cells through edge detection in fluorescence. We used the Marr-Hildreth edge detector which identifies edges by taking the second derivative of a lightly Gaussian blurred image. Edges are identified as those regions which cross from highly negative to highly positive values or vice-versa within a specified neighborhood. Bacterial cells were defined as regions within an intact and closed identified edge. All segmented objects were then labeled and passed through a series of filtering steps.

To ensure that primarily single cells were segmented, we imposed area and eccentricity bounds. We assumed that single cells projected into two dimensions are roughly  $2 \mu\text{m}$  long and  $1 \mu\text{m}$  wide, so that cells are likely to have an area between  $0.5 \mu\text{m}^2$  and  $6 \mu\text{m}$ . To determine the eccentricity bounds, we assumed that the a single cell can be approximated by an ellipse with semi-major ( $a$ ) and semi-minor ( $b$ ) axis lengths of  $0.5 \mu\text{m}$  and  $0.25 \mu\text{m}$ , respectively. The eccentricity of this hypothetical cell can be computed as

$$\text{eccentricity} = \sqrt{1 - \left(\frac{b}{a}\right)^2}, \quad (6.21)$$

yielding a value of approximately 0.8. Any objects with an eccentricity below this value were not considered to be single cells. After imposing both an area and eccentricity filter, the remaining objects were considered cells of interest and the mean fluorescence intensity of each cell was extracted.

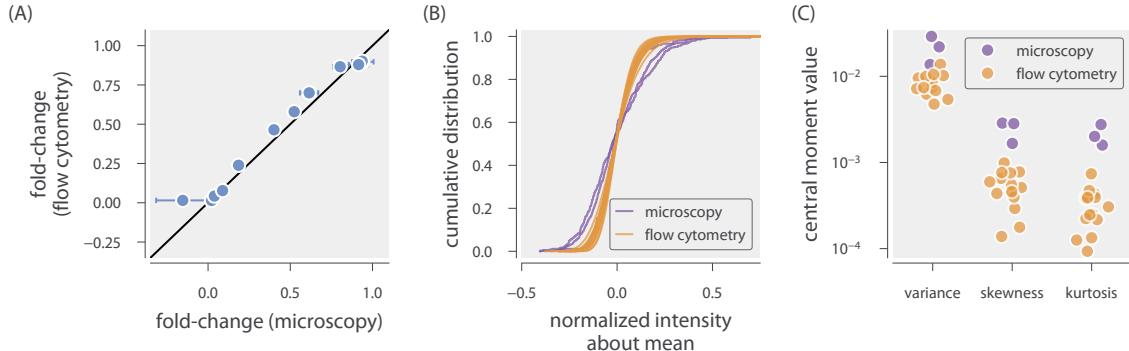
### Calculation of Fold-Change and Empirical Comparison

Cells exhibited background fluorescence even in the absence of an expressed fluorophore. We corrected for this autofluorescence contribution to the fold-change calculation by subtracting the mean YFP fluorescence of cells expressing only the mCherry volume marker from each experimental measurement. The fold-change in gene expression was therefore calculated as

$$\text{fold-change} = \frac{\langle I_{R>0} \rangle - \langle I_{\text{auto}} \rangle}{\langle I_{R=0} \rangle - \langle I_{\text{auto}} \rangle}, \quad (6.22)$$

where  $\langle I_{R>0} \rangle$  is the mean fluorescence intensity of cells expressing LacI repressors,  $\langle I_{\text{auto}} \rangle$  is the mean intensity of cells expressing only the mCherry volume marker, and  $\langle I_{R=0} \rangle$  is the mean fluorescence intensity of cells in the absence of LacI.

The agreement in the fold-change in gene expression between single-cell microscopy and flow cytometry can be seen in Fig. 6.9 (A) where the two methods have been plotted against each other. At this level, we see near perfect agreement between the methods when examining the mean level of gene expression. However, there is a distinct difference in higher moments of the gene expression distributions. Empirical cumulative distributions for a maximally-induced ( $5000 \mu\text{M}$  IPTG,  $R = 160$ ,  $\Delta\varepsilon_{RA} = -13.9 k_B T$ ) sample are shown as purple and orange lines in Fig. 6.9 (B), respectively. To make the different methods directly comparable, the expressions distributions were normalized to range between 0 and 1 and then centered about the mean of the distribution. While the means agree between the methods, it is immediately obvious that the width of the distributions are different with microscopy yielding distributions with a higher variance. To compare the distributions more quantitatively, we computed the central moment values for the variance, skewness, and kurtosis of the distributions (Fig. 6.9 (C)). This quantitative comparison reveals that the value of the moments can differ by close to an order of magnitude between the methods with flow cytometry systematically lower than the same distribution measured via microscopy. These results show that in terms of measuring the mean level of gene expression, the two methods can be used interchangeably. However, if one is interested in the higher moments



**Figure 6.9: Empirical comparison of flow cytometry and single-cell microscopy.** (A) The observed fold-change in gene expression for the IPTG titration of a strain with  $R = 260$  and  $\Delta\epsilon_{RA} = -13.9 k_B T$  using both microscopy (x-axis) and flow cytometry (y-axis). Points and errors represent the mean and standard error of 3 (microscopy) or 10 (flow cytometry) biological replicates. Black line indicates perfect agreement. (B) Empirical cumulative distributions of expression intensity for the strain used in (A) maximally induced with  $5000 \mu\text{M}$  IPTG. Purple and orange lines correspond to measurements with microscopy and flow cytometry, respectively. Fluorescence was normalized between 0 and 1 and centered about the observed mean. (C) Central moments of the distributions shown in (B) for microscopy and flow cytometry. Each point represents a single biological replicate. The Python code (`ch6_figS10.py`) used to generate this figure can be found on the thesis GitHub repository.

of the distribution, the choice of method does matter.

## 6.8 Fold-Change Sensitivity Analysis

In we found that the width of the credible regions varied widely depending on the repressor copy number  $R$  and repressor operator binding energy  $\Delta\epsilon_{RA}$ . More precisely, the credible regions were much narrower for low repressor copy numbers  $R$  and weak binding energy  $\Delta\epsilon_{RA}$ . In this section, we explain how this behavior comes about. We focus our attention on the maximum fold-change in the presence of saturating inducer. While it is straightforward to consider the width of the credible regions at any other inducer concentration, shows that the credible region are widest at saturation.

The width of the credible regions corresponds to how sensitive the fold-change

is to the fit values of the dissociation constants  $K_A$  and  $K_I$ . To be quantitative, we define

$$\Delta\text{fold-change}_{K_A} \equiv \text{fold-change}(K_A, K_I^{\text{fit}}) - \text{fold-change}(K_A^{\text{fit}}, K_I^{\text{fit}}), \quad (6.23)$$

the difference between the fold-change at a particular  $K_A$  value relative to the best-fit dissociation constant  $K_A^{\text{fit}} = 139 \mu\text{M}$ . For simplicity, we keep the inactive state dissociation constant fixed at its best-fit value  $K_I^{\text{fit}} = 0.53 \mu\text{M}$ . A larger difference  $\Delta\text{fold-change}_{K_A}$  implies a wider credible region. Similarly, we define the analogous quantity

$$\Delta\text{fold-change}_{K_I} = \text{fold-change}(K_A^{\text{fit}}, K_I) - \text{fold-change}(K_A^{\text{fit}}, K_I^{\text{fit}}) \quad (6.24)$$

to measure the sensitivity of the fold-change to  $K_I$  at a fixed  $K_A^{\text{fit}}$ . Fig. 6.10 shows both of these quantities in the limit  $c \rightarrow \infty$  for different repressor-DNA binding energies  $\Delta\varepsilon_{RA}$  and repressor copy numbers  $R$ .

To understand how the width of the credible region scales with  $\Delta\varepsilon_{RA}$  and  $R$ , we can Taylor expand the difference in fold-change to first order,  $\Delta\text{fold-change}_{K_A} \approx (\partial\text{fold-change}/\partial K_A)(K_A - K_A^{\text{fit}})$ , where the partial derivative has the form

$$\frac{\partial\text{fold-change}}{\partial K_A} = \frac{e^{-\beta\Delta\varepsilon_{AI}} \frac{n}{K_I} \left(\frac{K_A}{K_I}\right)^{n-1}}{\left(1 + e^{-\beta\Delta\varepsilon_{AI}} \left(\frac{K_A}{K_I}\right)^n\right)^2} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}} \left(1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}} \left(\frac{K_A}{K_I}\right)^n} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right)^{-2}. \quad (6.25)$$

Similarly, the Taylor expansion  $\Delta\text{fold-change}_{K_I} \approx (\partial\text{fold-change}/\partial K_I)(K_I - K_I^{\text{fit}})$  features the partial derivative

$$\frac{\partial\text{fold-change}}{\partial K_I} = -\frac{e^{-\beta\Delta\varepsilon_{AI}} \frac{n}{K_I} \left(\frac{K_A}{K_I}\right)^n}{\left(1 + e^{-\beta\Delta\varepsilon_{AI}} \left(\frac{K_A}{K_I}\right)^n\right)^2} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}} \left(1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}} \left(\frac{K_A}{K_I}\right)^n} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right)^{-2}. \quad (6.26)$$

From Eq. 6.25 and Eq. 6.26, we find that both  $\Delta\text{fold-change}_{K_A}$  and  $\Delta\text{fold-change}_{K_I}$  increase in magnitude with  $R$  and decrease in magnitude with  $\Delta\varepsilon_{RA}$ . Accordingly, we expect that the O3 strains (with the least negative  $\Delta\varepsilon_{RA}$ ) and the strains with the smallest repressor copy number will lead to partial derivatives with smaller

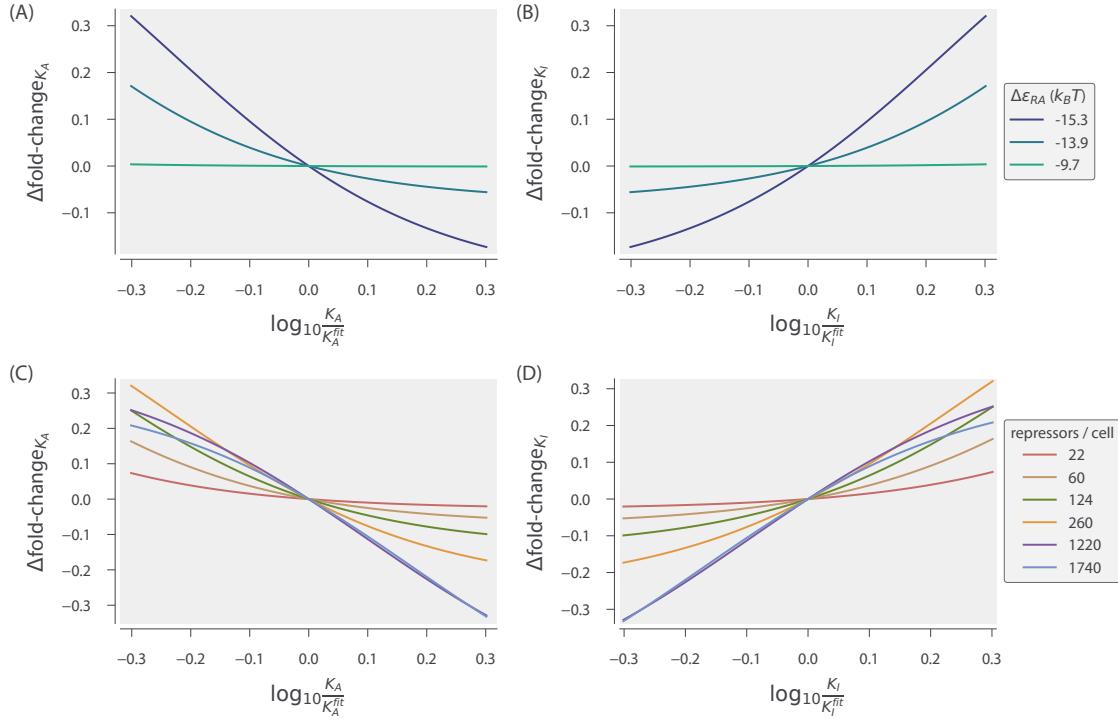
magnitude and hence to tighter credible regions. Indeed, this prediction is carried out in Fig. 6.10.

Lastly, we note that Eq. 6.25 and Eq. 6.26 enable us to quantify the scaling relationship between the width of the credible region and the two quantities  $R$  and  $\Delta\varepsilon_{RA}$ . For example, for the O3 strains, where the fold-change at saturating inducer concentration is  $\approx 1$ , the right-most term in both equations which equals the fold-change squared is roughly 1. Therefore, we find that both  $\frac{\partial \text{fold-change}}{\partial K_A}$  and  $\frac{\partial \text{fold-change}}{\partial K_I}$  scale linearly with  $R$  and  $e^{-\beta\Delta\varepsilon_{RA}}$ . Thus the width of the  $R = 22$  strain will be roughly 1/1000 as large as that of the  $R = 1740$  strain; similarly, the width of the O3 curves will be roughly 1/1000 the width of the O1 curves.

## 6.9 Global Fit of All Parameters

In Chapter 2, we used the repressor copy numbers  $R$  and repressor-DNA binding energies  $\Delta\varepsilon_{RA}$  as reported by Garcia and Phillips (2011). However, any error in these previous measurements of  $R$  and  $\Delta\varepsilon_{RA}$  will necessarily propagate into our own fold-change predictions. In this section we take an alternative approach to fitting the physical parameters of the system to that used in the main text. First, rather than fitting only a single strain, we fit the entire data set in along with microscopy data for the synthetic operator Oid. In addition, we also simultaneously fit the parameters  $R$  and  $\Delta\varepsilon_{RA}$  using the prior information given by the previous measurements. By using the entire data set and fitting all of the parameters, we obtain the best possible characterization of the statistical mechanical parameters of the system given our current state of knowledge.

To fit all of the parameters simultaneously, we follow a similar approach to the one detailed in the Materials & Methods of Chapter 2. Briefly, we perform a Bayesian parameter estimation of the dissociation constants  $K_A$  and  $K_I$ , the six different repressor copy numbers  $R$  corresponding to the six *lacI* ribosomal binding sites used in our work, and the four different binding energies  $\Delta\varepsilon_{RA}$  characterizing the four distinct operators used to make the experimental strains. As in the main text, we fit the logarithms  $\tilde{k}_A = -\log \frac{K_A}{1M}$  and  $\tilde{k}_I = -\log \frac{K_I}{1M}$  of the dissociation



**Figure 6.10: Determining how sensitive the fold-change values are to the fit values of the dissociation constants.** The difference  $\Delta\text{fold-change}_{K_A}$  in fold change when the dissociation constant  $K_A$  is slightly offset from its best-fit value  $K_A = 139^{+29}_{-22}\mu\text{M}$ , as given by . Fold-change is computed in the limit of saturating inducer concentration ( $c \rightarrow \infty$ , see ) where the credible regions in are widest. The O3 strain ( $\Delta\varepsilon_{RA} = -9.7 k_B T$ ) is about 1/1000 as sensitive as the O1 operator to perturbations in the parameter values, and hence its credible region is roughly 1/1000 as wide. All curves were made using  $R = 260$ . As in (A), but plotting the sensitivity of fold-change to the  $K_I$  parameter relative to the best-fit value  $K_I = 0.53^{+0.04}_{-0.04}\mu\text{M}$ . Note that only the magnitude, and not the sign, of this difference describes the sensitivity of each parameter. Hence, the O3 strain is again less sensitive than the O1 and O2 strains. As in (A), but showing how the fold-change sensitivity for different repressor copy numbers. The strains with lower repressor copy number are less sensitive to changes in the dissociation constants, and hence their corresponding curves in have tighter credible regions. All curves were made using  $\Delta\varepsilon_{RA} = -13.9 k_B T$ . As in (C), the sensitivity of fold-change with respect to  $K_I$  is again smallest (in magnitude) for the low repressor copy number strains. The Python code (ch6\_figS11.py) used to generate this figure can be found on the thesis GitHub repository.

constants which grants better numerical stability.

We assume that deviations of the experimental fold-change from the theoretical predictions are normally distributed with mean zero and standard deviation  $\sigma$ . We begin by writing Bayes' theorem,

$$g(\tilde{k}_A, \tilde{k}_I, \mathbf{R}, \Delta\epsilon_{RA}, \sigma | D) = \frac{f(D | \tilde{k}_A, \tilde{k}_I, \mathbf{R}, \Delta\epsilon_{RA}, \sigma) g(\tilde{k}_A, \tilde{k}_I, \mathbf{R}, \Delta\epsilon_{RA}, \sigma)}{f(D)}, \quad (6.27)$$

where  $\mathbf{R}$  is an array containing the six different repressor copy numbers to be fit,  $\Delta\epsilon_{RA}$  is an array containing the four binding energies to be fit, and  $D$  is the experimental fold-change data. The term  $P(\tilde{k}_A, \tilde{k}_I, \mathbf{R}, \Delta\epsilon_{RA}, \sigma | D)$  gives the probability distributions of all of the parameters given the data. The prefixes  $g$  and  $f$  denote probability densities of parameters and data, respectively. The term  $f(D | \tilde{k}_A, \tilde{k}_I, \mathbf{R}, \Delta\epsilon_{RA}, \sigma)$  represents the likelihood of having observed our experimental data given some value for each parameter.  $g(\tilde{k}_A, \tilde{k}_I, \mathbf{R}, \Delta\epsilon_{RA}, \sigma)$  contains all the prior information on the values of these parameters. Lastly,  $f(D)$  serves as a normalization constant and is neglected.

Given  $n$  independent measurements of the fold-change, the first term in can be written as

$$f(D | \tilde{k}_A, \tilde{k}_I, \mathbf{R}, \Delta\epsilon_{RA}, \sigma) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \prod_{i=1}^n \exp \left[ -\frac{(\text{fc}_{\text{exp}}^{(i)} - \text{fc}(\tilde{k}_A, \tilde{k}_I, R^{(i)}, \Delta\epsilon_{RA}^{(i)}, c^{(i)}))^2}{2\sigma^2} \right], \quad (6.28)$$

where  $\text{fc}_{\text{exp}}^{(i)}$  is the  $i^{\text{th}}$  experimental fold-change and  $\text{fc}(\dots)$  is the theoretical prediction. Note that the standard deviation  $\sigma$  of this distribution is not known and hence needs to be included as a parameter to be fit.

The second term in represents the prior information of the parameter values. We assume that all parameters are independent of each other, so that  $g(\tilde{k}_A, \tilde{k}_I, \mathbf{R}, \Delta\epsilon_{RA}, \sigma) = g(\tilde{k}_A) \cdot P(\tilde{k}_I) \cdot \prod_i P(R^{(i)}) \cdot \prod_j g(\Delta\epsilon_{RA}^{(j)}) \cdot g(\sigma)$ , where the superscript  $(i)$  indicates the repressor copy number of index  $i$  and the superscript  $(j)$  denotes the binding energy of index  $j$ . As above, we note that a prior must also be included for the unknown parameter  $\sigma$ .

Because we know nothing about the values of  $\tilde{k}_A$ ,  $\tilde{k}_I$ , and  $\sigma$  before performing the experiment, we assign maximally uninformative priors to each of these parameters. More specifically, we assign uniform priors to  $\tilde{k}_A$  and  $\tilde{k}_I$  and a Jeffreys prior to  $\sigma$ , indicating that  $K_A$ ,  $K_I$ , and  $\sigma$  are scale parameters (Sivia and Skilling, 2006). We do, however, have prior information for the repressor copy numbers and the repressor-DNA binding energies from Garcia and Phillips (2011). This prior knowledge is included within our model using an informative prior for these two parameters, which we assume to be Gaussian. Hence each of the  $R^{(i)}$  repressor copy numbers to be fit satisfies

$$g(R^{(i)}) = \frac{1}{\sqrt{2\pi\sigma_{R_i}^2}} \exp\left(-\frac{(R^{(i)} - \bar{R}^{(i)})^2}{2\sigma_{R_i}^2}\right), \quad (6.29)$$

where  $\bar{R}^{(i)}$  is the mean repressor copy number and  $\sigma_{R_i}$  is the variability associated with this parameter as reported in Garcia and Phillips (2011). Note that we use the given value of  $\sigma_{R_i}$  from previous measurements rather than leaving this as a free parameter.

Similarly, the binding energies  $\Delta\varepsilon_{RA}^{(j)}$  are also assumed to have a Gaussian informative prior of the same form. We write it as

$$g(\Delta\varepsilon_{RA}^{(j)}) = \frac{1}{\sqrt{2\pi\sigma_{\varepsilon_j}^2}} \exp\left(-\frac{(\Delta\varepsilon_{RA}^{(j)} - \bar{\Delta\varepsilon}_{RA}^{(j)})^2}{2\sigma_{\varepsilon_j}^2}\right), \quad (6.30)$$

where  $\bar{\Delta\varepsilon}_{RA}^{(j)}$  is the binding energy and  $\sigma_{\varepsilon_j}$  is the variability associated with that parameter around the mean value as reported in Garcia and Phillips (2011).

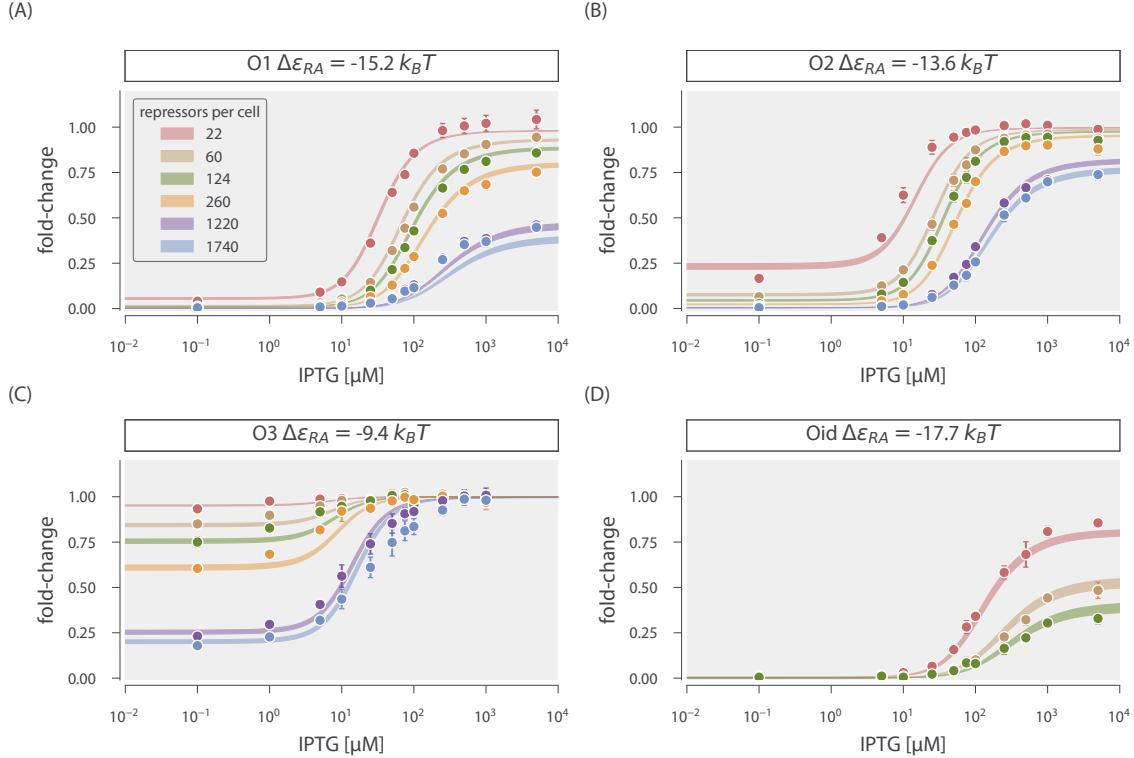
The  $\sigma_{R_i}$  and  $\sigma_{\varepsilon_j}$  parameters will constrain the range of values for  $R^{(i)}$  and  $\Delta\varepsilon_{RA}^{(j)}$  found from the fitting. For example, if for some  $i$  the standard deviation  $\sigma_{R_i}$  is very small, it implies a strong confidence in the previously reported value. Mathematically, the exponential in will ensure that the best-fit  $R^{(i)}$  lies within a few standard deviations of  $\bar{R}^{(i)}$ . Since we are interested in exploring which values could give the best fit, the errors are taken to be wide enough to allow the parameter estimation to freely explore parameter space in the vicinity of the best estimates. Putting

all these terms together, we use Markov chain Monte Carlo to sample the posterior distribution  $P(\tilde{k}_A, \tilde{k}_I, \mathbf{R}, \Delta\epsilon_{RA}, \sigma | D)$ , enabling us to determine both the most likely value for each physical parameter as well as its associated credible region.

Fig. 6.11 shows the result of this global fit. When compared with we can see that fitting for the binding energies and the repressor copy numbers improves the agreement between the theory and the data. Table 6.2 summarizes the values of the parameters as obtained with this MCMC parameter inference. We note that even though we allowed the repressor copy numbers and repressor-DNA binding energies to vary, the resulting fit values were very close to the previously reported values. The fit values of the repressor copy numbers were all within one standard deviation of the previous reported values provided in Garcia and Phillips (2011). And although some of the repressor-DNA binding energies differed by a few standard deviations from the reported values, the differences were always less than  $1 k_B T$ , which represents a small change in the biological scales we are considering. The biggest discrepancy between our fit values and the previous measurements arose for the synthetic Oid operator, which we discuss in more detail in the following section of this chapter.

Table 6.2: Global parameter estimates and comparison to previously reported values.

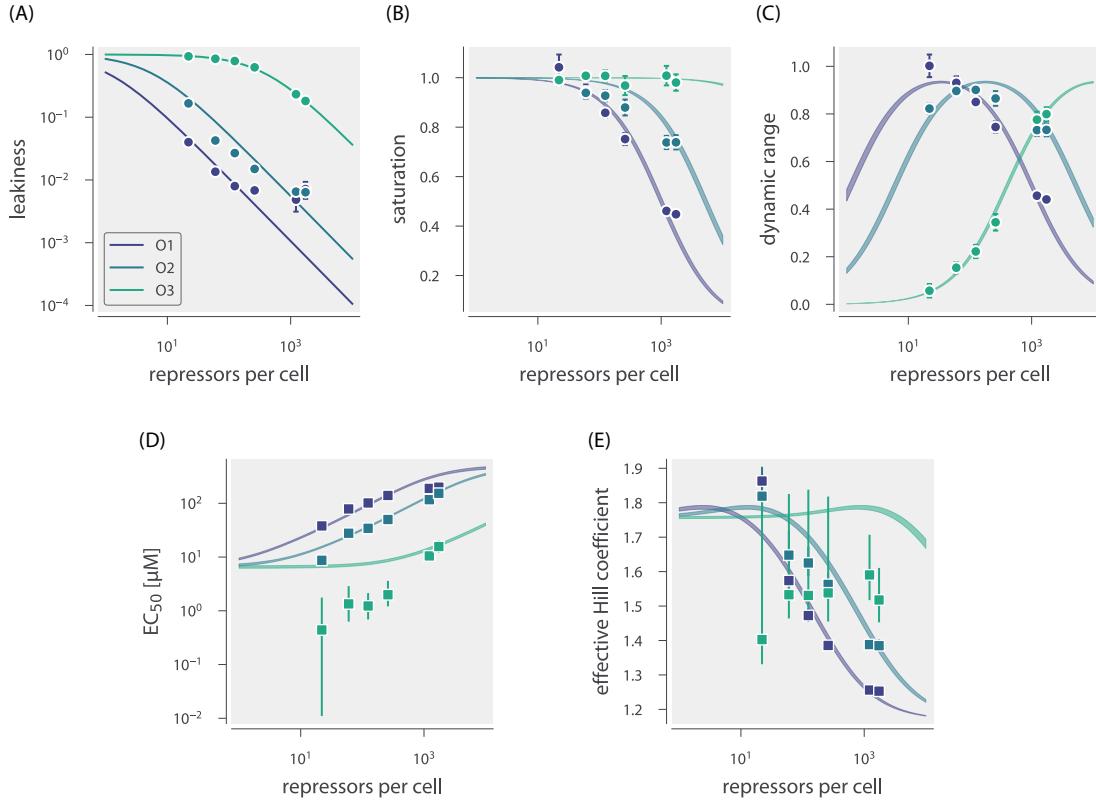
Parameter	Reported Values (Garcia and Phillips, 2011)	Global Fit
$K_A$	-	$205^{+11}_{-12} \mu\text{M}$
$K_I$	-	$0.73^{+0.04}_{-0.04} \mu\text{M}$
$R_{22}$	$22 \pm 4$ per cell	$20^{+1}_{-1}$ per cell
$R_{60}$	$60 \pm 20$ per cell	$74^{+4}_{-3}$ per cell
$R_{124}$	$124 \pm 30$ per cell	$130^{+6}_{-6}$ per cell
$R_{260}$	$260 \pm 40$ per cell	$257^{+9}_{-11}$ per cell
$R_{1220}$	$1220 \pm 160$ per cell	$1191^{+32}_{-55}$ per cell
$R_{1740}$	$1740 \pm 340$ per cell	$1599^{+75}_{-87}$ per cell
O1 $\Delta\epsilon_{RA}$	$-15.3 \pm 0.2 k_B T$	$-15.22^{+0.1}_{-0.1} k_B T$



**Figure 6.11: Global fit of dissociation constants, repressor copy numbers, and binding energies.** Theoretical prediction resulting from simultaneous estimation of the dissociation constants  $K_A$  and  $K_I$ , the six repressor copy numbers  $R$ , and the four repressor-DNA binding energies  $\Delta\epsilon_{RA}$  using the entire dataset. Points and errors represent the mean and standard error of  $\sim 10$  biological replicates for O1, O2, and O3 and 3 biological replicates for Oid. The Python code (ch6\_fig12.py) used to generate this figure can be found on the thesis GitHub repository.

Parameter	Reported Values (Garcia and Phillips, 2011)	Global Fit
O2 $\Delta\epsilon_{RA}$	$-13.9 \pm 0.2 k_B T$	$-13.06^{+0.1}_{-0.1} k_B T$
O3 $\Delta\epsilon_{RA}$	$-9.7 \pm 0.1 k_B T$	$-9.4^{+0.1}_{-0.1} k_B T$
Oid $\Delta\epsilon_{RA}$	$-17.0 \pm 0.2 k_B T$	$-17.7^{+0.2}_{-0.1} k_B T$

Fig. 6.12 shows the same key properties as in Fig. 2.7 , but uses the parameters obtained from this global fitting approach. We note that even by increasing the number of degrees of freedom in our fit, the result does not change substantially,



**Figure 6.12: Key properties of induction profiles as predicted with a global fit using all data.** Data for (A) leakiness, (B) saturation, and (C) dynamic range are computed directly from measured fold-change. Points and errors correspond to the mean and standard error of 10 - 11 biological replicates. Points in (D) and (E) for the  $[EC_{50}]$  and the effective Hill coefficient, respectively, represent the estimated value using parameter estimates of  $K_A$  and  $K_I$  for that particular strain. Errors represent the width of the 95% credible region. In all plots, curves represent the theoretical predictions given the parameter estimates conditioned on all data sets. The Python code (ch6\_figS13.py) used to generate this figure can be found on the thesis GitHub repository.

due to in general, only minor improvements between the theoretical curves and data. For the O3 operator data, again, agreement between the predicted  $[EC_{50}]$  and the effective Hill coefficient remain poor due the theory being unable to capture the steepness of the response curves.

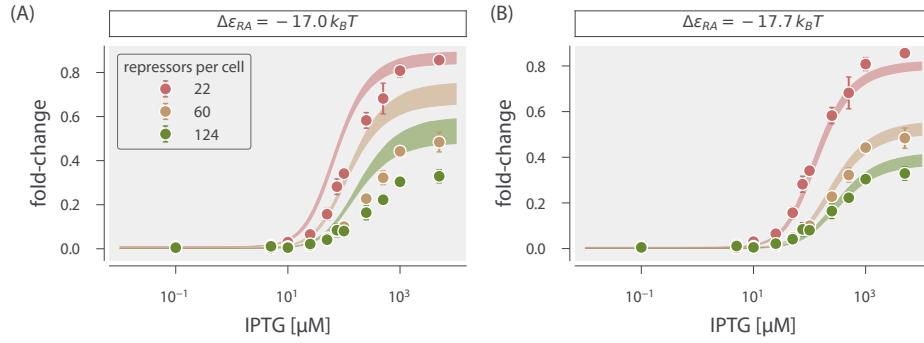
## 6.10 Applicability of Theory to the Oid Operator Sequence

In addition to the native operator sequences (O1, O2, and O3) considered in the main text, we were also interested in testing our model predictions against the synthetic Oid operator. In contrast to the other operators, Oid is one base pair shorter in length (20 bp), is fully symmetric, and is known to provide stronger repression than the native operator sequences considered so far. While the theory should be similarly applicable, measuring the lower fold-changes associated with this YFP construct was expected to be near the sensitivity limit for our flow cytometer, due to the especially strong binding energy of Oid ( $\Delta\epsilon_{RA} = -17.0 k_B T$ ) (Garcia and Phillips, 2011). Accordingly, fluorescence data for Oid were obtained using microscopy, which is more sensitive than flow cytometry.

We follow the approach of the main text and make fold-change predictions based on the parameter estimates from our strain with  $R = 260$  and an O2 operator. These predictions are shown in Fig. 6.13, where we also plot data taken in triplicate for strains containing  $R = 22, 60$ , and  $124$ , obtained by single-cell microscopy. We find that the data are systematically below the theoretical predictions. We also considered our global fitting approach (see previous section) to see whether we might find better agreement with the observed data. Interestingly, we find that the majority of the parameters remain largely unchanged, but our estimate for the Oid binding energy  $\Delta\epsilon_{RA}$  is shifted to  $-17.7 k_B T$  instead of the value  $-17.0 k_B T$  found in Garcia and Phillips (2011). In Fig. 6.13 we again plot the Oid fold-change data but with theoretical predictions using the new estimate for the Oid binding energy from our global fit and find substantially better agreement.

## 6.11 Comparison of Parameter Estimation and Fold-Change Predictions across Strains

The inferred parameter values for  $K_A$  and  $K_I$  in Chapter 2 were determined by fitting to induction fold-change measurements from a single strain ( $R = 260$ ,  $\Delta\epsilon_{RA} = -13.9 k_B T$ ,  $n = 2$ , and  $\Delta\epsilon_{AI} = 4.5 k_B T$ ). After determining these parameters, we were able to predict the fold-change of the remaining strains without



**Figure 6.13: Predictions of fold-change for strains with an Oid binding sequence versus experimental measurements with different repressor copy numbers.** Experimental data is plotted against the parameter-free predictions that are based on our fit to the O2 strain with  $R = 260$ . Here we use the previously measured binding energy  $\Delta\epsilon_{RA} = -17.0 \text{ } k_B T$  (Garcia and Phillips, 2011). The same experimental data is plotted against the best-fit parameters using the complete O1, O2, O3, and Oid data sets to infer  $K_A$ ,  $K_I$ , repressor copy numbers, and the binding energies of all operators. Here the major difference in the inferred parameters is a shift in the binding energy for Oid from  $\Delta\epsilon_{RA} = -17.0 \text{ } k_B T$  to  $\Delta\epsilon_{RA} = -17.7 \text{ } k_B T$ , which now shows agreement between the theoretical predictions and experimental data. Shaded regions from the theoretical curves denote the 95% credible region. The Python code (`ch6_figS14.py`) used to generate this figure can be found on the thesis GitHub repository.

any additional fitting. However, the theory should be independent of the specific strain used to estimate  $K_A$  and  $K_I$ ; using any alternative strain to fit  $K_A$  and  $K_I$  should yield similar predictions. For the sake of completeness, here we discuss the values for  $K_A$  and  $K_I$  that are obtained by fitting to each of the induction data sets individually. These fit parameters are shown in Fig. 2.6 (D) of Chapter 2, where we find close agreement between strains, but with some deviation and poorer inferences observed with the O3 operator strains. Overall, we find that regardless of which strain is chosen to determine the unknown parameters, the predictions laid out by the theory closely match the experimental measurements. Here we present a comparison of the strain specific predictions and measured fold-change data for each of the three operators considered.

We follow the approach taken in Chapter 2 and infer values for  $K_A$  and  $K_I$  by

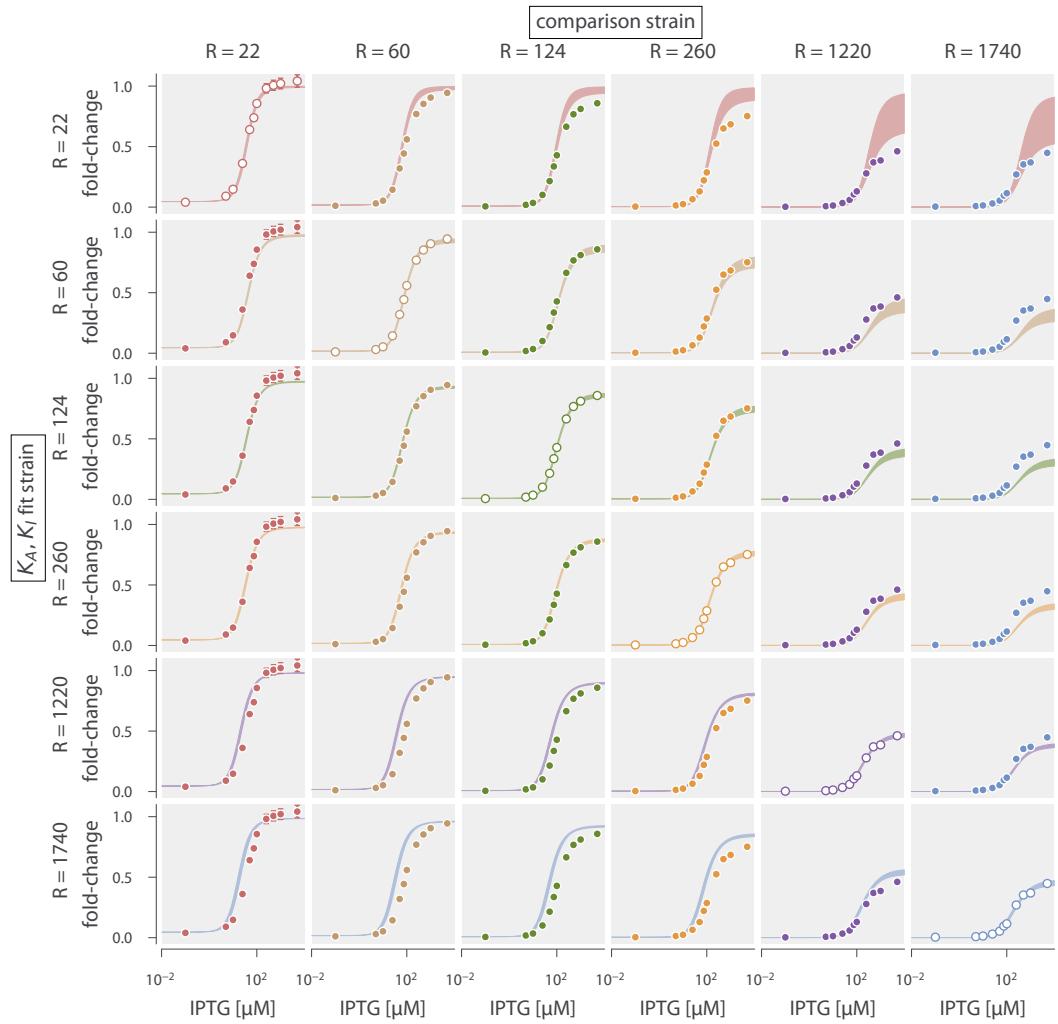
fitting to each combination of binding energy  $\Delta\varepsilon_{RA}$  and repressor copy number  $R$ . We then use these fitted parameters to predict the induction curves of all other strains. In Fig. 6.14, we plot these fold-change predictions along with experimental data for each of our strains that contains an O1 operator. To make sense of this plot we consider the first row as an example. In the first row,  $K_A$  and  $K_I$  were estimated using data from the strain containing  $R = 22$  and an O1 operator (top leftmost plot). The remaining plots in this row show the predicted fold-change using these values for  $K_A$  and  $K_I$ . In each row, we then infer  $K_A$  and  $K_I$  using data from a strain containing a different repressor copy number ( $R = 60$  in the second row,  $R = 124$  in the third row, and so on). In Fig. 6.15 and Fig. 6.16, we similarly apply this inference to our strains with O2 and O3 operators, respectively. We note that the overwhelming majority of predictions closely match the experimental data. The notable exception is that using the  $R = 22$  strain provides poor predictions for the strains with large copy numbers (especially  $R = 1220$  and  $R = 1740$ ), though it should be noted that predictions made from the  $R = 22$  strain have considerably broader credible regions. This loss in predictive power is due to the poorer estimates of  $K_A$  and  $K_I$  for the  $R = 22$  strain as shown in Fig. 2.6 (D).

## 6.12 Properties of Induction Titration Curves

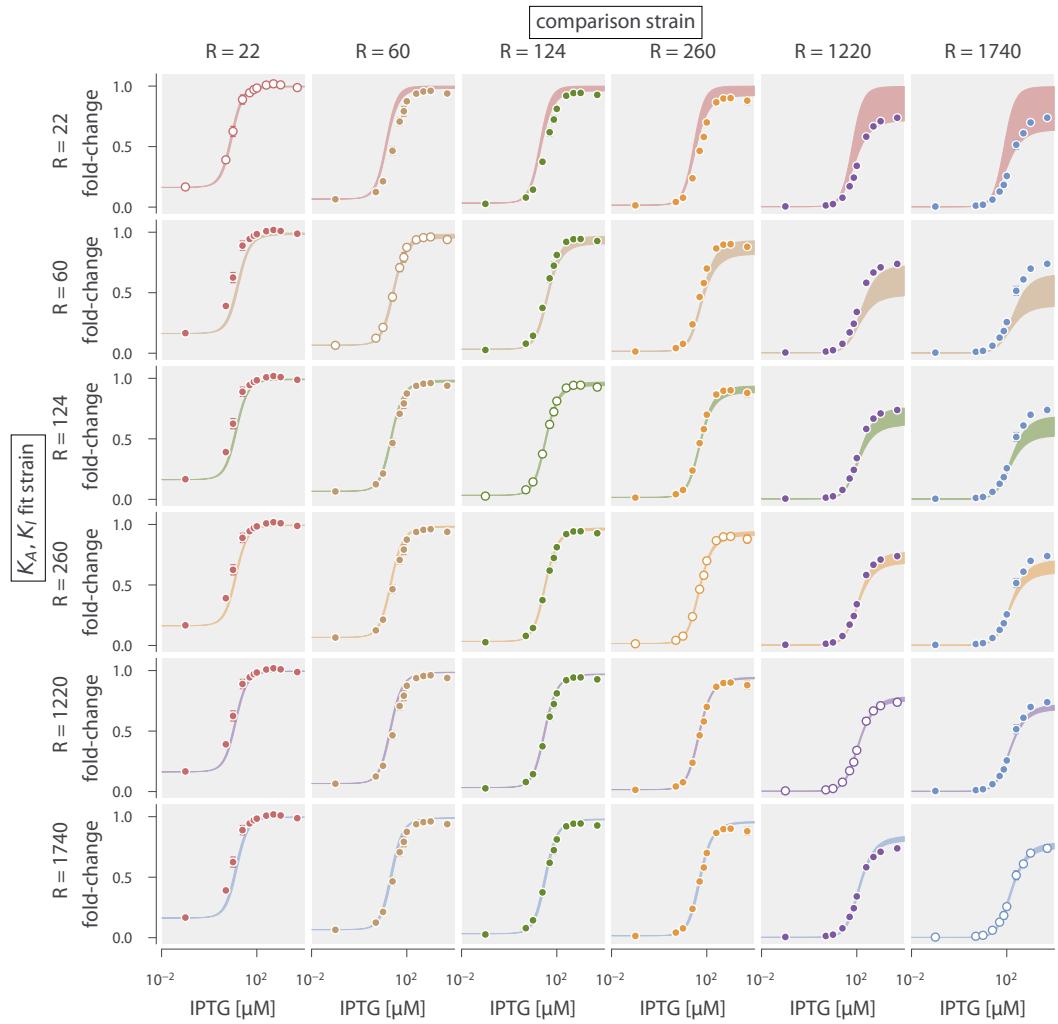
In this section, we expand on the phenotypic properties of the induction response that were explored in Chapter 2. We begin by expanding on our discussion of dynamic range and then show the analytic form of the  $[EC_{50}]$  for simple repression.

As stated in the main text, the dynamic range is defined as the difference between the maximum and minimum system response, or equivalently, as the difference between the saturation and leakiness of the system. The dynamic range is therefore given by

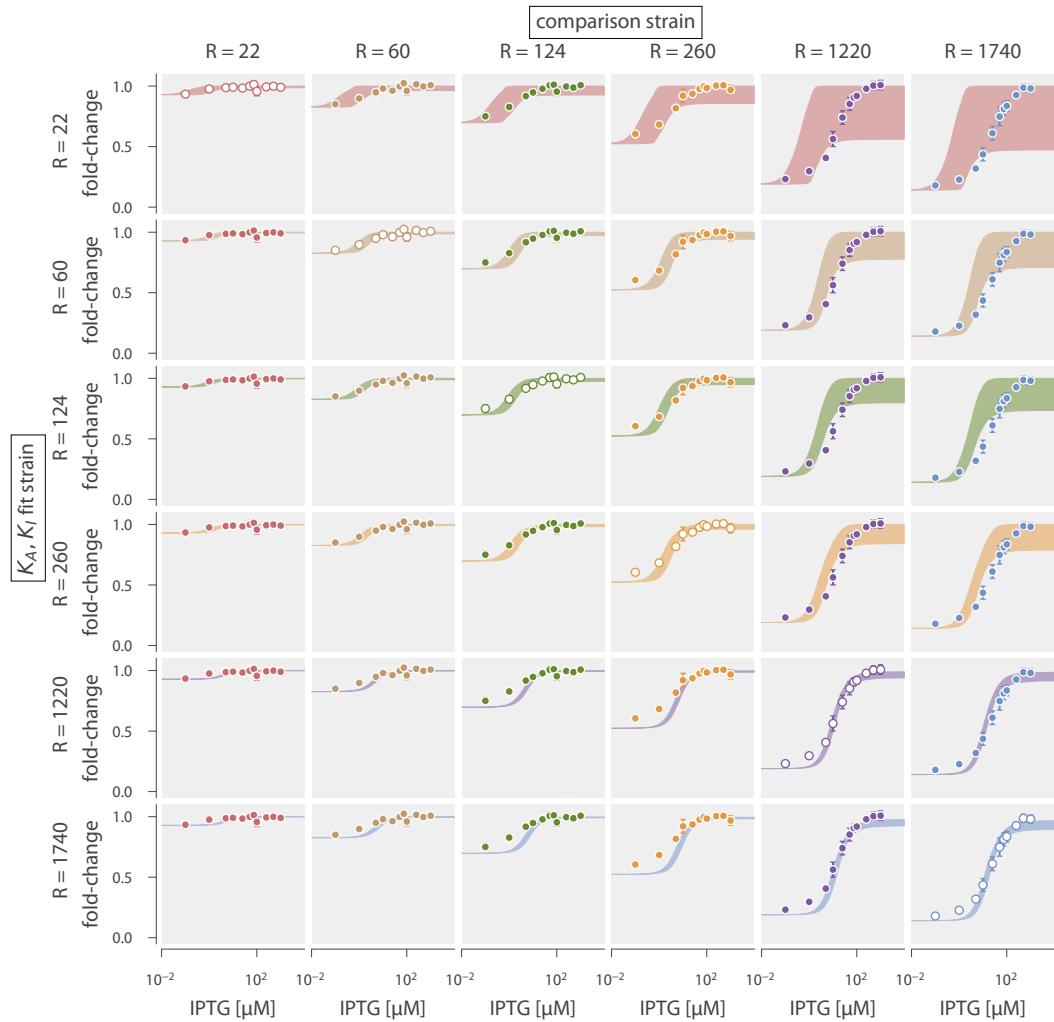
$$\text{dynamic range} = \left( 1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}} \left( \frac{K_A}{K_I} \right)^n N_{NS}} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}} \right)^{-1} - \left( 1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}} \right)^{-1}. \quad (6.31)$$



**Figure 6.14: O1 strain fold-change predictions based on strain-specific parameter estimation of  $K_A$  and  $K_I$ .** Fold-change in expression is plotted as a function of IPTG concentration for all strains containing an O1 operator. The Python code (`ch6_figS15-S17.py`) used to generate this figure can be found on the thesis GitHub repository.



**Figure 6.15: O2 strain fold-change predictions based on strain-specific parameter estimation of  $K_A$  and  $K_I$ .** Fold-change in expression is plotted as a function of IPTG concentration for all strains containing an O2 operator. The Python code (`ch6_figS15-S17.py`) used to generate this figure can be found on the thesis GitHub repository.



**Figure 6.16: O3 strain fold-change predictions based on strain-specific parameter estimation of  $K_A$  and  $K_I$ .** Fold-change in expression is plotted as a function of IPTG concentration for all strains containing an O3 operator. The Python code (`ch6_figS15-S17.py`) used to generate this figure can be found on the thesis GitHub repository.

The dynamic range, along with saturation and leakiness were plotted with our experimental data in Fig. 2.7(A-C) as a function of repressor copy number. Fig. 6.17 shows how these properties are expected to vary as a function of the repressor-operator binding energy. Note that the resulting curves for all three properties have the same shape as in Fig. 2.7 (A-C), since the dependence of the fold-change upon the repressor copy number and repressor-operator binding energy are both contained in a single multiplicative term,  $R e^{-\beta \Delta \varepsilon_{RA}}$ . Hence, increasing  $R$  on a logarithmic scale is equivalent to decreasing  $\Delta \varepsilon_{RA}$  on a linear scale.

An interesting aspect of the dynamic range is that it exhibits a peak as a function of either the repressor copy number (or equivalently of the repressor-operator binding energy). Differentiating the dynamic range (Eq. 6.31) and setting it equal to zero, we find that this peak occurs at

$$\frac{R^*}{N_{NS}} = e^{-\beta(\Delta \varepsilon_{AI} - \Delta \varepsilon_{RA})} \sqrt{e^{\Delta \varepsilon_{AI}} + 1} \sqrt{e^{\Delta \varepsilon_{AI}} + \left(\frac{K_A}{K_I}\right)^n}. \quad (6.32)$$

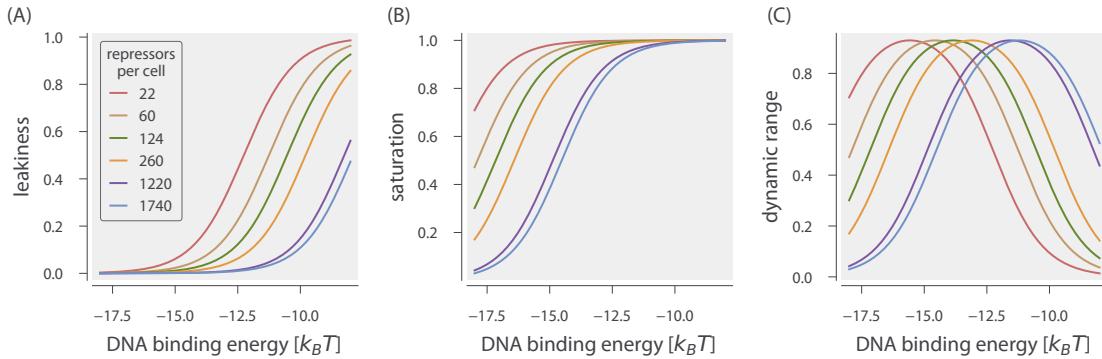
The magnitude of the peak is given by

$$\text{max dynamic range} = \frac{\left( \sqrt{e^{\Delta \varepsilon_{AI}} + 1} - \sqrt{e^{\Delta \varepsilon_{AI}} + \left(\frac{K_A}{K_I}\right)^n} \right)^2}{\left(\frac{K_A}{K_I}\right)^n - 1}, \quad (6.33)$$

which is independent of the repressor-operator binding energy  $\Delta \varepsilon_{RA}$  or  $R$ , and will only cause a shift in the location of the peak but not its magnitude.

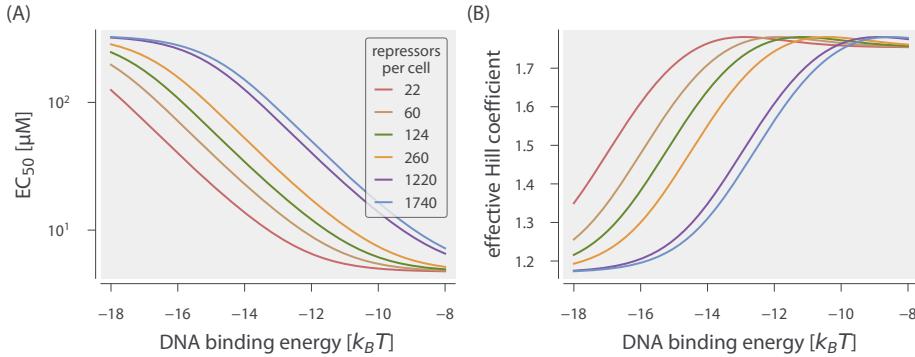
We now consider the two remaining properties, the  $[EC_{50}]$  and effective Hill coefficient, which determine the horizontal properties of a system - that is, they determine the range of inducer concentration in which the system's response goes from its minimum to maximum values. The  $[EC_{50}]$  denotes the inducer concentration required to generate fold-change halfway between its minimum and maximum value and was defined implicitly in Eq. 2.9. For the simple repression system, the  $[EC_{50}]$  is given by

$$\frac{[EC_{50}]}{K_A} = \frac{\frac{K_A}{K_I} - 1}{\frac{K_A}{K_I} - \left( \frac{\left(1 + \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_{RA}}\right) + \left(\frac{K_A}{K_I}\right)^n \left(2e^{-\beta \Delta \varepsilon_{AI}} + \left(1 + \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_{RA}}\right)\right)}{2\left(1 + \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_{RA}}\right) + e^{-\beta \Delta \varepsilon_{AI}} + \left(\frac{K_A}{K_I}\right)^n e^{-\beta \Delta \varepsilon_{AI}}} \right)^{\frac{1}{n}}} - 1. \quad (6.34)$$



**Figure 6.17: Dependence of leakiness, saturation, and dynamic range on the operator binding energy and repressor copy number.** Increasing repressor copy number or decreasing the repressor-operator binding energy suppresses gene expression and decreases both the (A) leakiness and (B) saturation. (C) The dynamic range retains its shape, but shifts right as the repressor copy number increases. The peak in the dynamic range can be understood by considering the two extremes for  $\Delta\epsilon_{RA}$ : for small repressor-operator binding energies, the leakiness is small but the saturation increases with  $\Delta\epsilon_{RA}$ , thereby decreasing the dynamic range. Repressor copy number does not affect the maximum dynamic range. The Python code (`ch6_figS18-S19.py`) used to generate this figure can be found on the thesis GitHub repository.

Using this expression, we can then find the effective Hill coefficient  $h$ , which equals twice the log-log slope of the normalized fold-change evaluated at  $c = [EC_{50}]$ . In Fig. 2.7 (D-E) we show how these two properties vary with repressor copy number, and in Fig. 6.18, we demonstrate how they depend on the repressor-operator binding energy. Both the  $[EC_{50}]$  and  $h$  vary significantly with repressor copy number for sufficiently strong operator binding energies. Interestingly, for weak operator binding energies on the order of the O3 operator, it is predicted that the effective Hill coefficient should not vary with repressor copy number. In addition, the maximum possible Hill coefficient is roughly 1.75, which stresses the point that the effective Hill coefficient should not be interpreted as the number of inducer binding sites, which is exactly 2.



**Figure 6.18: [EC<sub>50</sub>] and effective Hill coefficient depend strongly on repressor copy number and operator binding energy.** (A) [EC<sub>50</sub>] values from very small and tightly clustered to relatively large and expanded for stronger operator binding energies. (B) The effective Hill coefficient generally decreases with increasing repressor copy number, indicating a flatter normalized response. The maximum possible Hill coefficient is roughly 1.75 for all repressor-operator binding energies. The Python code (ch6\_figS18-S19.py) used to generate this figure can be found on the thesis GitHub repository.

### 6.13 Applications to Other Regulatory Architectures

In this section, we discuss how the theoretical framework presented in this work is sufficiently general to include a variety of regulatory architectures outside of simple repression by LacI. We begin by noting that the exact same formula for fold-change given in can also describe corepression. We then demonstrate how our model can be generalized to include other architectures, such as a coactivator binding to an activator to promote gene expression. In each case, we briefly describe the system and describe its corresponding theoretical description. For further details, we invite the interested reader to read Bintu et al. (2005a) and Marzen et al. (2013).

#### Corepression

Consider a regulatory architecture where binding of a transcriptional repressor occludes the binding of RNAP to the DNA. A corepressor molecule binds to the repressor and shifts its allosteric equilibrium towards the active state in which it binds more tightly to the DNA, thereby decreasing gene expression (in contrast,

an inducer shifts the allosteric equilibrium towards the inactive state where the repressor binds more weakly to the DNA). As in the main text, we can enumerate the states and statistical weights of the promoter and the allosteric states of the repressor. We note that these states and weights exactly match those in Fig. 2.2 and yield the same fold-change equation,

$$\text{fold-change} \approx \left( 1 + \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{\beta\Delta\varepsilon_{AI}} \left(1 + \frac{c}{K_I}\right)^n} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}} \right)^{-1}, \quad (6.35)$$

where  $c$  now represents the concentration of the corepressor molecule. Mathematically, the difference between these two architectures can be seen in the relative sizes of the dissociation constants  $K_A$  and  $K_I$  between the inducer and repressor in the active and inactive states, respectively. The corepressor is defined by  $K_A < K_I$ , since the corepressor favors binding to the repressor's active state; an inducer must satisfy  $K_I < K_A$ , as was found in Chapter 2. Much as was performed in Chapter 2, we can make some predictions about the how the response of a corepressor. In Fig. ?? (A), we show how varying the repressor copy number  $R$  and the repressor-DNA binding energy  $\Delta\varepsilon_{RA}$  influence the response. We draw the reader's attention to the decrease in fold-change as the concentration of effector is increased.

## Activation

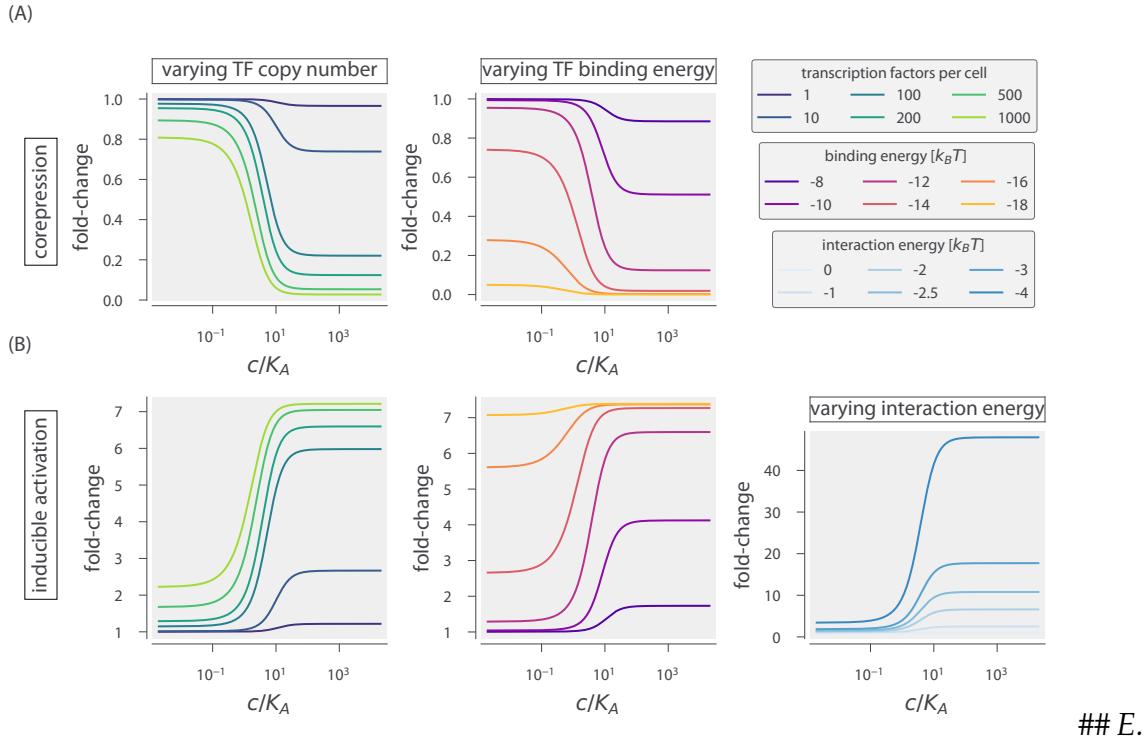
We now turn to the case of activation. While this architecture was not studied in this work, we wish to demonstrate how the framework presented here can be extended to include transcription factors other than repressors. To that end, we consider a transcriptional activator which binds to DNA and aids in the binding of RNAP through energetic interaction term  $\varepsilon_{AP}$ . Note that in this architecture, binding of the activator does not occlude binding of the polymerase. Binding of a coactivator molecule shifts its allosteric equilibrium towards the active state ( $K_A < K_I$ ), where the activator is more likely to be bound to the DNA and promote expression. Enumerating all of the states and statistical weights of this architecture and making the approximation that the promoter is weak generates a fold-change equation

of the form

$$\text{fold-change} = \frac{1 + \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{\beta\Delta\varepsilon_{AI}} \left(1 + \frac{c}{K_I}\right)^n} \frac{A}{N_{NS}} e^{-\beta\Delta\varepsilon_{AA}} e^{-\beta\varepsilon_{AP}}}{1 + \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{\beta\Delta\varepsilon_{AI}} \left(1 + \frac{c}{K_I}\right)^n} \frac{A}{N_{NS}} e^{-\beta\Delta\varepsilon_{AA}}}, \quad (6.36)$$

where  $A$  is the total number of activators per cell,  $c$  is the concentration of a coactivator molecule,  $\Delta\varepsilon_{AA}$  is the binding energy of the activator to the DNA in the active allosteric state, and  $\varepsilon_{AP}$  is the interaction energy between the activator and the RNAP. Unlike in the cases of induction and corepression, the fold-change formula for activation includes terms from when the RNAP is bound by itself on the DNA as well as when both RNAP and the activator are simultaneously bound to the DNA. explores predictions of the fold-change in gene expression by manipulating the activator copy number, DNA binding energy, and the polymerase-activator interaction energy. Note that with this activation scheme, the fold-change must necessarily be greater than one. An interesting feature of these predictions is the observation that even small changes in the interaction energy ( $< 0.5 k_B T$ ) can result in dramatic increase in fold-change.

As in the case of induction, the is straightforward to generalize. For example, the relative values of  $K_I$  and  $K_A$  can be switched such that  $K_I < K_A$  in which the secondary molecule drives the activator to assume the inactive state represents induction of an activator. While these cases might be viewed as separate biological phenomena, mathematically they can all be described by the same underlying formalism.



## E.

*coli* Primer and Strain List Here we provide additional details about the genotypes of the strains used, as well as the primer sequences used to generate them. *E. coli* strains were derived from K12 MG1655. For those containing  $R = 22$ , we used strain HG104 which additionally has the *lacYZA* operon deleted (positions 360,483 to 365,579) but still contains the native *lacI* locus. All other strains used strain HG105, where both the *lacYZA* and *lacI* operons have both been deleted (positions 360,483 to 366,637).

All 25x+11-yfp expression constructs were integrated at the *galK* locus (between positions 1,504,078 and 1,505,112) while the 3\*1x-lacI constructs were integrated at the *ybcN* locus (between positions 1,287,628 and 1,288,047). Integration was performed with  $\lambda$  Red recombineering (Sharan et al., 2009) as described in Garcia and Phillips (2011) using the primers listed in Table 6.3. We follow the notation of Lutz and Bujard (Lutz and Bujard, 1997) for the nomenclature of the different constructs used. Specifically, the first number refers to the antibiotic resistance cassette that is present for selection (2 = kanamycin, 3 = chloramphenicol, and 4 = spectinomycin) and the second number refers to the promoter used to drive expression of either

YFP or LacI ( $1 = P_{LtetO-1}$ , and  $5 = lacUV5$ ). Note that in  $25x+11\text{-yfp}$ ,  $x$  refers to the LacI operator used, which is centered at +11 (or alternatively, begins at the transcription start site). For the different LacI constructs,  $3^*\text{x-lacI}$ ,  $x$  refers to the different ribosomal binding site modifications that provide different repressor copy numbers and follows from Garcia and Phillips (2011). The asterisk refers to the presence of FLP recombinase sites flanking the chloramphenicol resistance gene that can be used to lose this resistance. However, we maintained the resistance gene in our constructs. A summary of the final genotypes of each strain is listed in Table 6.4. In addition each strain also contained the plasmid pZS4\*1-mCherry and provided constitutive expression of the mCherry fluorescent protein. This pZS plasmid is a low copy (SC101 origin of replication) where like with  $3^*\text{x-lacI}$ , mCherry is driven by a  $P_{LtetO-1}$  promoter.

Table 6.3: Primers used in this work.

Primer	Sequence (5' → 3')	Notes
pZSForwSeq2	TTCCCAACC	Forward sequencing
	TTACCAGAGG GC	primer for $3^*\text{x-lacI}$
251 F	CCTTTCGTCT	Forward sequencing
	TCACCTCGA	primer for $25x+11\text{-YFP}$
YFP1	ACTAGCAACAC	Reverse sequencing
	CAGAACAGCCC	primer for $3^*\text{x-lacI}$ and $25x+11\text{-YFP}$
HG 6.1 ( <i>galK</i> )	gtttgcgcgc agtcagcgat	Reverse primer for
	atccattttc gcgaatccg	$25x+11\text{-YFP}$ integration
	gagtgtaa	in to the <i>galK</i> locus
	aaACTAGCAAC	(lowercase).
	ACCAGAACAA GCC	

Primer	Sequence (5' → 3')	Notes
HG 6.3 ( <i>galK</i> )	ttcatattgt tcagcgacag cttgctgtac ggcaggcac cagctcttc cgGGCTAATGC ACCCAGTAA GG	Forward integration primer for 25x+11-YFP with homology to the <i>galK</i> locus (lowercase).
HG11.1 ( <i>ybcN</i> )	acctctgcgg aggggaagcg tgaacctctc acaagacgg catcaaatt acACTAGCAAC ACCAGAACAA GCC	Reverse integration primer for 3*1x-lacI with homology to the <i>ybcN</i> locus (lowercase).
HG11.3 ( <i>ybcN</i> )	ctgttagatgtg tccgttcatg acacgaataa gcgggtgtag ccattacgc cGGCTAATGCA CCCAGTAAG G	Forward integration primer for 3*1x-lacI with homology to the <i>ybcN</i> locus (lowercase).
ybcN-control-upstream-1	AGCGTTTGA CCTCTGCGGA	Sequencing primer to verify integration
ybcN-control-downstream-1	GCTCAGGTT TACGCTTAC GACG	Sequencing primer to verify integration

Table 6.4: *E. coli* strains used in this work.

Strain	Genotype
O1, R = 0	HG105:: <i>galK</i> <>25O1+11-YFP
O1, R = 22	HG104:: <i>galK</i> <>25O1+11-YFP
O1, R = 60	HG105:: <i>galK</i> <>25O1+11-YFP, <i>ybcN</i> <>3*1RBS1147-lacI
O1, R = 124	HG105:: <i>galK</i> <>25O1+11-YFP, <i>ybcN</i> <>3*1RBS446-lacI
O1, R = 260	HG105:: <i>galK</i> <>25O1+11-YFP, <i>ybcN</i> <>3*1RBS1027-lacI
O1, R = 1220	HG105:: <i>galK</i> <>25O1+11-YFP, <i>ybcN</i> <>3*1RBS1-lacI

---

<b>Strain</b>	<b>Genotype</b>
O1, R = 1740	HG105::galK<>25O1+11-YFP, <i>ybcN</i> <>3*1RBS1L-lacI
O2, R = 0	HG105::galK<>25O2+11-YFP
O2, R = 22	HG104::galK<>25O2+11-YFP
O2, R = 60	HG105::galK<>25O2+11-YFP, <i>ybcN</i> <>3*1RBS1147-lacI
O2, R = 124	HG105::galK<>25O2+11-YFP, <i>ybcN</i> <>3*1RBS446-lacI
O2, R = 260	HG105::galK<>25O2+11-YFP, <i>ybcN</i> <>3*1RBS1027-lacI
O2, R = 1220	HG105::galK<>25O2+11-YFP, <i>ybcN</i> <>3*1RBS1-lacI
O2, R = 1740	HG105::galK<>25O2+11-YFP, <i>ybcN</i> <>3*1RBS1L-lacI
O3, R = 0	HG105::galK<>25O3+11-YFP
O3, R = 22	HG104::galK<>25O3+11-YFP
O3, R = 60	HG105::galK<>25O3+11-YFP, <i>ybcN</i> <>3*1RBS1147-lacI
O3, R = 124	HG105::galK<>25O3+11-YFP, <i>ybcN</i> <>3*1RBS446-lacI
O3, R = 260	HG105::galK<>25O3+11-YFP, <i>ybcN</i> <>3*1RBS1027-lacI
O3, R = 1220	HG105::galK<>25O3+11-YFP, <i>ybcN</i> <>3*1RBS1-lacI
O3, R = 1740	HG105::galK<>25O3+11-YFP, <i>ybcN</i> <>3*1RBS1L-lacI
Oid, R = 22	HG104::galK<>25Oid+11-YFP
Oid, R = 60	HG105::galK<>25Oid+11-YFP, <i>ybcN</i> <>3*1RBS1147-lacI
Oid, R = 124	HG105::galK<>25Oid+11-YFP, <i>ybcN</i> <>3*1RBS446-lacI

---

*Chapter 7*

**SUPPLEMENTAL INFORMATION FOR CHAPTER III:  
PREDICTIVE SHIFTS IN FREE ENERGY COUPLE MUTATIONS  
TO THEIR PHENOTYPIC CONSEQUENCES**

A version of this chapter originally appeared as Chure, G; Razo-Mejia, M., Bel-liveau, N.M.; Kaczmarek, Zofii A.; Einav, T.; Barnes, Stephanie L.; Lewis, M., and Phillips, R. (2019). Predictive Shifts in Free Energy Couple Mutations to Their Phenotypic Consequences. PNAS 116(37), G.C., M.R.M, N.M.B., Z.A.K., and S.L.B designed the experiments and collected and analyzed data. G.C. developed theoretical treatment of free energy shifts. G.C., M.R.M, N.M.B., Z.A.K., T.E., S.L.B., and R.P. designed the research project. G.C. and R.P. wrote the paper. M.L. provided guidance and advice.

### **7.1 Non-Monotonic Behavior of $\Delta F$ Under Changing $K_A$ and $K_I$**

In Chapter 3, we illustrated that perturbations only to the allosteric parameters  $K_A$  and  $K_I$  relative to the wild-type values can result in a non-monotonic dependence of  $\Delta F$  on the inducer concentration  $c$ . In this section, we prove that when the ratio of  $K_A$  to  $K_I$  is the same between the mutant and wild-type proteins, the function must be monotonic. This section is paired with an interactive figure available on the paper website which illustrates how scaling  $K_A$  and  $K_I$  relative to the wild-type value results in non-monotonic behavior.

We define a monotonic function as a continuous function whose derivative does not change sign across the domain upon which it is defined. To show that  $\Delta F$  is non-monotonic when  $K_A$  and  $K_I$  are perturbed, we can compute the derivative of  $\Delta F$  with respect to the inducer concentration  $c$  and evaluate the sign of the derivative at the limits of inducer concentration. If the sign of the derivative is different at the limits of  $c = 0$  and  $c \gg 0$ , we can see that the function is non-monotonic. However, if the sign is the same in both limits, we can not say conclusively if it is

non-monotonic and must consider other diagnostics.

The free energy difference between a mutant and wild-type repressor when all parameters other than  $K_A$  and  $K_I$  are unperturbed can be written as

$$\beta\Delta F(c) = -\log \left( \frac{\left[ 1 + e^{-\beta\Delta\varepsilon_{AI}} \left( \frac{1 + \frac{c}{K_I^{(\text{mut})}}}{1 + \frac{c}{K_A^{(\text{mut})}}} \right)^2 \right]^{-1}}{\left[ 1 + e^{-\beta\Delta\varepsilon_{AI}} \left( \frac{1 + \frac{c}{K_I^{(\text{wt})}}}{1 + \frac{c}{K_A^{(\text{wt})}}} \right)^2 \right]^{-1}} \right), \quad (7.1)$$

in which  $\Delta\varepsilon_{AI}$  is the energy difference between the active and inactive states of the repressor,  $c$  is the inducer concentration, and  $\beta = 1/k_B T$  where  $k_B$  is the Boltzmann constant and  $T$  is the temperature. The derivative with respect to  $c$ , which we determined using Mathematica's (Wolfram Research, version 11.2) symbolic computing ability, is given as

$$\begin{aligned} \frac{\partial \beta\Delta F(c)}{\partial c} = & 2e^{-\beta\Delta\varepsilon_{AI}} \left( \frac{K_A^{(\text{mut})^2} (K_A^{(\text{mut})} - K_I^{(\text{mut})}) (c + K_I^{(\text{mut})})}{(c + K_A^{(\text{mut})}) \left[ (c + K_A^{(\text{mut})})^2 K_I^{(\text{mut})^2} + e^{-\beta\Delta\varepsilon_{AI}} K_A^{(\text{mut})^2} (c + K_I^{(\text{mut})})^2 \right]} \right. \\ & \left. - \frac{K_A^{(\text{wt})^2} (K_A^{(\text{wt})} - K_I^{(\text{wt})}) (c + K_I^{(\text{wt})})}{(c + K_A^{(\text{wt})}) \left[ (c + K_A^{(\text{wt})})^2 K_I^{(\text{wt})^2} + e^{-\beta\Delta\varepsilon_{AI}} K_A^{(\text{wt})^2} (c + K_I^{(\text{wt})})^2 \right]} \right). \end{aligned} \quad (7.2)$$

This unwieldy expression can be simplified by defining the values of  $K_A^{(\text{mut})} = \theta K_A^{(\text{wt})}$  and  $K_I^{(\text{mut})} = \theta K_I^{(\text{wt})}$  as relative changes to the wild-type values where  $\theta$  is a scaling parameter. While we can permit  $K_A^{(\text{mut})}$  and  $K_I^{(\text{mut})}$  to vary by different degrees, we will consider the case in which they are equally perturbed such that the ratio of  $K_A$  to  $K_I$  is the same between the mutant and wild-type versions of the repressor. While the equations become more cumbersome when one permits the dissociation constants to vary by different amounts (i.e.  $\theta_{K_A}, \theta_{K_I}$ ), one arrives at the

same conclusion. This definition allows us to rewrite Eq. 7.2 in the form of

$$\frac{\partial \beta\Delta F(c)}{\partial c} = 2K_A^{(\text{wt})}e^{-\beta\Delta\varepsilon_{AI}} \left( \frac{\theta^3(K_A^{(\text{wt})} - K_I^{(\text{wt})})(c + \theta K_I^{(\text{wt})})}{(c + \theta K_A^{(\text{wt})}) \left[ \theta^2 K_I^{(\text{wt})^2} (c + \theta K_A^{(\text{wt})})^2 + e^{-\beta\Delta\varepsilon_{AI}} \theta^2 K_A^{(\text{wt})^2} (c + \theta K_I^{(\text{wt})})^2 \right]} \right. \\ \left. - \frac{(K_A^{(\text{wt})} - K_I^{(\text{wt})})(c + K_I^{(\text{wt})})}{(c + K_A^{(\text{wt})}) \left[ (c + K_A^{(\text{wt})})^2 K_I^{(\text{wt})^2} + e^{-\beta\Delta\varepsilon_{AI}} K_A^{(\text{wt})^2} (c + K_I^{(\text{wt})})^2 \right]} \right). \quad (7.3)$$

With this derivative in hand, we can examine the limits of inducer concentration. As discussed in the main text, the free energy difference between the mutant and wild-type repressors when  $c = 0$  should be equal to 0. However, the derivative at  $c = 0$  will be different between the wild-type and the mutant. In this limit, Eq. 7.3 simplifies to

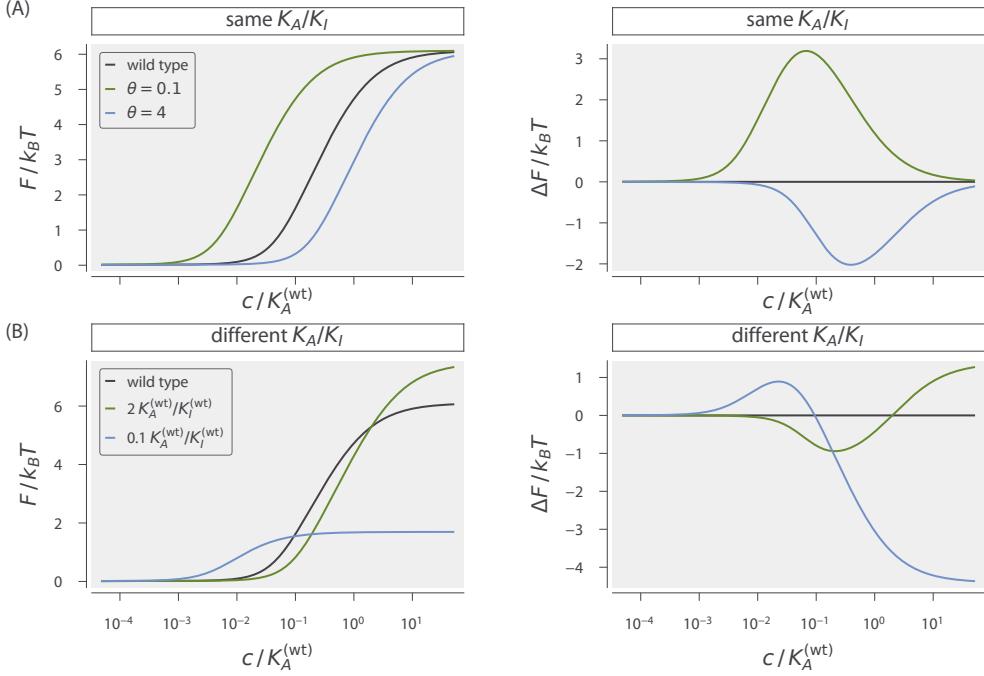
$$\left. \frac{\partial \beta\Delta F(c)}{\partial c} \right|_{c=0} = \frac{2e^{-\beta\Delta\varepsilon_{AI}} (K_A^{(\text{wt})} - K_I^{(\text{wt})})}{K_A^{(\text{wt})} K_I^{(\text{wt})} (1 + e^{-\beta\Delta\varepsilon_{AI}})} \left( \frac{1}{\theta} - 1 \right). \quad (7.4)$$

When  $\theta < 1$ , meaning that the affinity of the active and inactive states of the repressor to the inducer is increased relative to wild-type, the derivative is positive. Thus, the repressor bound state of the promoter becomes less energetically favorable than the repressor bound state. Similarly, if  $\theta > 1$ , binding of the inducer to the mutant repressor is weaker than the wild-type repressor, making  $\partial \beta\Delta F(c) / \partial c < 0$ , meaning the repressor bound state becomes more energetically favorable than the repressor unbound state of the promoter.

With an intuition for the sign of the derivative when  $c = 0$ , we can compute the derivative at another extreme where  $c \gg 0$ . Here, Eq. 7.3 reduces to

$$\left. \frac{\partial \beta\Delta F(c)}{\partial c} \right|_{c \gg 0} \approx \frac{2e^{-\beta\Delta\varepsilon_{AI}} K_A^{(\text{wt})^2} (K_A^{(\text{wt})} - K_I^{(\text{wt})})}{c^2 \left( K_I^{(\text{wt})^2} + e^{-\beta\Delta\varepsilon_{AI}} K_A^{(\text{wt})^2} \right)} (\theta - 1). \quad (7.5)$$

When  $\theta > 1$ , Eq. 7.5 is positive. This is the opposite sign of the derivative when  $c = 0$  when  $\theta > 1$ . When  $\theta < 1$ , Eq. 7.5 becomes negative whereas Eq. 7.4 is



**Figure 7.1: Non-monotonic behavior of  $\Delta F$  with changes in  $K_A$  and  $K_I$ .** Middle column shows the allosteric contribution of free energy  $F$  plotted as a function of the inducer concentration. Right column shows the free energy difference  $\Delta F$  as a function of inducer concentration, revealing non-monotonicity. (A) Behavior of  $F$  and  $\Delta F$  when the values of  $K_A$  and  $K_I$  change relative to wild-type, but maintain the same ratio.  $\theta$  is the scaling factor for both inducer dissociation constants. (B) Behavior of  $F$  and  $\Delta F$  when the values of  $K_A$  and  $K_I$  change relative to the wild-type, but by different factors. In both panels, the wild-type parameter values were taken to be  $K_A = 200 \mu\text{M}$ ,  $K_I = 1 \mu\text{M}$  and  $\Delta\varepsilon_{AI} = 4.5 k_B T$ . An interactive version of this figure is available on the paper website

positive. As the derivative of  $\Delta F$  with respect to  $c$  changes signs across the defined range of inducer concentrations, we can say the function is non-monotonic.

Fig. 7.1 shows the non-monotonic behavior of  $\Delta F$  when  $K_A$  and  $K_I$  change by the same factor  $\theta$  [maintaining the wild-type ratio, Fig. 7.1 (A) and when  $K_A$  and  $K_I$  change by different factors Fig. 7.1. In both cases, non-monotonic behavior is observed with the peak difference in the free energy covering several  $k_B T$ . We have hosted an interactive figure similar to Fig. 7.1 on the paper website where the reader can modify how  $K_A$  and  $K_I$  are affected by a mutation and examine how the active probability, free energy difference, and  $\partial\beta\Delta F/\partial c$  are tuned.

## 7.2 Bayesian Parameter Estimation For DNA Binding Mutants

In this section, we outline the statistical model used in this work to estimate the DNA binding energy for a given mutation in the DNA binding domain. The methodology presented here is similar to that performed in Chapter 2 and outlined in accompanying Chapter 6. In the following text, we take a very detailed approach to vetting the robustness of our statistical inference machinery as determination of parameter values is critical to assessing the effects of mutations. Similarly to what is presented in Chapter 6, we begin with a derivation of our statistical model using Bayes' theorem and then perform a series of principled steps to validate our choices of priors, ensure computational feasibility, and assess the validity of the model given the collected data. This work follows the analysis pipeline outlined by Michael Betancourt in his case-study entitled "Towards A Principled Bayesian Workflow."

The second subsection *Building a Generative Statistical Model* lays out the statistical model used in this work to estimate the DNA binding energy and the error term  $\sigma$ . The subsequent subsections – *Prior Predictive Checks*, *Simulation Based Calibration*, and *Posterior Predictive Checks* – define and summarize a series of tests that ensure that the parameters of the statistical model can be identified and are computationally tractable. To understand how we defined our statistical model, only the second subsection is needed.

### Calculation of the Fold-Change in Gene Expression

We appreciate the subtleties of the efficiency of photon detection in the flow cytometer, fluorophore maturation and folding, and autofluorescence correction, and we understand the importance in modeling the effects that these processes have on the reported value of the fold-change. However, in order to be consistent with the methods used in the literature, we took a more simplistic approach to calculate the fold-change. Given a set of fluorescence measurements of the constitutive expression control ( $R = 0$ ), an autofluorescence control (no YFP), and the experimental

strain ( $R > 0$ ), we calculate the fold-change as

$$\text{fold-change} = \frac{\langle I_{\text{cell}}(R > 0) \rangle - \langle I_{\text{autofluorescence}} \rangle}{\langle I_{\text{cell}}(R = 0) \rangle - \langle I_{\text{autofluorescence}} \rangle}. \quad (7.6)$$

It is important to note here that for a given biological replicate, we consider only a point estimate of the mean fluorescence for each sample and perform a simple subtraction to adjust for background fluorescence. For the analysis going forward, all mentions of measured fold-change are determined by this calculation.

### Building a Generative Statistical Model

To identify the minimal parameter set affected by a mutation, we assume that mutations in the DNA binding domain of the repressor alters only the DNA binding energy  $\Delta\varepsilon_{RA}$ , while the other parameters of the repressor are left unperturbed from their wild-type values. As a first approach, we can assume that all of the other parameters are known without error and can be taken as constants in our physical model. Ultimately, we want to know how probable a particular value of  $\Delta\varepsilon_{RA}$  is given a set of experimental measurements  $y$ . Bayes' theorem computes this distribution, termed the *posterior distribution* as

$$g(\Delta\varepsilon_{RA} | y) = \frac{f(y | \Delta\varepsilon_{RA})g(\Delta\varepsilon_{RA})}{f(y)} \quad (7.7)$$

where we have used  $g$  and  $f$  to represent probability densities over parameters and data, respectively. The expression  $f(y | \Delta\varepsilon_{RA})$  captures the likelihood of observing our data set  $y$  given a value for the DNA binding energy under our physical model. All knowledge we have of what the DNA binding energy *could* be, while remaining completely ignorant of the experimental measurements, is defined in  $g(\Delta\varepsilon_{RA})$ , referred to as the *prior distribution*. Finally, the likelihood that we would observe the data set  $y$  while being ignorant of our physical model is defined by the denominator  $f(y)$ . In this work, this term serves only as a normalization factor and as a result will be treated as a constant. We can therefore say that the posterior distribution of  $\Delta\varepsilon_{RA}$  is proportional to the joint distribution between the likelihood and

the prior,

$$g(\Delta\epsilon_{RA} | y) \propto f(y | \Delta\epsilon_{RA})g(\Delta\epsilon_{RA}). \quad (7.8)$$

We are now tasked with translating this generic notation into a concrete functional form. Our physical model derived in Chapter 2 given by computes the average fold-change in gene expression. Speaking practically, we make several replicate measurements of the fold-change to reduce the effects of random errors. As each replicate is independent of the others, it is reasonable to expect that these measurements will be normally distributed about the theoretical value of the fold-change  $\mu$ , computed for a given  $\Delta\epsilon_{RA}$ . We can write this mathematically for each measurement as

$$f(y | \Delta\epsilon_{RA}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_i^N \exp \left[ \frac{-(y_i - \mu(\Delta\epsilon_{RA}))^2}{2\sigma^2} \right], \quad (7.9)$$

where  $N$  is the number of measurements in  $y$  and  $y_i$  is the  $i^{\text{th}}$  experimental fold-change measurement. We can write this likelihood in shorthand as

$$f(y | \Delta\epsilon_{RA}) = \text{Normal}\{\mu(\Delta\epsilon_{RA}), \sigma\} \quad (7.10)$$

which we will use for the remainder of this section.

Using a normal distribution for our likelihood has introduced a new parameter  $\sigma$  which describes the spread of our measurements about the true value. We must therefore include it in our parameter estimation and assign an appropriate prior distribution such that the posterior distribution becomes

$$g(\Delta\epsilon_{RA}, \sigma | y) \propto f(y | \Delta\epsilon_{RA}, \sigma)g(\Delta\epsilon_{RA})g(\sigma). \quad (7.11)$$

We are now tasked with assigning functional forms to the priors  $g(\Delta\epsilon_{RA})$  and  $g(\sigma)$ . Though one hopes that the result of the inference is not too dependent on the choice of prior, it is important to choose one that is in agreement with our physical and physiological intuition of the system.

We can impose physically reasonable bounds on the possible values of the DNA binding energy  $\Delta\epsilon_{RA}$ . We can say that it is unlikely that any given mutation in

the DNA binding domain will result in an affinity greater than that of biotin to streptavidin [1 fM  $\approx -35 \text{ k}_\text{B}T$ , BNID 107139 (??)], one of the strongest known non-covalent bonds. Similarly, it's unlikely that a given mutation will result in a large, positive binding energy, indicating non-specific binding is preferable to specific binding ( $\sim 1$  to  $10 \text{ k}_\text{B}T$ ). While it is unlikely for the DNA binding energy to exceed these bounds, it's not impossible, meaning we should not impose these limits as hard boundaries. Rather, we can define a weakly informative prior as a normal distribution with a mean and standard deviation as the average of these bounds,

$$g(\Delta\epsilon_{RA}) \sim \text{Normal}\{-12, 12\}$$

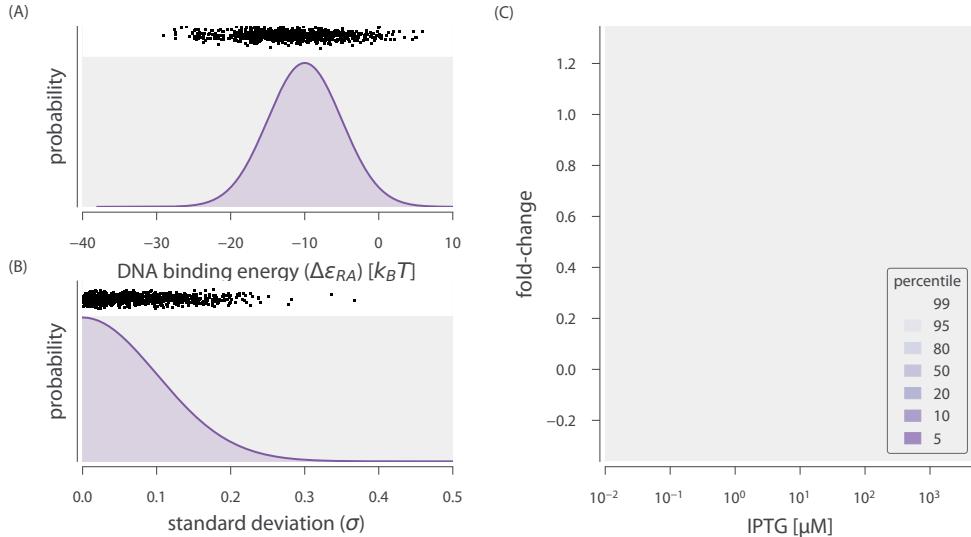
{eq:epRA\_prior} whose probability density function is shown in Fig. 7.2 (A).

By definition, fold-change is restricted to the bounds  $[0, 1]$ . Measurement noise and fluctuations in autofluorescence background subtraction means that experimental measurements of fold-change can extend beyond these bounds, though not substantially. By definition, the scale parameter  $\sigma$  must be positive and greater than zero. We also know that for the measurements to be of any use, the error should be less than the available range of fold-change, 1.0. We can choose such a prior as a half normal distribution

$$g(\sigma) = \frac{1}{\phi} \sqrt{\frac{2}{\pi}} \exp\left[-\frac{\sigma^2}{2\phi^2}\right]; \forall \sigma \geq 0 \quad (7.12)$$

where  $\phi$  is the standard deviation. By choosing  $\phi = 0.1$ , it is unlikely that  $\sigma \geq 1$  yet not impossible, permitting the occasional measurement significantly outside of the theoretical bounds. The probability density function for this prior is shown in Fig. 7.2 (B).

While these choices for the priors seem reasonable, we can check their appropriateness by using them to simulate a data set and checking that the hypothetical fold-change measurements obey our physical and physiological intuition.



**Figure 7.2: Prior distributions and prior predictive check for estimation of the DNA binding energy.** (A) Prior probability density function for DNA binding energy  $\Delta\epsilon_{RA}$  as  $\sim \text{Normal}(-12, 12)$ . (B) Prior probability density function for the standard deviation in measurement noise  $\sigma \sim \text{HalfNormal}(0, 0.1)$ . (C) Percentiles of values drawn from the likelihood distribution given draws from prior distributions given  $R = 260$ ,  $K_A = 139 \mu\text{M}$ ,  $K_I = 0.53 \mu\text{M}$ , and  $\Delta\epsilon_{AI} = 4.5 k_B T$ , which match the parameters used in Razo-Mejia et al. (2018). Black points at top of (A) and (B) represent draws used to generate fold-change measurements from the likelihood distribution. Percentiles in (C) generated from 800 draws from the prior distributions. For each draw from the prior distributions, a data set of 70 measurements over 12 IPTG concentrations (ranging from 0 to 5000  $\mu\text{M}$ ) were generated from the likelihood.

### Prior Predictive Checks

If our choice of prior distribution for each parameter is appropriate, we should be able to simulate data sets using these priors that match our expectations. In essence, we would hope that these prior choices would generate some data sets with fold-change measurements above 1 or below zero, but they should be infrequent. If we end up getting primarily negative values for fold-change, for example, then we can surmise that there is something wrong in our definition of the prior distribution. This method, coined a *prior predictive check*, was first put forward in Good (1950) and has received newfound attention in computational statistics.

We perform the simulation in the following manner. We first draw a random

value for  $\Delta\epsilon_{RA}$  out of its prior distribution stated in Eq. ?? and calculate what the mean fold-change should be given our physical model. With this in hand, we draw a random value for  $\sigma$  from its prior distribution, specified in Eq. 7.12. We then generate a simulated data set by drawing  $\approx 70$  fold-change values across twelve inducer concentrations from the likelihood distribution which we defined in Eq. 7.10. This roughly matches the number of measurements made for each mutant in this work. We repeat this procedure for 800 draws from the prior distributions, which is enough to observe the occasional extreme fold-change value from the likelihood. As the DNA binding energy is the only parameter of our physical model that we are estimating, we had to choose values for the others. We kept the values of the inducer binding constants  $K_A$  and  $K_I$  the same as the wild-type repressor ( $139 \mu\text{M}$  and  $0.53 \mu\text{M}$ , respectively). We chose to use  $R = 260$  repressors per cell as this is the repressor copy number we used in the main text to estimate the DNA binding energies of the three mutants.

The draws from the priors are shown in Fig. 7.2 (A) and (B) as black points above the corresponding distribution. To display the results, we computed the percentiles of the simulated data sets at each inducer concentration. These percentiles are shown as red shaded regions in Fig. 7.2 (C). The 5th percentile (dark purple band) has the characteristic profile of an induction curve. Given that the prior distribution for  $\Delta\epsilon_{RA}$  is centered at  $-12 k_B T$  and we chose  $R = 260$ , we expect the generated data sets to cluster about the induction profile defined by these values. More importantly, approximately 95% of the generated data sets fall between fold-change values of -0.1 and 1.1, which is within the realm of possibility given the systematic and biological noise in our experiments. The 99<sup>th</sup> percentile maximum is approximately 1.3 and the minimum approximately -0.3. While we could tune our choice of prior further to minimize draws this far from the theoretical bounds, we err on the side of caution and accept these values as it is possible that fold-change measurements this high or low can be observed, albeit rarely. Through these prior predictive checks, we feel confident that these choices of priors are appropriate for the parameters we wish to estimate. We can now move

forward and make sure that the statistical model as a whole is valid and computationally tractable.

### Sensitivity Analysis and Simulation Based Calibration

Satisfied with our choice of prior distributions, we can proceed to check other properties of the statistical model and root out any pathologies lurking in our model assumptions.

To build trust in our model, we could generate a data set  $\tilde{y}$  with a *known* value for  $\sigma$  and  $\Delta\varepsilon_{RA}$ , estimate the posterior distribution  $g(\Delta\varepsilon_{RA}, \sigma | \tilde{y})$ , and determine how well we were able to retrieve the true value of the parameters. However, running this once or twice for handpicked values of  $\sigma$  and  $\Delta\varepsilon_{RA}$  won't reveal edge-cases in which the inference fails, some of which may exist in our data. Rather than performing this operation once, we can run this process over a variety of data sets where the ground truth parameter value is drawn from the prior distribution (as we did for the prior predictive checks). For an arbitrary parameter  $\theta$ , the joint distribution between the ground truth value  $\tilde{\theta}$ , the inferred value  $\theta$ , and the simulated data set  $\tilde{y}$  can be written as

$$\pi(\theta, \tilde{y}, \tilde{\theta}) = g(\theta | \tilde{y}) f(\tilde{y} | \tilde{\theta}) g(\tilde{\theta}). \quad (7.13)$$

If this process is run for a large number of simulations, Eq. 7.13 can be marginalized over all data sets  $\tilde{y}$  and all ground truth values  $\tilde{\theta}$  to yield the original prior distribution,

$$\int d\tilde{\theta} \int d\tilde{y} \pi(\theta, \tilde{y}, \tilde{\theta}) = g(\theta). \quad (7.14)$$

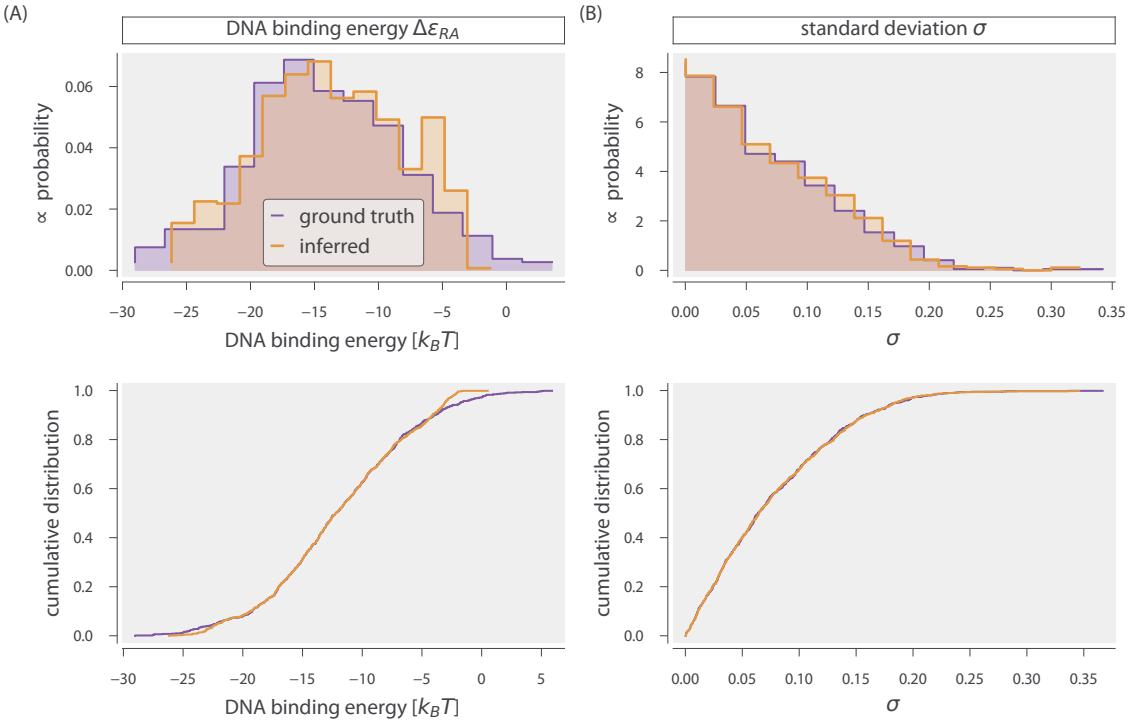
This result, described by Talts et al. (2018), holds true for *any* statistical model and is a natural self consistency property of Bayesian inference. Any deviation between the distribution of our inferred values for  $\theta$  and the original prior distribution  $g(\theta)$  indicates that either our statistical model is malformed or the computational method is not behaving as expected. There are a variety of ways we can ensure that this condition is satisfied, which we outline below.

Using the data set generated for the prior predictive checks [shown in Fig. 7.2 (C)], we sampled the posterior distribution and compute  $\Delta\varepsilon_{RA}$  and  $\sigma$  for each simulation and checked that they matched the original prior distribution. To perform the inference, we use Markov chain Monte Carlo (MCMC) to sample the posterior distribution. Specifically, we use the Hamiltonian Monte Carlo algorithm implemented in the Stan probabilistic programming language (Carpenter et al., 2017). The specific code files can be accessed through the paper website or the associated GitHub repository. The original prior distribution and the distribution of inferred parameter values can be seen in Fig. Fig. 7.3 (A) and (B). For both  $\Delta\varepsilon_{RA}$  and  $\sigma$ , we can accurately recover the ground truth distribution (purple) via sampling with MCMC (orange). For  $\Delta\varepsilon_{RA}$ , there appears to be an upper and lower limit past which we are unable to accurately infer the binding energy. This can be seen in both the histogram Fig. 7.3 and the empirical cumulative distribution Fig. 7.3 as deviations from the ground truth when DNA binding is below  $\approx -25k_B T$  or above  $\approx -5k_B T$ . These limits hinder our ability to comment on exceptionally strong or weak binding affinities. However, as all mutants queried in this work exhibited binding energies between these limits, we surmise that the inferential scheme permits us to draw conclusions about the inferred DNA binding strengths.

Rather than examining the agreement of the data-averaged posterior and the ground truth prior distribution solely by eye, we can compute summary statistics using the mean  $\mu$  and standard deviation  $\sigma$  of the posterior and prior distributions which permit easier identification of pathologies in the inference. One such quantity is the posterior  $z$ -score, which is defined as

$$z = \frac{\mu_{\text{posterior}} - \tilde{\theta}}{\sigma_{\text{posterior}}}. \quad (7.15)$$

This statistic summarizes how accurately the posterior recovers the ground truth value beyond simply reporting the mean, median, or mode of the posterior distribution.  $Z$ -scores around 0 indicate that the posterior is concentrating tightly about the true value of the parameter whereas large values (either positive or negative) indicate that the posterior is concentrating elsewhere. A useful feature of



**Figure 7.3: Comparison of averaged posterior and prior distributions for  $\Delta\epsilon_{RA}$  and  $\sigma$ .** (A) Distribution of the average values for the DNA binding energy  $\Delta\epsilon_{RA}$  (red) overlaid with the ground truth distribution (blue). (B) Data averaged posterior (red) for the standard deviation of fold-change measurements overlaid with the ground truth distribution (blue). Top and bottom show the same data with different visualizations.

this metric is that the width of the posterior is also considered. It is possible that the posterior could have a mean very close to the ground truth value, but have an incredibly narrow distribution/spread such that it does not overlap with the ground-truth. Only comparing the mean value to the ground truth would suggest that the inference “worked”. However with a small standard deviation generates a very large  $z$ -score, telling us that something has gone awry.

If our inferential model is behaving properly, the width of the posterior distribution should be significantly smaller than the width of the prior, meaning that the posterior is being informed by the data. The level to which the posterior is being informed by the data can be easily calculated given knowledge of both the prior and posterior distribution. This quantity, aptly named the shrinkage  $s$ , can

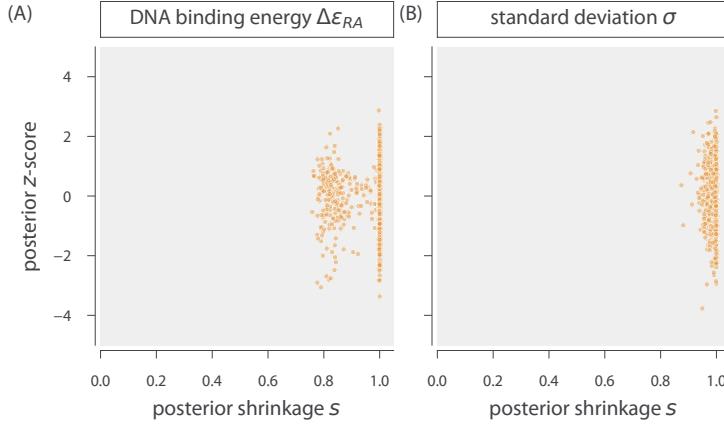
be computed as

$$s = 1 - \frac{\sigma_{\text{posterior}}^2}{\sigma_{\text{prior}}^2}. \quad (7.16)$$

When the shrinkage is close to zero, the variance of the posterior is approximately the same as the variance of the prior, model is not being properly informed by the data. When  $s \approx 1$ , the variance of the posterior is much smaller than the variance of the prior, indicating that the it is being highly informed by the data. A shrinkage less than 0 indicates that the posterior is wider than the prior distribution, revealing a severe pathology in either the model itself or the implementation.

In Fig. 7.4), we compute these summary statistics for each parameter. For both  $\Delta\varepsilon_{RA}$  and  $\sigma$ , we see clustering of the  $z$ -score about 0 with the extrema reaching  $\approx \pm 3$ . This suggests that for the vast majority of our simulated data sets, the posterior distribution concentrated about the ground truth value. We also see that for both parameters, the posterior shrinkage  $s$  is  $\approx 1$ , indicating that the posterior is being highly informed by the data. There is a second distribution centered  $\approx 0.8$  for  $\Delta\varepsilon_{RA}$ , indicating that for a subset of the data sets, the posterior is only  $\approx 80\%$  narrower than the prior distribution. These samples are those that were drawn outside of the limits of  $\approx -25$  to  $-5 k_B T$  where the inferential power is limited. Nevertheless, the posterior still significantly shrank, indicating that the data strongly informs the posterior.

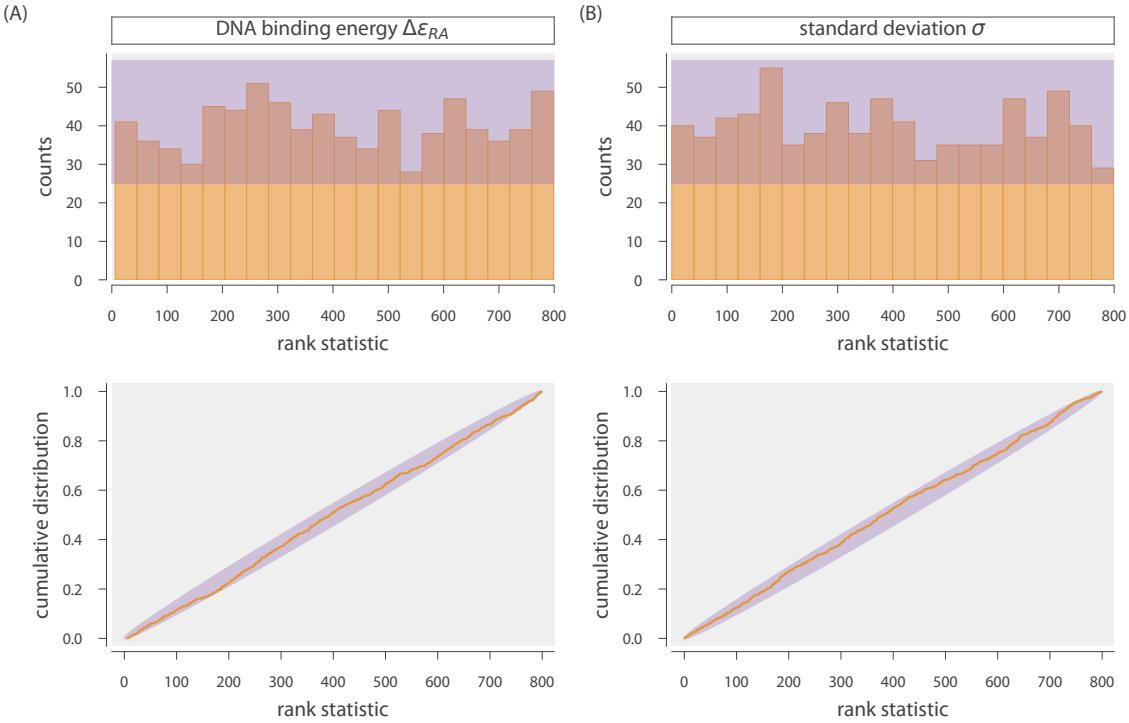
The general self-consistency condition given by Eq. 7.14 provides another route to ensure that the model is computationally tractable. Say that we draw a value for the DNA binding energy from the prior distribution, simulate a data set, and sample the posterior using MCMC. The result of this sampling is a collection of  $N$  values of the parameter which may be above, below, or equal to the ground-truth value. From this set of values, we select  $L$  of them and rank order them by their value. Talts and colleagues (Talts et al., 2018) derived a general theorem which states that the number of samples less than the ground truth value of the parameter (termed the rank statistic) is uniformly distributed over the interval  $[0, L]$ . As Eq. 7.14) *must* hold true for any statistical model, deviations from uniformity sig-



**Figure 7.4: Inferential sensitivity for estimation of  $\Delta\epsilon_{RA}$  and  $\sigma$ .** The posterior z-score for each posterior distribution inferred from a simulated data set is plotted against the shrinkage for (A) the DNA binding energy  $\Delta\epsilon_{RA}$  and (B) the standard deviation of fold-change measurements  $\sigma$

nal that there is a problem in the implementation of the statistical model. How the distribution deviates is also informative as different types of failures result in different distributions. The nature of these deviations, along with a more formal proof of the uniform distribution of rank statistics can be found in Talts et al. (2018) where it was originally derived.

Given the sampling statistics for each of the simulated data sets, we took 800 of the MCMC samples of the posterior distribution for each of the 800 simulated data sets and computed the rank statistic. The distributions are shown in Fig. 7.5 as both histograms and ECDFs for the DNA binding energy and standard deviation. The distribution of rank statistics for both parameters appears to be uniform. The purple band overlaying the histograms (top row) as well as the purple envelopes overlaying the ECDFs (bottom row) represent the 99<sup>th</sup> percentile expected from a true uniform distribution. The uniformity of this distribution, along with the well-behaved  $z$ -scores and shrinkage for each parameter, tells us that there are no underlying pathologies in our statistical model and that it is computationally tractable. However, this does not mean that it is correct. Whether this model is valid for the actual observed data is the topic of the next section.



**Figure 7.5: Rank distribution of the posterior samples from simulated data.** Top row shows a histogram of the rank distribution with  $n = 20$  bins. Bottom row is the cumulative distribution for the same data. Purple bands correspond to the 99th percentile of expected variation from a uniform distribution. (A) Distribution for the DNA binding energy  $\Delta\epsilon_{RA}$  and (B) for the standard deviation  $\sigma$ .

### Parameter Estimation and Posterior Predictive Checks

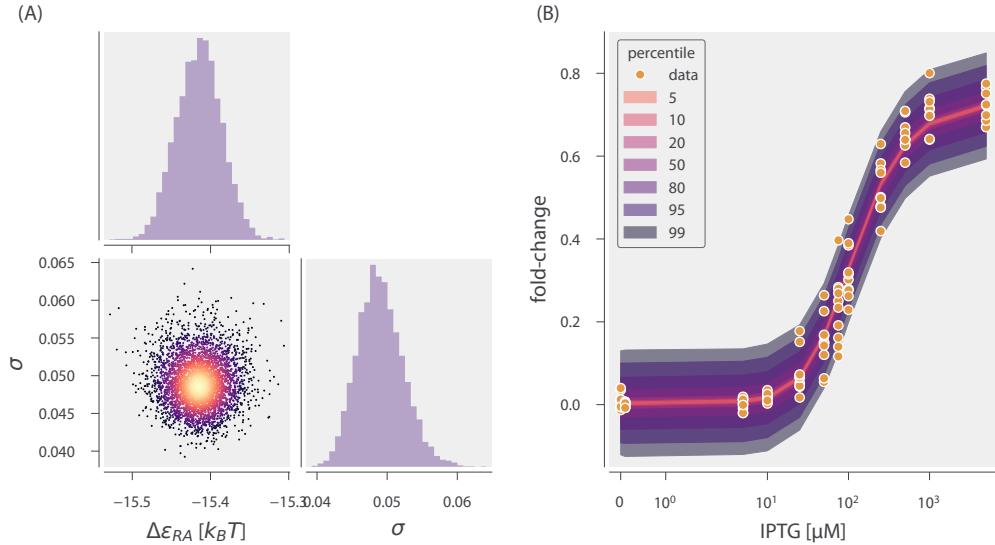
We now turn to applying our vetted statistical model to experimental measurements. While the same statistical model was applied to all three DNA binding mutants, here we only focus on the mutant Q18M for brevity.

Using a single induction profile, we sampled the posterior distribution over both the DNA binding energy  $\Delta\epsilon_{RA}$  and the standard deviation  $\sigma$  using MCMC implemented in the Stan programming language. The output of this process is a set of 4000 samples of both parameters along with the value of their log posterior probabilities, which serves as an approximate measure of the probability of each value. The individual samples are shown in Fig. 7.6. The joint distribution between  $\Delta\epsilon_{RA}$  and  $\sigma$  is shown in the lower left hand corner, and the marginal distributions

for each parameter are shown above and to the right of the joint distribution, respectively. The joint distribution is color coded by the value of the log posterior, with yellow and blue corresponding to high and low probability, respectively. The symmetric shape of the joint distribution is a telling sign that there is no correlation between two parameters. The marginal distributions for each parameter are also relatively narrow, with the DNA binding energy covering a range of  $\approx 0.6 k_B T$  and  $\sigma$  spanning  $\approx 0.02$ . To more precisely quantify the uncertainty, we computed the shortest interval of the marginal distribution for each parameter contains 95% of the probability. The bounds of this interval, coined the Bayesian credible region, can accommodate asymmetry in the marginal distribution since the upper and lower bounds of the estimate are reported. In the main text, we reported the DNA binding energy estimated from these data to be  $15.43^{+0.06}_{-0.06} k_B T$ , where the first value is the median of the distribution and the super- and subscripts correspond to the upper and lower bounds of the credible region, respectively.

While looking at the shape of the posterior distribution can be illuminating, it is not enough to tell us if the parameter values extracted make sense or accurately describe the data on which they were conditioned. To assess the validity of the statistical model in describing actual data, we again turn to simulation, this time using the posterior distributions for each parameter rather than the prior distributions. The likelihood of our statistical model assumes that across the entire induction profile, the observed fold-change is normally distributed about the theoretical prediction with a standard deviation  $\sigma$ . If this is an accurate depiction of the generative process, we should be able to draw values from the likelihood using the sampled values for  $\Delta\epsilon_{RA}$  and  $\sigma$  that are indistinguishable from the actual experimental measurements. This process is known as a *posterior predictive check* and is a Bayesian method of assessing goodness-of-fit.

For each sample from the posterior, we computed the theoretical mean fold-change given the sampled value for  $\Delta\epsilon_{RA}$ . With this mean in hand, we used the corresponding sample for  $\sigma$  and drew a data set from the likelihood distribution



**Figure 7.6: Markov Chain Monte Carlo (MCMC) samples and posterior predictive check for DNA binding mutant Q18M.** (A) Marginal and joint sampling distributions for DNA binding energy  $\Delta\epsilon_{RA}$  and  $\sigma$ . Each point in the joint distribution is a single sample. Marginal distributions for each parameter are shown adjacent to joint distribution. Color in the joint distribution corresponds to the value of the log posterior with the progression of blue to yellow corresponding to increasing probability. (B) The posterior predictive check of model. The measurements of the fold-change in gene expression are shown as black open-faced circles. The percentiles are shown as colored bands and indicate the fraction of simulated data drawn from the likelihood that fall within the shaded region.

the same size as the real data set used for the inference. As we did this for every sample of our MCMC output (a total of  $\approx 4000$ ), it is more instructive to compute the percentiles of the generated data than to show the entire output. In Fig. 7.6 (B), the percentiles of the generated data sets are shown overlaid with the data used for the inference. We see that all of the data points fall within the 99<sup>th</sup> percentile of simulated data sets with the 5<sup>th</sup> percentile tracking the mean of the data at each inducer concentration. As there are no systematic deviations or experimental observations that fall far outside those generated from the statistical model, we can safely say that the statistical model derived here accurately describes the observed data.

### 7.3 Inferring the Free Energy From Fold-Change Measurements

In this section, we describe the statistical model to infer the free energy  $F$  from a set of fold-change measurements. We follow the same principled workflow as described previously for the DNA binding estimation, including declaration of the generative model, prior predictive checks, simulation based calibration, and posterior predictive checks. Finally, we determine an empirical limit in our ability to infer the free energy and define a heuristic which can be used to identify measurements that are likely inaccurate. To understand the statistical model and the empirical limits of detection, only the subsections *Building A Generative Model* and *Sensitivity Limits and Systematic Errors in Inference* are necessary.

#### Building A Generative Model

In Chapter 2, we showed that the fold-change equation can be rewritten in the form of a Fermi function,

$$\text{fold-change} = \frac{1}{1 + e^{-F/k_B T}}, \quad (7.17)$$

where  $F$  corresponds to the free energy difference between the repressor bound and unbound states of the promoter. While the theory prescribes a way for us to calculate the free energy based on our knowledge of the biophysical parameters, we can directly calculate the free energy of a measurement of fold-change by simply rearranging Eq. 7.17 as

$$F = -k_B T \log \left( \frac{1}{\text{fold-change}} - 1 \right). \quad (7.18)$$

With perfect measurement of the fold-change in gene expression (assuming no experimental or measurement noise), the free energy can be directly calculated. However, actual measurements of the fold-change in gene expression can extend beyond the theoretical bounds of 0 and 1, for which the free energy is mathematically undefined.

As the fold-change measurements between biological replicates are independent, it is reasonable to assume that they are normally distributed about a mean

value  $\mu$  with a standard deviation  $\sigma$ . While the mean value is restricted to the bounds of  $[0, 1]$ , fold-change measurements outside of these bounds are still possible given that they are distributed about the mean with a scale of  $\sigma$ . Thus, if we have knowledge of the mean fold-change in gene expression about which the observed fold-change is distributed, we can calculate the mean free energy as

$$F = -k_B T \log \left( \frac{1}{\mu} - 1 \right). \quad (7.19)$$

For a given set of fold-change measurements  $y$ , we wish to infer the posterior probability distribution for  $\mu$  and  $\sigma$ , given by Bayes' theorem as

$$g(\mu, \sigma | y) \propto f(y | \mu, \sigma) g(\mu) g(\sigma), \quad (7.20)$$

where we have dropped the normalization constant  $f(y)$  and assigned a proportionality between the posterior and joint probability distribution. Given that the measurements are independent, we define the likelihood  $f(y | \mu, \sigma)$  as a normal distribution,

$$f(y | \mu, \sigma) \sim \text{Normal}\{\mu, \sigma\}. \quad (7.21)$$

While the mean  $\mu$  is restricted to the interval  $[0, 1]$ , there is no reason *a priori* to think that it is more likely to be closer to either bound. To remain uninformative and be as permissive as possible, we define a prior distribution for  $\mu$  as a Uniform distribution between 0 and 1,

$$g(\mu) = \begin{cases} \frac{1}{\mu_{\max} - \mu_{\min}} & \mu_{\min} < \mu < \mu_{\max} \\ 0 & \text{otherwise} \end{cases}. \quad (7.22)$$

Here,  $\mu_{\min} = 0$  and  $\mu_{\max} = 1$ , reducing  $g(\mu)$  to 1. For  $\sigma$ , we can again assume a half-normal distribution with a standard deviation of 0.1 as was used for estimating the DNA binding energy Eq. 7.12,

$$g(\sigma) = \text{HalfNormal}\{0, 0.1\}. \quad (7.23)$$

With a full generative model defined, we can now use prior predictive checks to ensure that our choices of prior are appropriate for the inference.

## Prior Predictive Checks

To check the validity of the chosen priors, we pulled 1000 combinations of  $\mu$  and  $\sigma$  from their respective distributions Fig. 7.7 and subsequently drew a set of 10 fold-change values (a number comparable to the number of biological replicates used in this work) from a normal distribution defined by  $\mu$  and  $\sigma$ . To visualize the range of values generated from these checks, we computed the percentiles of the empirical cumulative distributions of the fold-change values, as can be seen in Fig. 7.7 (C). Approximately 95% of the the generated fold-change measurements were between the theoretical bounds of [0, 1] whereas 5% of the data sets fell outside with the maximum and minimum values extending to  $\approx 1.2$  and  $-0.2$ , respectively. Given our familiarity with these experimental strains and the detection sensitivity of the flow cytometer, these excursions beyond the theoretical bounds agree with our intuition. Satisfied with our choice of prior distributions, we can proceed to check the sensitivity and computational tractability of our model through simulation based calibration.

## Simulation Based Calibration

To ensure that the parameters can be estimated with confidence, we sampled the posterior distribution of  $\mu$  and  $\sigma$  for each data set generated from the prior predictive checks. For each inference, we computed the z-score and shrinkage for each parameter, shown in Fig. Fig. 7.8(A). For both parameters, the z-scores are approximately centered about zero, indicating that the posteriors concentrate about the ground truth value of the parameter. The z-scores for  $\sigma$  green points in Fig. 7.8 appear to be slightly off centered with more negative values than positive. This suggests that  $\sigma$  is more likely to be slightly overestimated in some cases. The shrinkage parameter for  $\mu$  (red points) is very tightly distributed about 1.0, indicating that the prior is being strongly informed by the data. The shrinkage is more broadly distributed for for  $\sigma$  with a minimum value of  $\approx 0.5$ . However, the median shrinkage for  $\sigma$  is  $\approx 0.9$ , indicating that half of the inferences shrank the prior distribution by at least 90%. While we could revisit the model to try and improve

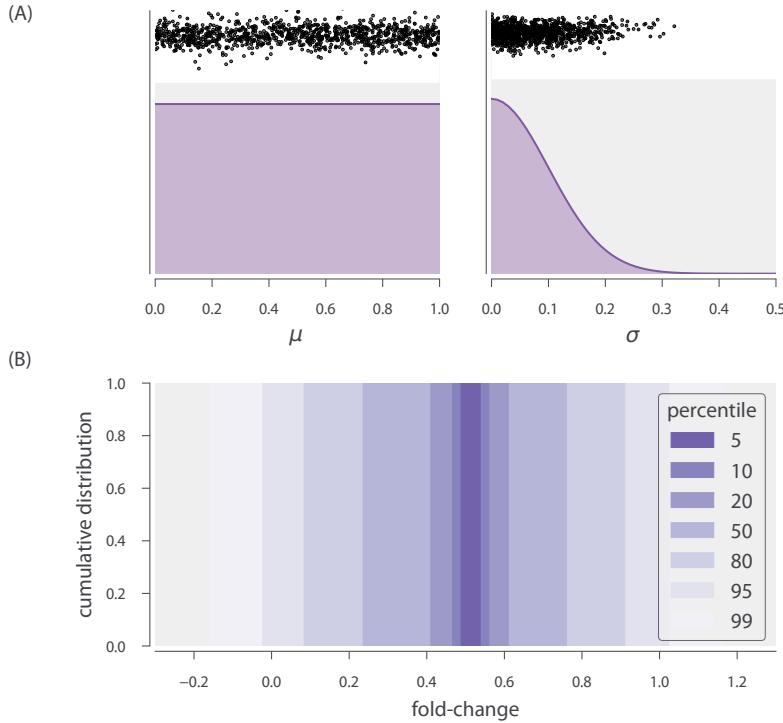
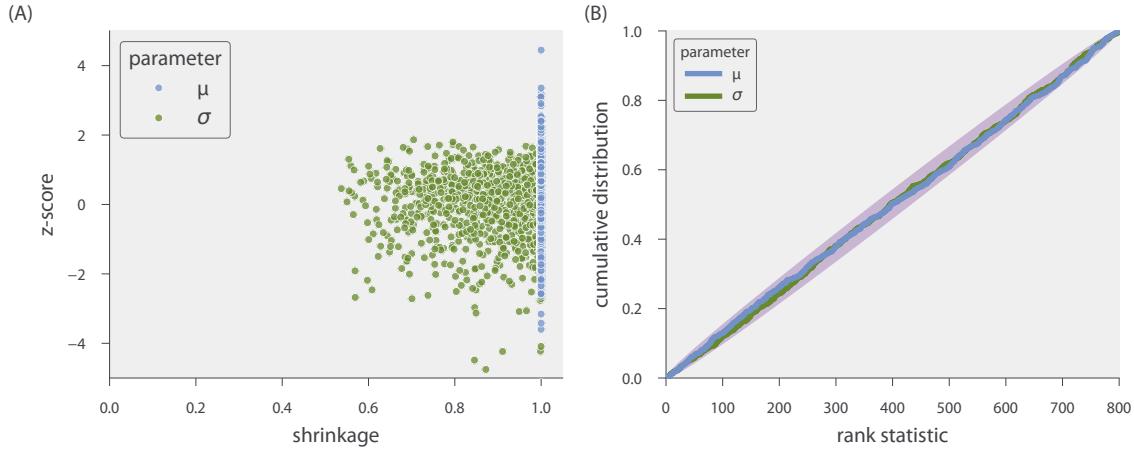


Figure 7.7: \*\*Prior predictive checks for inference of the mean fold-change. (A) The prior distributions for  $\mu$  (left) and  $\sigma$  (right). The vertical axis is proportional to the probability of the value. Black points above distributions correspond to the values used to perform the prior predictive checks. (B) Percentiles of the data generated for each draw from the prior distributions shown as a cumulative distribution. Percentiles were calculated for 1000 generated data sets, each with 10 fold-change measurements drawn from the likelihood given the drawn values of  $\mu$  and  $\sigma$ .

the shrinkage values, we are more concerned with  $\mu$  which shows high shrinkage and zero-centered  $z$ -scores.

To ensure that the model is computationally tractable, we computed the rank statistic of each parameter for each inference. The empirical cumulative distributions for  $\mu$  (black) and  $\sigma$  (red) can be seen in Fig. 7.8 (B). Both distributions appear to be uniform, falling within the 99<sup>th</sup> percentile of the variation expected from a true uniform distribution. This indicates that the self-consistency relation defined by Eq. 7.14. holds for this statistical model. With a computationally tractable model in hand, we can now apply the statistical model to our data and verify that data sets drawn from the data-conditioned posterior are indistinguishable from the ex-



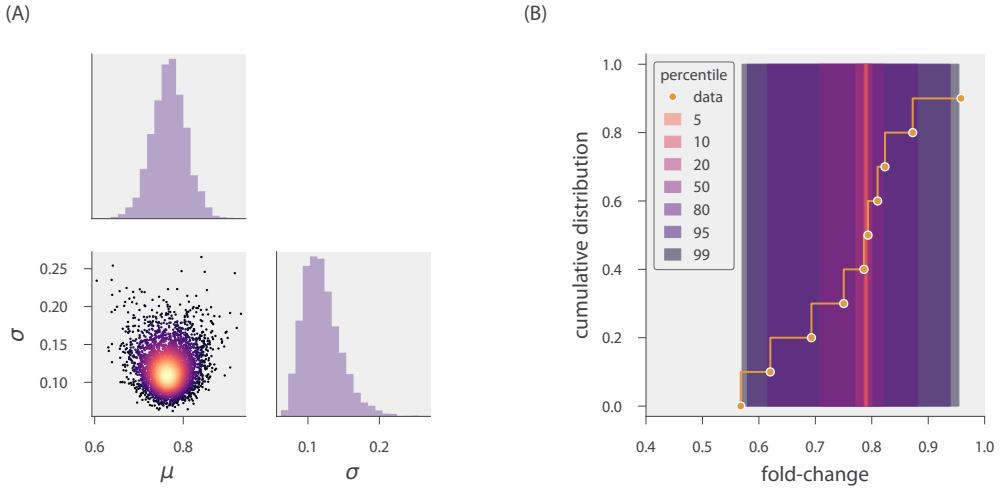
**Figure 7.8: Sensitivity measurements and rank statistic distribution of the statistical model estimating  $\mu$  and  $\sigma$ .** (A) Posterior z-score of each inference plotted against the posterior shrinkage factor for the parameters  $\mu$  (red points) and  $\sigma$  (black points). (B) Distribution of rank statistics for  $\mu$  (red) and  $\sigma$  (black). Gray envelope represents the 99<sup>th</sup> percentile of a true uniform distribution.

perimental measurements.

### Posterior Predictive Checks

The same statistical model was applied to every unique set of fold-change measurements used in this work. Here, we focus only on the set of fold-change measurements for the double mutant Y17I-Q291V at 50  $\mu$ M IPTG. The samples from the posterior distribution conditioned on this dataset can be seen in Fig. 7.9 (A). The joint distribution, shown in the lower left-hand corner, appears fairly symmetric, indicating that  $\mu$  and  $\sigma$  are independent. There is a slight asymmetry in the sampling of  $\sigma$ , which can be more clearly seen in the corresponding marginal distribution to the right of the joint distribution.

For each MCMC sample of  $\mu$  and  $\sigma$ , we drew 10 samples from a normal distribution defined by these parameters. From this collection of data sets, we computed the percentiles of the empirical cumulative distribution and plotted them over the data, as can be seen in Fig. Fig. 7.9 (B). We find that the observed data falls within the 99<sup>th</sup> percentile of the generated data sets. This illustrates that the model can



**Figure 7.9: MCMC sampling output and posterior predictive checks of the statistical model for the mean fold-change  $\mu$  and standard deviation  $\sigma$ .** (A) Corner plot of sampling output. The joint distribution between  $\sigma$  and  $\mu$  is shown in the lower left hand corner. Each point is an individual sample. Points are colored by the value of the log posterior with increasing probability corresponding to transitions from purple to orange. Marginal distributions for each parameter are shown adjacent to the joint distribution. (B) Percentiles of the cumulative distributions from the posterior predictive checks are shown as shaded bars. Data on which the posterior was conditioned are shown as white orange circles and lines.

produce data which is identically distributed to the actual experimental measurements, validating our choice of statistical model.

### Sensitivity Limits and Systematic Errors in Inference

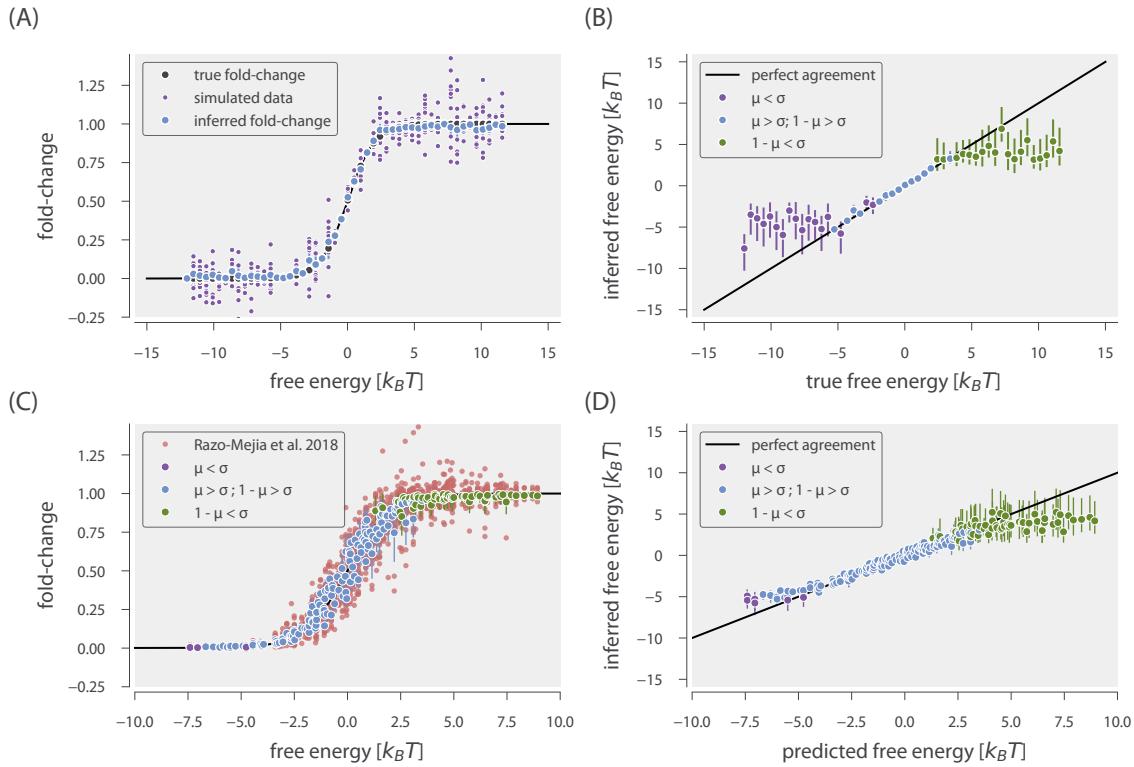
Considering the results from the prior predictive checks, simulation based calibration, and posterior predictive checks, we can say that the statistical model for inferring  $\mu$  and  $\sigma$  fold-change from a collection of noisy fold-change measurements is valid and computationally tractable. Upon applying this model to the experimental data of the wild-type strain (where the free energy is theoretically known), we observed that systematic errors arise when the fold-change is exceptionally high or low, making the resulting inference of the free energy inaccurate.

To elucidate the source of this systematic error, we return to a simulation based approach in which the true free energy is known black points in ???. For a range of

free energies, we computed the theoretical fold-change prescribed by Eq. ??). For each free energy value, we pulled a value for  $\sigma$  from the prior distribution defined in Eq. 7.12 and generated a data set of 10 measurements by drawing values from a normal distribution defined by the true fold-change and the drawn value of  $\sigma$  purple points in ?. We then sampled the statistical model over these data and inferred the mean fold-change  $\mu$  orange points in ?. By eye, the inferred points appear to collapse onto the master curve, in many cases overlapping the true values. However, the points with a free energy less than  $\approx -2 k_B T$  and greater than  $\approx 2 k_B T$  are slightly above or below the master curve, respectively. This becomes more obvious when the inferred free energy is plotted as a function of the true free energy, shown in Fig. ?? (B). Points in which the difference between  $\mu$  and the nearest boundary (0 or 1) is less than the value of  $\sigma$  are shown as purple or green. When this condition is met, the inferred mean free energy strays from the true value, introducing a systematic error. This suggests that the spread of the fold-change measurements sets the detection limit of fold-change close to either boundary. Thus, the narrower the spread in the fold-change the better the estimate of the fold-change near the boundaries.

These systematic errors can be seen in experimental measurements of the wild-type repressor. Data from Razo-Mejia et al. (2018) in which the IPTG titration profiles of seventeen different bacterial strains were measured is shown collapsed onto the master curve in Fig. Fig. ?? (C) as red points. Here, each point corresponds to a single biological replicate. The inferred mean fold-change  $\mu$  and 95% credible regions are shown as purple, blue, or green points. The color of these points correspond to the relative value of  $\mu$  or  $1 - \mu$  to  $\sigma$ . The discrepancy between the predicted and inferred free energy of each measurement set can be seen in Fig. ??(D). The significant deviation from the predicted and inferred free energy occurs past the detection limit set by  $\sigma$ . In this work, we therefore opted to not display inferred free energies at the extrema where the inferred fold-change was closer to the boundaries than the corresponding standard deviation, as it reflects limitations in our measurement rather than a deviation from the theoretical pre-

dictions.



## Additional Characterization of DNA Binding Mutants In Chapter 3, we estimated the DNA binding energy of each mutant using the mutant strains that had approximately 260 repressors per cell. In this section, we examine the effect of the choice of fit strain on the predictions of both the induction profiles and  $\Delta F$  for each DNA binding domain mutant.

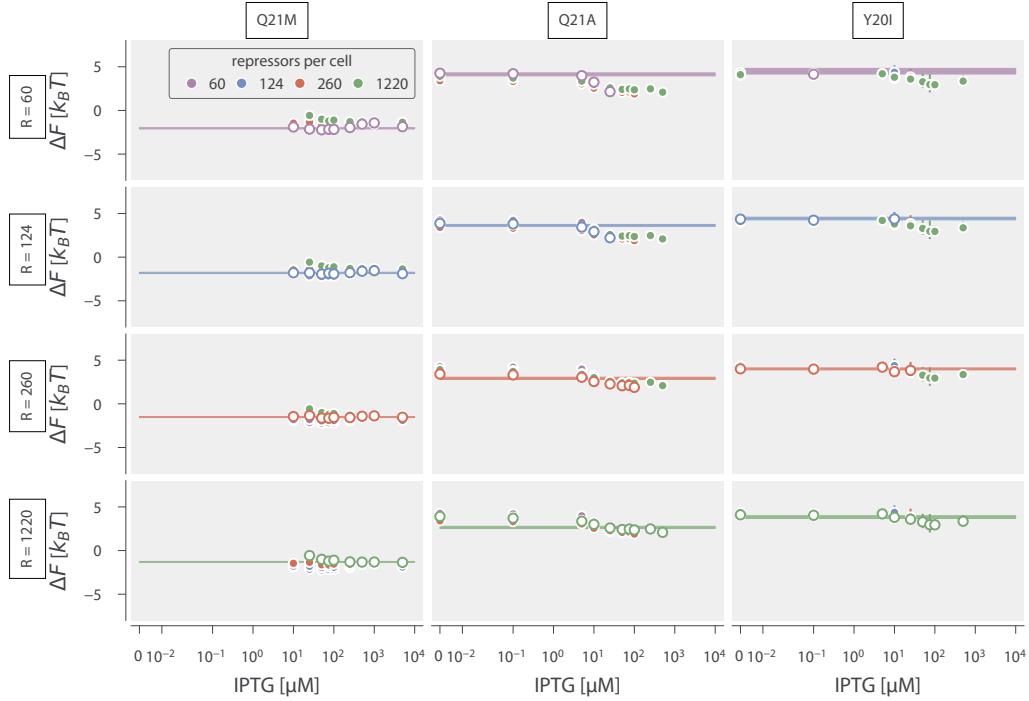
We applied the statistical model derived in Section 2 of this chapter for each unique strain of the DNA binding mutants and estimated the DNA binding energy. The median of the posterior distribution along with the upper and lower bounds of the 95% credible region are reported in Table 7.1. We found that the choice of fitting strain did not strongly influence the estimate of the DNA binding energy. The largest deviations appear for the weakest binding mutants paired with the lowest repressor copy number. In these cases, such as for Q18A, the difference in binding energy between the repressor copy numbers is  $\approx 1 k_B T$  which is small compared to the overall DNA binding energy. Using these energies, we computed the predicted

induction profiles of each mutant with different repressor copy numbers, shown in Fig. 7.10. In this plot, the rows correspond to the repressor copy number of the strain used to estimate the DNA binding energy. The columns correspond to the repressor copy number of the predicted strains. The diagonals, shaded in grey, show the induction profile of the fit strain along with the corresponding data. In all cases, we find that the predicted profiles are relatively accurate with the largest deviations resulting from using the lowest repressor copy number as the fit strain.

Table 7.1: Estimated DNA binding energy for DNA binding domain mutants with different repressor copy numbers. Reported values are the median of the posterior distribution with the upper and lower bounds of the 95% credible regions.

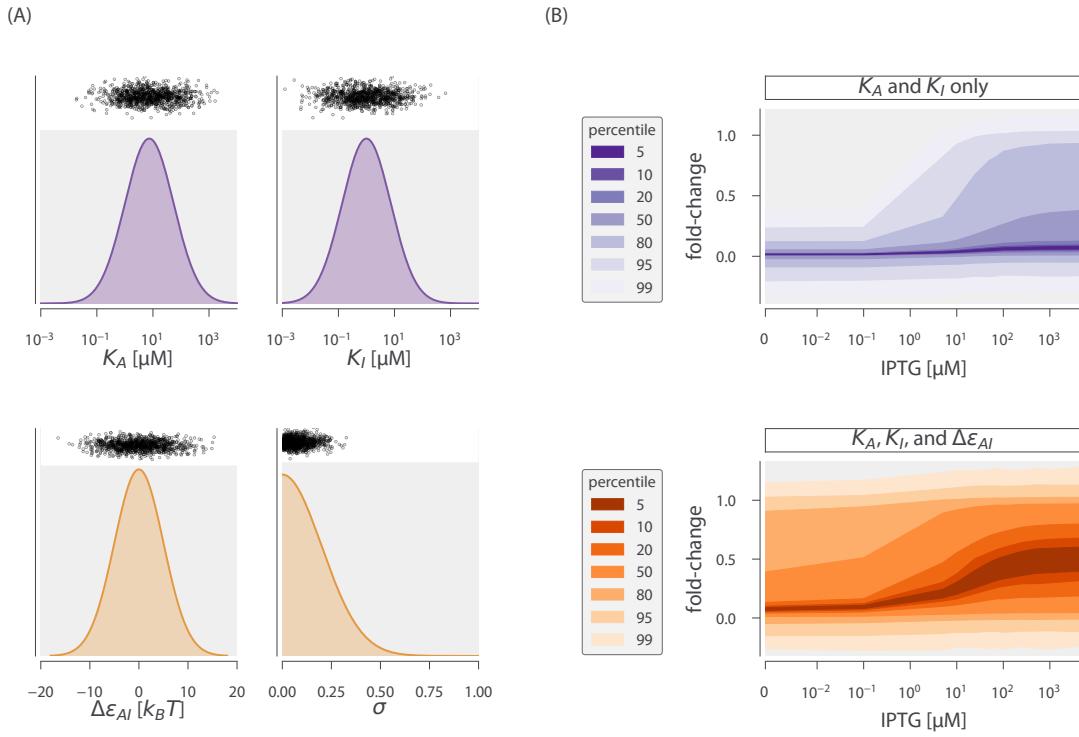
Mutant	Repressors	DNA Binding Energy [ $k_B T$ ]
Q18A	60	$-9.8^{+0.2}_{-0.2}$
	124	$-10.3^{+0.1}_{-0.1}$
	260	$-11.0^{+0.1}_{-0.1}$
	1220	$-11.3^{+0.1}_{-0.1}$
Q18M	60	$-15.83^{+0.08}_{-0.08}$
	124	$-15.7^{+0.1}_{-0.1}$
	260	$-15.43^{+0.07}_{-0.06}$
	1220	$-15.27^{+0.07}_{-0.07}$
Y17I	60	$-9.4^{+0.3}_{-0.3}$
	124	$-9.5^{+0.1}_{-0.1}$
	260	$-9.9^{+0.1}_{-0.1}$
	1220	$-10.1^{+0.2}_{-0.2}$

The predicted change in free energy  $\Delta F$  using each fit strain can be seen in Fig. ???. In this figure, the rows represent the repressor copy number of the strain to which the DNA binding energy was fit whereas the columns correspond to each



**Figure 7.10: Pairwise comparisons of DNA binding mutant induction profiles.** Rows correspond to the repressor copy number of the strain used to estimate the DNA binding energy for each mutant. Columns correspond to the repressor copy number of the strains that are predicted. Diagonals in which the data used to estimate the DNA binding energy are shown with a gray background.

mutant. In each plot, we have shown the data for all repressor copy numbers with the fit strain represented by white filled circles. Much as for the induction profiles, we see little difference in the predicted  $\Delta F$  for each strain, all of which accurately describe the inferred free energies. The ability to accurately predict the majority of the induction profiles of each mutant with repressor copy numbers ranging over two orders of magnitude strengthens our assessment that for these DNA binding domain mutations, only the DNA binding energy is modified.



{#fig:DNA\_delF\_pa}

short-caption="Dependence of fitting strain on  $\Delta F$  predictions of DNA binding domain mutants.\*\*}

## 7.4 Bayesian Parameter Estimation for Inducer Binding Domain Mutants

In Chapter 3, we put forward two naïve hypotheses for which parameters of our fold-change equation are affected by mutations in the inducer binding domain of the repressor. The first hypothesis was that only the inducer dissociation constants,  $K_A$  and  $K_I$ , were perturbed from their wild-type values. Another hypothesis was that the inducer dissociation constants were affected in addition to the energetic difference between the active and inactive states of the repressor,  $\Delta\varepsilon_{AI}$ .

In this section, we first derive the statistical model for each hypothesis and then perform a series of diagnostic tests that expose the inferential limitations of each model. With well calibrated statistical models, we then apply each to an induction profile of the inducer binding mutant Q291K and assess the validity of each hypothesis. To understand the statistical models for each hypothesis, only the subsection *Building A Generative Statistical Model* is necessary.

### Building a Generative Statistical Model

For both hypotheses, we assume that the underlying physical model is the same while a subset of the parameters are modified. As the fold-change measurements for each biological replicate are statistically independent, we can assume that they are normally distributed about the theoretical fold-change value. Thus, for each model, we must include a parameter  $\sigma$  which is the standard deviation of the distribution of fold-change measurements. For the first hypothesis, in which only  $K_A$  and  $K_I$  are changed, we are interested in sampling the posterior distribution

$$g(K_A, K_I, \sigma | y) \propto f(y | K_A, K_I, \sigma) g(K_A) g(K_I) g(\sigma), \quad (7.24)$$

where  $y$  corresponds to the set of fold-change measurements. In the above model, we have assumed that the priors for  $K_A$  and  $K_I$  are independent. It is possible that it is more appropriate to assume that they are dependent and that a single prior distribution captures both parameters,  $g(K_A, K_I)$ . However, assigning this prior is more difficult and requires strong knowledge *a priori* about the relationship between them. Therefore, we continue under the assumption that the priors are independent.

The generic posterior given in Eq. 7.24 can be extended to evaluate the second hypothesis in which  $\Delta\varepsilon_{AI}$  is also modified,

$$g(K_A, K_I, \Delta\varepsilon_{AI}, \sigma | y) \propto f(y | K_A, K_I, \Delta\varepsilon_{AI}, \sigma) g(K_A) g(K_I) g(\Delta\varepsilon_{AI}) g(\sigma) \quad (7.25)$$

where we have included  $\Delta\varepsilon_{AI}$  as an estimated parameter and assigned a prior distribution.

As we have assumed that the fold-change measurements across replicates are independent and normally distributed, the likelihoods for each hypothesis can be written as

$$f(y | K_A, K_I, \sigma) \sim \text{Normal}\{\mu(K_A, K_I), \sigma\}, \quad (7.26)$$

for the first hypothesis and

$$f(y | K_A, K_I, \Delta\varepsilon_{AI}, \sigma) \sim \text{Normal}\{\mu(K_A, K_I, \Delta\varepsilon_{AI}), \sigma\}, \quad (7.27)$$

for the second. Here, we have assigned  $\mu(\dots)$  as the mean of the normal distribution as a function of the parameters defined by our fold-change equation.

With a likelihood distribution in hand, we now turn toward assigning functional forms to each prior distribution. As we have used in the previous sections of this chapter, we can assign a half-normal prior for  $\sigma$  with a standard deviation of 0.1, namely,

$$g(\sigma) \sim \text{HalfNormal}\{0, 0.1\}. \quad (7.28)$$

It is important to note that the inducer dissociation constants  $K_A$  and  $K_I$  are scale invariant, meaning that a change from  $0.1 \mu\text{M}$  to  $1 \mu\text{M}$  yields a decrease in affinity equal to a change from  $10 \mu\text{M}$  to  $100 \mu\text{M}$ . As such, it is better to sample the dissociation constants on a logarithmic scale. We can assign a log normal prior for each dissociation constant as

$$g(K_A) = \frac{1}{K_A \sqrt{2\pi\phi^2}} \exp\left[-\frac{(\log \frac{K_A}{1\mu\text{M}} - \mu_{K_A})^2}{2\phi^2}\right], \quad (7.29)$$

or with the short-hand notion of

$$g(K_A) \sim \text{LogNormal}\{\mu_{K_A}, \phi\} \quad (7.30)$$

For  $K_A$ , we assigned a mean  $\mu_{K_A} = 2$  and a standard deviation  $\phi = 2$ . For  $K_I$ , we chose a mean of  $\mu_{K_I} = 0$  and  $\phi = 2$ , capturing our prior knowledge that  $K_A > K_I$  for the wild-type LacI. While the prior distributions are centered differently, they both show extensive overlap, permitting mutations in which  $K_A < K_I$ . For  $\Delta\varepsilon_{AI}$ , we assign a normal distribution of the prior centered at 0 with a standard deviation of  $5 k_B T$ ,

$$g(\Delta\varepsilon_{AI}) \sim \text{Normal}\{0, 5\}. \quad (7.31)$$

This permits values of  $\Delta\varepsilon_{AI}$  that are above or below zero, meaning that the inactive state of the repressor can be either more or less energetically favorable to the active state. A standard deviation of  $5 k_B T$  permits a wide range of energies with  $+5 k_B T$  and  $-5 k_B T$  corresponding to  $\approx 99.5\%$  and  $\approx 0.5\%$  of the repressors being active in the absence of inducer, respectively.

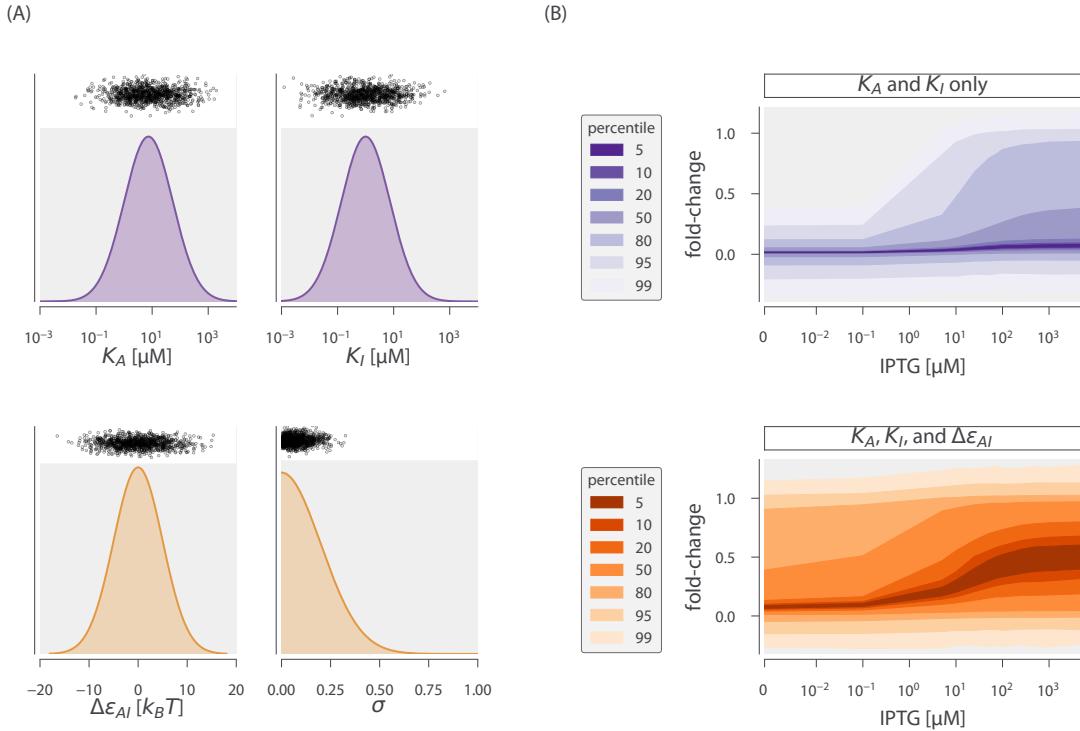
### Prior predictive checks

To ensure that these choices of prior distributions are appropriate, we performed prior predictive checks for each hypothesis as previously described in the second section of this chapter. We drew 1000 values from the prior distributions shown in Fig. 7.11 (A) for  $K_A$ ,  $K_I$ , and  $\Delta\varepsilon_{AI}$ . Using the draws from the  $K_A$ , and  $K_I$  priors alone, we generated data sets of  $\approx 70$  measurements. The percentiles of the fold-change values drawn for the 1000 simulations is shown in the top panel of Fig. 7.11 (B).

It can be seen that in the absence of inducer, the fold-change values are close to zero and are distributed about the leakiness value due to  $\sigma$ . This is in contrast to the data sets generated when  $\Delta\varepsilon_{AI}$  is permitted to vary along with  $K_A$  and  $K_I$ . In the bottom panel of Fig. 7.11 (B), the fold-change when  $c = 0$  can extend above 1.0 which is possible only when  $\Delta\varepsilon_{AI}$  is included, which sets what fraction of the repressors is active. Under both hypotheses, the 99<sup>th</sup> percentile of the fold-change extends to just above 1 or just below 0, which matches our intuition of how the data should behave. Given these results, we are satisfied with these choices of priors and continue onto the next level of calibration of our model.

### Simulation Based Calibration

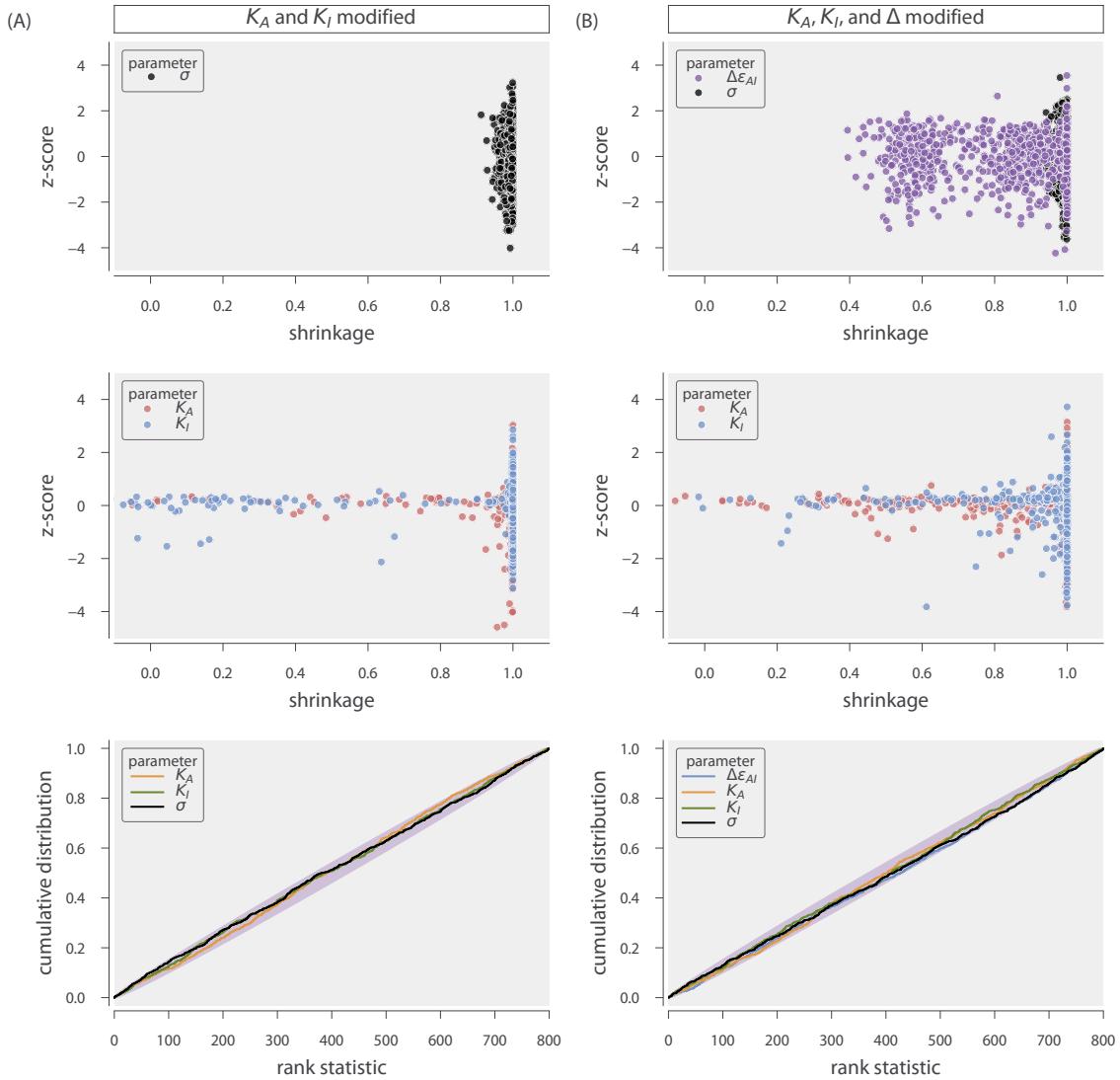
With an appropriate choice of priors, we turn to simulation based calibration to root out any pathologies lurking in the model itself or the implementation through MCMC. For each parameter under each model, we compute the z-score and shrinkage of each inference, shown in Fig. 7.12. Under the first hypothesis in which  $K_A$  and  $K_I$  are the only perturbed parameters Fig. 7.12, we see all parameters have z-scores clustered around 0, indicating that the value of the ground-truth is being accurately estimated through the inference. While the shrinkage for  $\sigma$  is close to 1 (indicating the prior is being informed by the data), the shrinkage for  $K_A$  and  $K_I$  is heavily tailed with some values approaching zero. This is true for both statistical models, indicating that for some values of  $K_A$  and  $K_I$ , the parameters are difficult to pin down with high certainty. In the application of these models to data, this



**Figure 7.11: Prior predictive checks for two hypotheses of inducer binding domain mutants.** (A) Probability density functions for  $K_A$ ,  $K_I$ ,  $\Delta\epsilon_{AI}$ , and  $\sigma$ . Black points correspond to draws from the distributions used for prior predictive checks. (B) Percentiles of the simulated data sets using draws from the  $K_A$  and  $K_I$  distributions only (top, red bands) and using draws from  $K_A$ ,  $K_I$ , and  $\Delta\epsilon_{AI}$  (bottom, blue bands).

will be revealed as large credible regions in the reported parameters. Under the second hypothesis in which all allosteric parameters are allowed to change, we see moderate shrinkage for  $\Delta\epsilon_{AI}$  purple points in 7.12 with the minimum shrinkage being around 0.5. The samples resulting in low shrinkage correspond to values of  $\Delta\epsilon_{AI}$  that are highly positive or highly negative, in which small changes in the active fraction of repressors cannot be accurately measured through our model. However, the median shrinkage for  $\Delta\epsilon_{AI}$  is approximately 0.92, meaning that the data highly informed the prior distributions for the majority of the inferences. The rank distributions for all parameters under each model appear to be highly uniform, indicating that both statistical models are computationally tractable.

With knowledge of the caveats of estimating  $K_A$  and  $K_I$  for both models, we



**Figure 7.12: Simulation based calibration of statistical models for inducer binding domain mutants.** (A) Sensitivity statistics and rank distribution for a statistical model in which  $K_A$  and  $K_I$  are the only parameters permitted to vary. (B) Sensitivity statistics and rank distribution for a model in which all allosteric parameters  $K_A$ ,  $K_I$ , and  $\Delta\epsilon_{AI}$  are allowed to be modified by the mutation. Gray envelope in the bottom plots correspond to the 99<sup>th</sup> percentile of variation expected from a true uniform distribution.

proceed with our analysis and examine how accurately these models can capture the phenomenology of the data.

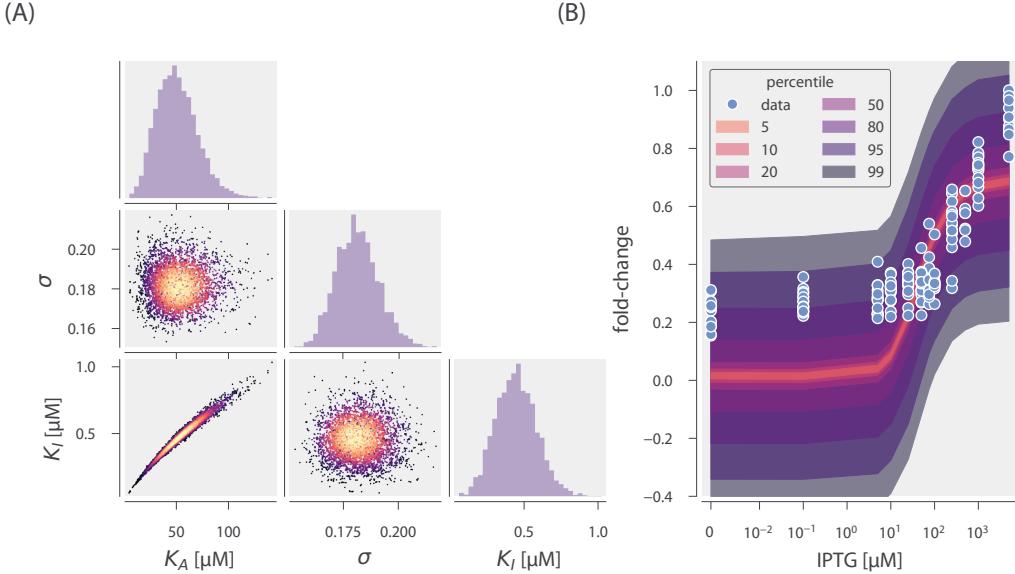
## Posterior Predictive Checks

With a properly calibrated statistical model for each hypothesis, we now apply it to a representative dataset. While each model was applied to each inducer binding domain mutant, we only show the application to the mutant Q291K with 260 repressors per cell paired with the native *lac* operator O2.

The results from applying the statistical model in which only  $K_A$  and  $K_I$  can change is shown in Fig. 7.13. The joint and marginal distributions for each parameter Fig. 7.13 reveal a strong correlation between  $K_A$  and  $K_I$  whereas all other parameters are symmetric and independent. While the joint and marginal distributions look well behaved, the percentiles of the posterior predictive checks Fig. 7.13 are more suspect. While all data falls within the 95<sup>th</sup> percentile, the overall trend of the data is not well predicted. Furthermore, the percentiles expand far below zero, indicating that the sampling of  $\sigma$  is compensating for the leakiness in the data being larger than it should be if only  $K_A$  and  $K_I$  were the changing parameters.

We see significant improvement when  $\Delta\varepsilon_{AI}$  is permitted to vary in addition to  $K_A$  and  $K_I$ . Fig. 7.14 (A) shows the joint and marginal distributions between all parameters from the MCMC sampling. We still see correlation between  $K_A$  and  $K_I$ , although it is not as strong as in the case where they are the only parameters allowed to change due to the mutation. We also see that the marginal distribution for  $\sigma$  has shrunk significantly compared to the marginal distribution in Fig. 7.13 (A). The percentiles of the posterior predictive checks, shown in Fig. 7.14 (B) are much more in line with the experimental measurements, with the 5th percentile following the data for the entire induction profile.

In this section we have presented two hypotheses for the minimal parameter set needed to describe the inducer binding mutations, derived a statistical model for each, thoroughly calibrated its behavior, and applied it to a representative data set. The posterior predictive checks (Fig. 7.13 and Fig. 7.14) help us understand which hypothesis is more appropriate for that particular mutant. The incredibly wide percentiles and significant change in the leakiness that result from a model in



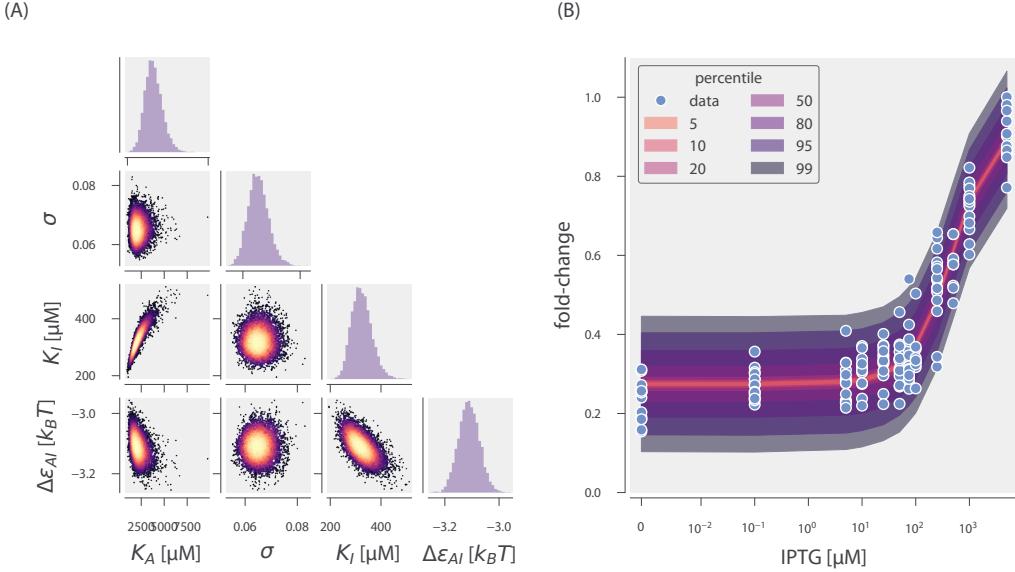
**Figure 7.13: Posterior predictive checks for inducer binding domain mutants where only  $K_A$  and  $K_I$  are changed.** (A) MCMC sampling output for each parameter. Joint distributions are colored by the value of the log posterior with increasing probability corresponding to transition from blue to yellow. (B) Percentiles of the data generated from the likelihood distribution for each sample of  $K_A$ ,  $K_I$ , and  $\sigma$ . Overlaid points are the experimentally observed measurements.

which only  $K_A$  and  $K_I$  are perturbed suggests that more than those two parameters should be changing. We see significant improvement in the description of the data when  $\Delta\varepsilon_{AI}$  is altered, indicating that it is the more appropriate hypothesis of the two.

## 7.5 Additional Characterization of Inducer Binding Domain Mutants

To predict the induction profiles of the inducer binding mutants, we used only the induction profile of each mutant paired with the native O2 lac operator to infer the parameters. Here, we examine the influence the choice of fit strain has on the predictions of the induction profiles and  $\Delta F$  for each mutant.

In Chapter 3, we dismissed the hypothesis that only  $K_A$  and  $K_I$  were changing due to the mutation and based the fit to a single induction profile. In Fig. 7.15, the fits and predictions for each mutant paired with each operator sequence queried. Here, the rows correspond to the operator sequence of the fit strain while the



**Figure 7.14: Posterior predictive checks for inducer binding domain mutants where all allosteric parameters can change.** (A) MCMC sampling output for all parameters. Joint distributions are colored by the value of the log posterior with increasing probability corresponding to the transition from blue to yellow. Marginal distributions are shown adjacent to each joint distribution. (B) Percentiles of the data generated from the likelihood for each sample of  $K_A$ ,  $K_I$ ,  $\Delta\varepsilon_{AI}$ , and  $\sigma$ . The corresponding experimental data for Q291K are shown as black open-faced circles.

columns correspond to the operator sequence of the predicted strain. The diagonals show the fit induction profiles and the corresponding data. Regardless of the choice of fit strain, the predicted induction profiles of the repressor paired with the O3 operator are poor, with the leakiness in each case being significantly underestimated. We also see that fitting to O3 results in poor predictions with incredibly wide credible regions for the other two operators. In Razo-Mejia et al. (2018), we also found that fitting  $K_A$  and  $K_I$  to the induction profile of O3 generally resulted in poor predictions of the other strains with comparably wide credible regions.

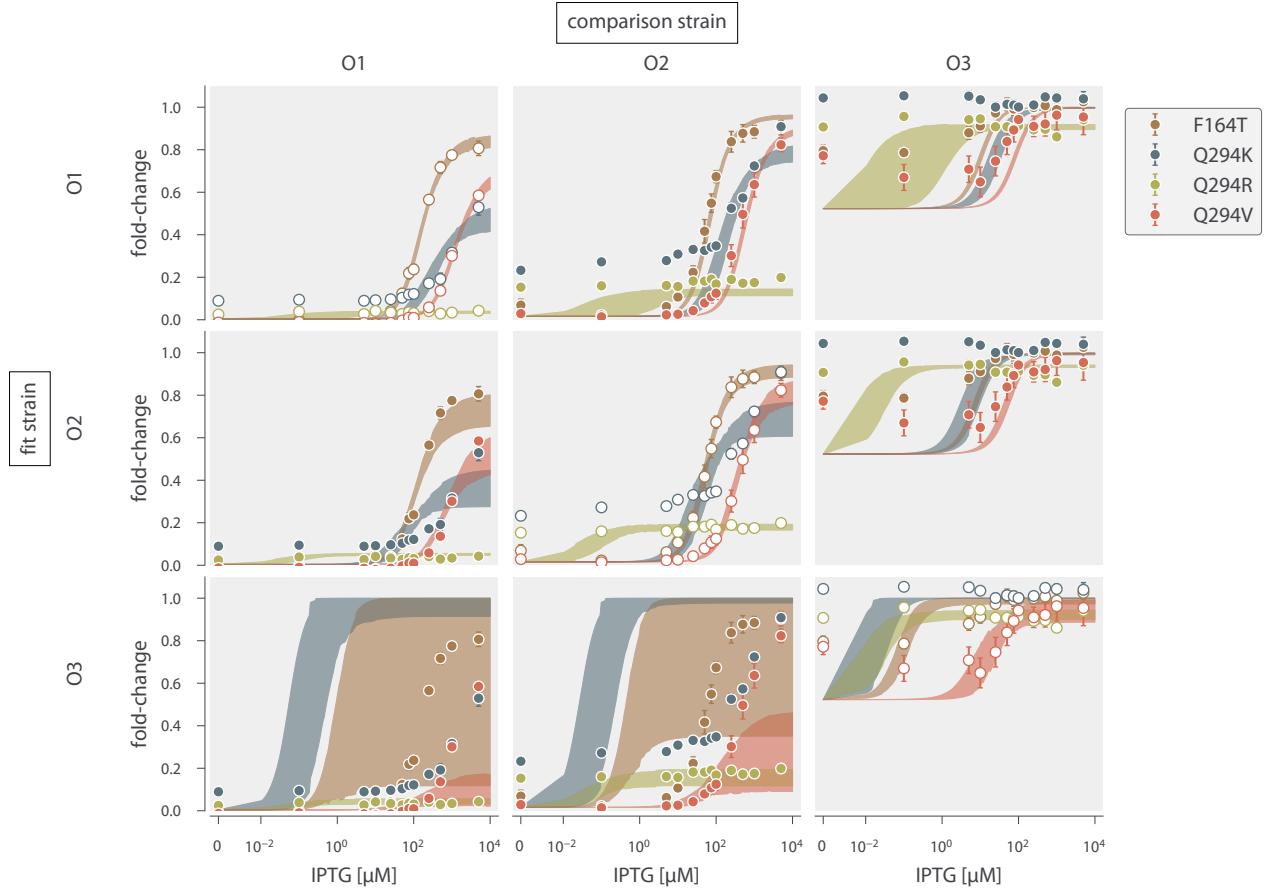
When  $\Delta\varepsilon_{AI}$  is included as a parameter, however, the predictive power is improved for all three operators, as can be seen in Fig. 7.16. While the credible regions are still wide when fit to the O3 operator, they are much narrower than under the first hypothesis. We emphasize that we are able to accurately predict the leakiness of nearly every strain by redetermining  $\Delta\varepsilon_{AI}$  whereas the leakiness was not pre-

dicted when only  $K_A$  and  $K_I$  were considered. Thus, we conclude that all three allosteric parameters  $K_A$ ,  $K_I$ , and  $\Delta\varepsilon_{AI}$  are modified for these four inducer binding domain mutations. The values of the inferred parameters are reported in Table 7.3.

We also examined the effect the choice of fit strain has on the predicted  $\Delta F$ , shown in Fig. 7.17. We find that the predictions agree with the data regardless of the choice of fit strain. One exception is the prediction of the Q291K  $\Delta F$  when the parameters fit to the O3 induction profile are used. As the induction profile for Q291K paired with O3 is effectively flat at a fold-change of 1, it is difficult to properly estimate the parameters of our sigmoidal function. We note all measurements of  $\Delta F$  for Q291K are described by using either the parameters fit to either O1 or O3 induction profiles, suggesting that the choice of fit strain makes little difference.

Table 7.2: Inferred values of  $K_A$ ,  $K_I$ , and  $\Delta\varepsilon_{AI}$  for inducer binding domain mutants. Values reported are the mean of the posterior distribution with the upper and lower bounds of the 95% credible region.

Mutant	Operator	$K_A$ [ $\mu\text{M}$ ]	$K_I$ [ $\mu\text{M}$ ]	$\Delta\varepsilon_{AI}$ [ $\text{k}_\text{B}\text{T}$ ]
F164T	O1	$290^{+60}_{-56}$	$1^{+4}_{-0.98}$	$4^{+5}_{-3}$
	O2	$165^{+90}_{-65}$	$3^{+6}_{-3}$	$1^{+5}_{-2}$
	O3	$110^{+700}_{-105}$	$7^{+5}_{-4}$	$-0.9^{+0.4}_{-0.3}$
Q291K	O1	$> 1000$	$410^{+150}_{-100}$	$-3.2^{+0.1}_{-0.1}$
	O2	$> 1000$	$310^{+70}_{-60}$	$-3.11^{+0.07}_{-0.07}$
	O3	$10^{+200}_{-10}$	$1^{+9}_{-1}$	$-7^{+3}_{-5}$
Q291R	O1	$3^{+27}_{-3}$	$2^{+20}_{-2}$	$-1.9^{+0.4}_{-0.3}$
	O2	$9^{+20}_{-9}$	$8^{+20}_{-8}$	$-2.32^{+0.01}_{-0.09}$
	O3	$6^{+24}_{-6}$	$9^{+30}_{-9}$	$-2.6^{+0.4}_{-0.5}$
Q291V	O1	$> 1000$	$3^{+13}_{-3}$	$6^{+4}_{-4}$
	O2	$650^{+450}_{-250}$	$8^{+8}_{-8}$	$3^{+6}_{-3}$
	O3	$100^{+400}_{-90}$	$22^{+33}_{-18}$	$0.1^{+0.8}_{-0.6}$

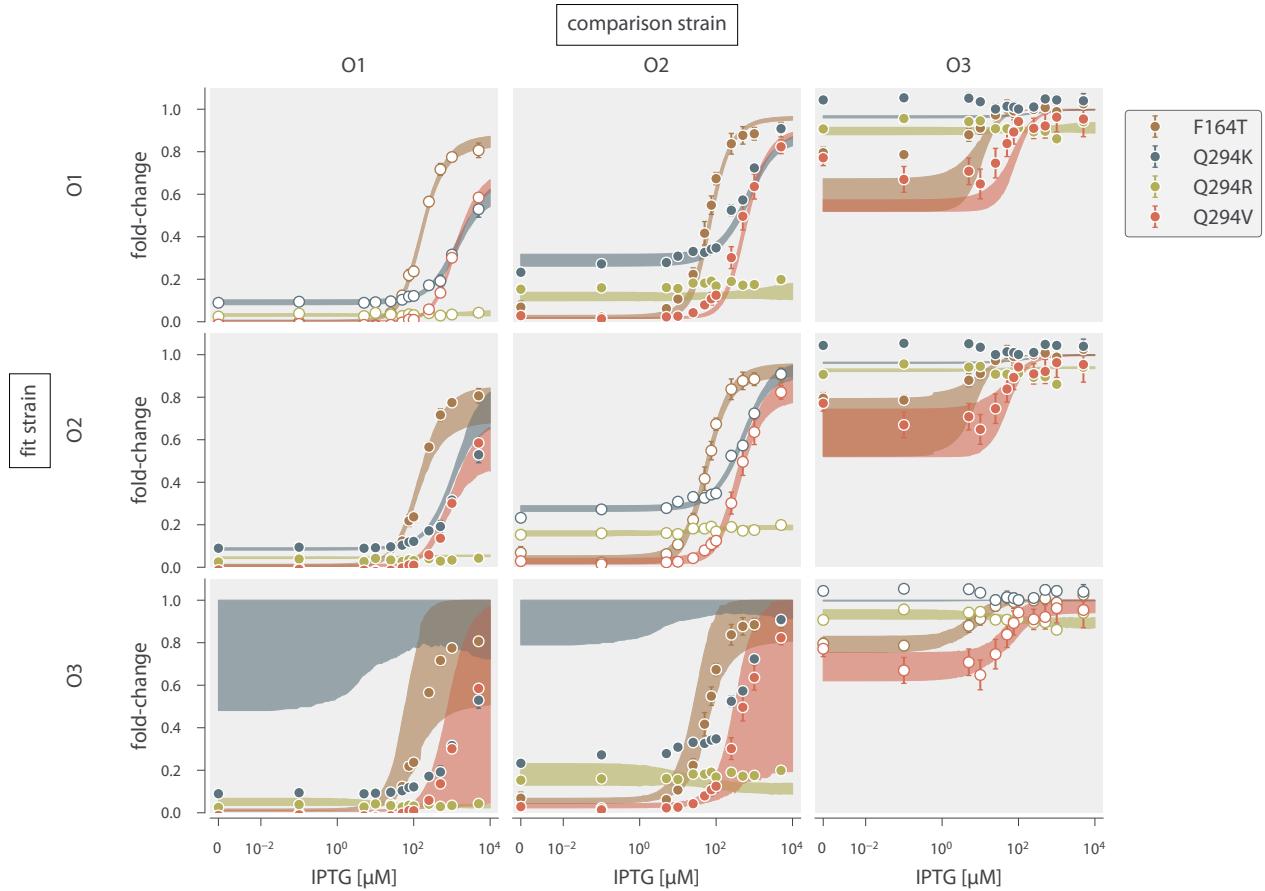


**Figure 7.15: Pairwise comparison of fit strain versus predictions assuming only  $K_A$  and  $K_I$  are influenced by the mutation.** Rows correspond to the operator sequence of the strain used for the parameter inference. Columns correspond to the operator sequence of the predicted strain. Colors identify the mutation. Diagonal positions (gray background) show the induction fit strain and profiles.

## 7.6 Comparing Parameter Values To The Literature

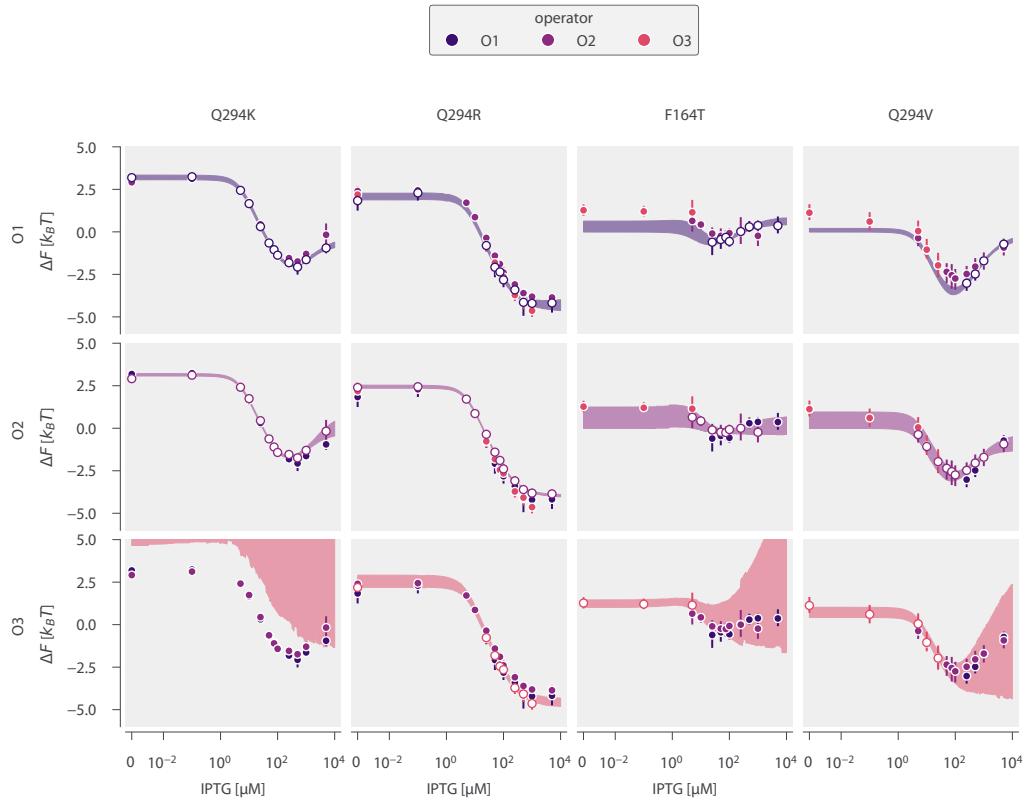
In this section, we compare and contrast the biophysical parameter values we use to characterize the wild-type Lac repressor with the rich literature that consists of *in vitro* and *in vivo* experiments. This section has an accompanying interactive figure available on the paper website which allows the reader to examine different combinations of parameter values and their agreement or disagreement with data taken from Garcia and Phillips (2011); Brewster et al. (2014); Razo-Mejia et al. (2018).

While the mutations used in this work and those in Daber et al. (2011) are



**Figure 7.16: Pairwise comparison of fit strain versus predictions assuming all allosteric parameters are affected by the mutation.** Rows correspond to the operator of the strain used to fit the parameters. Columns correspond to the operator of the strains whose induction profile is predicted. Mutants are identified by color. Diagonals (gray background) show the induction profiles of the strain to which the parameters were fit.

the same, we report significantly different values for the inducer binding, DNA binding parameters, and the relative energy difference between active and inactive states of the mutant repressors. The apparent disagreement of parameter values between the present work and those presented in Daber et al. (2011) in part stem from different treatments of the values for the wild-type Lac repressor. Since its isolation by Gilbert and Müller-Hill in the 1960's Gilbert and Müller-Hill (1966), the Lac repressor has been the subject of intense biochemical and structural study. Many measurements of the inducer and DNA binding kinetics of the repressor



**Figure 7.17: Comparison of choice of fit strain on predicted  $\Delta F$  profiles.** Rows correspond to the operator of the strain to which the parameters were fit. Columns correspond to mutations. Points are colored by their operator sequence. The data corresponding to the operator of the fit strain are shown as white-faced points.

*in vitro* (such as O’Gorman et al., 1980) and their values have informed the fitting of other parameters from measurements *in vivo* (such as Daber et al., 2011; ???). All of these measurements, however, do not *directly* measure the DNA- or inducer-binding kinetics nor the equilibrium constant between the active and inactive states of the repressor. To properly estimate the parameters, one must either have direct measurement of a subset of the parameters or make assumptions regarding the states of the system. Examples of the estimated allosteric parameter values of the wild-type LacI repressor from our previous work (Razo-Mejia et al., 2018), that of Daber et al. (2011), and *in vitro* measurements from O’Gorman et al. (1980) are given in Table 7.3. The theoretical predictions for the fold-change in gene expression, along with values reported in (???), can be seen using the interactive

figure on the paper website, where the reader can also enter their own parameter values and independently assess the agreement or lack thereof with the data.

It is notable that differences between the various references shown in Table 7.3 can be drastic, in some cases differing by almost an order of magnitude. Of particular note is the disagreement in the energy difference between the active and inactive states of the repressor,  $\Delta\epsilon_{AI}$ . Daber et al. (2011) determines a negative value of  $\Delta\epsilon_{AI}$  meaning that the inactive state of the repressor is energetically favorable to the active state. In stark contrast is the value reported in our previous work of  $+4.5 k_B T$ , implying that the active state is significantly more stable than the inactive state. The now seminal *in vitro* measurements reported in O’Gorman et al. (1980) suggest that the two states are nearly energetically equivalent.

The wide range of these reported values demonstrate that such thermodynamic models are highly degenerate, meaning that many combinations of parameter values can result in nearly equally good descriptions of the data. To illustrate this point, we estimated  $K_A$ ,  $K_I$ , and DNA binding energy  $\Delta\epsilon_{RA}$  for each operator to the data reported in Razo-Mejia et al. (2018); Garcia and Phillips (2011); Brewster et al. (2014), using the three values of  $\Delta\epsilon_{AI}$  shown in Table 7.4. The resulting fits can be seen in Fig. 7.18. Despite the drastically different values of  $\Delta\epsilon_{AI}$  it is possible to generate adequate fits by modulating the other parameters. The parameter values of these fits, reported in Table 7.3 further illustrate this point as they differ significantly from one another.

The theoretically predicted fold-change in the presence of multiple promoters, shown in Fig. 7.18 (E) is perhaps the most informative test of the parameter values. The theoretical advancements made in Weinert et al. (2014) provide a means to mathematically grasp the intricacies of plasmid-borne expression through calculation of the chemical potential. This formalism transforms the intimidating combinatorics of  $R$  repressors and  $N$  specific binding sites (i.e. plasmid reporter genes) into a two-state system where one can compute an *effective* repressor copy number regulating a single promoter. Using this formalism, the input-output function for

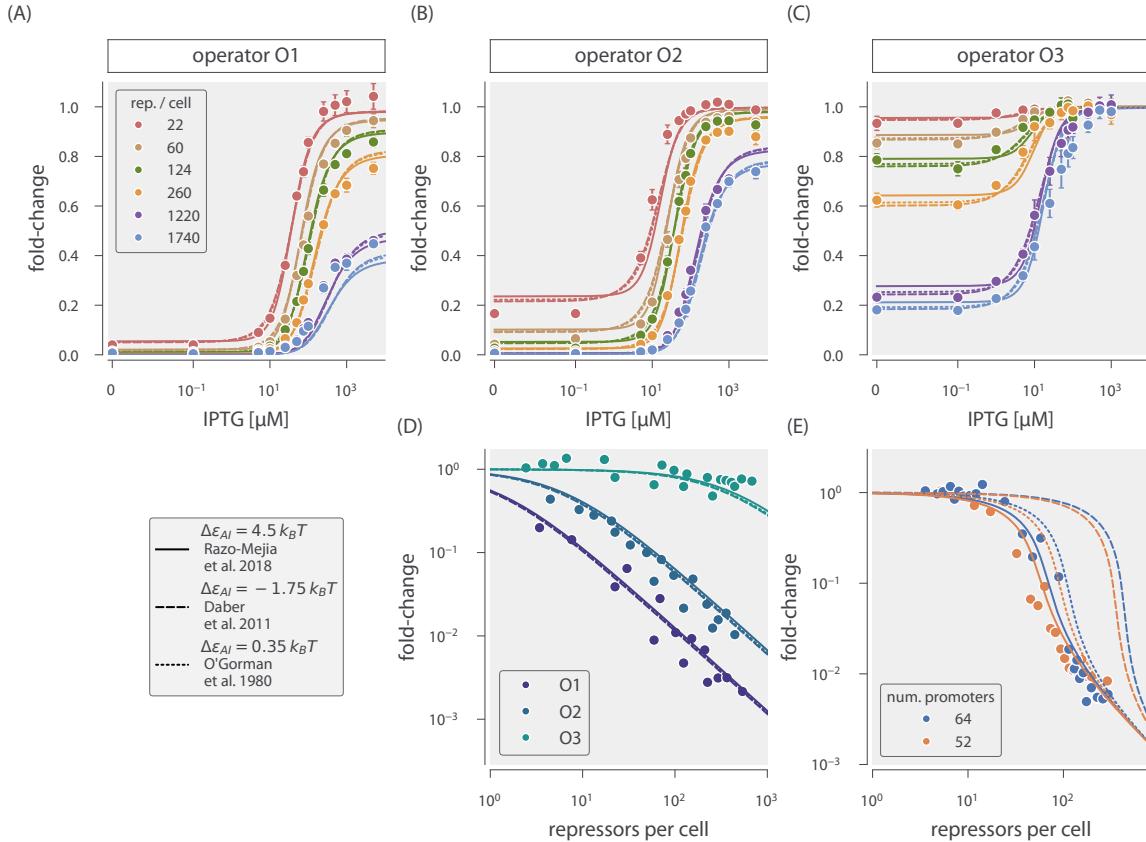
the fold-change in gene expression can be written as

$$\text{fold-change} = \frac{1}{1 + \lambda_R(c)e^{-\beta\Delta\varepsilon_{RA}}}, \quad (7.32)$$

where the effective repressor copy number (termed the fugacity) is denoted as  $\lambda_R(c)$ . In Chapter 6 of this thesis, we show that the inflection point of Eq. 7.32, whose curves shown in the bottom right-hand plot of Fig. 7.18, is located where the number of specific binding sites for the repressor is approximately equal to the number of repressors in the active state. Using this key feature, one can infer  $\Delta\varepsilon_{AI}$  given prior knowledge of  $\Delta\varepsilon_{RA}$  and the total number of repressors per cell. As shown in Fig. 7.18, using values of  $\Delta\varepsilon_{AI}$  from Razo-Mejia et al. (2018); O’Gorman et al. (1980) approximately agree with the measurements whereas the predicted curves using  $\Delta\varepsilon_{AI}$  from Daber et al. (2011) overestimates the fold-change, even though these values accurately describe the simple-repression data shown in the other panels of Fig. 7.18.

Without some direct *in vivo* measurements of these parameters, one must make assumptions about the system to make any quantitative progress. We chose to use the parameter values defined in our laboratory as the repressor fugacity provides us with an independent, albeit indirect, measurement of  $\Delta\varepsilon_{AI}$  which other works such as Daber et al. (2011) and O’Gorman et al. (1980) do not. Both of these works determine all of parameter values simultaneously by fitting to a single set of measurements. While these measurements accurately describe their data, their parameter values are less successful in accounting for data from Brewster et al. (2014); Garcia and Phillips (2011); Razo-Mejia et al. (2018). In the context of this work, we emphasize that we make many of the same qualitative conclusions as in Daber et al. (2011) with respect to the effects of the mutations using our particular set of parameter values.

While we use different values for  $\Delta\varepsilon_{AI}$ , the qualitative results between this work and that of Daber et al. (2011) are in agreement. For example, both works determine that mutations in the DNA binding domain alter only the DNA binding affinity whereas the mutations in the inducer binding pocket affect only the allosteric



**Figure 7.18: Degenerate fits of data from Razo-Mejia et al. (2018); Brewster et al. (2014); Garcia and Phillips (2011) using different values for active/inactive state energy difference  $\Delta\epsilon_{AI}$ .** In all plots, the solid, dashed, and dotted lines correspond to the best-fit curves conditioned on the data using parameter values for  $\Delta\epsilon_{AI}$  of  $4.5 k_B T$ ,  $-1.75 k_B T$ , and  $0.35 k_B T$ , respectively. Induction profiles from Razo-Mejia et al. (2018) for operators O1 (A), O2 (B), and O3 (C) are shown as points and errors which correspond to the mean and standard error of at least 10 biological replicates. (D) Leakiness measurements of the simple repression motif with one unique regulated reporter gene. Data shown are from Garcia and Phillips (2011); Brewster et al. (2014). (E) The transcription factor titration effect. For gene expression measurements on plasmids, the fold-change as a function of repressor copy number exhibits strong nonlinearities. We used this effect as a way to independently infer the parameter  $\Delta\epsilon_{AI}$  and it can be seen that this breaks the degeneracy between different parameters.

parameters. Because the biological variables such as repressor and reporter gene copy number are tightly controlled in our system, we are able to more precisely measure features of the induction profiles such as the leakiness in gene expression. This ability allows us to detect changes in the active/inactive equilibrium which were masked in Daber et al. (2011) by the experimental design. While the precise parameter values may be different between publications, the exploration of free energy differences resulting from mutations are parameter-value independent and any parameter disagreements do not change our theoretical model. Fig. 2 of the main text is presented with no knowledge of parameter values – it simply shows the mathematics of the model. While the value of  $\Delta F$  will ultimately depend on the parameter values, the formalism of this work remains agnostic to the parameter values and can be a useful tool for classifying mutations and couple the sequence-level variation to the systems-level response.

Table 7.3: Thermodynamic parameter values of wild-type LacI from the literature.

Parameter	Value	Reference
$\Delta\epsilon_{AI}$	$\geq 4.5 k_B T$	Razo-Mejia et al. (2018)
	$-1.7 k_B T$	Daber et al. (2011)
	$0.35 k_B T$	O'Gorman et al. (1980)
$K_A$	$139^{+22}_{-20} \mu\text{M}$	Razo-Mejia et al. (2018)
	$16 \mu\text{M}$	Daber et al. (2011)
	$133 \mu\text{M}$	O'Gorman et al. (1980)
$K_I$	$0.53^{+0.01}_{-0.01} \mu\text{M}$	Razo-Mejia et al. (2018)
	$2 \mu\text{M}$	Daber et al. (2011)
	$4 \mu\text{M}$	O'Gorman et al. (1980)

Table 7.4: Estimated parameters from global fits of data from literature.

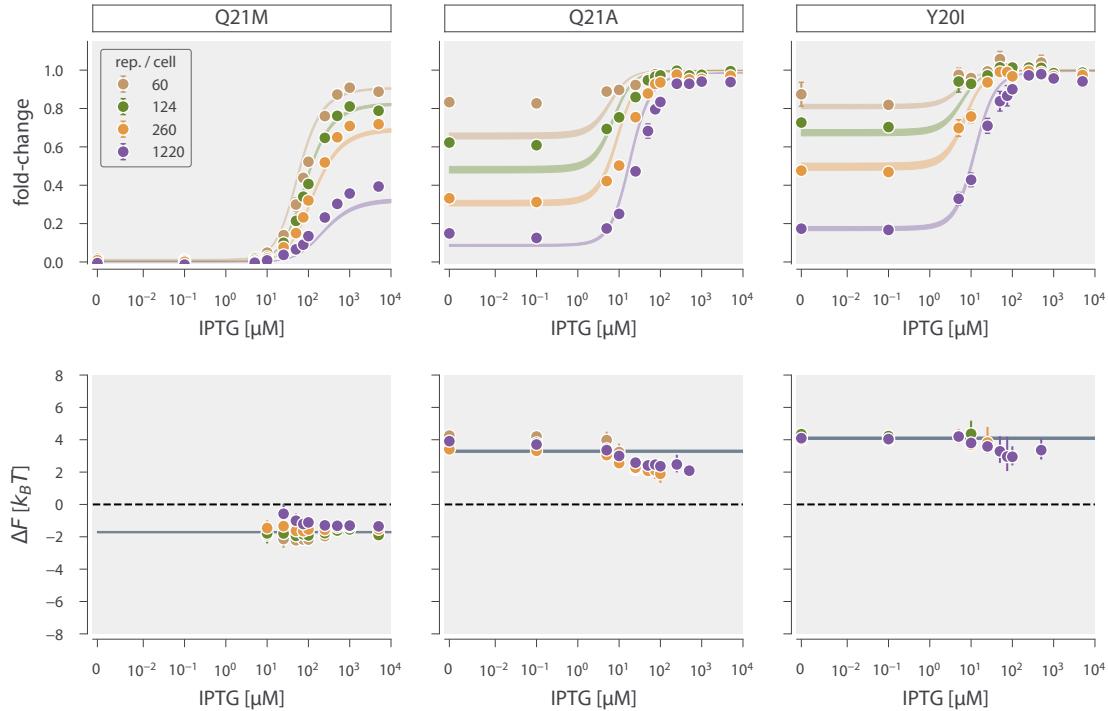
Parameter	Value	$\Delta\varepsilon_{AI}$ Reference Value
$\Delta\varepsilon_{RA}$ (O1 operator)	$-15.1^{+0.1}_{-0.1} k_B T$	$4.5 k_B T$ (Razo-Mejia et al., 2018)
	$-17.1^{+0.1}_{-0.1} k_B T$	$-1.75 k_B T$ (Daber et al., 2011)
	$-15.7^{+0.1}_{-0.1} k_B T$	$0.35 k_B T$ (O'Gorman et al., 1980)
$\Delta\varepsilon_{RA}$ (O1 operator)	$-15.1^{+0.1}_{-0.1} k_B T$	$4.5 k_B T$ (???)
	$-17.1^{+0.1}_{-0.1} k_B T$	$-1.75 k_B T$ (???)
	$-15.7^{+0.1}_{-0.1} k_B T$	$0.35 k_B T$ (???)
$\Delta\varepsilon_{RA}$ (O2 operator)	$-13.4^{+0.1}_{-0.1} k_B T$	$4.5 k_B T$ (Razo-Mejia et al., 2018)
	$-15.4^{+0.1}_{-0.1} k_B T$	$-1.75 k_B T$ (Daber et al., 2011)
	$-14.0^{+0.1}_{-0.1} k_B T$	$0.35 k_B T$ (O'Gorman et al., 1980)
$\Delta\varepsilon_{RA}$ (O3 operator)	$-9.21^{+0.06}_{-0.06} k_B T$	$4.5 k_B T$ (Razo-Mejia et al., 2018)
	$-11.29^{+0.06}_{-0.06} k_B T$	$-1.75 k_B T$ (Daber et al., 2011)
	$-9.85^{+0.06}_{-0.05} k_B T$	$0.35 k_B T$ (O'Gorman et al., 1980)
$K_A$	$225^{+10}_{-10} \mu\text{M}$	$4.5 k_B T$ (Razo-Mejia et al., 2018)
	$290^{+20}_{-20} \mu\text{M}$	$-1.75 k_B T$ (Daber et al., 2011)

Parameter	Value	$\Delta\epsilon_{AI}$ Reference Value
	$270^{+20}_{-20} \mu\text{M}$	$0.35 k_B T$ (O'Gorman et al., 1980)
$K_I$	$0.81^{+0.05}_{-0.05} \mu\text{M}$	$4.5 k_B T$ (Razo-Mejia et al., 2018)
	$8.2^{+0.5}_{-0.5} \mu\text{M}$	$-1.75 k_B T$ (Daber et al., 2011)
	$5.5^{+0.5}_{-0.3} \mu\text{M}$	$0.35 k_B T$ (O'Gorman et al., 1980)

## 7.7 Parameter Estimation Using All Induction Profiles

In Chapter 3 and Sec. 7.1 and 7.2 of this chapter, we have laid out our strategy for inferring the the various parameters of our model to a single induction profile and using the resulting values to predict the free energy and induction profiles of other strains. In this section, we estimate the parameters using all induction profiles of a single mutant and using the estimated values to predict the free energy profiles.

The inferred DNA binding energies considering induction profiles of all repressor copy numbers for the three DNA binding mutants are reported in Table 7.5. These parameters are close to those reported in Table 7.1 for each repressor copy number with Q18A showing the largest differences. The resulting induction profiles and predicted change in free energy for these mutants can be seen in Fig. 7.19. Overall, the induction profiles match the data to an appreciable agree. We acknowledge that even when using *all* repressor copy numbers, the fit to Q18A remains imperfect. However we contend that this disagreement is comparable to that observed in Razo-Mejia et al. (2018) which described the induction profile of the wild-type repressor. We find that the predicted change in free energy bottom row in 7.19 narrows compared to that in Fig. ?? and Fig. 3.3 of Chapter 3, confirming that considering all induction profiles improves our inference of the most-likely DNA binding energy. There appears to be a very slight trend in the



**Figure 7.19: Induction profiles and predicted change in free energy using parameters estimated from the complete data sets.** Top row shows fold-change measurements (points) as mean and standard error with ten to fifteen biological replicates. Shaded lines correspond to the 95% credible regions of the induction profiles using the estimated values of the DNA binding energies reported in Table 7.5. Bottom row shows the 95% credible regions of the predicted change in free energy (shaded lines) along with the inferred free energy of data shown in the top row. In all plots, the inducer concentration is shown on a symmetric log scale with linear scaling between 0 and 10<sup>-2</sup> μM and log scaling elsewhere.

$\Delta F$  for Q18A at higher inducer concentrations, though the overall change in free energy from 0 to 5000 μM IPTG is small.

**Table 7.5: Estimated DNA binding energies for each DNA binding domain mutant using all repressor copy numbers**

Mutant	$\Delta \varepsilon_{RA}$ [k <sub>B</sub> T]
Y17I	$-9.81^{+0.04}_{-0.08}$
Q18A	$-10.60^{+0.07}_{-0.07}$

Mutant	$\Delta\varepsilon_{RA}$ [ $k_B T$ ]
Q18M	$-15.61^{+0.05}_{-0.05}$

We also estimated the allosteric parameters ( $K_A$ ,  $K_I$ , and  $\Delta\varepsilon_{AI}$ ) for all inducer binding domain mutations using the induction profiles of all three operator sequences. The values, reported in Table 7.6 are very similar to those estimated from a single induction profile (Table 7.3). We note that for Q291R, it is difficult to properly estimate the values for  $K_A$  and  $K_I$  as the observed induction profile is approximately flat. The induction profiles and predicted change in free energy for each inducer binding mutant is shown in Fig. 7.20. We see notable improvement in the agreement between the induction profiles and the observed data, indicating that considering all data significantly shrinks the uncertainty of each parameter. The predicted change in free energy is also improved compared to that shown in Fig. 7.17. We emphasize that the observed free energy difference for each point assumes no knowledge of the underlying parameters and comes directly from measurements. The remarkable agreement between the predicted free energy and the observations illustrates that redetermining the allosteric parameters is sufficient to describe how the free energy changes as a result of the mutation.

Table 7.6: Estimated values for  $K_A$ ,  $K_I$ , and  $\Delta\varepsilon_{AI}$  for inducer binding domain mutations using induction profiles of all operator sequences.

Mutant	$K_A$ [ $\mu\text{M}$ ]	$K_I$ [ $\mu\text{M}$ ]	$\Delta\varepsilon_{AI}$ [ $k_B T$ ]
F161T	$300^{+60}_{-60}$	$12.7^{+0.1}_{-0.1}$	$-0.9^{+0.3}_{-0.3}$
Q291K	$> 1 \text{ mM}$	$330^{+60}_{-70}$	$-3.17^{+0.07}_{-0.07}$
Q291R	$> 1 \text{ mM}$	$> 1 \text{ mM}$	$-2.4^{+0.2}_{-0.2}$
Q291V	$> 1 \text{ mM}$	$53^{+17}_{-13}$	$0^{+0.3}_{-0.3}$

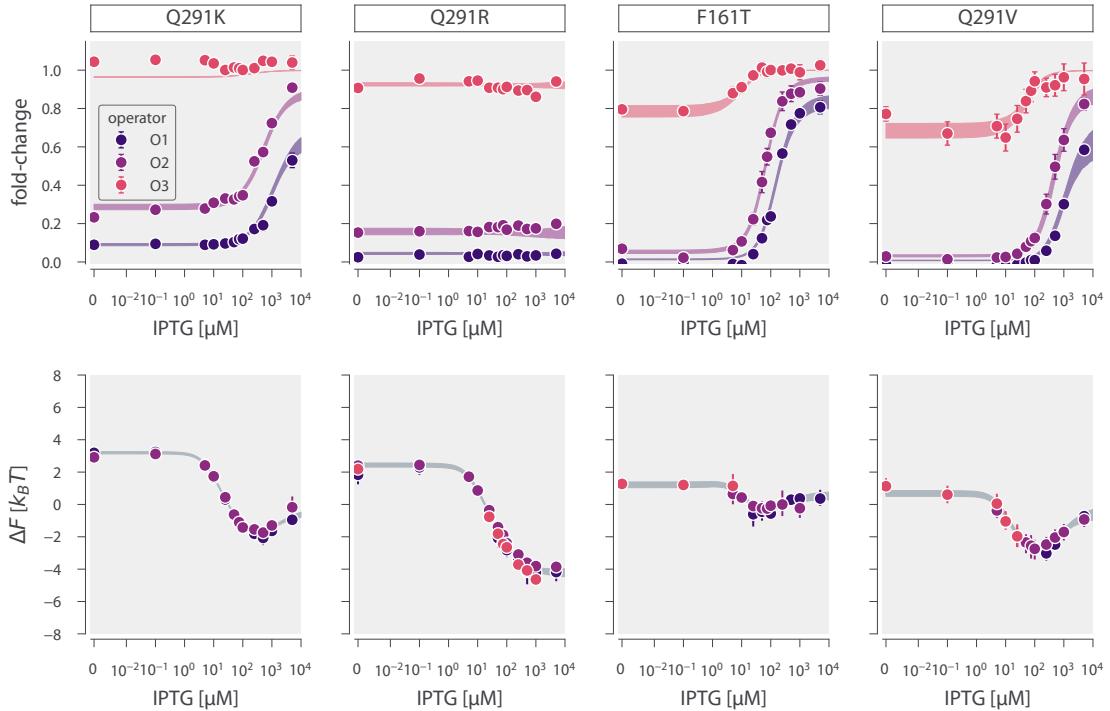


Figure 7.20: \*\*Induction profiles and predicted change in free energy using parameters estimated from the complete data sets for inducer binding domain mutants. Top row shows fold-change measurements (points) as mean and standard error with ten to fifteen biological replicates. Shaded lines correspond to the 95% credible regions of the induction profiles using the estimated values of the allosteric parameters reported in Table 7.6. Bottom row shows the 95% credible regions of the predicted change in free energy (shaded lines) along with the inferred free energy of data shown in the top row. In all plots, the inducer concentration is shown on a symmetric log scale with linear scaling between  $0$  and  $10^{-2} \mu\text{M}$  and log scaling elsewhere.

## 7.8 Generalizability of Data Collapse To Other Regulatory Architectures

In Chapters 2, 3, and 4, we have stated that the input-output function for the fold-change in gene expression can be rewritten in the form of a Fermi function. However, this result can be derived directly by coarse graining the transcription factor bound and unbound states of the systems' partition function. In this section, we show how coarse graining of promoter occupancy states results in a general Fermi function so long as the transcription factor bound states and transcriptionally active states do not overlap.

As shown in Chapter 2, the partition function  $\mathcal{Z}$  for the simple repression motif (ignoring allosteric control) can be enumerated as

$$\mathcal{Z} = \underbrace{1 + \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_P}}_{\text{repressor unbound states}} + \underbrace{\frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_R}}_{\text{repressor bound state}} = \mathcal{Z}_{-r} + \mathcal{Z}_r,$$

{eq:z\_simple\_rep} where  $R$  is the number of repressors,  $P$  is the number of polymerases,  $N_{NS}$  is the number of nonspecific binding sites and  $\Delta \varepsilon_P$  and  $\Delta \varepsilon_R$  are the binding energies of the polymerase and repressor to the DNA, respectively. The states can be grouped into either repressor bound states (right-hand terms) or repressor unbound states (left hand terms), denoted as  $\mathcal{Z}_r$  and  $\mathcal{Z}_{-r}$ , respectively. The probability of the repressor *not* being bound to the promoter can be computed as

$$P(\neg r) = \frac{\mathcal{Z}_{-r}}{\mathcal{Z}_{-r} + \mathcal{Z}_r}. \quad (7.33)$$

In this coarse-grained description, the transcriptionally active states are separate from the repressor bound states. Thus, to calculate the fold-change in gene expression, we can compute the ratio of  $P(\neg r | R > 0)$  when repressor is present to  $P(\neg r | R = 0)$  when repressor is absent. As the latter term is equal to 1, the fold-change in gene expression is equivalent to Eq. 7.33. We can compute an *effective* free energy  $\tilde{F}$  of each coarse-grained state as

$$\tilde{F}_{-r} = -k_B T \log \mathcal{Z}_{-r}; \quad \tilde{F}_r = -k_B T \log \mathcal{Z}_r. \quad (7.34)$$

With an effective energy in hand, we can write Eq. 7.33 (which is equivalent to the fold-change) in terms of the Boltzmann weights of these two coarse-grained states as

$$\text{fold-change} = \frac{e^{-\beta \tilde{F}_{-r}}}{e^{-\beta \tilde{F}_{-r}} + e^{-\beta \tilde{F}_r}}. \quad (7.35)$$

A simple rearrangement of this result produces a Fermi function

$$\text{fold-change} = \frac{1}{1 + e^{-\beta \tilde{F}}}, \quad (7.36)$$

where  $\tilde{F}$  is the effective free energy of the system defined as

$$\tilde{F} = -k_B T (\log \mathcal{Z}_r - \log \mathcal{Z}_{-r}). \quad (7.37)$$

In the case of the simple repression motif, we apply the weak promoter approximation which is based on the fact that  $(P/N_{NS})e^{-\beta\Delta\varepsilon_P} \ll 1$ , reducing  $\mathcal{Z}_{\neg r} = 1$ . With this approximation, the effective free energy  $\tilde{F}$  defined in Eq. 7.37 can be interpreted as the free energy between the repressed and transcriptionally active states. In general, this coarse-graining approach can be applied to any regulatory architecture in which the transcription factor bound states and transcriptionally active states share no overlapping substates. Fig. 7.21 shows various repression based architectures along with the coarse graining of states and the effective free energy.

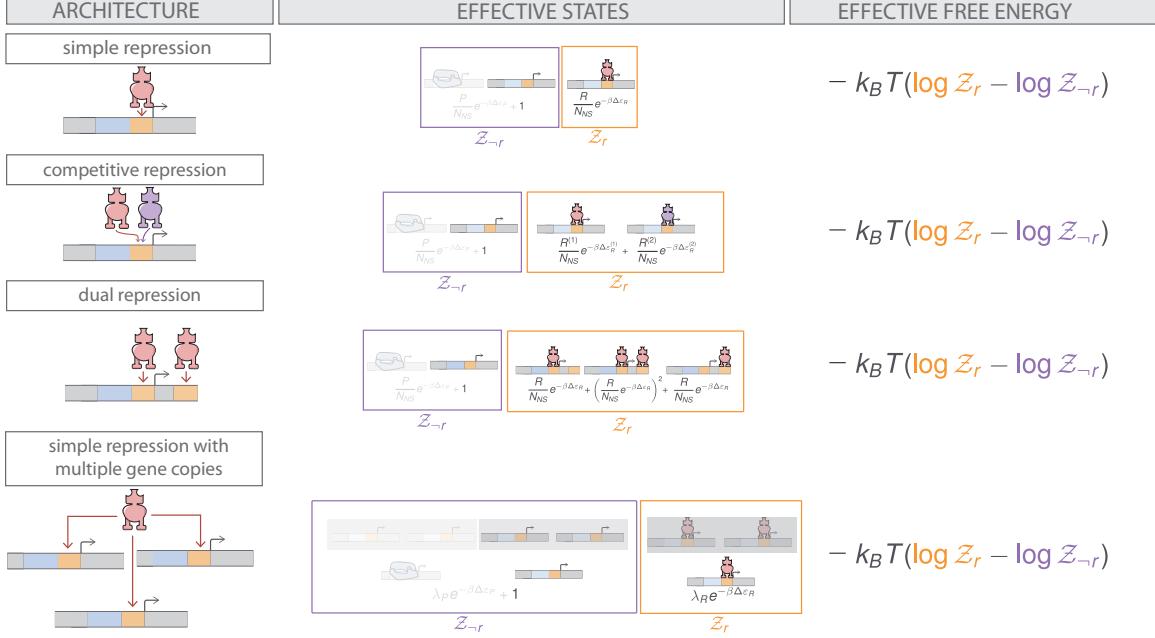
This approach cannot be applied to architectures in which the transcription factor bound and transcriptionally active states cannot be separated. One such example is the case of simple activation in which the binding of an activator increases gene expression. For this architecture, the total partition function (as derived in Bintu et al., 2005a) is

$$\mathcal{Z} = \underbrace{1 + \frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_P}}_{\text{activator unbound states}} + \underbrace{\frac{A}{N_{NS}}e^{-\beta\Delta\varepsilon_A} + \frac{A}{N_{NS}}\frac{P}{N_{NS}}e^{-\beta(\Delta\varepsilon_A + \Delta\varepsilon_P + \varepsilon_{\text{interaction}})}}_{\text{activator bound states}} = \mathcal{Z}_{\neg a} + \mathcal{Z}_a. \quad (7.38)$$

Here we've used  $A$  to denote the number of activators,  $\Delta\varepsilon_A$  is the binding energy of the activator to its specific binding site, and  $\varepsilon_{\text{interaction}}$  is the interaction energy between the activator and polymerase which makes the activator-polymerase-DNA state more energetically favorable to the polymerase-DNA state. Unlike in the case of repression, the transcriptionally active states are present in the  $\mathcal{Z}_a$ . While we can compute the probability of the activator not being bound  $P(\neg a)$ , this is not equivalent to the fold-change in gene expression as was the case for simple repression. Rather, the fold-change in gene expression can be written as

$$\text{fold-change} = \frac{\mathcal{Z}_a - 1 + \mathcal{Z}_{\neg a} - \frac{A}{N_{NS}}e^{-\beta\Delta\varepsilon_A}}{\mathcal{Z}_a + \mathcal{Z}_{\neg a}} \frac{\mathcal{Z}_a}{\mathcal{Z}_a - 1}. \quad (7.39)$$

This is a more cumbersome expression than in the case for simple repression, and cannot be massaged into a one-parameter description of the fold-change. Thus,



**Figure 7.21: Various repression-based regulatory architectures and their coarse-grained states.** The left column shows a schematic of the regulatory architecture. The middle column shows the equilibrium states of the system coarse grained into repressor bound (red boxes) and repressor unbound (blue boxes) states. States in which the polymerase is bound is shaded in white and are neglected by the weak promoter approximation. The right column illustrates that the formulation of the effective free energy is the same for all architectures, although the formulation of  $\mathcal{Z}_{-r}$  and  $\mathcal{Z}_r$  are different between the examples. The bottom row illustrates the coarse graining of a simple repression motif in which the same repressor regulates multiple copies of the same gene. Rather than considering all states, the gene expression can calculated by considering one promoter and an effective copy number, described by the repressor fugacity  $\lambda_r$  and is derived in Weinert et al. (2014). Grayed-out states illustrate this further level of coarse graining.

the analysis presented here is only applicable to cases in which the occupancy of the promoter by either the polymerase or the transcription factor are mutually exclusive.

## 7.9 Strain and Oligonucleotide Information

Table 7.7: *Escherichia coli* strains used in this work

Class	LacI Mutant	Operator	Rep. per	Genotype
			Cell	
-	-	-	22	MG1655:: $\Delta lacZYA$
-	-	O1	0	MG1655:: $\Delta lacIZYA$ ; $galK<>25O1+11-$ YFP
-	-	O2	0	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11-$ YFP
-	-	O3	0	MG1655:: $\Delta lacIZYA$ ; $galK<>25O3+11-$ YFP
WT	WT	O2	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11-$ YFP; $ybcN<>3^1$ - RBS1027LacI
DNA	Y17I	O2	60	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11-$ YFP; $ybcN<>3^1$ - RBS1147LacI(Y17I)
DNA	Y17I	O2	124	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11-$ YFP; $ybcN<>3^1$ - RBS446LacI(Y17I)

<b>Class</b>	<b>LacI Mutant</b>	<b>Operator</b>	<b>Cell</b>	<b>Rep. per</b>
				<b>Genotype</b>
DNA	Y17I	O2	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11-$ YFP; $ybcN<>3^1-$ RBS1027LacI(Y17I)
DNA	Y17I	O2	1220	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11-$ YFP; $ybcN<>3^1-$ RBS1LacI(Y17I)
DNA	Q18A	O2	60	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11-$ YFP; $ybcN<>3^1-$ RBS1147LacI(Q18A)
DNA	Q18A	O2	124	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11-$ YFP; $ybcN<>3^1-$ RBS446LacI(Q18A)
DNA	Q18A	O2	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11-$ YFP; $ybcN<>3^1-$ RBS1027LacI(Q18A)

<b>Class</b>	<b>LacI Mutant</b>	<b>Operator</b>	<b>Cell</b>	<b>Rep. per</b>
				<b>Genotype</b>
DNA	Q18A	O2	1220	MG1655:: <i>ΔlacIZYA</i> ; <i>galK</i> <>25O2+11- YFP; <i>ybcN</i> <>3*1- RBS1LacI(Q18A)
DNA	Q18M	O2	60	MG1655:: <i>ΔlacIZYA</i> ; <i>galK</i> <>25O2+11- YFP; <i>ybcN</i> <>3*1- RBS1147LacI(Q18M)
DNA	Q18M	O2	124	MG1655:: <i>ΔlacIZYA</i> ; <i>galK</i> <>25O2+11- YFP; <i>ybcN</i> <>3*1- RBS446LacI(Q18M)
DNA	Q18M	O2	260	MG1655:: <i>ΔlacIZYA</i> ; <i>galK</i> <>25O2+11- YFP; <i>ybcN</i> <>3*1- RBS1027LacI(Q18M)
DNA	Q18M	O2	1220	MG1655:: <i>ΔlacIZYA</i> ; <i>galK</i> <>25O2+11- YFP; <i>ybcN</i> <>3*1- RBS1LacI(Q18M)

<b>Class</b>	<b>LacI Mutant</b>	<b>Operator</b>	<b>Cell</b>	<b>Rep. per</b>
				<b>Genotype</b>
IND	F161T	O1	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O1+11-$ YFP; $ybcN<>3^1-$ RBS1027LacI(F161T)
IND	F161T	O2	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11-$ YFP; $ybcN<>3^1-$ RBS1027LacI(F161T)
IND	F161T	O3	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O3+11-$ YFP; $ybcN<>3^1-$ RBS1027LacI(F161T)
IND	Q291V	O1	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O1+11-$ YFP; $ybcN<>3^1-$ RBS1027LacI(Q291V)
IND	Q291V	O2	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11-$ YFP; $ybcN<>3^1-$ RBS1027LacI(Q291V)

<b>Class</b>	<b>LacI Mutant</b>	<b>Operator</b>	<b>Cell</b>	<b>Rep. per</b>
				<b>Genotype</b>
IND	Q291V	O3	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O3+11-$ YFP; $ybcN<>3^1-$ RBS1027LacI(Q291V)
IND	Q291K	O1	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O1+11-$ YFP; $ybcN<>3^1-$ RBS1027LacI(Q291K)
IND	Q291K	O2	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11-$ YFP; $ybcN<>3^1-$ RBS1027LacI(Q291K)
IND	Q291K	O3	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O3+11-$ YFP; $ybcN<>3^1-$ RBS1027LacI(Q291K)
IND	Q291R	O1	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O1+11-$ YFP; $ybcN<>3^1-$ RBS1027LacI(Q291R)

<b>Class</b>	<b>LacI Mutant</b>	<b>Operator</b>	<b>Cell</b>	<b>Rep. per</b>
				<b>Genotype</b>
IND	Q291R	O2	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11-$ YFP; $ybcN<>3^1-$ RBS1027LacI(Q291R)
IND	Q291R	O3	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O3+11-$ YFP; $ybcN<>3^1-$ RBS1027LacI(Q291R)
DBL	Y17I-F161T	O2	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11-$ YFP; $ybcN<>3^1-$ RBS1027LacI(Y17IF161T)
DBL	Y17I-Q291V	O2	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11-$ YFP; $ybcN<>3^1-$ RBS1027LacI(Y17IQ291V)
DBL	Y17I-Q291K	O2	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11-$ YFP; $ybcN<>3^1-$ RBS1027LacI(Y17IQ291K)

<b>Class</b>	<b>LacI Mutant</b>	<b>Operator</b>	<b>Cell</b>	<b>Rep. per</b>
				<b>Genotype</b>
DBL	Q18A-F161T	O2	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11$ -YFP; $ybcN<>3^1$ -RBS1027LacI(Q18AF161T)
DBL	Q18A-Q291V	O2	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11$ -YFP; $ybcN<>3^1$ -RBS1027LacI(Q18AQ291V)
DBL	Q18A-Q291K	O2	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11$ -YFP; $ybcN<>3^1$ -RBS1027LacI(Q18AQ291K)
DBL	Q18M-F161T	O2	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11$ -YFP; $ybcN<>3^1$ -RBS1027LacI(Q18MF161T)
DBL	Q18M-Q291V	O2	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11$ -YFP; $ybcN<>3^1$ -RBS1027LacI(Q18MQ291V)

Class	LacI Mutant	Operator	Cell	Rep. per
				Genotype
DBL	Q18M-Q291K	O2	260	MG1655:: $\Delta lacIZYA$ ; $galK<>25O2+11$ -YFP; $ybcN<>3^*1$ -RBS1027LacI(Q18MQ291K)

Table 7.8: Oligonucleotides used for mutant generation.

Primer Name	Sequence (5' →			
	3')	Description	Method	
10.1	acctctgcg	Integration into	$\lambda$ -Red	
	gaggggaag	$ybcN$ locus	Recombineering	
	cgtgaacctc			
	tcacaagacg			
	gcatcaaatt			
	acactagca			
	acaccagaac agc			
10.3	ctgttagatgt	Integration into	$\lambda$ -Red	
	gtccgttca	$ybcN$ locus	Recombineering	
	tgacacgaat			
	aagcggtgta			
	gccattacg			
	ccggctaatt			
	gcacccagt aagg			
GCMWC-001	ccggcatac	Amplification	QuickChange	
	tctgcgaca	of plasmid	Mutagenesis	

<b>Primer Name</b>	<b>Sequence (5' → 3')</b>		
		<b>Description</b>	<b>Method</b>
GCMWC-002	gtgtcttta	Q18M Mutation	QuickChange
	tATGaccgt	(CAG→ATG)	Mutagenesis
	ttcccg		
GCMWC-003	tgtctttat	Q18A Mutation	QuickChange
	GCGaccgttt	(CAG→GCG)	Mutagenesis
	cccg		
GCMWC-004	gttaacggcg	Amplification	QuickChange
	ggtatataac	of plasmid	Mutagenesis
GCMWC-005	caccatcaaa	Q291V	QuickChange
	GTGgatttt	Mutation (CAG	Mutagenesis
	cgcctgc	→ GTG)	
GCMWC-006	caccatcaaa	Q291K	QuickChange
	AAGgattttcg	Mutation	Mutagenesis
	cc	(CAG→AAG)	
GCMWC-007	cagtattatt	F161T Mutation	QuickChange
	ACCtccccatga	(TTC→ACC)	Mutagenesis
	agacgg		
GCMWC-008	ttgatgggtg	Amplification	QuickChange
	tctggtcag	of plasmid	Mutagenesis
GCMWC-009	gcataactctg	Amplification	QuickChange
	cgacatcgta	taa of plasmid	Mutagenesis
	tttc		
GCMWC-010	cggtgtctct	Y17I Mutation	QuickChange
	ATTcagaccg	(TAT→ATT)	Mutagenesis
GCMWC-017	ccatcaaaAG	Q291R	Gibson
	Ggattttcgc	Mutation	Assembly
	ctgctggg	(CAG→AGG)	
	gcaaaccag		

<b>Primer Name</b>	<b>Sequence (5' →</b>		
	<b>3')</b>	<b>Description</b>	<b>Method</b>
GCMWC-018	ggcgaaaatc	Q291R	Gibson
	CCTtttgatg	Mutation	Assembly
	gtggtaa	(CTG→CCT)	
	cggcggg		

*Chapter 8*

## SUPPLEMENTAL INFORMATION FOR CHAPTER IV: THE PHYSIOLOGICAL ADAPTABILITY OF A SIMPLE GENETIC CIRCUIT

A version of this chapter originally appeared as Chure, G; Kaczmarek, Z.A.; and Phillips, R (2019). "Physiological Adaptability and Parametric Versatility In A Simple Genetic Circuit. Currently in revision and available on the bioRxiv. G.C. and Z.A.K. designed experiments and collected/analyzed data. G.C. developed theoretical models and statistical inference techniques. G.C. and R.P. designed the research and wrote the paper.

### **8.1 Non-parametric Inference of Growth Rates**

In this section, we discuss the measurement of the bacterial growth curves as well as our strategy for estimating the growth rates for all experimental conditions in this work.

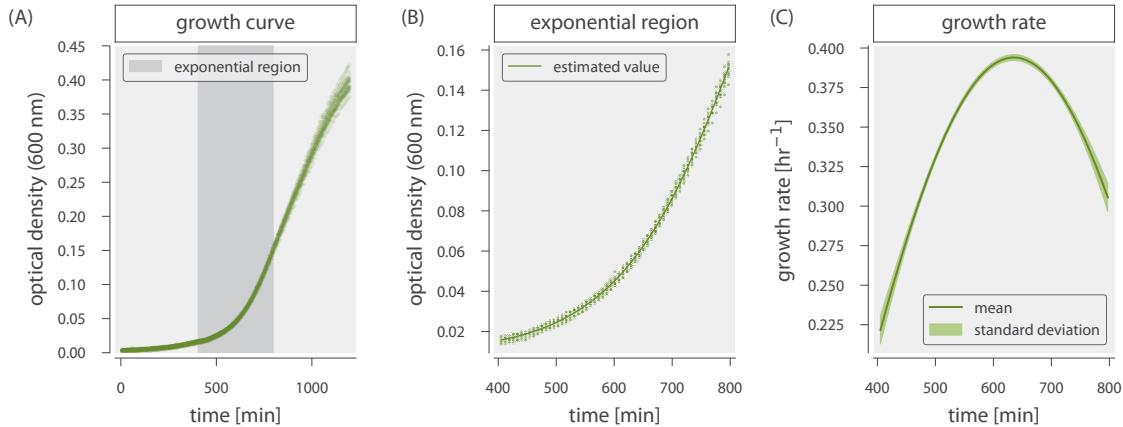
#### **Experimental growth curves**

As is described in the Materials and Methods section of Chapter 4, we measured the growth of *E. coli* strains constitutively expressing YFP using a BioTek Cytation 5 96-well plate reader generously provided by the lab of Prof. David Van Valen at Caltech. Briefly, cells were grown overnight in a nutrient rich LB medium to saturation and were subsequently diluted 1000 fold into the appropriate growth medium. Once these diluted cultures reached an OD<sub>600nm</sub> of  $\approx 0.3$ , the cells were again diluted 100 fold into fresh medium preheated to the appropriate temperature . Aliquots of 300  $\mu$ L of this dilution were then transferred to a 96-well plate leaving two-rows on all sides of the plate filled with sterile media to serve as blank measurements. Once prepared, the plate was transferred to the plate reader and OD<sub>600nm</sub> measurements were measured every  $\approx 5 - 10$  min for 12 to 18 hours. Be-

tween measurements, the plates were agitated with a linear shaking mode to avoid sedimentation of the culture. A series of technical replicates of the growth curve in glycerol supplemented medium at 37° C is shown in Fig. 8.1 (A).

### Inference of maximum growth rate

The phenomenon of collective bacterial growth has been the subject of intense research for the better part of a century (Jun et al., 2018; Schaechter et al., 1958) yielding many parametric descriptions of the bulk growth rates each with varying degrees of detail (Allen and Waclaw, 2018; Jun et al., 2018). For this scope of this work, we are not particularly interested in estimating numerous parameters that describe the phenomenology of the growth curves. Rather, we are interested in knowing a single quantity – the maximum growth rate – and the degree to which it is tuned across the different experimental conditions. To avoid forcing the bacterial growth curves into a parametric form, we treated the observed growth curves using Gaussian process modeling using as implemented in the FitDeriv software described in Ref. (Swain et al., 2016). This method permits an estimation of the most-likely  $OD_{600nm}$  value at each point in time given knowledge of the adjacent measurements. The weighting given to all points in the series of measurements is defined by the covariance kernel function and we direct the reader to Ref. (Swain et al., 2016) for a more detailed discussion on this kernel choice and overall implementation of Gaussian process modeling of time-series data. As this approach estimates the probability of a given  $OD_{600nm}$  at each time point, we can compute the mean and standard deviation. Fig. 8.1 (B) show the raw measurements and the mean estimated value in green and light, respectively. With the high temporal resolution of the  $OD_{600nm}$  measurements, modeling the entire growth curve becomes a computationally intensive task. Furthermore, as we are interested in only the maximum growth rate, there is no benefit in analyzing the latter portion of the experiment where growth slows and the population reaches saturation. We therefore manually restricted each analyzed growth curve to a region capturing early and mid exponential phase growth, illustrated by the shaded region in Fig. 8.1 (A).



**Figure 8.1: Non-parametric characterization of bacterial growth curves and estimation of the growth rate.** (A) Representative biological replicate growth curve of a  $\Delta lacIZYA$  *E. coli* strain in glycerol supplemented M9 minimal medium at 37° C. Points represent individual optical density measurements across eight technical replicates. The shaded region illustrates the time window from which the maximum growth rate was inferred. (B) Green points are those from the shaded region in (A). Line is the estimated value of the optical density resulting from the gaussian process modeling. (C) The value of the first derivative of the optical density as a function of time estimated via gaussian process modeling. The first derivative is taken as the growth rate at each time point. The solid line is the estimated mean value and the shaded region represents the standard deviation of the posterior distribution.

With a smooth description of the  $OD_{600nm}$  measurements as a function of time with an appropriate measure of uncertainty, we can easily compute the time derivative which is equivalent to the growth rate. A representative time derivative is shown in Fig. 8.1 (C). Here, the dark green curve is the mean value of the time derivative and the shaded region is  $\pm$  one standard deviation. The maximum value of this inferred derivative is the reported maximum growth rate of that experimental condition.

## 8.2 Approximating Cell Volume

In Fig. 4.2 of Chapter 4, we make reference to the volume of the cells grown in various conditions. Here, we illustrate how we approximated this estimate using measurements of the individual cell segmentation masks.

Estimation of bacterial cell volume and its dependence on the total growth rate has been the target of numerous quantitative studies using a variety of methods including microscopy (Pilizota and Shaevitz, 2012, 2014; Schaechter et al., 1958; Schmidt et al., 2016; Taheri-Araghi et al., 2015a) and microfluidics (???) revealing fascinating phenomenology of growth at the single cell level . Despite the high precision and extensive calibration of these methods, it is not uncommon to have different methods yield different estimates, indicating that it is not a trivial measurement to make. In the present work, we sought to estimate the cell volume and compare it to the well-established empirical results of the field of bacterial physiology to ensure that our experimental protocol does not alter the physiology beyond expectations. As the bulk of this work is performed using single-cell microscopy, we chose to infer the approximate the cell volume from the segmentation masks produced by the SuperSegger MATLAB software (???) which reported the cell length and width in units of pixels which can be converted to meaningful units given knowledge of the camera interpixel distance.

We approximated each segmented cell as a cylinder of length  $a$  and radius  $r$  capped on each end by hemispheres with a radius  $r$ . With these measurements in hand, the total cell volume was computed as

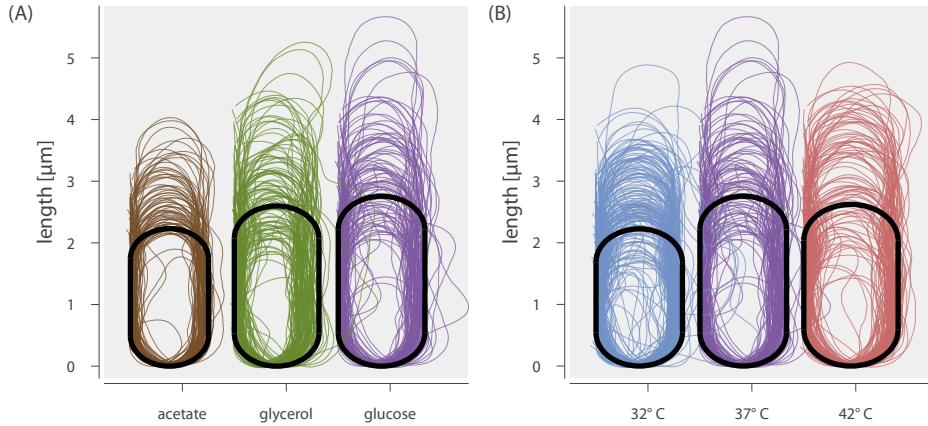
$$V_{\text{cell}} = \pi r^2 \left( a + \frac{4}{3}r \right). \quad (8.1)$$

The output of the SuperSegger segmentation process is an individual matrix for each cell with a variety of fluorescence statistics and information regarding the cell shape. Of the latter category, the software reports in pixels the total length  $\ell$  and width  $w$  of the cell segmentation mask. Given these measurements, we computed the radius  $r$  of the spherocylinder as

$$r = \frac{w}{2} \quad (8.2)$$

and the cylinder length  $a$  as

$$a = \ell - w. \quad (8.3)$$



**Figure 8.2: Growth-rate dependence of cell shape and spherocylinder approximation.** Contours of segmentation masks for a single experiment of each condition are shown as thin colored lines for (A) carbon quality variation and (B) temperature variation.

Fig. 8.2 shows the validity of modeling the segmentation masks as a spherocylinder in two dimensions. Here, the thin colored lines are the contours of a collection of segmentation masks and the black solid lines are spherocylinders using the average length and width of the segmentation masks, as calculated by Eq. 8.2 and Eq. 8.3. It appears that this simple approximation is reasonable for the purposes of this work.

### 8.3 Counting Repressors

In this section, we expand upon the theoretical and experimental implementation of the fluorescence calibration method derived in Ref. (Rosenfeld et al., 2005). We cover several experimental data validation steps as well as details regarding the parameter inference. Finally, we comment on the presence of a systematic error in the repressor counts due to continued asynchronous division between sample preparation and imaging.

#### Theoretical Background of the Binomial Partitioning Method

A key component of this work is the direct measurement of the repressor copy number in each growth condition using fluorescence microscopy. To translate be-

tween absolute fluorescence and protein copy number, we must be able to estimate the average brightness of a single fluorophore or, in other words, determine a calibration factor  $\alpha$  that permits translation from copy number to intensity or vice versa. Several methods have been used over the past decade to estimate this factor, such as measuring single-molecule photobleaching steps (Bialecka-Fornal et al., 2012; Garcia et al., 2011b), measurement of *in vivo* photobleaching rates (Kim et al., 2016; Nayak and Rutenberg, 2011), and through measuring the partitioning of fluorescent molecules between sibling cells after cell division (Brewster et al., 2014; Rosenfeld et al., 2005, 2006). In this work, we used the latter method to estimate the brightness of a single LacI-mCherry dimer. Here, we derive a simple expression which allows the determination of  $\alpha$  from measurements of the fluorescence intensities of a collection of sibling cells.

In the absence of measurement error, the fluorescence intensity of a given cell is proportional to the total number of fluorescent proteins  $N_{\text{prot}}$  by some factor  $\alpha$ ,

$$I_{\text{cell}} = \alpha N_{\text{prot}}. \quad (8.4)$$

Assuming that no fluorophores are produced or degraded over the course of the cell cycle, the fluorescent proteins will be partitioned into the two sibling cells such that the intensity of each sibling can be computed as

$$I_1 = \alpha N_1; I_2 = \alpha(N_{\text{tot}} - N_1), \quad (8.5)$$

where  $N_{\text{tot}}$  is the total number of proteins in the parent cell and  $N_1$  and  $N_2$  correspond to the number of proteins in sibling cell 1 and 2, respectively. This explicitly states that fluorescence is conserved upon a division,

$$I_{\text{tot}} = I_1 + I_2. \quad (8.6)$$

As the observed intensity is directly proportional to the number of proteins per cell, measuring the variance in intensity between sibling cells provides some information as to how many proteins were there to begin with. We can compute these fluctuations as the squared intensity difference between the two siblings as

$$\langle (I_1 - I_2)^2 \rangle = \langle (2I_1 - I_{\text{tot}})^2 \rangle. \quad (8.7)$$

We can relate Eq. 8.7 in terms of the number of proteins using Eq. 8.4 as

$$\langle (I_1 - I_2)^2 \rangle = 4\alpha^2 \langle N_1^2 \rangle - 4\alpha^2 \langle N_1 \rangle N_{\text{tot}} + \alpha^2 N_{\text{tot}}^2, \quad (8.8)$$

where the squared fluctuations are now cast in terms of the first and second moment of the probability distribution for  $N_1$ .

Without any active partitioning of the proteins into the sibling cells, one can model the probability distribution  $g(N_1)$  of finding  $N_1$  proteins in sibling cell 1 as a binomial distribution,

$$g(N_1 | N_{\text{tot}}, p) = \frac{N_{\text{tot}}!}{N_1!(N_{\text{tot}} - N_1)!} p^{N_1} (1-p)^{N_{\text{tot}} - N_1}, \quad (8.9)$$

where  $p$  is the probability of a protein being partitioned into one sibling over the other. With a probability distribution for  $N_1$  in hand, we can begin to simplify Eq. 8.8. Recall that the mean and variance of a binomial distribution are

$$\langle N_1 \rangle = N_{\text{tot}}p, \quad (8.10)$$

and

$$\langle N_1^2 \rangle - \langle N_1 \rangle^2 = N_{\text{tot}}p(1-p), \quad (8.11)$$

respectively. With knowledge of the mean and variance, we can solve for the second moment as

$$\langle N_1^2 \rangle = N_{\text{tot}}p(1-p) + N_{\text{tot}}^2 p^2 = N_{\text{tot}}p(1-p + N_{\text{tot}}p). \quad (8.12)$$

By plugging Eq. 8.10 and Eq. 8.12 into our expression for the fluctuations (Eq. 8.8), we arrive at

$$\langle (I_1 - I_2)^2 \rangle = 4\alpha^2 \left[ (N_{\text{tot}}p[1-p - N_{\text{tot}}p]) - N_{\text{tot}}^2 \left( p + \frac{1}{4} \right) \right], \quad (8.13)$$

which is now defined in terms of the total number of proteins present in the parent cell. Assuming that the proteins are equally partitioned  $p = 1/2$ , Eq. 8.13 reduces to

$$\langle (I_1 - I_2)^2 \rangle = \alpha^2 N_{\text{tot}} = \alpha I_{\text{tot}}. \quad (8.14)$$

Invoking our assertion that fluorescence is conserved (Eq. 8.6),  $I_{tot}$  is equivalent to the sum total fluorescence of the siblings,

$$\langle (I_1 - I_2)^2 \rangle = \alpha(I_1 + I_2). \quad (8.15)$$

Thus, given snapshots of cell intensities and information of their lineage, one can compute how many arbitrary fluorescence units correspond to a single fluorescent protein.

### Cell Husbandry and Time-Lapse Microscopy

The fluorescence calibration method was first described and implemented by Rosenfeld et al. (2005) followed by a more in-depth approach on the statistical inference of the calibration factor in 2006 (Rosenfeld et al., 2006). In both of these works, the partitioning of a fluorescent protein was tracked across many generations from a single parent cell, permitting inference of a calibration factor from a single lineage. Brewster et al. (2014) applied this method in a slightly different manner by quantifying the fluorescence across a large number of *single* division events. Thus, rather than examining the partitioning of fluorescence down many branches of a single family tree, it was estimated from an array of single division events where the fluorescence intensity of the parent cell was variable. In the present work, we take a similar approach to that of Brewster et al. (2014) and examine the partitioning of fluorescence among a large number of independent cell divisions.

A typical experimental work-flow is shown in Fig. 8.3. For each experiment, the strains were grown in varying concentrations of ATC to tune the expression of the repressor. Once the cells had reached exponential phase growth ( $OD_{600nm} \approx 0.3$ ), the cells were harvested and prepared for imaging. This involved two separate sample handling procedures, one for preparing samples for lineage tracking and estimation of the calibration factor and another for taking snapshots of cells from each ATC induction condition for the calculation of fold-change.

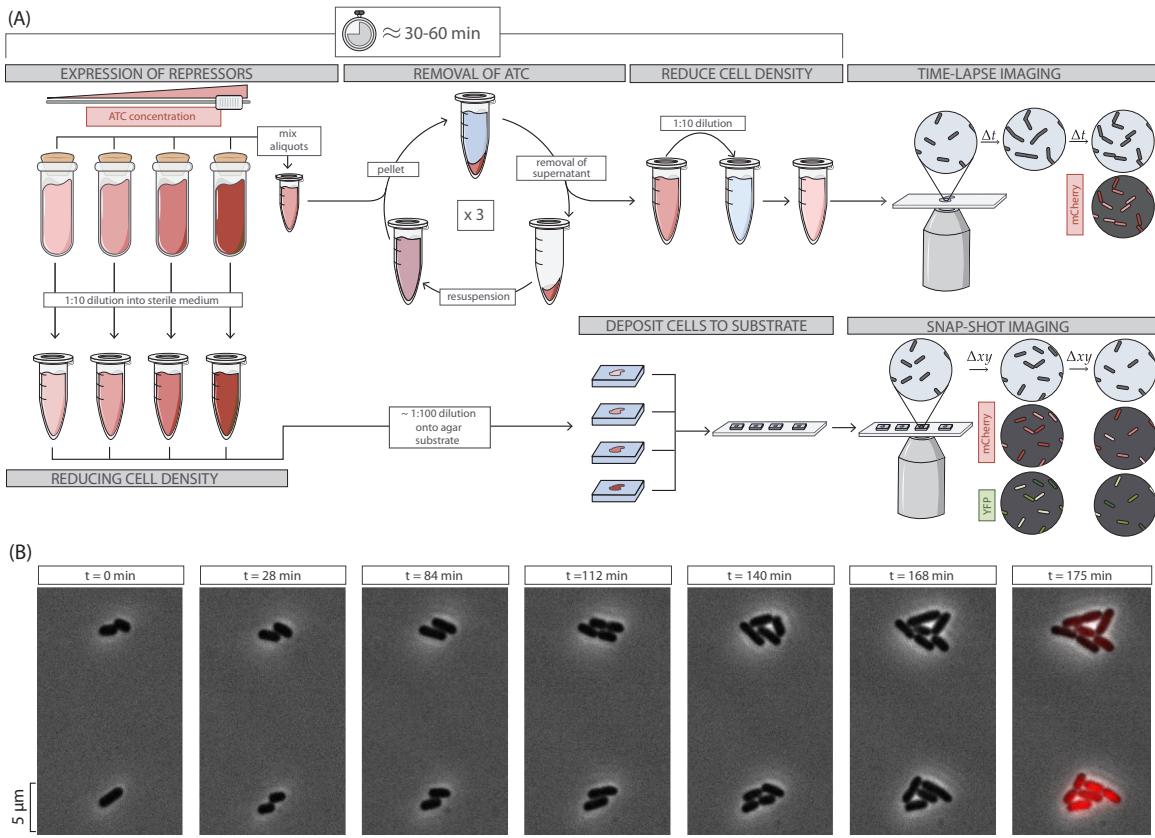
To prepare cells for the calibration factor measurement, a 100  $\mu\text{L}$  aliquot of each ATC induction condition was combined and mixed in a 1.5 mL centrifuge

tube. This cell mixture was then pelleted at  $13000 \times g$  for 1 – 2 minutes. The supernatant containing ATC was then aspirated and the pellet was resuspended in an equal volume of sterile growth media without ATC. This washing procedure was repeated three times to ensure that any residual ATC had been removed from the culture and that expression of LacI-mCherry had ceased. Once washed and resuspended, the cells were diluted ten fold into sterile M9 medium and then imaged on a rigid agarose substrate. Depending on the precise growth condition, a variety of positions were imaged for 1.5 to 4 hours with a phase contrast image acquired every 5 to 15 minutes to facilitate lineage tracking. On the final image of the experiment, an mCherry fluorescence image was acquired of every position. The experiments were then transferred to a computing cluster and the images were computationally analyzed, as described in the next section.

During the washing steps, the remaining ATC induced samples were prepared for snapshot imaging to determine the repressor copy number and fold-change in gene expression for each ATC induction condition. Without mixing the induction conditions together, each ATC induced sample was diluted 1:10 into sterile M9 minimal medium and vigorously mixed. Once mixed, a small aliquot of the samples were deposited onto rigid agarose substrates for later imaging. While this step of the experiment was relatively simple, the total preparation procedure typically lasted between 30 and 60 minutes. As is discussed later, the continued growth of the asynchronously growing culture upon dilution into the sterile medium results in a systematic error in the calculation of the repressor copy number.

### **Lineage Tracking and Fluorescence Quantification**

Segmentation and lineage tracking of both the fluorescence snap shots and time-lapse growth images were performed using the SuperSegger v1.03 (???) software using MATLAB R2017B (MathWorks, Inc). The result of this segmentation is a list of matrices for each unique imaged position with identifying data for each segmented cell such as an assigned ID number, the ID of the sibling cell, the ID of the parent cell, and various statistics. These files were then analyzed using Python



**Figure 8.3: An experimental workflow for time-lapse imaging.** (A) the series of steps followed in a given experiment. Cells are grown in various concentrations of ATC (shaded red cultures) to an  $OD_{600nm} \approx 0.3$ . Equal aliquots of each ATC-induced culture are mixed into a single eppendorf tube and pelleted via centrifugation. The supernatant containing ATC is aspirated and replaced with an equal volume of sterile, ATC-free growth medium. This washing procedure is repeated three times to ensure that residual ATC is removed from the culture and expression of the LacI-mCherry fusion is ceased. After a final resuspension in sterile ATC-free medium, the cell mixture is diluted 10 fold to reduce cell density. small aliquot of this mixture is then mounted and imaged at 100x magnification until at least one cell division has occurred. (B) Two representative microcolonies from a time-lapse growth experiment. The time point is provided above each image. After at least one division has occurred, a final mCherry fluorescence image is acquired and quantified.

3.7. All scripts and software used to perform this analysis can be found on the associated paper website and GitHub repository.

Using the ID numbers assigned to each cell in a given position, we matched all sibling pairs present in the last frame of the growth movie when the final mCherry fluorescence image was acquired. These cells were then filtered to exclude segmentation artifacts (such as exceptionally large or small cells) as well as any cells which the SuperSegger software identified as having an error in segmentation. Given the large number of cells tracked in a given experiment, we could not manually correct these segmentation artifacts, even though it is possible using the software. To err on the side of caution, we did not consider these edge cases in our analysis.

With sibling cells identified, we performed a series of validation checks on the data to ensure that both the experiment and analysis behaved as expected. Three validation checks are illustrated in Fig. 8.4. To make sure that the computational pairing of sibling cells was correct, we examined the intensity distributions of each sibling pair. If siblings were being paired solely on their lineage history and not by other features (such as size, fluorescence, etc), one would expect the fluorescence distributions between the two sibling cells to be identical. Fig. 8.4(A) shows the nearly identical intensity distributions of all siblings from a single experiment, indicating that the pairing of siblings is independent of their fluorescence.

Furthermore, we examined the partitioning of the fluorescence between the siblings. In Eq. 8.10, we defined the mean number of proteins inherited by one sibling. This can be easily translated into the language of intensity as

$$\langle I_1 \rangle = \alpha (I_1 + I_2) p, \quad (8.16)$$

where we assume that fluorescence is conserved during a division event such that  $I_{\text{tot}} = I_1 + I_2$ . As described previously, we make the simplifying assumption that partitioning of the fluorophores between the two sibling cells is fair, meaning that  $p = 0.5$ . We can see if this approximation is valid by computing the fractional

intensity of each sibling cell as

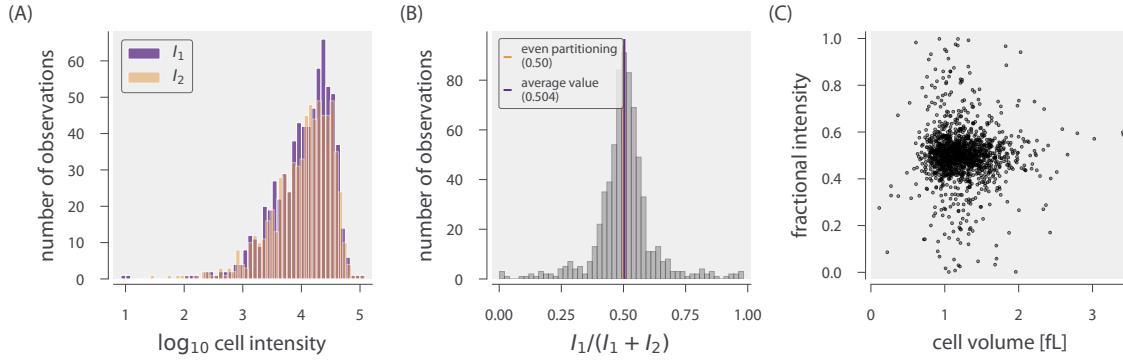
$$p = \frac{\langle I_1 \rangle}{\langle I_1 + I_2 \rangle}. \quad (8.17)$$

Fig. 8.4(B) shows the distribution of the fractional intensity for each sibling pair. The distribution is approximately symmetric about 0.5, indicating that siblings are correctly paired and that the partitioning of fluorescence is approximately equal between siblings. Furthermore, we see no correlation between the cell volume immediately after division and the observed fractional intensity. This suggests that the probability of partitioning to one sibling or the other is not dependent on the cell size. An assumption [backed by experimental measurements (Garcia and Phillips, 2011; Phillips et al., 2019)] in our thermodynamic model is that all repressors in a given cell are bound to the chromosome, either specifically or nonspecifically. As the chromosome is duplicated and partitioned into the two siblings without fail, our assumption of repressor adsorption implies that partitioning should be independent of the size of the respective sibling. The collection of these validation statistics give us confidence that both the experimentation and the analysis are properly implemented and not introducing bias into our estimation of the calibration factor.

### Statistical Inference of the Fluorescence Calibration Factor

As is outlined in the Materials and Methods section of the Chapter 4, we took a Bayesian approach towards our inference of the calibration factor given fluorescence measurements of sibling cells. Here, we expand in detail on this statistical model and its implementation.

To estimate the calibration factor  $\alpha$  from a set of lineage measurements, we assume that fluctuations in intensity resulting from measurement noise is negligible compared to that resulting from binomial partitioning of the repressors upon cell division. In the absence of measurement noise, the intensity of a given cell  $I_{cell}$  can be directly related to the total number of fluorophores  $N$  through a scaling factor



**Figure 8.4: Experimental sanity checks and inference of a fluorescence calibration factor.** (A) Intensity distributions of sibling cells after division. Arbitrarily labeled “sibling 1” and “sibling 2” distributions are shown in purple and orange, respectively. Similarity of the distributions illustrates lack of intensity bias on sibling pair assignment. (B) the fractional intensity of sibling 1 upon division. For each sibling pair, the fractional intensity is computed as the intensity of sibling 1  $I_1$  divided by the summed intensities of both siblings  $I_1 + I_2$ . (C) Partitioning intensity as a function of cell volume. The fractional intensity of every sibling cell is plotted against its estimated newborn volume.

$\alpha$ , such that

$$I_{cell} = \alpha N. \quad (8.18)$$

Assuming no fluorophores are produced or degraded over the course of a division cycle, the fluorescence of the parent cell before division is equal to the sum of the intensities of the sibling cells,

$$I_{parent} = I_1 + I_2 = \alpha (N_1 + N_2), \quad (8.19)$$

where subscripts 1 and 2 correspond to arbitrary labels of the two sibling cells. We are ultimately interested in knowing the probability of a given value of  $\alpha$  which, using Bayes’ theorem, can be written as

$$g(\alpha | I_1, I_2) = \frac{f(I_1, I_2 | \alpha)g(\alpha)}{f(I_1, I_2)}, \quad (8.20)$$

where we have used  $g$  and  $f$  to denote probability densities over parameters and data, respectively. The first quantity in the numerator  $f(I_1, I_2 | \alpha)$  describes the likelihood of observing the data  $I_1, I_2$  given a value for the calibration factor  $\alpha$ . The

term  $g(\alpha)$  captures all prior knowledge we have about what the calibration factor could be, remaining ignorant of the collected data. The denominator  $f(I_1, I_2)$  is the likelihood of observing our data  $I_1, I_2$  irrespective of the calibration factor and is a loose measure of how well our statistical model describes the data. As it is difficult to assign a functional form to this term and serves as a multiplicative constant, it can be neglected for the purposes of this work.

Knowing that the two observed sibling cell intensities are related, conditional probability allows us to rewrite the likelihood as

$$f(I_1 | I_2, \alpha) = f(I_1 | I_2, \alpha) f(I_2 | \alpha), \quad (8.21)$$

where  $f(I_2 | \alpha)$  describes the likelihood of observing  $I_2$  given a value of  $\alpha$ . As  $I_2$  can take any value with equal probability, this term can be treated as a constant. Through change of variables and noting that  $I_2 = \alpha N_2$ , we can cast the likelihood  $f(I_1 | I_2, \alpha)$  in terms of the number of proteins  $N_1$  and  $N_2$  as

$$f(I_1 | I_2, \alpha) = f(N_1 | N_2, \alpha) \left| \frac{dN_1}{dI_1} \right| = \frac{1}{\alpha} f(N_1 | N_2, \alpha). \quad (8.22)$$

Given that the proteins are binomially distributed between the sibling cells with a probability  $p$  and that the intensity is proportional to the number of fluorophores, this likelihood becomes

$$f(I_1 | I_2, \alpha, p) = \frac{1}{\alpha} \frac{\left( \frac{I_1 + I_2}{\alpha} \right)!}{\left( \frac{I_1}{\alpha} \right)! \left( \frac{I_2}{\alpha} \right)!} p^{\frac{I_1}{\alpha}} (1-p)^{\frac{I_2}{\alpha}} \quad (8.23)$$

However, The quantity  $I/\alpha$  is not exact, making calculation of its factorial undefined. These factorials can therefore be approximated by a gamma function as  $n! = \Gamma(n+1)$ .

Assuming that partitioning of a protein between the two sibling cells is a fair process ( $p = 1/2$ ), Eq. 8.23 can be generalized to a set of lineage measurements  $[I_1, I_2]$  as

$$f([I_1] | [I_2], \alpha) = \frac{1}{\alpha^k} \prod_i^k \frac{\Gamma \left( \frac{I_1 + I_2}{\alpha} + 1 \right)}{\Gamma \left( \frac{I_1}{\alpha} + 1 \right) \Gamma \left( \frac{I_2}{\alpha} + 1 \right)} 2^{-\frac{I_1 + I_2}{\alpha}}, \quad (8.24)$$

where  $k$  is the number of division events observed.

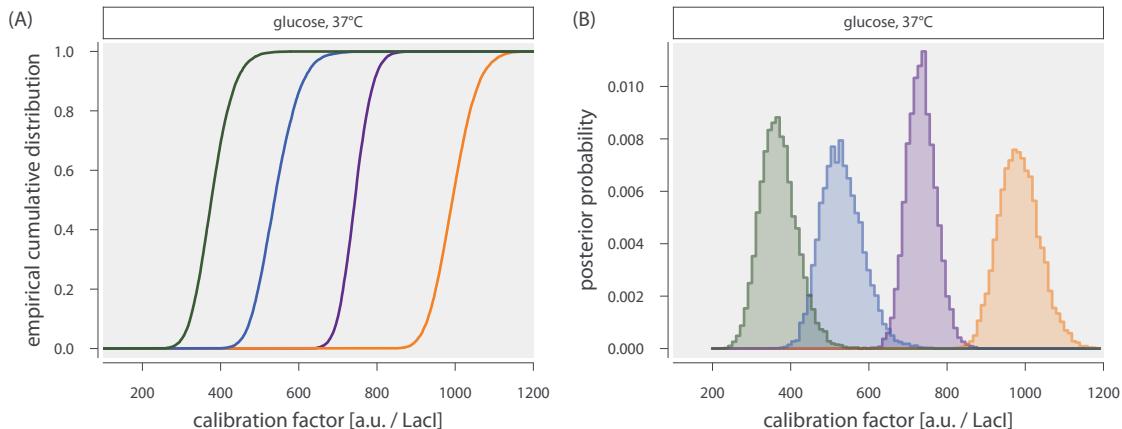
With a likelihood in place, we can now assign a functional form to the prior distribution for the calibration factor  $g(\alpha)$ . Though ignorant of data, the experimental design is such that imaging of a typical highly-expressing cell will occupy 2/3 of the dynamic range of the camera. We can assume it's more likely that the calibration factor will be closer to 0 a.u. than the bit depth of the camera (4095 a.u.) or larger. We also know that it is physically impossible for the fluorophore to be less than 0 a.u., providing a hard lower-bound on its value. We can therefore impose a weakly informative prior distribution as a half normal distribution,

$$g(\alpha) = \sqrt{\frac{2}{\pi\sigma^2}} \exp\left[\frac{-\alpha^2}{2\sigma^2}\right]; \forall \alpha > 0. \quad (8.25)$$

where the standard deviation is large, for example,  $\sigma = 500$  a.u. / fluorophore. We evaluated the posterior distribution using Markov chain Monte Carlo (MCMC) as is implemented in the Stan probabilistic programming language (Carpenter et al., 2017). The .stan file associated with this model along with the Python code used to execute it can be accessed on the paper website and GitHub repository. Fig. 8.5 shows the posterior probability distributions of the calibration factor estimated for several biological replicates of the glucose growth condition at 37° C. For each posterior distribution, the mean and standard deviation was used as the calibration factor and uncertainty for the corresponding data set.

### Correcting for Systematic Experimental Error

While determination of the calibration factor relies on time-resolved measurement of fluorescence partitioning, we computed the repressor copy number and fold-change in gene expression from still snapshots of each ATC induction condition and two control samples, as is illustrated in Fig. 8.3 Given these snapshots, individual cells were segmented again using the SuperSegger software in MATLAB R2017b. The result of this analysis is an array of single-cell measurements of the YFP and mCherry fluorescence intensities. With these values and a calibration factor estimated by Eq. 8.24 and Eq. 9.24, we can compute the estimated number of

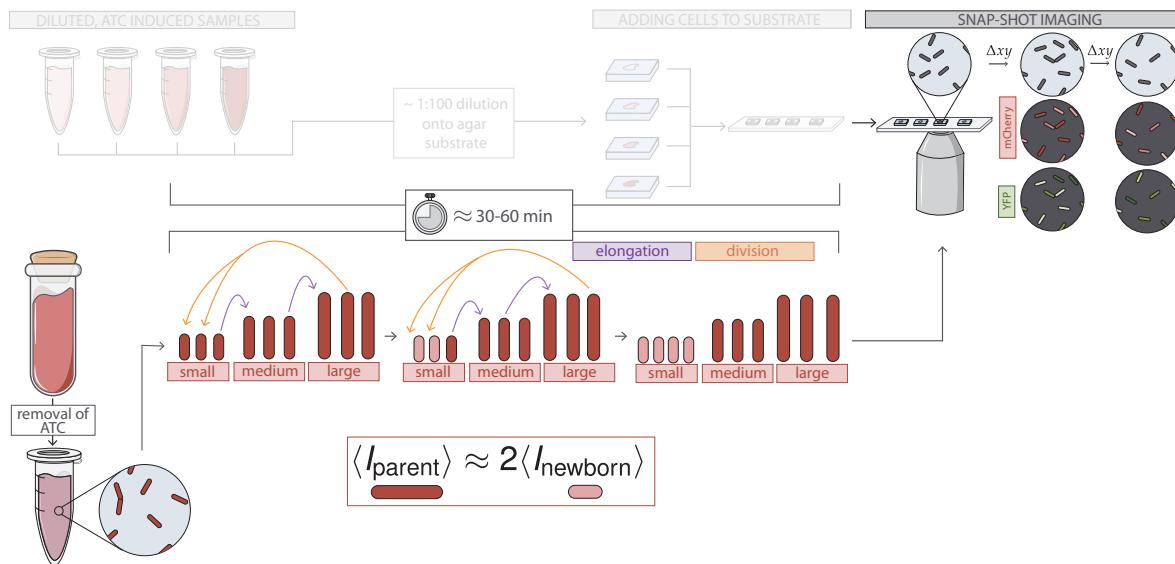


**Figure 8.5: Representative posterior distributions for the calibration factor.** (A) Empirical cumulative distribution functions and (B) histograms of the posterior distribution for the calibration factor  $\alpha$  for several biological replicates. Different colors indicate different biological replicates of experiments performed in glucose supplemented medium at 37° C.

repressors per cell in every condition.

However, as outlined previously and in Fig. 8.3, the sample preparation steps for these experiments involve several steps which require careful manual labor. This results in an approximately 30 to 60 minute delay from when production of the LacI-mCherry construct is halted by removal of ATC to actual imaging on the microscope. During this time, the diluted cultures are asynchronously growing, meaning that the cells of the culture are at different steps in the cell cycle. Thus, at any point in time, a subset of cells will be on the precipice of undergoing division, partitioning the cytosolic milieu between the two progeny. As LacI-mCherry is no longer being produced, the cells that divided during the dwell time from cell harvesting to imaging will have reduced the number of repressors by a factor of 2 on average. This principle of continued cell division is shown Fig. 8.6.

How does this partitioning affect our calculation and interpretation of the fold-change in gene expression? Like the LacI-mCherry fusion, the YFP reporter proteins are also partitioned between the progeny after a division event such that, on average, the total YFP signal of the newborn cells is one-half that of the parent



**Figure 8.6: Continued division results in a systematic error in repressor counts.** During the sample preparation steps, the asynchronous culture continues to divide though the length of the sample preparation step is less than a cell doubling time. Because production of LacI-mCherry is halted, cells that complete a division cycle during this time will partition their repressors between the progeny. The total number of repressors in parent cells just before division is, on average, twice that of the newborn cells.

cell. As the maturation time of the mCherry and YFP variants used in this work are relatively long in *E. coli* (Balleza et al., 2018; Nagai et al., 2002), we can make the assumption any newly-expressed YFP molecules after cells have divided are not yet visible in our experiments. Thus, the fold-change in gene expression of the average parent cell can be calculated given knowledge of the average expression of the progeny.

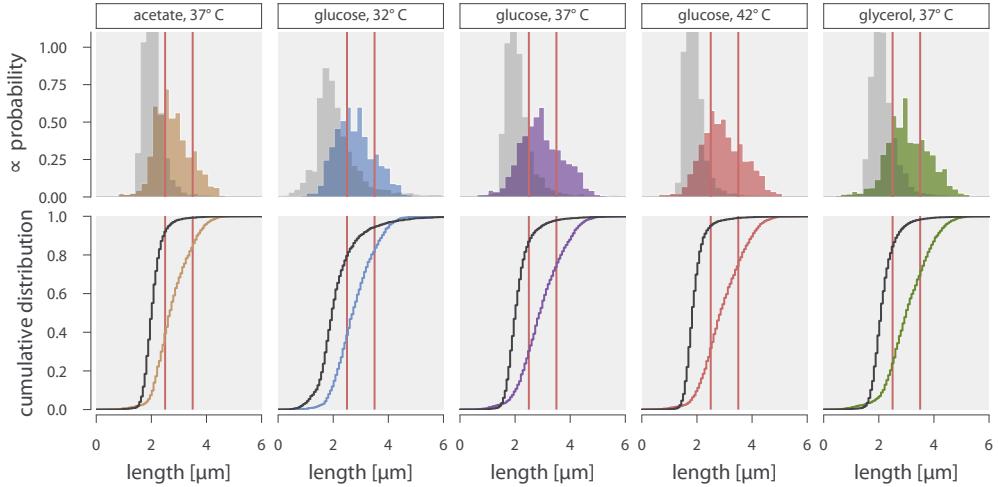
The fold-change in gene expression is a relative measurement to a control which is constitutively expressing YFP. As the latter control sample is also asynchronously dividing, the measured YFP intensity of the newborn constitutively expressing cells is on average 1/2 that of the parent cell. Therefore, the fold-change in gene expression can be calculated as

$$\langle \text{fold-change}_{\text{parent}} \rangle = \frac{2 \times \langle I_{\text{newborn}}^{(\text{YFP})} (R > 0) \rangle}{2 \times \langle I_{\text{newborn}}^{(\text{YFP})} (R = 0) \rangle} = \frac{\langle I_{\text{newborn}}^{(\text{YFP})} (R > 0) \rangle}{\langle I_{\text{newborn}}^{(\text{YFP})} (R = 0) \rangle}. \quad (8.26)$$

Thus, when calculating the fold-change in gene expression one does not need to correct for any cell division that occurs between sample harvesting and imaging as it is a relative measurement. However, the determination of the repressor copy number is a *direct* measurement and requires a consideration of unknown division events.

To address this source of systematic experimental error, we examined the cell length distributions of all segmented cells from the snapshots as well as the distribution of newborn cell lengths from the time-lapse measurements. Fig. 8.7 shows that a significant portion of the cells from the snapshots (colored distributions) overlap with the distribution of newborn cell lengths (grey distributions). For each condition, we partitioned the cells from the snapshots into three bins based on their lengths – “small” cells had a cell length less than  $2.5 \mu\text{m}$ , “medium” cells had lengths between  $2.5 \mu\text{m}$  and  $3.5 \mu\text{m}$ , and “large” cells being longer than  $3.5 \mu\text{m}$ . These thresholds were chosen manually by examining the newborn cell-size distributions and are shown as red vertical lines in Fig. 8.7. Under this partitioning, we consider all “small” cells to have divided between cessation of LacI-mCherry production and imaging, “medium”-length cells to have a mixture of long newborn cells (from the tail of the newborn cell length distribution) and cells that haven’t divided, and cells in the “long” group to be composed entirely of cells which did not undergo a division over the course of sample preparation.

Given this coarse delineation of cell age by length, we examined how correction factors could be applied to correct for the the undesired systematic error due to dilution of repressors. We took the data collected in this work and compared the results to the fold-change in gene expression reported in the literature for the same regulatory architecture. Without correcting for undesired cell division, the observed fold-change in gene expression falls below the prediction and does not overlap with data from the literature (Fig. 8.8(A), light purple). Using the uncorrected measurements, we estimated the DNA binding energy to be  $\Delta\epsilon_R \approx -15 k_B T$  which does not agree with the value for the O2 operator reported in Garcia and



**Figure 8.7: Cell length distributions of fluorescence snapshots and newborn cells.** Top row shows the distribution of cell lengths (pole-to-pole, colored distributions) off cells imaged for calculation of the repressor copy number and fold-change in gene expression. The distribution of newborn cell lengths for that given condition is shown in grey. The vertical red lines correspond to the cell length threshold of  $2.5\mu\text{m}$  and  $3.5\mu\text{m}$ , from left to right, respectively. Cells to the left of the first vertical line were identified as “small”, cells in between the two vertical lines to be “medium” sized, and “long” cells to the right of the second vertical line. Cells below the  $2.4\mu\text{m}$  threshold were treated as cells who divided after production of lacI-mCherry had been halted. Bottom row shows the same data as the top row but as the empirical cumulative distribution.

Phillips (2011) or with the inferred DNA binding energies from the other data sources (Fig. 8.8 (B)).

These results emphasize the need to correct for undesired dilution of repressors through cell division during the sample preparation period, and we now consider several different manners of applying this correction. We first consider the extreme case where all cells of the culture underwent an undesired division after LacI-mCherry production was halted. This means that the average repressor copy number measured from all cells is off by a factor of 2. The result of applying a factor of 2 correction to all measurements can be seen in Fig. 8.8 as dark red points. Upon applying this correction, we find that the observed fold-change in gene expression agrees with the prediction and data from other sources in the literature. The estimated DNA binding energy  $\Delta\epsilon_R$  from these data is also in agreement with

other data sources Fig. 8.8. This result suggests that over the course of sample preparation, a non-negligible fraction of the diluted culture undergoes a division event before being imaged.

We now begin to relax assumptions as to what fraction of the measured cells underwent a division event before imaging. As described above, in drawing distinctions between “small”, “medium”, and “large” cells, we assume the latter represent cells which *did not* undergo a division between the harvesting and imaging of the samples. Thus, the repressor counts of these cells should require no correction. The white-faced points in Fig. 8.8(A) shows the fold-change in gene expression of *only* the large cell fraction, which falls within error of the theoretical prediction. Furthermore, the inferred DNA binding energy falls within error of that inferred from data of Garcia and Phillips (2011) and that inferred from data assuming all cells underwent a division event Fig. 8.8, though it does not fall within error of the binding energy reported in Garcia and Phillips (2011).

The most realistic approach that can be taken to avoid using only the “large” bin of cells is to assume that all cells with a length  $\ell < 2.5 \mu\text{m}$  have undergone a division, requiring a two-fold correction to their average repressor copy number. The result of this approach can be seen in Fig. 8.8 as purple points. The inferred DNA binding energy from this correction approach falls within error of that inferred from only the large cells white-faced points in 8.8 as well as overlapping with the estimate treating all cells as having undergone a division (red).

There are several experimental techniques that could be implemented to avoid needing to apply a correction factor as described here. In Brewster et al. (2014), the fold-change in gene expression was measured by tracking the production rate of a YFP reporter before and after a single cell division event coupled with direct measurement of the repressor copy number using the same binomial partitioning method. This implementation required an extensive degree of manual curation of segmentation as well as correcting for photobleaching of the reporter, which in itself is a non-trivial correction (Garcia et al., 2011b). The experimental approach

presented here sacrifices a direct measure of the repressor copy number for each cell via the binomial partitioning method, but permits the much higher throughput needed to assay the variety of environmental conditions. Ultimately, the inferred DNA binding energy for all of the scenarios described above agree within  $1 k_B T$ , a value smaller than the natural variation in the DNA binding energies of the three native *lacO*, which is  $\approx 6 k_B T$ . For the purposes of this work, we erred on the side of caution and only used the cells deemed “large” for the measurements reported in Chapter 4 and the remaining sections of this chapter.

#### 8.4 Parameter Estimation of DNA Binding Energies and Comparison Across Carbon Source

In the main text, we conclude that the biophysical parameters defining the fold-change input-output function are unperturbed between different carbon sources. This conclusion is reached primarily by comparing how well the fold-change and the free energy shift  $\Delta F$  is predicted using the parameter values determined in glucose supported medium at 37° C. In this section, we redetermine the DNA binding energy for each carbon source condition and test its ability to predict the fold-change of the other conditions.

##### Reparameterizing the fold-change input-output function

As described previously, the fold-change in gene expression is defined by the total repressor copy number  $R$ , the energetic difference between the active and inactive states of the repressor  $\Delta\varepsilon_{AI}$ , and the binding energy of the repressor to the DNA  $\Delta\varepsilon_{RA}$ . Using fluorescence microscopy, we can directly measure the average repressor copy number per cell, reducing the number of variable parameters to only the energetic terms.

Estimating both  $\Delta\varepsilon_{RA}$  and  $\Delta\varepsilon_{AI}$  simultaneously is fraught with difficulty as the parameters are highly degenerate (Razo-Mejia et al., 2018). We can avoid this degeneracy by reparameterizing the fold-change input-output function as

$$\text{fold-change} = \left(1 + \frac{R}{N_{NS}} e^{-\beta\varepsilon}\right)^{-1}, \quad (8.27)$$

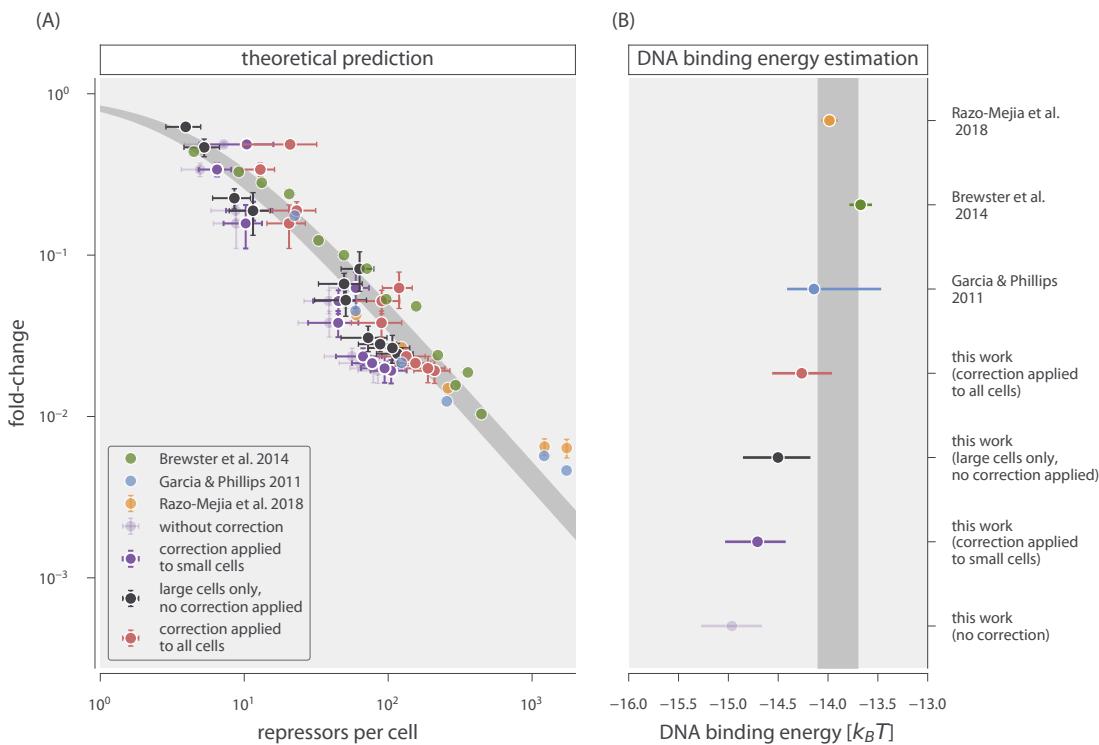


Figure 8.8: \*\*Influence of a correction factor on fold-change and the DNA binding energy. (A) Fold-change in gene expression measurements from (Brewster et al., 2014; Garcia and Phillips, 2011; Razo-Mejia et al., 2018) along with data from this work. Data from this work shown are with no correction (purple), correcting for small cells only (dark purple), using only the large cell fraction (black faced point), and treating all cells as newborn cells (red). Where visible, errors correspond to the standard error of 5 to 10 biological replicates. (B) estimated DNA binding energy from each data set. Points are the median of the posterior distribution over the DNA binding energy. Horizontal lines indicate the width of the 95% credible region of the posterior distribution. Grey lines in (A) and (B) correspond to the theoretical prediction and estimated binding energy from Garcia and Phillips (2011), respectively.

where  $\epsilon$  is the effective energetic parameter

$$\epsilon = \Delta\epsilon_R - k_B T \log \left( 1 + e^{-\beta\Delta\epsilon_{AI}} \right). \quad (8.28)$$

Thus, to further elucidate any changes to the parameter values due to changing the carbon source, we can infer the best-fit value of  $\epsilon$  for each condition and explore how well it predicts the fold-change in other conditions.

### Statistical Inference of $\epsilon$

We are interested in the probability distribution of the parameter  $\epsilon$  given knowledge of the repressor copy number  $R$  and a collection of fold-change measurements  $\mathbf{fc}$  can be calculated via Bayes' theorem as

$$g(\epsilon | R, \mathbf{fc}) = \frac{f(\mathbf{fc} | R, \epsilon)g(\epsilon)}{f(\mathbf{fc})}, \quad (8.29)$$

where  $g$  and  $f$  represent probability densities of parameters and data, respectively. In this context, the denominator term  $f(\mathbf{fc})$  serves only as normalization constant and can be neglected. As is described in detail in Refs. (Chure et al., 2019; Razo-Mejia et al., 2018), we assume that a given set of fold-change measurements are normally distributed about the theoretical value  $\mu$  defined by the fold-change function. The likelihood can be mathematically defined as

$$f(\mathbf{fc} | \epsilon, R) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_i^N \exp \left[ -\frac{(\mathbf{fc}_i - \mu(\epsilon, R))^2}{2\sigma^2} \right], \quad (8.30)$$

where  $N$  is the total number of fold-change measurements, and  $\sigma$  is the standard deviation of the observations about the true mean and is another parameter that must be included in the estimation. As the fold-change in gene expression in this work covers several orders of magnitude (from  $\approx 10^{-3} - 10^0$ ) it is better to condition the parameters on the log fold-change rather than linear scaling, translating Eq. 8.30 to

$$f(\mathbf{fc}^* | \epsilon, R) = \frac{1}{(2\pi\sigma)^{N/2}} \prod_i^N \exp \left[ -\frac{(\mathbf{fc}_i^* - \mu^*(\epsilon, R))^2}{2\sigma^2} \right], \quad (8.31)$$

where  $\mathbf{fc}^*$  and  $\mu^*(\epsilon, R)$  are the transformations

$$\mathbf{fc}^* = \log(\mathbf{fc})$$

and

$$\mu^*(\epsilon, R) = -\log \left( 1 + \frac{R}{N_{NS}} e^{-\beta\epsilon} \right), \quad (8.32)$$

respectively.

With a likelihood in hand, we can now turn towards defining functional forms for the prior distributions  $g(\sigma)$  and  $g(\epsilon)$ . For these definitions, we can turn to those used in Chure et al. (2019),

$$g(\epsilon) \sim \text{Normal}(\mu = -12, \sigma = 6) \quad (8.33)$$

and

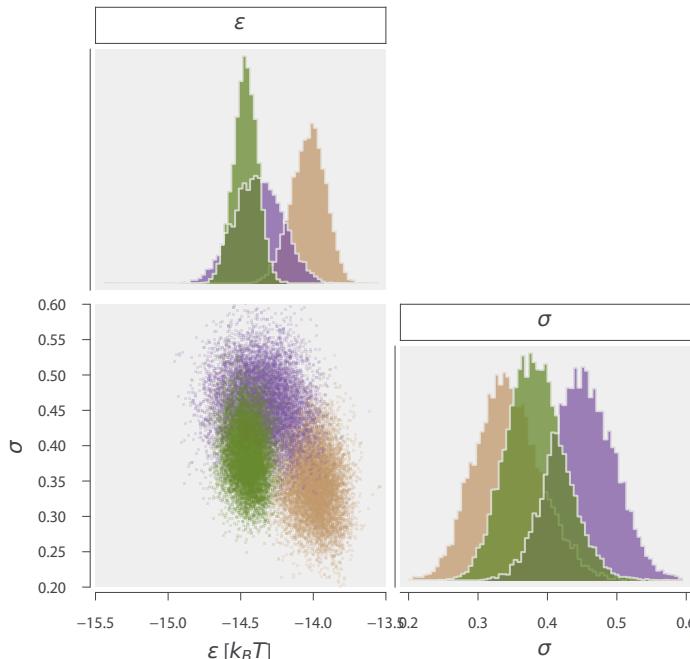
$$g(\sigma) \sim \text{HalfNormal}(\phi = 0.1), \quad (8.34)$$

where we introduce the shorthand notation of “Normal” and “HalfNormal”. Combining Eq. 8.31 and Eq. 8.34 - Eq. 8.33 yields the complete posterior distribution for estimating the DNA binding energy for each carbon source medium. The complete posterior distribution was sampled using Markov chain Monte Carlo in the Stan probabilistic programming language (Carpenter et al., 2017).

The sampled posterior distributions for  $\epsilon$  and  $\sigma$  for each carbon source condition shown in Fig. ?? and are summarized in Table 8.1. The posterior distributions of  $\epsilon$  across the conditions are approximately equal with highly overlapping 95% credible regions. The predictive capacity of each estimate of  $\epsilon$  is shown in Fig. 8.9 where all fold-change measurements fall upon the theoretical prediction regardless of which carbon source that the parameter value was conditioned upon. With this analysis, we can say with quantitative confidence that the biophysical parameters are indifferent to the physiological changes resulting from variation in carbon quality.

Table 8.1: Summarized parameter estimates of  $\epsilon$  and  $\sigma$  given a single growth condition. reported as median and upper/lower bounds of 95% credible region.

Growth Condition	Parameter	Value
Glucose, 37° C	$\epsilon$	$-14.5^{+0.2}_{-0.3} k_B T$
	$\sigma$	$0.3^{+0.1}_{-0.1}$
Glycerol, 37° C	$\epsilon$	$-14.6^{+0.1}_{-0.1} k_B T$
	$\sigma$	$0.39^{+0.10}_{-0.08}$
Acetate, 37° C	$\epsilon$	$-14.1^{+0.2}_{-0.3} k_B T$
	$\sigma$	$0.35^{+0.1}_{-0.1}$

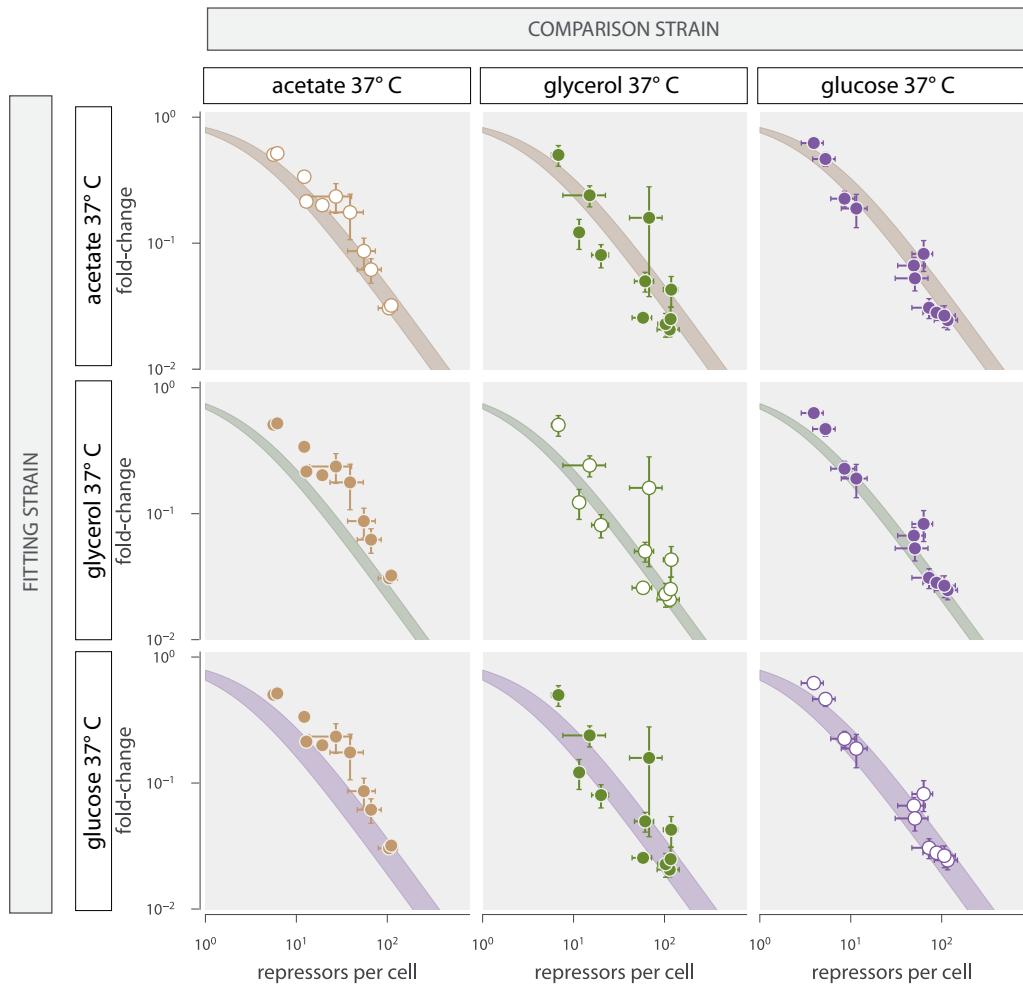


{#fig:carbon\_cornerplot short-

caption="Posterior probability distributions of effective DNA binding and standard deviation for different carbon sources.}

## 8.5 Statistical Inference of Entropic Costs

In the main text, we describe how a simple rescaling of the energetic parameters  $\Delta\epsilon_R$  and  $\Delta\epsilon_{AI}$  is not sufficient to describe the fold-change in gene expression when



**Figure 8.9: Pairwise estimation and prediction of DNA binding energies.** Rows indicate the strain to which the effective DNA binding energy  $\epsilon$  was estimated and columns are the strains whose fold-change is predicted. Shaded lines represent the 95% credible region of the prediction given the estimated value of  $\epsilon$ . Points and error correspond to the median and standard error of fit to eight biological replicates.

the growth temperature is changed from 37° C. In this section, we describe the inference of hidden entropic parameters to phenomenologically describe the temperature dependence of the fold-change in gene expression.

### Definition of hidden entropic costs

The values of the energetic parameters  $\Delta\epsilon_R$  and  $\Delta\epsilon_{AI}$  were determined in a glucose supplemented medium held at 37° C which we denote as  $T_{ref}$ . A null model to describe temperature dependence of these parameters is to rescale them to the changed temperature  $T_{exp}$  as

$$\Delta\epsilon^* = \frac{T_{ref}}{T_{exp}} \Delta\epsilon, \quad (8.35)$$

where  $\Delta\epsilon^*$  is either  $\Delta\epsilon_R$  or  $\Delta\epsilon_{AI}$ . However, we found that this null model was not sufficient to describe the fold-change in gene expression, prompting the formulation of a new phenomenological description.

Our thermodynamic model for the fold-change in gene expression coarse-grains the regulatory architecture to a two state model, meaning many of the rich features of regulation such as vibrational entropy, the material properties of DNA, and the occupancy of the repressor to the DNA are swept into the effective energetic parameters. As temperature was never perturbed when this model was developed, modeling these features was not necessary. However, we must now return to these features to consider what may be affected.

Without assigning a specific mechanism, we can say that there is a temperature-dependent entropic parameter that was neglected in the estimation of the energetic parameters in Garcia and Phillips 2011 (Garcia and Phillips, 2011) and Razo-Mejia *et al.* 2018 (Razo-Mejia et al., 2018). In this case, the inferred energetic parameter  $\Delta\epsilon$  is composed of enthalpic ( $\Delta H$ ) and entropic ( $\Delta S$ ) parameters,

$$\Delta\epsilon^* = \Delta H - T\Delta S.$$

For a set of fold-change measurements at a temperature  $T_{exp}$ , we are interested in estimating values for  $\Delta H$  and  $\Delta S$  for each energetic parameter. Given measurements from Refs. (Garcia and Phillips, 2011; Razo-Mejia et al., 2018), we know at

37°C what  $\Delta\varepsilon_{RA}$  and  $\Delta\varepsilon_{AI}$  are inferred to be, placing a constraint on the possible values of  $\Delta H$  and  $\Delta S$ ,

$$\Delta H_R = \Delta\varepsilon_R + T_{ref}\Delta S_R = T_{ref}\Delta S_R - 13.9 k_B T, \quad (8.36)$$

and

$$\Delta H_{AI} = \Delta\varepsilon_{AI} + T_{ref}\Delta S_{AI} = T_{ref}\Delta S_{AI} + 4.5 k_B T \quad (8.37)$$

for the DNA binding energy and allosteric state energy difference, respectively.

### Statistical inference of $\Delta S_R$ and $\Delta S_{AI}$

Given the constraints from Eq. 8.36 and Eq. 8.37, we are interested in inferring the entropic parameters  $\Delta S_R$  and  $\Delta S_{AI}$  given literature values for  $\Delta\varepsilon_{RA}$  and  $\Delta\varepsilon_{AI}$  and the set of fold-change measurements  $\mathbf{fc}$  at a given temperature  $T_{exp}$ . The posterior probability distribution for the entropic parameters can be enumerated via Bayes' theorem as

$$g(\Delta S_R, \Delta S_{AI} | \mathbf{fc}) = \frac{f(\mathbf{fc} | \Delta S_R, \Delta S_{AI})g(\Delta S_R, \Delta S_{AI})}{f(\mathbf{fc})}, \quad (8.38)$$

where  $g$  and  $f$  are used to denote probability densities over parameters and data, respectively. As we have done elsewhere in this SI text, we treat  $f(\mathbf{fc})$  as a normalization constant and neglect it in our estimation of  $g(\Delta S_R, \Delta S_{AI} | \mathbf{fc})$ . Additionally, as is discussed in detail earlier in this chapter, we consider the log fold-change measurements to be normally distributed about a mean  $fc^*$  defined by the fold-change input-output function and a standard deviation  $\sigma$ . Thus, the likelihood for the fold-change in gene expression is

$$f(\mathbf{fc} | \Delta S_R, \Delta S_{AI}, \sigma, T_{exp}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_i^N \exp \left[ -\frac{[\log fc_i - \log fc^*(\Delta S_R, \Delta S_{AI}, T_{exp})]^2}{2\sigma^2} \right]. \quad (8.39)$$

In calculating the mean  $fc^*$ , the effective energetic parameters  $\Delta\varepsilon_R^*$  and  $\Delta\varepsilon_{AI}^*$  can be defined and constrained using Eq. 8.36 and Eq. 8.37 as

$$\Delta\varepsilon_R^* = \Delta H_R - T_{exp}\Delta S_R = \Delta S_R(T_{ref} - T_{exp}) - 13.9 k_B T_{ref}, \quad (8.40)$$

and

$$\Delta\varepsilon_{AI}^* = \Delta H_{AI} - T_{exp}\Delta S_{AI} = \Delta S_{AI}(T_{ref} - T_{exp}) + 4.5k_B T_{ref}. \quad (8.41)$$

As the enthalpic parameters are calculated directly from the constraints of Eq. 8.36 and Eq. 8.37, we must only estimate three parameters,  $\Delta S_R$ ,  $\Delta S_{AI}$ , and  $\sigma$ , each of which need a functional form for the prior distribution.

*A priori*, we know that both  $\Delta S_R$  and  $\Delta S_{AI}$  must be small because  $T_{ref}$  and  $T_{exp}$  are defined in K. As these entropic parameters can be either positive or negative, we can define the prior distributions  $g(\Delta S_R)$  and  $g(\Delta S_{AI})$  as a normal distribution centered at zero with a small standard deviation,

$$g(\Delta S_R) \sim Normal(\mu = 0, \sigma = 0.1), \quad (8.42)$$

and

$$g(\Delta S_{AI}) \sim Normal(\mu = 0, \sigma = 0.1). \quad (8.43)$$

The standard deviation  $\sigma$  can be defined as a half-normal distribution centered at 0 with a small standard deviation  $\phi$ ,

$$g(\sigma) \sim HalfNormal(\phi = 0.1). \quad (8.44)$$

With the priors and likelihood functions in hand, we sampled the posterior distribution using Markov chain Monte Carlo as implemented in the Stan probabilistic programming language (Carpenter et al., 2017). We performed three different estimations – one inferring the parameters using only data at 32° C, one using only data from 42° C, and one using data sets from both temperatures pooled together.

The sampling results can be seen in Fig. 8.10. The estimation of  $\Delta S_R$  is distinct for each condition where as the sampling for  $\Delta S_{AI}$  is the same for all conditions and is centered about 0. The latter suggests that the value of  $\Delta\varepsilon_{AI}$  determined at 37° C is not dependent on temperature within the resolution of our experiments. The difference between the estimated value of  $\Delta S_R$  between temperatures suggests that there is another component of the temperature dependence that is not captured by

the inclusion of a single entropic parameter. Fig. 8.11 shows that estimating  $\Delta S$  from one temperature is not sufficient to predict the fold-change in gene expression at another temperature. The addition of the entropic parameter leads to better fit of the 32° C condition than the simple rescaling of the energy as described by Eq. 8.35 [Fig. 8.11, dashed line], but poorly predicts the behavior at 42° C. Performing the inference on the combined 32° C and 42° C data strikes a middle ground between the predictions resulting from the two temperatures alone Fig. 8.11.

### Entropy as a function of temperature

In the previous section, we made an approximation of the energetic parameters  $\Delta\varepsilon_R$  and  $\Delta\varepsilon_{AI}$  to be defined by an enthalpic and entropic term, both of which being independent of temperature. However, entropy can be (and in many cases is) dependent on the system temperature, often in a non-trivial manner.

To explore the effects of a temperature-dependent entropy in our prediction of the fold-change, we perform a Taylor expansion of Eq. 8.5 about the entropic parameter with respect to temperature keeping only the first order term such that

$$\Delta S = \Delta S_0 + \Delta S_1 T, \quad (8.45)$$

where  $\Delta S_0$  is a constant, temperature-independent entropic term and  $\Delta S_1$  is the entropic contribution per degree Kelvin. With this simple relationship enumerated, we can now define the temperature-dependent effective free energy parameters  $\Delta\varepsilon_R^*$  and  $\Delta\varepsilon_{AI}^*$  as

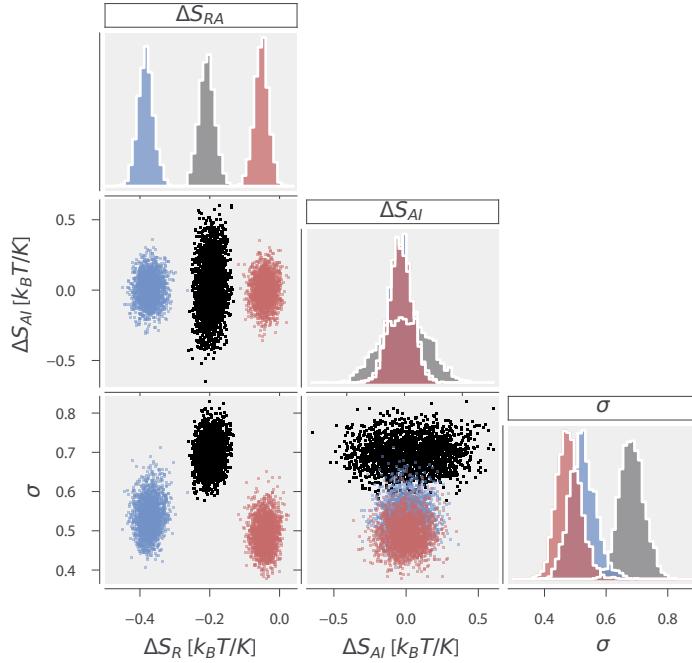
$$\Delta\varepsilon_R^* = (S_{0R} + S_{1R}T_{exp})(T_{ref} - T_{exp}) - 13.9k_B T, \quad (8.46)$$

and

$$\Delta\varepsilon_{AI}^* = (S_{0AI} + S_{1AI}T_{exp})(T_{ref} - T_{exp}) + 4.5k_B T, \quad (8.47)$$

respectively, again relying on the constraints defined by Eq. 8.40 and Eq. 8.41.

Using a similar inferential approach as described in the previous section, we sample the posterior distribution of these parameters using Markov chain Monte Carlo and compute the fold-change and shift in free energy for each temperature.



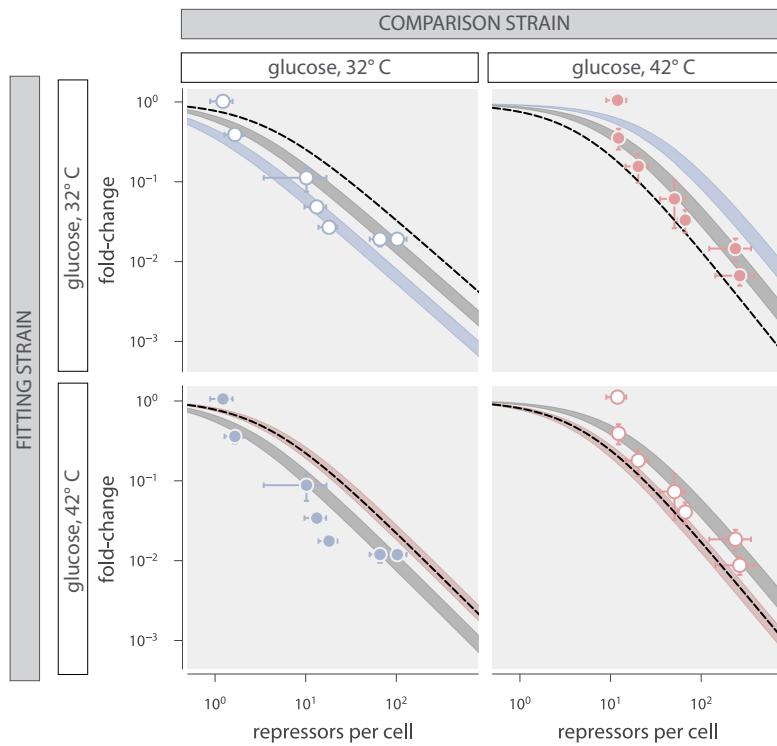
**Figure 8.10: Sampled posterior probability distributions of entropic penalty parameter inference.** Marginal and joint distributions conditioned only on data collected at 32° C, only on 42 ° C, or on both temperatures are shown in blue, red, and black, respectively. The value  $\Delta S_R$  and  $\Delta S_{AI}$  are given in  $k_B T/K$  where K is 1 degree Kelvin.

As seen in Fig. 8.12, there is a negligible improvement in the description of the data by including this temperature dependent entropic parameter.

These results together suggest that our understanding of temperature dependence in this regulatory architecture is incomplete and requires further research from both theoretical and experimental standpoints.

## 8.6 Media Recipes and Bacterial Strains

The primary interest in varying the available carbon source in growth media was to modulate the quality of the carbon rather than the quantity. With this in mind, we developed the various growth media to contain the same net number of carbon atoms per cell. The standard reference was 0.5% (w/v) glucose (Garcia and Phillips, 2011), which results in  $10^8$  carbon atoms per  $10^{-15}$  L. The base recipe is given in Table 8.2. The bacterial strains used in this work are given in Table 8.3.



**Figure 8.11: Pairwise predictions of fold-change in gene expression at different temperatures.** Each row represents the condition used to infer the entropic parameters and columns are the conditions that are being predicted. color shaded regions represent the 95% credible region of the predicted fold-change given the inferred values of  $\Delta S_R$  and  $\Delta S_{AI}$  for each condition. The black dashed line represents the predicted fold-change in gene expression by a simple rescaling of the binding energy determined at 37°C. The grey shaded region is the 95% credible region of the fold-change given estimation of  $\Delta S_R$  and  $\Delta S_{AI}$  conditioned on both temperatures pooled together.

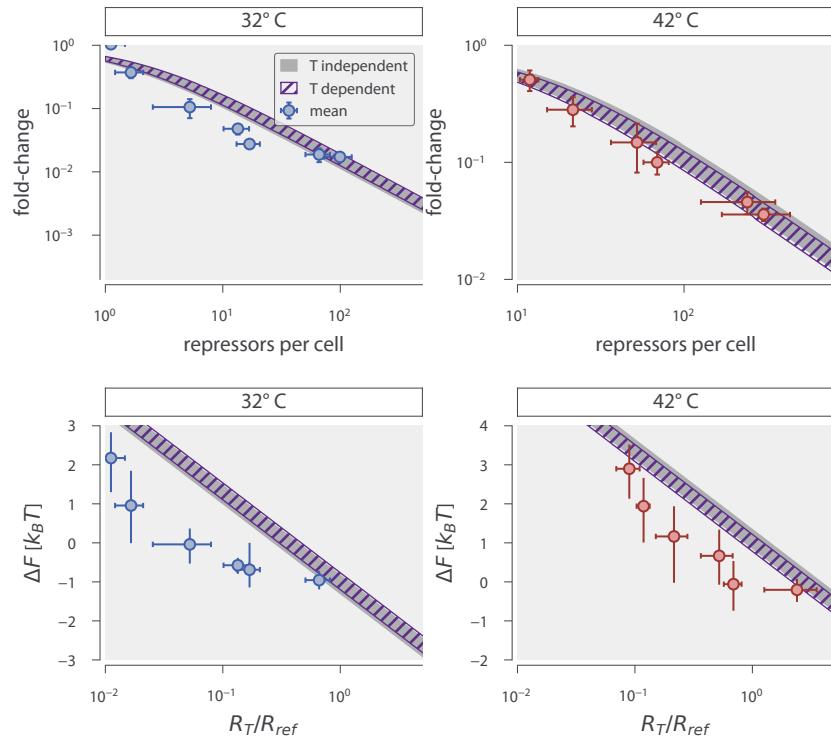
Table 8.2: M9 minimal medium recipe for each carbon-supplemented medium.

Ingredient	Stock Concentration	Volume	Final Concentration
ddH <sub>2</sub> O	-	773 mL	-
CaCl <sub>2</sub>	1M	100 μL	100 μM
MgSO <sub>4</sub>	1M	2mL	2mM
M9 Salts	5X	200 mL	-
(BD Medical, Cat. No. 248510)			

Ingredient	Stock Concentration	Volume	Final Concentration
<i>Carbon Source</i>			$10^8$ C / fL
Glucose	20% (w/v)	25 mL	0.5% (w/v)
Glycerol	20% (w/v)	25 mL	0.5% (w/v)
Acetate	20% (w/v)	25 mL	0.5% (w/v)

Table 8.3: Bacterial strains used in various physiological states.

Genotype	Plasmid	Notes
MG1655: $\Delta$ lacZYA; <i>intC</i> <>4*CFP	–	Autofluorescence control
MG1655: $\Delta$ lacZYA; <i>intC</i> <>4*CFP <i>galK</i> <>25-O2+11-YFP	–	Constitutive expression control
MG1655: $\Delta$ lacZYA; <i>intC</i> <>4*CFP <i>galK</i> <>25-O2+11-YFP <i>ybcN</i> <>1-lacI( $\Delta$ 353-363)-mCherry	pZS3P <sub>N25</sub> -tetR	Strain with ATC inducible lacI-mCherry



**Figure 8.12: Fold-change and shift in free energy including a temperature-dependent entropic contribution.** Top row illustrates the estimated fold-change in gene expression at 32° C (left) and 42° C (right). Bottom row shows the estimated shift in free energy for each temperature. The grey transparent line in all plots shows the relevant quantity including only a temperature-independent entropic parameter. Purple hashed line illustrates the relevant quantity including a temperature-dependent entropic contribution. The width of each curve indicates the 95% credible region of the relevant quantity. Points in each correspond to the experimental measurements. The points in the top row represent the mean and standard error of five to eight biological replicates. Points in the bottom row correspond to the median value of the inferred free energy shift and the mean of five to eight biological replicates for the repressor copy number. Vertical error represents the upper and lower bounds of the 95% credible region of the parameter. Horizontal error bars correspond to the standard error of five to eight biological replicates.

## SUPPLEMENTAL INFORMATION FOR CHAPTER 5: HOW BACTERIA ADAPT TO CHANGES IN OSMOLARITY

A version of this chapter was published as Chure, G.\* , Lee, H.J.\* , Rasmussen, A., and Phillips, R. (2018). *Connecting the Dots between Mechanosensitive Channel Abundance, Osmotic Shock, and Survival at Single-Cell Resolution*. Journal of Bacteriology 200. (\* contributed equally). G.C., H.J.L, and R.P. designed and planned experiments. G.C. and H.J.L performed experiments. H.J.L constructed bacterial strains. A.R. performed electrophysiology experiments. G.C. performed data analysis and figure generation. G.C. and R.P. wrote the manuscript. ## Experimental validation of MscL-sfGFP Despite revolutionizing modern cell biology, tagging proteins with fluorophores can lead to myriad deleterious effects such as mislocalization, abrogation of functionality, or even cytotoxicity. In this section, we examine the stability and functionality of the MscL-sfGFP construct used in this work.

### Comparing functionality of wild-type and fluorescently tagged MscL

To quantitatively compare the functionality between the wild-type MscL and MscL-sfGFP, patch-clamp electrophysiology experiments were conducted on each channel. Patch-clamp recordings were performed on membrane patches derived from giant protoplasts which were prepared as previously described (Blount et al., 1999). In brief, cells were grown in Luria-Bertani (LB) medium with 0.06 mg/ml cephalexin for 2.5 hours. The elongated cells were then collected by centrifugation and digested by 0.2 mg/ml lysozyme to form giant protoplasts.

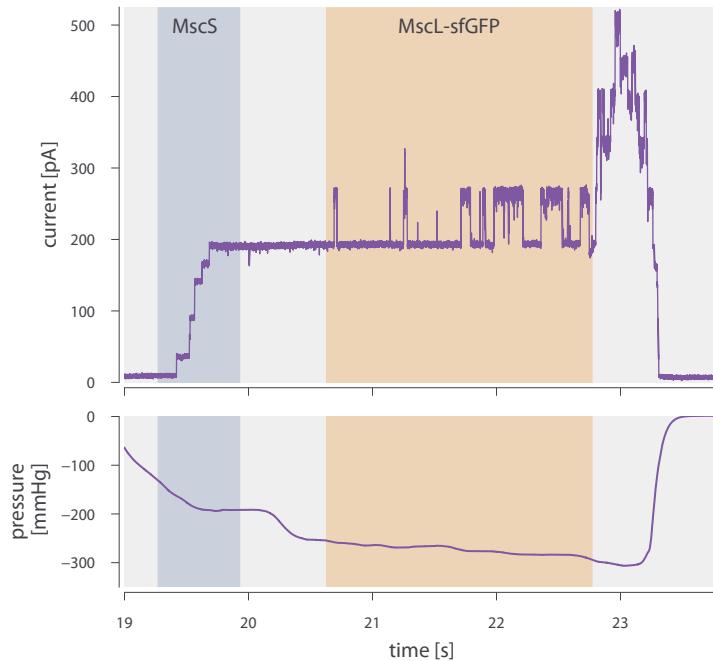
Excised, inside-out patches were analyzed at a membrane potential of -20 mV with pipette and bath solutions containing 200 mM KCl, 90 mM MgCl<sub>2</sub>, 10 mM CaCl<sub>2</sub>, and 5 mM HEPES buffer at pH 7. All data were acquired at a sampling rate of 50 kHz with 5 kHz filtration using an AxoPatch 200B amplifier and pClamp software (Molecular Devices). The pressure threshold for activation a single MscS

channel (blue stripe in Fig. 9.1) was compared to that of single MscL channels (yellow strip in Fig. 9.1). The pressure threshold for activation of the MscL channels was referenced against the activation threshold of MscS to determine the pressure ratio (PL:PS) for gating as previously described (Blount et al., 1996). Recordings of the transmembrane current were made of three individual patches with an average PL:PS ratio of 1.56 for MscL-sfGFP. This ratio quantitatively agrees with the PL:PS ratio of 1.54 measured in a strain (MJF429 from the Booth laboratory) which expresses the wild-type MscL protein from the chromosome. The average transient current change from MscL openings (Fig. S1 shaded yellow region ) is 75 pA, corresponding to a single channel conductance of 3.7 nS, comparable to the reported values of wild-type MscL. The agreement between these two strains indicates that there is negligible difference in functionality between MscL and MscL-sfGFP, allowing us to make physiological conclusions of the wild-type channel from our experiments.

### **Maturation time of MscL-sfGFP**

Reliable quantification of the channel copy number is paramount to this work. As such, it is important to verify that the detected fluorescence per cell accurately represents the total cellular MscL copy number. We have made the assumption that the total fluorescence per represents all MscL-sfGFP channels present. However, it is possible that there are more channels present per cell but are not detected as the fluorophores have not properly matured. This potential error becomes more significant with longer maturation times of the fluorophore as the mean expression level changes with the growth phase of the culture. With a maturation time much longer than the typical cell division time, it is possible that the measured channel copy number represents only a fraction of the total number inherited over generations.

In our earlier work, we quantified the MscL-sfGFP channel copy number using fluorescence microscopy as well as with quantitative Western blotting. We found that these two methods agreed within 20% of the mean value, often with



**Figure 9.1: Characteristic MscL-sfGFP conductance obtained through patch-clamp electrophysiology.** Top panel presents a characteristic measurement of channel current obtained through a patch-clamp electrophysiology measurement of bacterial protoplasts. The bottom panel shows the applied pressure through the micropipette to facilitate opening of the mechanosensitive channels. The blue shaded region indicates opening of the mechanosensitive channel of small conductance (MscS). The shaded yellow region represents opening of single MscL channels. These regions were used to compute the PL:PS ratio.

the counts resulting from microscopy being slightly larger than those measured through Western blotting (Bialecka-Fornal et al., 2012). This strongly suggests that a negligible amount of channels are not observed due to inactive fluorophores.

Despite these suggestive data, we directly measured the maturation time of the superfolder GFP protein. We constructed a chromosomal integration of sfGFP expressed from a promoter under regulation from plasmid-borne TetR (*E. coli* MG1655 K12  $\Delta lacIZYA$   $ybcN::sfGFP$ ). These cells were allowed to grow in LB supplemented with 500 mM NaCl held at 37°C to an OD<sub>600nm</sub> of approximately 0.3. At this time, transcription and translation of the sfGFP gene was induced by addition of 10 ng/mL of anhydrous tetracycline. This expression was allowed to occur for three minutes before the addition of 100  $\mu$ g/mL of kanamycin, ceasing proper protein

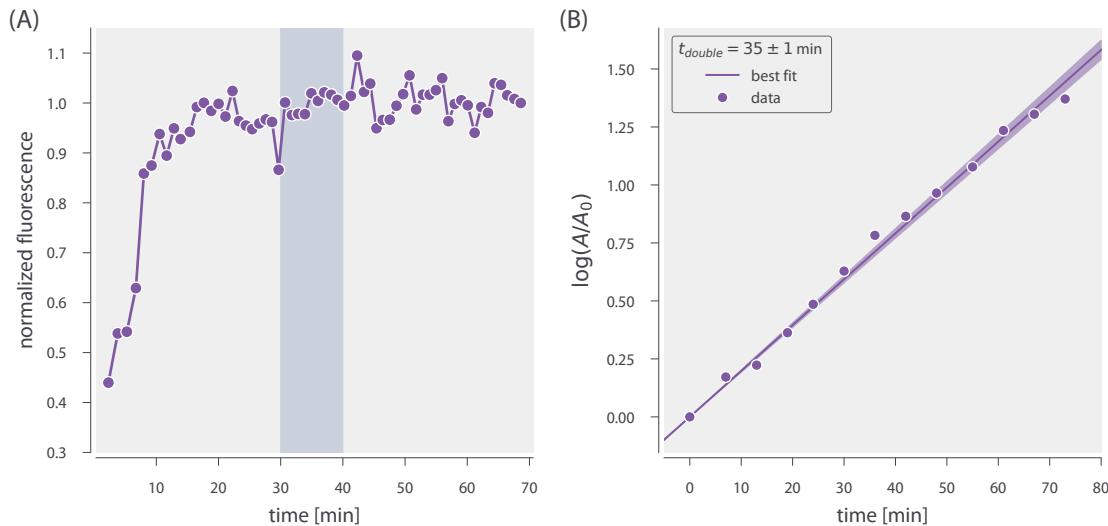
synthesis. Three minutes of expression was chosen to provide enough time for transcription and translation. The sfGFP variant used in this work is 1155 base pairs. We can assume that the rate for transcription is 42 nucleotides per second (BNID 108488)(Milo et al., 2010), meaning approximately 28 seconds are needed to transcribe the gene. The translation rate is on the order of 10 amino acids per second, (12 - 42 amino acids / s, BNID 100059)(Milo et al., 2010). This means that 39 seconds are needed to complete translation. In total, approximately one minute is needed to complete expression of the genes. These numbers are not known for LB supplemented with 500 mM NaCl but may be reduced. For this reason, we extended the length of induction to three minutes before translation was ceased.

The excess anhydrous tetracycline was removed from the culture through centrifugation and washing with one volume of LB supplemented with 500 mM NaCl and 100  $\mu$ g/mL kanamycin at 37°C. The maturation of sfGFP was then monitored through flow cytometry by measuring the mean expression of 100,000 cells every 60 to 90 seconds. The result of these measurements are shown in Fig. 9.2.

We observe complete maturation of the protein within 20 minutes after translation of the sfGFP gene was ceased. While the growth rate in LB + 500mM NaCl varies depending on the expression of MscL-sfGFP, we typically observe doubling times between 30 and 40 minutes, as indicated by a yellow stripe in Fig. 9.2A. To examine the “best case” scenario for cell growth in this medium, we measured the growth rate of the same *E. coli* strain used to measure the fluorophore maturation time (Fig. 9.2 B). We observed a doubling time of  $35 \pm 1$  min, which falls in the middle of the yellow stripe shown in Fig. 9.2 A. These data, coupled with our previous quantification of MscL copy number using independent methods, suggests that the fluorescence measurements made in this work reflect the total amount of MscL protein expressed.

## 9.1 Calibration of a Standard Candle

To estimate the single-cell MscL abundance via microscopy, we needed to determine a calibration factor that could translate arbitrary fluorescence units to protein



**Figure 9.2: Measurement of sfGFP maturation as a function of time through flow cytometry.** (A) Measurement of sfGFP fluorescence intensity as a function of time after cessation of protein translation. Points and connected lines indicate means of gated flow cytometry intensity distributions. Yellow stripe indicates the range of doubling times observed for the various RBS mutant strains described in this work (B) Growth curve of *E. coli* MG1655 cells in LB + 500mM NaCl. Red points indicate individual absorbance measurements. Line of best fit is shown in black with the uncertainty shown in shaded gray. The measured doubling time was  $35 \pm 1 \text{ min}$ .

copy number. To compute this calibration factor, we relied on *a priori* knowledge of the mean copy number of MscL-sfGFP for a particular bacterial strain in specific growth conditions. In Bialecka-Fornal et al. 2012 (Bialecka-Fornal et al., 2012), the average MscL copy number for a population of cells expressing an MscL-sfGFP fusion (*E. coli* K-12 MG1655  $\phi(mscL\text{-}sfGFP)$ ) cells was measured using quantitative Western blotting and single-molecule photobleaching assays. By growing this strain in identical growth and imaging conditions, we can make an approximate measure of this calibration factor. In this section, we derive a statistical model for estimating the most-likely value of this calibration factor and its associated error.

### Definition of a calibration factor

We assume that all detected fluorescence signal from a particular cell is derived from the MscL-sfGFP protein, after background subtraction and correction for auto-

ofluorescence. The arbitrary units of fluorescence can be directly related to the protein copy number via a calibration factor  $\alpha$ ,

$$I_{\text{tot}} = \alpha N_{\text{tot}}, \quad (9.1)$$

where  $I_{\text{tot}}$  is the total cell fluorescence and  $N_{\text{tot}}$  is the total number of MscL proteins per cell. Bialecka-Fornal et al. (2012) et al. report the average cell Mscl copy number for the population rather than the distribution. Knowing only the mean, we can rewrite Eq. 9.1 as

$$\langle I_{\text{tot}} \rangle = \alpha \langle N_{\text{tot}} \rangle, \quad (9.2)$$

assuming that  $\alpha$  is a constant value that does not change from cell to cell or fluorophore to fluorophore.

The experiments presented in this work were performed using non-synchronously growing cultures. As there is a uniform distribution of growth phases in the culture, the cell size distribution is broad the the extremes being small, newborn cells and large cells in the process of division. As described in the main text, the cell size distribution of a population is broadened further by modulating the MscL copy number with low copy numbers resulting in aberrant cell morphology. To speak in the terms of an effective channel copy number, we relate the average areal intensity of the population to the average cell size,

$$\langle I_{\text{tot}} \rangle = \langle I_A \rangle \langle A \rangle = \alpha \langle N_{\text{tot}} \rangle, \quad (9.3)$$

where  $\langle I_A \rangle$  is the average areal intensity in arbitrary units per pixel of the population and  $\langle A \rangle$  is the average area of a segmented cell. As only one focal plane was imaged in these experiments, we could not compute an appropriate volume for each cell given the highly aberrant morphology. We therefore opted to use the projected two-dimensional area of each cell as a proxy for cell size. Given this set of measurements, the calibration factor can be computed as

$$\alpha = \frac{\langle I_A \rangle \langle A \rangle}{\langle N_{\text{tot}} \rangle}. \quad (9.4)$$

While it is tempting to use Eq. 9.4 directly, there are multiple sources of error that are important to propagate through the final calculation. The most obvious error to include is the measurement error reported in Bialecka-Fornal et al. 2012 for the average MscL channel count (Bialecka-Fornal et al., 2012). There are also slight variations in expression across biological replicates that arise from a myriad of day-to-day differences. Rather than abstracting all sources of error away into a systematic error budget, we used an inferential model derived from Bayes' theorem that allows for the computation of the probability distribution of  $\alpha$ .

### Estimation of $\alpha$ for a single biological replicate

A single data set consists of several hundred single-cell measurements of intensity, area of the segmentation mask, and other morphological quantities. The areal density  $I_A$  is computed by dividing the total cell fluorescence by the cell area  $A$ . We are interested in computing the probability distributions for the calibration factor  $\alpha$ , the average cell area  $\langle A \rangle$ , and the mean number of channels per cell  $\langle N_{\text{tot}} \rangle$  for the data set as a whole given only  $I_A$  and  $A$ . Using Bayes' theorem, the probability distribution for these parameters given a single cell measurement, hereafter called the posterior distribution, can be written as

$$g(\alpha, \langle A \rangle, \langle N_{\text{tot}} \rangle | A, I_A) = \frac{f(A, I_A | \alpha, \langle A \rangle, \langle N_{\text{tot}} \rangle)g(\alpha, \langle A \rangle, \langle N_{\text{tot}} \rangle)}{f(\alpha, I_A)}, \quad (9.5)$$

where  $g$  and  $f$  represent probability density functions over parameters and data, respectively. The term  $f(A, I_A | \alpha, \langle A \rangle, \langle N_{\text{tot}} \rangle)$  in the numerator represents the likelihood of observing the areal intensity  $I_A$  and area  $A$  of a cell for a given values of  $\alpha$ ,  $\langle A \rangle$ , and  $\langle N_{\text{tot}} \rangle$ . The second term in the numerator  $g(\alpha, \langle A \rangle, \langle N_{\text{tot}} \rangle)$  captures all prior knowledge we have regarding the possible values of these parameters knowing nothing about the measured data. The denominator,  $f(I_A, A)$  captures the probability of observing the data knowing nothing about the parameter values. This term, in our case, serves simply as a normalization constant and is neglected for the remainder of this section.

To determine the appropriate functional form for the likelihood and prior, we

must make some assumptions regarding the biological processes that generate them. As there are many independent processes that regulate the timing of cell division and cell growth, such as DNA replication and peptidoglycan synthesis, it is reasonable to assume that for a given culture the distribution of cell size would be normally distributed with a mean of  $\langle A \rangle$  and a variance  $\sigma_{\langle A \rangle}$ . Mathematically, we can write this as

$$f(A | \langle A \rangle, \sigma_{\langle A \rangle}) \propto \frac{1}{\sigma_{\langle A \rangle}} \exp \left[ -\frac{(A - \langle A \rangle)^2}{2\sigma_{\langle A \rangle}^2} \right], \quad (9.6)$$

where the proportionality results from dropping normalization constants for notational simplicity.

While total cell intensity is intrinsically dependent on the cell area the areal intensity  $I_A$  is independent of cell size. The myriad processes leading to the detected fluorescence, such as translation and proper protein folding, are largely independent, allowing us to assume a normal distribution for  $I_A$  as well with a mean  $\langle I_A \rangle$  and a variance  $\sigma_{I_A}^2$ . However, we do not have knowledge of the average areal intensity for the standard candle strain *a priori*. This can be calculated knowing the calibration factor, total MscL channel copy number, and the average cell area as

$$I_A = \frac{\alpha \langle N_{\text{tot}} \rangle}{\langle A \rangle}. \quad (9.7)$$

Using Eq. 9.7 to calculate the expected areal intensity for the population, we can write the likelihood as a Gaussian distribution,

$$f(I_A | \alpha, \langle A \rangle, \langle N_{\text{tot}} \rangle, \sigma_{I_A}) \propto \frac{1}{\sigma_{I_A}} \exp \left[ -\frac{\left( I_A - \frac{\alpha \langle N_{\text{tot}} \rangle}{\langle A \rangle} \right)^2}{2\sigma_{I_A}^2} \right]. \quad (9.8)$$

With these two likelihoods in hand, we are tasked with determining the appropriate priors. As we have assumed normal distributions for the likelihoods of  $\langle A \rangle$  and  $I_A$ , we have included two additional parameters,  $\sigma_{\langle A \rangle}$  and  $\sigma_{I_A}$ , each requiring their own prior probability distribution. It is common practice to assume maximum ignorance for these variances and use a Jeffreys prior (Sivia and Skilling,

2006),

$$g(\sigma_{\langle A \rangle}, \sigma_{I_A}) = \frac{1}{\sigma_{\langle A \rangle} \sigma_{I_A}}. \quad (9.9)$$

The next obvious prior to consider is for the average channel copy number  $\langle N_{\text{tot}} \rangle$ , which comes from Bialecka-Fornal et al. 2012. In this work, they report a mean  $\mu_N$  and variance  $\sigma_N^2$ , allowing us to assume a normal distribution for the prior,

$$g(\langle N_{\text{tot}} \rangle | \mu_N, \sigma_N) \propto \frac{1}{\sigma_N} \exp \left[ -\frac{(\langle N_{\text{tot}} \rangle - \mu_N)^2}{2\sigma_N^2} \right]. \quad (9.10)$$

For  $\alpha$  and  $\langle A \rangle$ , we have some knowledge of what these parameters can and cannot be. For example, we know that neither of these parameters can be negative. As we have been careful to not overexpose the microscopy images, we can say that the maximum value of  $\alpha$  would be the bit-depth of our camera. Similarly, it is impossible to segment a single cell with an area larger than our camera's field of view (although there are biological limitations to size below this extreme). To remain maximally uninformative, we can assume that the parameter values are uniformly distributed between these bounds, allowing us to state

$$g(\alpha) = \begin{cases} \frac{1}{\alpha_{\max} - \alpha_{\min}} & \alpha_{\min} \leq \alpha \leq \alpha_{\max} \\ 0 & \text{otherwise} \end{cases}, \quad (9.11)$$

for  $\alpha$  and

$$g(\langle A \rangle) = \begin{cases} \frac{1}{\langle A \rangle_{\max} - \langle A \rangle_{\min}} & \langle A \rangle_{\min} \leq \langle A \rangle \leq \langle A \rangle_{\max} \\ 0 & \text{otherwise} \end{cases} \quad (9.12)$$

for  $\langle A \rangle$ .

Piecing Eq. 9.6 through Eq. 9.12 together generates a complete posterior probability distribution for the parameters given a single cell measurement. This can be

generalized to a set of  $k$  single cell measurements as

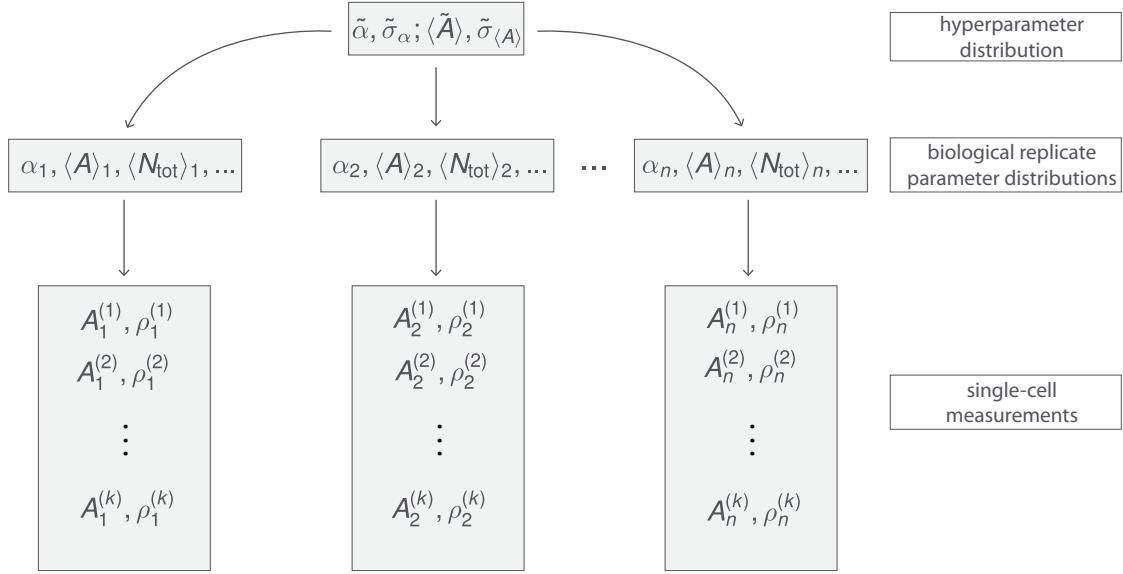
$$g(\alpha, \langle A \rangle, \langle N_{\text{tot}} \rangle, \sigma_{I_A}, \sigma_{\langle A \rangle} \mid [I_A, A], \mu_N, \sigma_N) \propto \frac{1}{(\alpha_{\max} - \alpha_{\min})(\langle A \rangle_{\max} - \langle A \rangle_{\min})} \frac{1}{(\sigma_{I_A} \sigma_{\langle A \rangle})^{k+1}} \times \\ \frac{1}{\sigma_N} \exp \left[ -\frac{(\langle N_{\text{tot}} \rangle - \mu_N)^2}{2\sigma_N^2} \right] \prod_i^k \exp \left[ -\frac{(A^{(i)} - \langle A \rangle)^2}{2\sigma_{\langle A \rangle}^2} - \frac{\left( I_A^{(i)} - \frac{\alpha \langle N_{\text{tot}} \rangle}{\langle A \rangle} \right)^2}{2\sigma_{I_A}^2} \right] \quad (9.13)$$

where  $[I_A, A]$  represents the set of  $k$  single-cell measurements.\*\*\*\*

As small variations in the day-to-day details of cell growth and sample preparation can alter the final channel count of the standard candle strain, it is imperative to perform more than a single biological replicate. However, properly propagating the error across replicates is non trivial. One option would be to pool together all measurements of  $n$  biological replicates and evaluate the posterior given in Eq. 9.13. However, this by definition assumes that there is no difference between replicates. Another option would be to perform this analysis on each biological replicate individually and then compute a mean and standard deviation of the resulting most-likely parameter estimates for  $\alpha$  and  $\langle A \rangle$ . While this is a better approach than simply pooling all data together, it suffers a bias from giving each replicate equal weight, skewing the estimate of the most-likely parameter value if one replicate is markedly brighter or dimmer than the others. Given this type of data and a limited number of biological replicates ( $n = 6$  in this work), we chose to extend the Bayesian analysis presented in this section to model the posterior probability distribution for  $\alpha$  and  $\langle A \rangle$  as a hierarchical process in which  $\alpha$  and  $\langle A \rangle$  for each replicate is drawn from the same distribution.

### A hierarchical model for estimating $\alpha$

In the previous section, we assumed maximally uninformative priors for the most-likely values of  $\alpha$  and  $\langle A \rangle$ . While this is a fair approach to take, we are not completely ignorant with regard to how these values are distributed across biological replicates. A major assumption of our model is that the most-likely value



**Figure 9.3: Schematic of hierarchical model structure.** The hyper-parameter probability distributions (top panel) are used as an informative prior for the most-likely parameter values for each biological replicate (middle panel). The single-cell measurements of cell area and areal intensity (bottom panel) are used as data in the evaluation of the likelihood.

of  $\alpha$  and  $\langle A \rangle$  for each biological replicate are comparable, so long as the experimental error between them is minimized. In other words, we are assuming that the most-likely value for each parameter for each replicate is drawn from the same distribution. While each replicate may have a unique value, they are all related to one another. Unfortunately, proper sampling of this distribution requires an extensive amount of experimental work, making inferential approaches more attractive.

This approach, often called a multi-level or hierarchical model, is schematized in Fig. 9.3. Here, we use an informative prior for  $\alpha$  and  $\langle A \rangle$  for each biological replicate. This informative prior probability distribution can be described by summary statistics, often called hyper-parameters, which are then treated as the “true” value and are used to calculate the channel copy number. As this approach allows us to get a picture of the probability distribution of the hyper-parameters, we are able to report a point estimate for the most-likely value along with an error estimate that captures all known sources of variation.

The choice for the functional form for the informative prior is often not obvi-

ous and can require other experimental approaches or back-of-the-envelope estimates to approximate. Each experiment in this work was carefully constructed to minimize the day-to-day variation. This involved adhering to well-controlled growth temperatures and media composition, harvesting of cells at comparable optical densities, and ensuring identical imaging parameters. As the experimental variation is minimized, we can use our knowledge of the underlying biological processes to guess at the approximate functional form. For similar reasons presented in the previous section, cell size is controlled by a myriad of independent processes. As each replicate is independent of another, it is reasonable to assume a normal distribution for the average cell area for each replicate. This normal distribution is described by a mean  $\langle \tilde{A} \rangle$  and variance  $\tilde{\sigma}_{\langle A \rangle}$ . Therefore, the prior for  $\langle A \rangle$  for  $n$  biological replicates can be written as

$$g(\langle A \rangle | \langle \tilde{A} \rangle, \tilde{\sigma}_{\langle A \rangle}) \propto \frac{1}{\tilde{\sigma}_{\langle A \rangle}^n} \prod_{j=1}^n \exp \left[ -\frac{(\langle A \rangle_j - \langle \tilde{A} \rangle)^2}{2\tilde{\sigma}_{\langle A \rangle}^2} \right]. \quad (9.14)$$

In a similar manner, we can assume that the calibration factor for each replicate is normally distributed with a mean  $\tilde{\alpha}$  and variance  $\tilde{\sigma}_\alpha$ ,

$$g(\alpha | \tilde{\alpha}, \tilde{\sigma}_\alpha) \propto \frac{1}{\tilde{\sigma}_\alpha^n} \prod_{j=1}^n \exp \left[ -\frac{(\alpha_j - \tilde{\alpha})^2}{2\tilde{\sigma}_\alpha^2} \right]. \quad (9.15)$$

With the inclusion of two more normal distributions, we have introduced four new parameters, each of which needing their own prior. However, our knowledge of the reasonable values for the hyper-parameters has not changed from those described for a single replicate. We can therefore use the same maximally uninformative Jeffreys priors given in Eq. 9.9 for the variances and the uniform distributions given in Eq. 9.11 and Eq. 9.12 for the means. Stitching all of this work together generates the full posterior probability distribution for the best-estimate of  $\tilde{\alpha}$  and

$\langle \tilde{A} \rangle$  shown in Eq. 9.2 given  $n$  replicates of  $k$  single cell measurements,

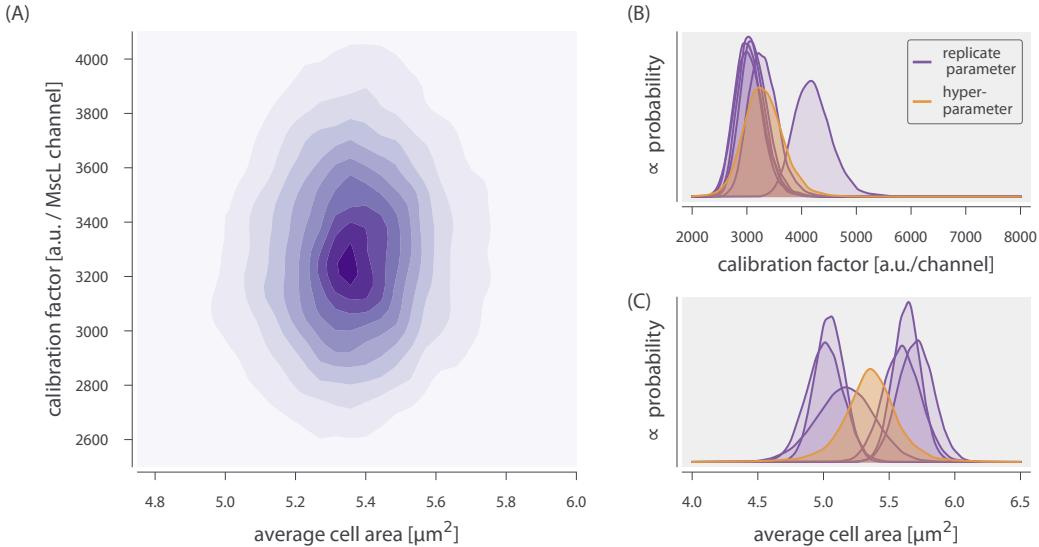
$$\begin{aligned}
g(\tilde{\alpha}, \tilde{\sigma}_\alpha, \langle \tilde{A} \rangle, \tilde{\sigma}_{\langle A \rangle}, \{ \langle N_{\text{tot}} \rangle, \langle A \rangle, \alpha, \sigma_{I_A} \} | [I_A, A], \mu_N, \sigma_N) \propto \\
\frac{1}{(\tilde{\alpha}_{\max} - \tilde{\alpha}_{\min})(\langle \tilde{A} \rangle_{\max} - \langle \tilde{A} \rangle_{\min})\sigma_N^n (\tilde{\sigma}_\alpha \tilde{\sigma}_{\langle A \rangle})^{n+1}} \times \\
\prod_{j=1}^n \exp \left[ -\frac{(\langle N \rangle_j^{(i)} - \mu_N)^2}{2\sigma_N^2} - \frac{(\alpha_j - \tilde{\alpha})^2}{2\tilde{\sigma}_\alpha^2} - \frac{(\langle A \rangle_j - \langle \tilde{A} \rangle)^2}{2\tilde{\sigma}_{\langle A \rangle}^2} \right] \times \\
\frac{1}{(\sigma_{I_{A_j}} \sigma_{\langle A \rangle_j})^{k_j+1}} \prod_{i=1}^{k_j} \exp \left[ -\frac{(A_j^{(i)} - \langle A \rangle_j)^2}{2\sigma_{\langle A \rangle_j}^{(i)2}} - \frac{\left( I_{A_j}^{(i)} - \frac{\alpha_j \langle N_{\text{tot}} \rangle_j}{\langle A \rangle_j} \right)^2}{2\sigma_{I_{A_j}}^{(i)2}} \right]
\end{aligned} \tag{9.16}$$

where the braces  $\{ \dots \}$  represent the set of parameters for biological replicates and the brackets  $[ \dots ]$  correspond to the set of single-cell measurements for a given replicate.

While Eq. 9.16 is not analytically solvable, it can be easily sampled using Markov chain Monte Carlo (MCMC). The results of the MCMC sampling for  $\tilde{\alpha}$  and  $\langle \tilde{A} \rangle$  can be seen in Fig. 9.4. From this approach, we found the most-likely parameter values of  $3300^{+700}_{-700}$  a.u. per MscL channel and  $5.4^{+0.4}_{-0.5} \mu\text{m}^2$  for  $\tilde{\alpha}$  and  $\langle \tilde{A} \rangle$ , respectively. Here, we've reported the median value of the posterior distribution for each parameter with the upper and lower bound of the 95% credible region as superscript and subscript, respectively. These values and associated errors were used in the calculation of channel copy number.

### Effect of correction

The posterior distributions for  $\alpha$  and  $\langle A \rangle$  shown in Fig. ?? were used directly to compute the most-likely channel copy number for each measurement of the Shine-Dalgarno mutant strains, as is described in the coming section. The importance of this correction can be seen in Fig. 9.5. Cells with low abundance of MscL channels exhibit notable morphological defects, as illustrated in Fig. 9.5 (A). While these would all be considered single cells, the two-dimensional area of each may be comparable to two or three wild-type cells. For all of the Shine-Dalgarno mutants, the distribution of projected cell area has a long tail, with the extremes reaching  $35 \mu\text{m}^2$

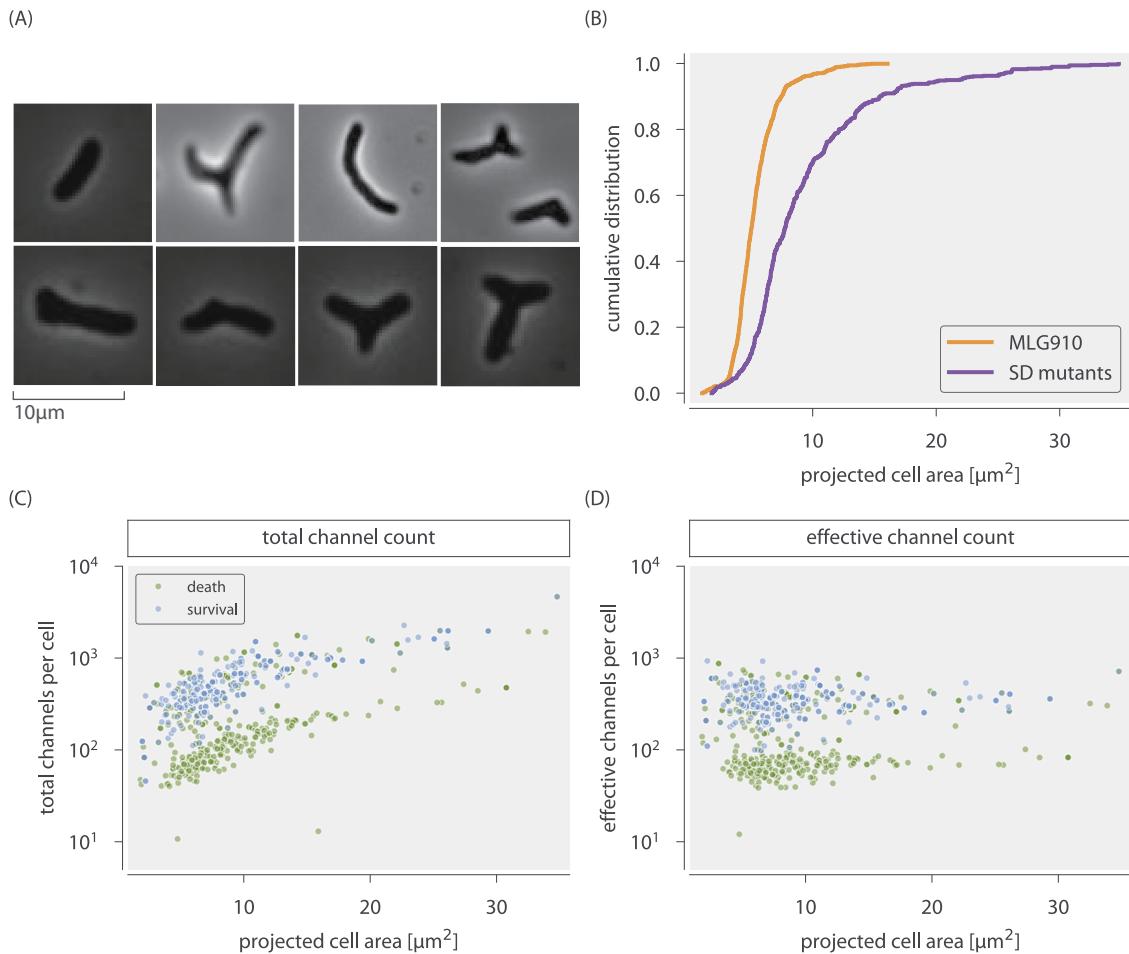


**Figure 9.4: Posterior distributions for hyper-parameters and replicate parameters.** (A) The posterior probability distribution for  $\tilde{\alpha}$  and  $\langle \tilde{A} \rangle$ . Probability increases from light to dark red. The replicate parameter (blue) and hyper-parameter (red) marginalized posterior probability distributions for  $\alpha$  (B) and  $\langle A \rangle$  (C).

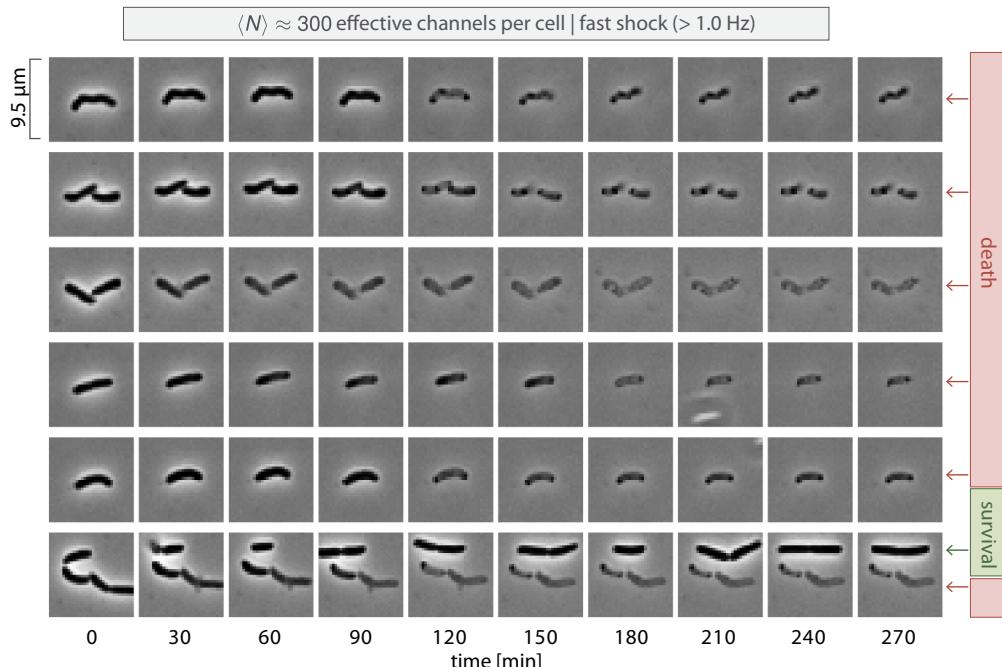
per cell Fig. 9.5. Calculating the total number of channels per cell does nothing to decouple this correlation between cell area and measured cell intensity. Fig. 9.5 (C) shows the correlation between cell area and the total number of channels without normalizing to an average cell size  $\langle A \rangle$  differentiated by their survival after an osmotic down-shock. This correlation is removed by calculating an effective channel copy number shown in Fig. 9.5 (D).

## 9.2 Classification of Cell Fates

We defined a survival event as a cell that went on to divide at least twice in the several hours following the applied osmotic shock. In nearly all of our experiments, cells which did not survive an osmotic shock exhibited necrosis with loss of phase contrast, extensive blebbing and bursting of the membrane, and the presence of dark aggregates at the cell poles. An example field across time is shown below in Fig. 9.6 where the cells are necrotic. On occasion, we observed cells which did not obviously display the aforementioned death criteria yet did not undergo one or two division events. These cells were not counted in our experiments and



**Figure 9.5: Influence of area correction for Shine-Dalgarno mutants.** (A) Representative images of aberrant cell morphologies found in low-expressing Shine-Dalgarno mutants. (B) Empirical cumulative distribution of two-dimensional projected cell area for the standard candle strain MLG910 (gray line) and for all Shine-Dalgarno mutants (red line). (C) The correlation between channel copy number and cell area without the area correction. (D) The correlation between effective channel copy number and cell area with the area correction applied.



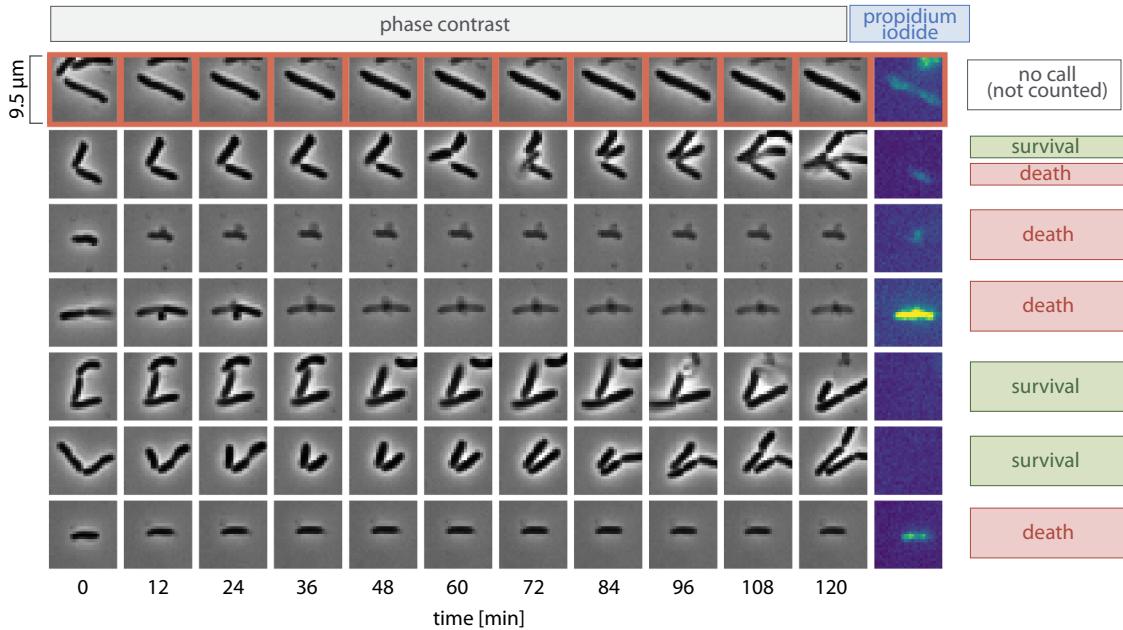
**Figure 9.6: Time lapse of a representative field after osmotic shock and the resulting classifications.** Each row shows an individual cell or pair of neighboring cells over time after the application of a fast osmotic shock. Cells classified as dead are denoted by red arrows. The lone surviving cell in this field (bottom row, top 1/4 of image) is marked in green.

were not included in the final tally of survival versus death. Across our 2822 single cell measurements, such “no call” classifications were observed only 83 times, constituting only 3% of the total cell measurements. A breakdown of all classification types and their respective abundances can be seen in Table S1.

Table 9.1: Cell fate classifications and their relative abundances in the complete data set.

Classification	Number of Observations	Percentage of Measurements
Dead-On-Arrival	11	0.4%
No Call	83	3%
Death	1246	44%
Survival	1482	53%

To assess the validity of our morphology-based classification scheme, we performed a subset of the osmotic shock experiments described in the manuscript using propidium iodide staining to mark cells which had compromised membranes,



**Figure 9.7: Representative images of propidium iodide staining after a strong osmotic shock.** Phase contrast images of individual or pairs of cells as a function of time (columns). The final column corresponds to fluorescence from propidium iodide. Bright fluorescence indicates intercalation with DNA indicating cell death. Classification of survival based only from morphology is shown as text in the final column. Highlighted row indicates a “no call” event where morphology alone could not be used to determine survival or death.

confirming their viability and effectiveness of the stain itself. Given this data set, we compared the classification breakdown using our morphology-based method with the conclusive results from the propidium iodide staining (Table 9.2). We found that the two approaches to defining death agreed within 1%. This agreement leads us to believe that our definition of cell survival as morphological regularity and sustained cell growth is sufficiently accurate to draw physiological conclusions from our experiments.

Table 9.2: Comparison of morphology-based and dye-based survival classification.

<b>Classification</b>	<b>Observations via Morphology</b>	<b>Observations via Propidium Iodide Staining</b>
Dead-On-Arrival	184	185
No Call	2	1
Survival	5	5

### 9.3 Logistic Regression

In this work, we were interested in computing the survival probability under a large hypo-osmotic shock as a function of MscL channel number. As the channel copy number distributions for each Shine-Dalgarno sequence mutant were broad and overlapping, we chose to calculate the survival probability through logistic regression – a method that requires no binning of the data providing the least biased estimate of survival probability. Logistic regression is a technique that has been used in medical statistics since the late 1950’s to describe diverse phenomena such as dose response curves, criminal recidivism, and survival probabilities for patients after treatment (Anderson et al., 2003; Mishra et al., 2016; Stahler et al., 2013). It has also found much use in machine learning to tune a binary or categorical response given a continuous input (Cheng and Hüllermeier, 2009; Dreiseitl and Ohno-Machado, 2002).

In this section, we derive a statistical model for estimating the most-likely values for the coefficients  $\beta_0$  and  $\beta_1$  and use Bayes’ theorem to provide an interpretation for the statistical meaning.

#### Bayesian parameter estimation of $\beta_0$ and $\beta_1$

The central challenge of this work is to estimate the probability of survival  $p_s$  given only a measure of the total number of MscL channels in that cell. In other

words, for a given measurement of  $N_c$  channels, we want to know likelihood that a cell would survive an osmotic shock. Using Bayes' theorem, we can write a statistical model for the survival probability as

$$g(p_s | N_c) = \frac{f(N_c | p_s)g(p_s)}{f(N_c)}, \quad (9.17)$$

where  $g$  and  $f$  represent probability density functions over parameters and data, respectively. The posterior probability distribution  $g(p_s | N_c)$  describes the probability of  $p_s$  given a specific number of channels  $N_c$ . This distribution is dependent on the likelihood of observing  $N_c$  channels assuming a value of  $p_s$  multiplied by all prior knowledge we have about knowing nothing about the data,  $g(s)$ . The denominator  $f(N_c)$  in Eq. 9.17 captures all knowledge we have about the available values of  $N_c$ , knowing nothing about the true survival probability. As this term acts as a normalization constant, we will neglect it in the following calculations for convenience. To begin, we must come up with a statistical model that describes the experimental measurable in our experiment – survival or death. As this is a binary response, we can consider each measurement as a Bernoulli trial with a probability of success matching our probability of survival  $p_s$ ,

$$f(s | p_s) = p_s^s(1 - p_s)^{1-s}, \quad (9.18)$$

where  $s$  is the binary response of 1 or 0 for survival and death, respectively. As is stated in the introduction to this section, we decided to use a logistic function to describe the survival probability. We assume that the log-odds of survival is linear with respect to the effective channel copy number  $N_c$  as

$$\log \frac{p_s}{1 - p_s} = \beta_0 + \beta_1 N_c, \quad (9.19)$$

where  $\beta_0$  and  $\beta_1$  are coefficients which describe the survival probability in the absence of channels and the increase in log-odds of survival conveyed by a single channel. The rationale behind this interpretation is presented in the following section, *A Bayesian interpretation of  $\beta_0$  and  $\beta_1$* . Using this assumption, we can solve for the survival probability  $p_s$  as,

$$p_s = \frac{1}{1 + e^{-\beta_0 - \beta_1 N_c}}. \quad (9.20)$$

With a functional form for the survival probability, the likelihood stated in Eq. 9.17 can be restated as

$$f(N_c, s | \beta_0, \beta_1) = \left( \frac{1}{1 + e^{-\beta_0 - \beta_1 N_c}} \right)^s \left( 1 - \frac{1}{1 + e^{-\beta_0 - \beta_1 N_c}} \right)^{1-s}. \quad (9.21)$$

As we have now introduced two parameters,  $\beta_0$ , and  $\beta_1$ , we must provide some description of our prior knowledge regarding their values. As is typically the case, we know nothing about the values for  $\beta_0$  and  $\beta_1$ . These parameters are allowed to take any value, so long as it is a real number. Since all values are allowable, we can assume a flat distribution where any value has an equally likely probability. This value of this constant probability is not necessary for our calculation and is ignored. For a set of  $k$  single-cell measurements, we can write the posterior probability distribution stated in Eq. 9.17 as

$$g(\beta_0, \beta_1 | N_c, s) = \prod_{i=1}^n \left( \frac{1}{1 + e^{-\beta_0 - \beta_1 N_c^{(i)}}} \right)^{s^{(i)}} \left( 1 - \frac{1}{1 + e^{-\beta_0 - \beta_1 N_c^{(i)}}} \right)^{1-s^{(i)}} \quad (9.22)$$

Implicitly stated in Eq. 9.22 is absolute knowledge of the channel copy number  $N_c$ . However, as is described in *Standard Candle Calibration*, we must convert from a measured areal sfGFP intensity  $I_A$  to a effective channel copy number,

$$N_c = \frac{I_A \langle \tilde{A} \rangle}{\tilde{\alpha}}, \quad (9.23)$$

where  $\langle \tilde{A} \rangle$  is the average cell area of the standard candle strain and  $\tilde{\alpha}$  is the most-likely value for the calibration factor between arbitrary units and protein copy number. In *Standard Candle Calibration*, we detailed a process for generating an estimate for the most-likely value of  $\langle \tilde{A} \rangle$  and  $\tilde{\alpha}$ . Given these estimates, we can include an informative prior for each value. From the Markov chain Monte Carlo samples shown in Fig. ??, the posterior distribution for each parameter is approximately Gaussian. By approximating them as Gaussian distributions, we can assign an informative prior for each as

$$g(\alpha | \tilde{\alpha}, \tilde{\sigma}_\alpha) \propto \frac{1}{\tilde{\sigma}_\alpha^k} \prod_{i=1}^k \exp \left[ -\frac{(\alpha_i - \tilde{\alpha})^2}{2\tilde{\sigma}_\alpha^2} \right] \quad (9.24)$$

**Posterior distributions for logistic regression coefficients evaluated for fast and slow shock rates.** (A) Kernel density estimates of the posterior distribution for  $\beta_0$  for fast (blue) and slow (purple) shock rates. (B) Kernel density estimates of posterior distribution for  $\beta_1$ .

Figure 9.8: **Posterior distributions for logistic regression coefficients evaluated for fast and slow shock rates.** (A) Kernel density estimates of the posterior distribution for  $\beta_0$  for fast (blue) and slow (purple) shock rates. (B) Kernel density estimates of posterior distribution for  $\beta_1$ .

for the calibration factor for each cell and

$$g(\langle A \rangle | \langle \tilde{A} \rangle, \tilde{\sigma}_{\langle A \rangle}) = \frac{1}{\tilde{\sigma}_{\langle A \rangle}^k} \prod_{i=1}^k \exp \left[ -\frac{(\langle A \rangle_i - \langle \tilde{A} \rangle)^2}{2\tilde{\sigma}_{\langle A \rangle}^2} \right], \quad (9.25)$$

where  $\tilde{\sigma}_\alpha$  and  $\tilde{\sigma}_{\langle A \rangle}$  represent the variance from approximating each posterior as a Gaussian. The proportionality for each prior arises from the neglecting of normalization constants for notational convenience.

Given Eq. 9.21 through Eq. 9.25, the complete posterior distribution for estimating the most likely values of  $\beta_0$  and  $\beta_1$  can be written as

$$\begin{aligned} g(\beta_0, \beta_1 | [I_A, s], \langle \tilde{A} \rangle, \tilde{\sigma}_{\langle A \rangle}, \tilde{\alpha}, \tilde{\sigma}_\alpha) \propto & \frac{1}{(\tilde{\sigma}_\alpha \tilde{\sigma}_{\langle A \rangle})^k} \prod_{i=1}^k \left( 1 + \exp \left[ -\beta_0 - \beta_1 \frac{I_{Ai} \langle A \rangle_i}{\alpha_i} \right] \right)^{-s_i} \times \\ & \left( 1 - \left( 1 + \exp \left[ -\beta_0 - \beta_1 \frac{I_{Ai} \langle A \rangle_i}{\alpha_i} \right] \right)^{-1} \right)^{1-s_i} \exp \left[ -\frac{(\langle A \rangle_i - \langle \tilde{A} \rangle)^2}{2\tilde{\sigma}_{\langle A \rangle}^2} - \frac{(\alpha_i - \tilde{\alpha})^2}{2\tilde{\sigma}_\alpha^2} \right]. \end{aligned} \quad (9.26)$$

As this posterior distribution is not solvable analytically, we used Markov chain Monte Carlo to draw samples out of this distribution, using the log of the effective channel number as described in the main text. The posterior distributions for  $\beta_0$  and  $\beta_1$  for both slow and fast shock rate data can be seen in Fig. 9.8

### A Bayesian interpretation of $\beta_0$ and $\beta_1$

The assumption of a linear relationship between the log-odds of survival and the predictor variable  $N_c$  appears to be arbitrary and is presented without justification. However, this relationship is directly connected to the manner in which Bayes' theorem updates the posterior probability distribution upon the observation of new

data. In following section, we will demonstrate this connection using the relationship between survival and channel copy number. However, this description is general and can be applied to any logistic regression model so long as the response variable is binary. This connection was shown briefly by Allen Downey in 2014 and has been expanded upon in this work (Downey, 2014).

The probability of observing a survival event  $s$  given a measurement of  $N_c$  channels can be stated using Bayes' theorem as

$$g(s | N_c) = \frac{f(N_c | s)g(s)}{f(N_c)}. \quad (9.27)$$

where  $g$  and  $f$  represent probability density functions over parameters and data respectively. The posterior distribution  $g(s | N_c)$  is the quantity of interest and implicitly related to the probability of survival. The likelihood  $g(N_c | s)$  tells us the probability of observing  $N_c$  channels in this cell given that it survives. The quantity  $g(s)$  captures all *a priori* knowledge we have regarding the probability of this cell surviving and the denominator  $f(N_c)$  tells us the converse – the probability of observing  $N_c$  cells irrespective of the survival outcome.

Proper calculation of Eq. 9.27 requires that we have knowledge of  $f(N_c)$ , which is difficult to estimate. While we are able to give appropriate bounds on this term, such as a requirement of positivity and some knowledge of the maximum membrane packing density, it is not so obvious to determine the distribution between these bounds. Given this difficulty, it's easier to compute the odds of survival  $\mathcal{O}(s | N_c)$ , the probability of survival  $s$  relative to death  $d$ ,

$$\mathcal{O}(s | N_c) = \frac{g(s | N_c)}{g(d | N_c)} = \frac{f(N_c | s)g(s)}{f(N_c | d)g(d)}, \quad (9.28)$$

where  $f(N_c)$  is cancelled. The only stipulation on the possible value of the odds is that it must be a positive value. As we would like to equally weigh odds less than one as those of several hundred or thousand, it is more convenient to compute the log-odds, given as

$$\log \mathcal{O}(s | N_c) = \log \frac{g(s)}{g(d)} + \log \frac{f(N_c | s)}{f(N_c | d)}. \quad (9.29)$$

Computing the log-transform reveals two interesting quantities. The first term is the ratio of the priors and tells us the *a priori* knowledge of the odds of survival irrespective of the number of channels. As we have no reason to think that survival is more likely than death, this ratio goes to unity. The second term is the log likelihood ratio and tells us how likely we are to observe a given channel copy number  $N_c$  given the cell survives relative to when it dies.

For each channel copy number, we can evaluate Eq. 9.29 to measure the log-odds of survival. If we start with zero channels per cell, we can write the log-odds of survival as

$$\log \mathcal{O}(s | N_c = 0) = \log \frac{g(s)}{g(d)} + \log \frac{f(N_c = 0 | s)}{f(N_c = 0 | d)}. \quad (9.30)$$

For a channel copy number of one, the odds of survival is

$$\log \mathcal{O}(s | N_c = 1) = \log \frac{g(s)}{g(d)} + \log \frac{f(N_c = 1 | s)}{f(N_c = 1 | d)}. \quad (9.31)$$

In both Eq. 9.30 and Eq. 9.31, the log of our *a priori* knowledge of survival versus death remains. The only factor that is changing is log likelihood ratio. We can be more general in our language and say that the log-odds of survival is increased by the difference in the log-odds conveyed by addition of a single channel. We can rewrite the log likelihood ratio in a more general form as

$$\log \frac{f(N_c | s)}{f(N_c | d)} = \log \frac{f(N_c = 0 | s)}{f(N_c = 0 | d)} + N_c \left[ \log \frac{f(N_c = 1 | s)}{f(N_c = 1 | d)} - \log \frac{f(N_c = 0 | s)}{f(N_c = 0 | d)} \right], \quad (9.32)$$

where we are now only considering the case in which  $N_c \in [0, 1]$ . The bracketed term in Eq. 9.32 is the log of the odds of survival given a single channel relative to the odds of survival given no channels. Mathematically, this odds-ratio can be expressed as

$$\log \mathcal{OR}_{N_c}(s) = \log \frac{\frac{f(N_c=1 | s)g(s)}{f(N_c=1 | d)g(d)}}{\frac{f(N_c=0 | s)g(s)}{f(N_c=0 | d)g(d)}} = \log \frac{f(N_c = 1 | s)}{f(N_c = 1 | d)} - \log \frac{f(N_c = 0 | s)}{f(N_c = 0 | d)}. \quad (9.33)$$

Eq. 9.33 is mathematically equivalent to the bracketed term shown in Eq. 9.32.

We can now begin to staple these pieces together to arrive at an expression for the log odds of survival. Combining Eq. 9.32 with Eq. 9.29 yields

$$\log \mathcal{O}(s | N_c) = \log \frac{g(s)}{g(d)} + \log \frac{f(N_c = 0 | s)}{f(N_c = 0 | d)} + N_c \left[ \frac{f(N_c = 1 | s)}{f(N_c = 1 | d)} - \log \frac{f(N_c = 0 | s)}{f(N_c = 0 | d)} \right]. \quad (9.34)$$

Using our knowledge that the bracketed term is the log odds-ratio and the first two times represents the log-odds of survival with  $N_c = 0$ , we conclude with

$$\log \mathcal{O}(s | N_c) = \log \mathcal{O}(s | N_c = 0) + N_c \log \mathcal{OR}_{N_c}(s). \quad (9.35)$$

This result can be directly compared to Eq. 1 presented in the main text,

$$\log \frac{p_s}{1 - p_s} = \beta_0 + \beta_1 N_c, \quad (9.36)$$

which allows for an interpretation of the seemingly arbitrary coefficients  $\beta_0$  and  $\beta_1$ . The intercept term,  $\beta_0$ , captures the log-odds of survival with no MscL channels. The slope,  $\beta_1$ , describes the log odds-ratio of survival which a single channel relative to the odds of survival with no channels at all. While we have examined this considering only two possible channel copy numbers (1 and 0), the relationship between them is linear. We can therefore generalize this for any MscL copy number as the increase in the log-odds of survival is constant for addition of a single channel.

### Other properties as predictor variables

The previous two sections discuss in detail the logic and practice behind the application of logistic regression to cell survival data using only the effective channel copy number as the predictor of survival. However, there are a variety of properties that could rightly be used as predictor variables, such as cell area and shock rate. As is stipulated in our standard candle calibration, there should be no correlation between survival and cell area. Fig. 9.9A and B show the logistic regression performed on the cell area. We see for both slow and fast shock groups, there is little change in survival probability with changing cell area and the wide credible regions allow for both positive and negative correlation between survival and

area. The appearance of a bottle neck in the notably wide credible regions is a result of a large fraction of the measurements being tightly distributed about a mean value. Fig. 9.9C shows the predicted survival probability as a function of the shock rate. There is a slight decrease in survivability as a function of increasing shock rate, however the width of the credible region allows for slightly positive or slightly negative correlation. While we have presented logistic regression in this section as a one-dimensional method, Eq. 9.19 can be generalized to  $n$  predictor variables  $x$  as

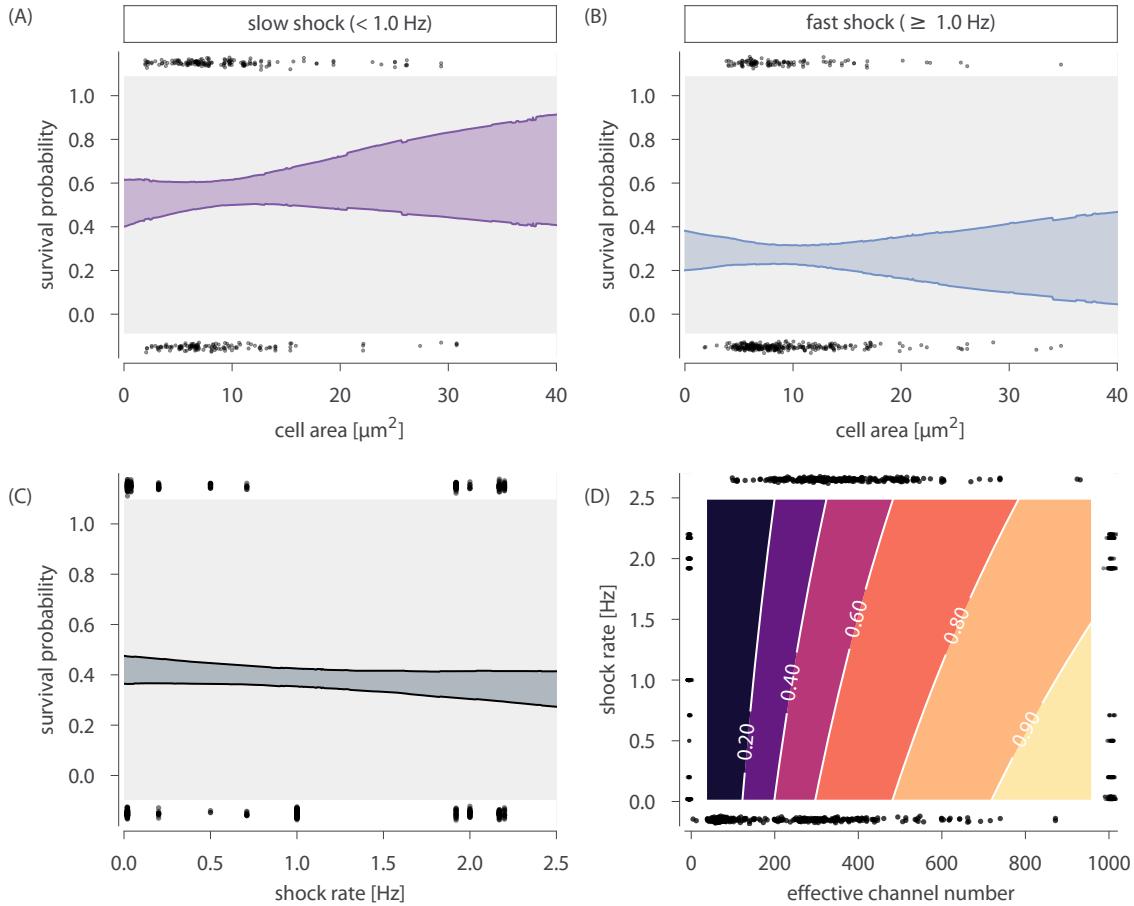
$$\log \frac{p_s}{1 - p_s} = \beta_0 + \sum_i^n \beta_i x_i. \quad (9.37)$$

Using this generalization, we can use both shock rate and the effective channel copy number as predictor variables. The resulting two-dimensional surface of survival probability is shown in Fig. 9.9 (D). As is suggested by Fig. 9.9 (C), the magnitude of change in survivability as the shock rate is increased is smaller than that along increasing channel copy number, supporting our conclusion that for MscL alone, the copy number is the most important variable in determining survival.

#### 9.4 Classification Of Shock Rate

Its been previously shown that the rate of hypo-osmotic shock dictates the survival probability (Bialecka-Fornal et al., 2015). To investigate how a single channel contributes to survival, we queried survival at several shock rates with varying MscL copy number. In the main text of this work, we separated our experiments into arbitrary bins of “fast” ( $\geq 1.0$  Hz) and “slow” ( $< 1.0$  Hz) shock rates. In this section, we discuss our rationale for coarse graining our data into these two groupings.

As is discussed in the main text and in the supplemental section *Logistic Regression*, we used a bin-free method of estimating the survival probability given the MscL channel copy number as a predictor variable. While this method requires no binning of the data, it requires a data set that sufficiently covers the physiological range of channel copy number to accurately allow prediction of survivability. Fig. 9.10 shows the results of the logistic regression treating each shock rate as an

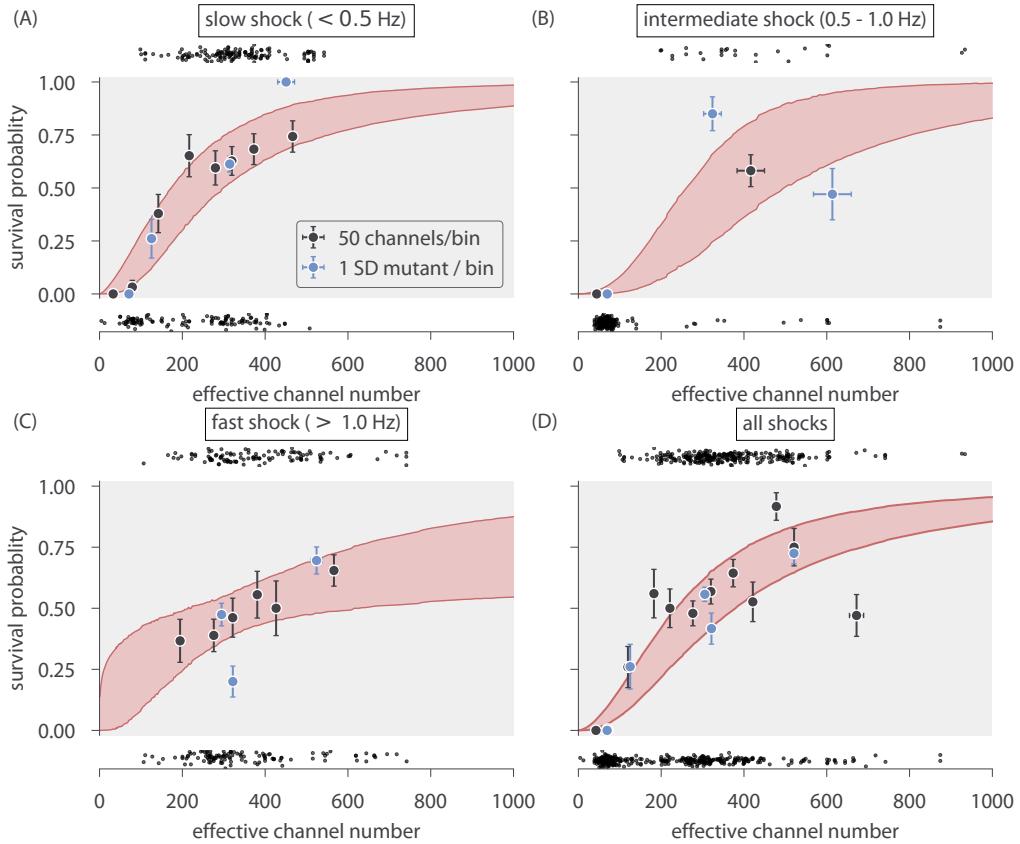


**Figure 9.9: Survival probability estimation using alternative predictor variables.** (A) Estimated survival probability as a function of cell area for the slow shock group. (B) Estimated survival probability as a function of cell area for the fast shock group. (C) Estimated survival probability as a function shock rate. Black points at top and bottom of plots represent single-cell measurements of cells who survived and perished, respectively. Shaded regions in (A) - (C) represent the 95% credible region. (D) Surface of estimated survival probability using both shock rate and effective channel number as predictor variables. Black points at left and right of plot represent single-cell measurements of cells which survived and died, respectively, sorted by shock rate. Points at top and bottom of plot represent survival and death sorted by their effective channel copy number. Labeled contours correspond to the survival probability.

individual data set. The most striking feature of the plots shown in Fig. 9.10 is the inconsistent behavior of the predicted survivability from shock rate to shock rate. The appearance of bottle necks in the credible regions for some shock rates (0.2Hz, 0.5Hz, 2.00Hz, and 2.20 Hz) appear due to a high density of measurements within a narrow range of the channel copy number at the narrowest point in the bottle neck. While this results in a seemingly accurate prediction of the survival probability at that point, the lack of data in other copy number regimes severely limits our extrapolation outside of the copy number range of that data set. Other shock rates (0.018 Hz, 0.04 Hz, and 1.00 Hz) demonstrate completely pathological survival probability curves due to either complete survival or complete death of the population.

Ideally, we would like to have a wide range of MscL channel copy numbers at each shock rate shown in Fig. 9.10. However, the low-throughput nature of these single-cell measurements prohibits completion of this within a reasonable time frame. It is also unlikely that thoroughly dissecting the shock rate dependence will change the overall finding from our work that several hundred MscL channels are needed to convey survival under hypo-osmotic stress.

Given the data shown in Fig. 9.10, we can try to combine the data sets into several bins. Fig. 9.11 shows the data presented in Fig. 9.10 separated into “slow” ( $< 0.5$  Hz, A), “intermediate” (0.5 - 1.0 Hz, B), and “fast” ( $> 1.0$  Hz, C) shock groups. Using these groupings, the full range of MscL channel copy numbers are covered for each case, with the intermediate shock rate sparsely sampling copy numbers greater than 200 channels per cell. In all three of these cases, the same qualitative story is told – several hundred channels per cell are necessary for an appreciable level of survival when subjected to an osmotic shock. This argument is strengthened when examining the predicted survival probability by considering all shock rates as a single group, shown in Fig. 9.11 (D). This treatment tells nearly the same quantitative and qualitative story as the three rate grouping shown in this section and the two rate grouping presented in the main text. While there



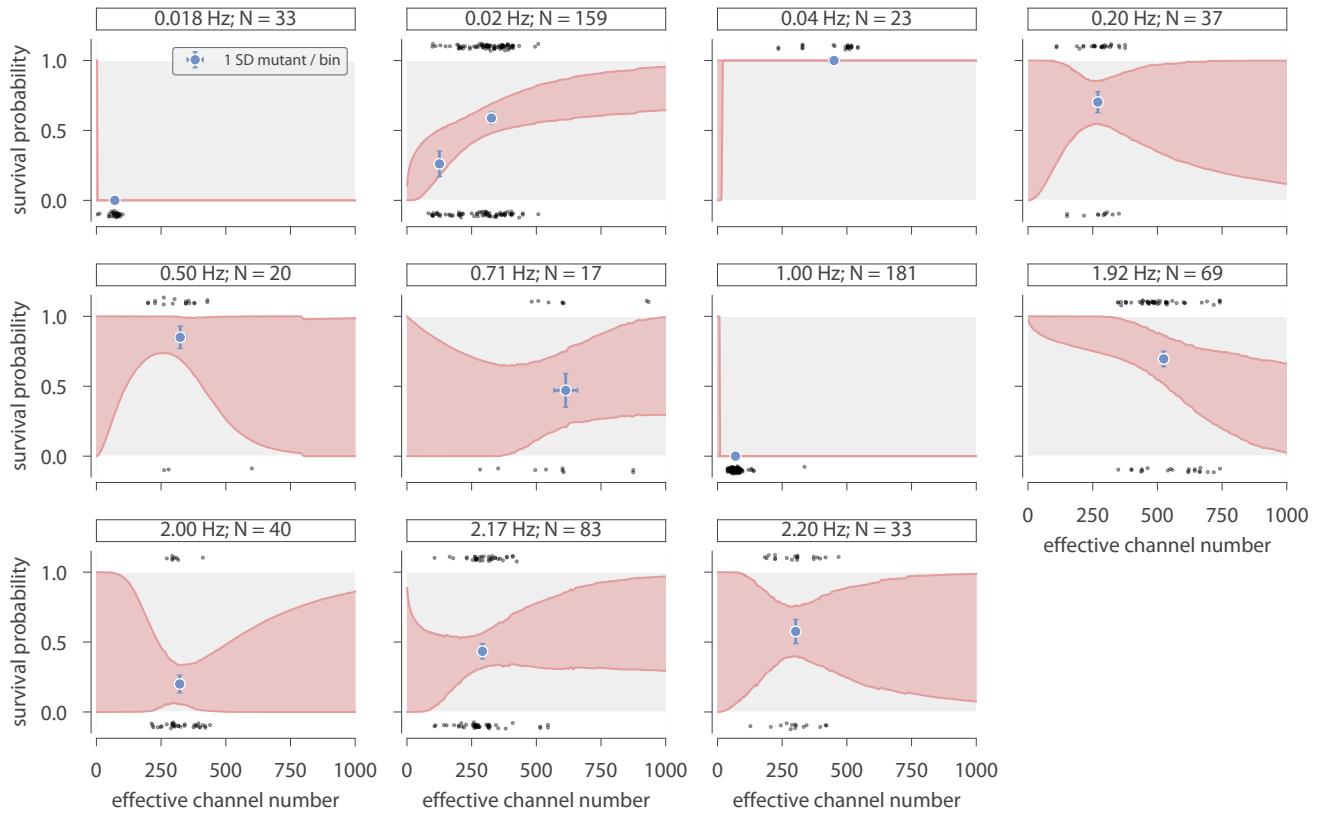
**Figure 9.10: Binning by individual shock rates.** Survival probability estimates from logistic regression (red lines) and the computed survival probability for all SD mutants subjected to that shock rate (blue points). Black points at top and bottom of each plot correspond to single cell measurements of survival (top) and death (bottom). Red shaded regions signify the 95% credible region of the logistic regression. Horizontal error bars of blue points are the standard error of the mean channel copy number. Vertical error bars of blue points correspond to the uncertainty in survival probability by observing  $n$  survival events from  $N$  single-cell measurements.

does appear to be a dependence on the shock rate for survival when only MscL is expressed, the effect is relatively weak with overlapping credible regions for the logistic regression across the all curves. To account for the sparse sampling of high copy numbers observed in the intermediate shock group, we split this set and partitioned the measurements into either the “slow” ( $< 1.0$  Hz) or “fast” ( $\geq 1.0$  Hz) groups presented in the main text of this work.

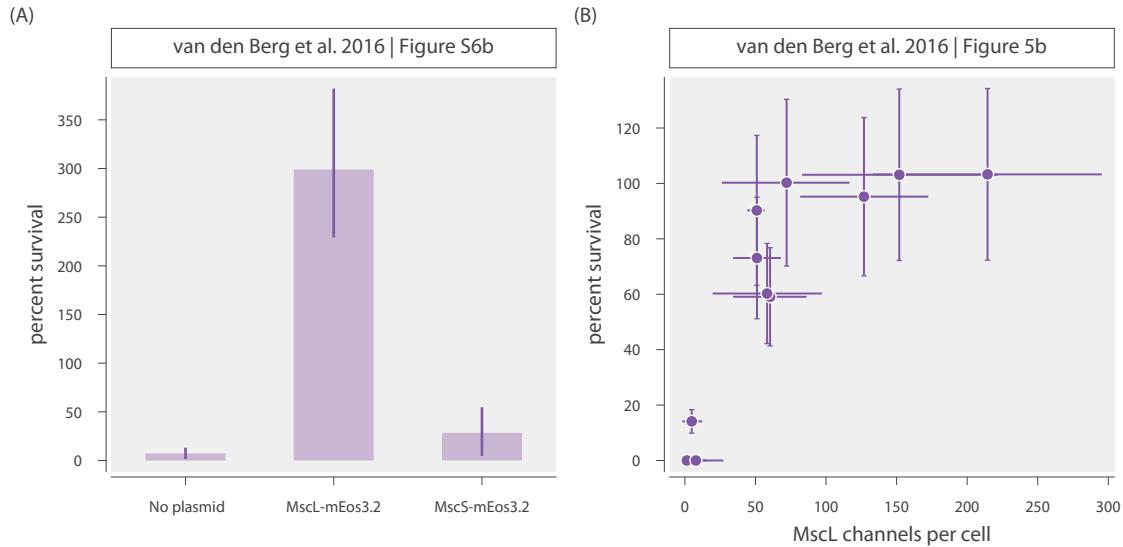
## 9.5 Comparison of Survival Probability with van den Berg et al. 2016

In van den Berg et al. 2016, the authors report a 100% survival rate at approximately 100 channels per cell. While the number of mechanosensitive channels per cell was quantified at the level of single cells, the survival probability was measured in bulk using ensemble plating assays. The results of these experiments considering the contribution of MscL to survival is shown in Figure 5 of their work, although without displayed uncertainty in the survival probability. Figure S6B of their work shows the approximate error in survival probability through ensemble plating assays for three different strains Fig. 9.12, which is approximately 30%. Using this approximate error and the data shown in their Figure 5B, we have reproduced this plot with error bars in both measured dimensions Fig. 9.12. This plot shows that even when the mean survival probability is 100%, the variation in the measured survival probability is large, extending as low as  $\approx 70\%$ . This variation is likely born from a multitude of experimental steps including time of outgrowth, variation in shock rate, plating efficiency, and counting errors. As our experimental approach directly measures the survival/death of individual cells, we remove many sources of error that would arise from an ensemble approach, albeit at lower throughput. While it is possible that the discrepancy between van den Berg et al. (2016) and the work presented in Chapter 5 could arise from other unknown factors, we believe that single-cell experiments introduce the fewest sources of error.

## 9.6 *E. coli* Strains



**Figure 9.11: Coarse graining shock rates into different groups.** Estimated survival probability curve for slow (A), intermediate (B), and fast (C) shock rates. (D) Estimated survival probability curve from pooling all data together, ignoring varying shock rates. Red shaded regions correspond to the 95% credible region of the survival probability estimated via logistic regression. Black points at top and bottom of each plot represent single-cell measurements of cells which survived and died, respectively. Black points and error bars represent survival probability calculations from bins of 50 channels per cell. Blue points represent the survival probability for a given Shine-Dalgarno mutant. Horizontal error bars are the standard error of the mean with at least 25 measurements and vertical error bars signifies the uncertainty in the survival probability from observing  $n$  survival events out of  $N$  total measurements.



**Figure 9.12: MscL abundance vs survival data reported in van den Berg et al. 2016 with included error.** (A) Reported survival probabilities of a strain lacking all mechanosensitive channels (“no plasmid”), plasmid borne MscL-mEos3.2, and plasmid borne MscS-mEos3.2. Approximate reported errors for MscL-mEos3.2 survival probability is 30%. (B) The measurement of survival probability as a function of MscL channel copy number was obtained from Figure 5B in van den Berg et al 2016. Errors in channel copy number represent the standard deviation of several biological replicates (present in original figure) while the error in survival probability is taken as  $\approx 30\%$ .

Table 9.3: *Escherichia coli* strains used in Chapters 5 and 9.

Strain name	Genotype	Reference
MJF641	Frag1, $\Delta mscL::cm$ , $\Delta mscS$ , $\Delta mscK::kan$ , $\Delta ybdG::apr$ , $\Delta ynaI$ , $\Delta yjeP$ , $\Delta ybiO$ , $ycjM::Tn10$	Edwards et al. (2012)
MLG910	MG1655, $\Delta mscL ::\phi mscL-sfGFP$ , $\Delta galK::kan$ , $\Delta lacI$ , $\Delta lacZYA$	Bialecka-Fornal et al. (2012)
D6LG-Tn10	Frag1, $\Delta mscL ::\phi mscL-sfGFP$ , $\Delta mscS$ , $\Delta mscK::kan$ , $\Delta ybdG::apr$ , $\Delta ynaI$ , $\Delta yjeP$ , $\Delta ybiO$ , $ycjM::Tn10$	Chure et al. (2018)

Strain name	Genotype	Reference
D6LG (SD0)	Frag1, <i>ΔmscL::φmscL-sfGFP, ΔmscS, ΔmscK::kan, ΔybdG::apr, ΔynaI, ΔyjeP, ΔybiO</i>	Chure et al. (2018)
XTL298	CC4231, <i>araD:: tetA-sacB-amp</i>	(Li et al., 2013)
D6LTetSac	Frag1, <i>mscL-sfGFP:: tetA-sacB, ΔmscS, ΔmscK::kan, ΔybdG::apr, ΔynaI, ΔyjeP, ΔybiO</i>	Chure et al. (2018)
D6LG (SD1)	Frag1, <i>ΔmscL ::φmscL-sfGFP, ΔmscS, ΔmscK::kan, ΔybdG::apr, ΔynaI, ΔyjeP, ΔybiO</i>	Chure et al. (2018)
D6LG (SD2)	Frag1, <i>ΔmscL ::φmscL-sfGFP, ΔmscS, ΔmscK::kan, ΔybdG::apr, ΔynaI, ΔyjeP, ΔybiO</i>	Chure et al. (2018)
D6LG (SD4)	Frag1, <i>ΔmscL ::φmscL-sfGFP, ΔmscS, ΔmscK::kan, ΔybdG::apr, ΔynaI, ΔyjeP, ΔybiO</i>	Chure et al. (2018)
D6LG (SD6)	Frag1, <i>ΔmscL ::φmscL-sfGFP, ΔmscS, ΔmscK::kan, ΔybdG::apr, ΔynaI, ΔyjeP, ΔybiO</i>	Chure et al. (2018)
D6LG (12SD2)	Frag1, <i>ΔmscL ::φmscL-sfGFP, ΔmscS, ΔmscK::kan, ΔybdG::apr, ΔynaI, ΔyjeP, ΔybiO</i>	Chure et al. (2018)
D6LG (16SD0)	Frag1, <i>ΔmscL ::φmscL-sfGFP, ΔmscS, ΔmscK::kan, ΔybdG::apr, ΔynaI, ΔyjeP, ΔybiO</i>	Chure et al. (2018)

Table 9.4: Oligonucleotide sequences used in Chapters 5 and 9. Bold and italics correspond to Shine-Dalgarno sequence modifications and AT hairpin insertion modifications, respectively. Double bar || indicates a transposon insertion site.

Primer Name	Sequence (5' → 3')
<i>Tn10delR</i>	taaaggccaa <del>cggcatccaggcgga</del> catactcagca   ccttcgcaaggtaacagagtaaaacatccaccat
<i>MscLSPSac</i>	gaaaatggcttaacattgttagacttatggttgcgg

---

Primer Name	Sequence (5' → 3')
	cttcatagggagTCCTAATTTTGTGACACTCTATC
<i>MscLSPSacR</i>	accacgttcccgcatcgcaaattcgcaaat
	tcttaataatgctcatATCAAAGGGAAAATGTCCATA
<i>MscL-SD1R</i>	atcgcaaattcgcaattttataatgctcat
	gttatttcctcatgaagccgacaaccataagtctaacaaa
<i>MscL-SD2R</i>	atcgcaaattcgcaattttataatgctcatgttatt
	tcccctatgaagccgacaaccataagtctaacaaa
<i>MscL-SD4R</i>	atcgcaaattcgcaattttataatgctcat
	gttatt cctgctatgaagccgacaaccataagtctaacaaa
<i>MscL-SD6R</i>	atcgcaaattcgcaattttataatgctcat
	gttatt gctcgttatgaagccgacaaccataagtctaacaaa
<i>MscL-12SD2R</i>	atcgcaaattcgcaattttataatgctcat
	atatatatatat tcccctatgaagccgacaaccataagtctaacaaa
<i>MscL-16SD0R</i>	atcgcaaattcgcaattttataatgctcat
	atatatatatatatat ctcccctatgaagccgacaaccataagtctaacaaa

---

## REFERENCES

- Ackers, G.K., and Johnson, A.D. (1982). Quantitative model for gene regulation by A phage repressor. *Proc. Natl. Acad. Sci. USA* 5.
- Aghaeepour, N., Finak, G., Hoos, H., Mosmann, T.R., Brinkman, R., Gottardo, R., and Scheuermann, R.H. (2013). Critical assessment of automated flow cytometry data analysis techniques. *Nature Methods* 10, 228–238.
- Allen, R.J., and Waclaw, B. (2018). Bacterial growth: A statistical physicist's guide. *Reports on Progress in Physics* 82, 016601.
- Anderson, R.P., Jin, R., and Grunkemeier, G.L. (2003). Understanding logistic regression analysis in clinical reports: An introduction. *The Annals of Thoracic Surgery* 75, 753–757.
- Auerbach, A. (2012). Thinking in cycles: MWC is a good model for acetylcholine receptor-channels. *The Journal of Physiology* 590, 93–98.
- Balleza, E., Kim, J.M., and Cluzel, P. (2018). Systematic characterization of maturation time of fluorescent proteins in living cells. *Nature Methods* 15, 47–51.
- Barnes, S.L., Belliveau, N.M., Ireland, W.T., Kinney, J.B., and Phillips, R. (2019). Mapping DNA sequence to transcription factor binding energy in vivo. *PLOS Computational Biology* 15, e1006226.
- Bavi, N., Cortes, D.M., Cox, C.D., Rohde, P.R., Liu, W., Deitmer, J.W., Bavi, O., Strop, P., Hill, A.P., Rees, D., et al. (2016). The role of MscL amphipathic N terminus indicates a blueprint for bilayer-mediated gating of mechanosensitive channels. *Nature Communications* 7, 11984.
- Berg, H.C., and Purcell, E.M. (1977). Physics of chemoreception. *Biophysical Journal* 20, 193–219.
- Berg, J., Willmann, S., and Lässig, M. (2004). Adaptive evolution of transcription factor binding sites. *BMC Evolutionary Biology* 4, 42.

- Bialecka-Fornal, M., Lee, H.J., DeBerg, H.A., Gandhi, C.S., and Phillips, R. (2012). Single-Cell Census of Mechanosensitive Channels in Living Bacteria. *PLoS ONE* 7, e33077.
- Bialecka-Fornal, M., Lee, H.J., and Phillips, R. (2015). The Rate of Osmotic Down-shock Determines the Survival Probability of Bacterial Mechanosensitive Channel Mutants. *Journal of Bacteriology* 197, 231–237.
- Bintu, L., Buchler, N.E., Garcia, H.G., Gerland, U., Hwa, T., Kondev, J., and Phillips, R. (2005a). Transcriptional regulation by the numbers: Models. *Current Opinion in Genetics & Development* 15, 116–124.
- Bintu, L., Buchler, N.E., Garcia, H.G., Gerland, U., Hwa, T., Kondev, J., Kuhlman, T., and Phillips, R. (2005b). Transcriptional regulation by the numbers: Applications. *Current Opinion in Genetics & Development* 15, 125–135.
- Blount, P., Sukharev, S.I., Schroeder, M.J., Nagle, S.K., and Kung, C. (1996). Single residue substitutions that change the gating properties of a mechanosensitive channel in *Escherichia Coli*. *Proc Natl Acad Sci U S A* 93, 11652–11657.
- Blount, P., Sukharev, S.I., Moe, P.C., Martinac, B., and Kung, C. (1999). Mechanosensitive channels of bacteria. *Methods in Enzymology* 294, 458–482.
- Bochner, B.R., Huang, H.-C., Schieven, G.L., and Ames, B.N. (1980). Positive selection for loss of tetracycline resistance. *Journal of Bacteriology* 143, 926–933.
- Boedicker, J.Q., Garcia, H.G., Johnson, S., and Phillips, R. (2013a). DNA sequence-dependent mechanics and protein-assisted bending in repressor-mediated loop formation. *Physical Biology* 10, 066005.
- Boedicker, J.Q., Garcia, H.G., and Phillips, R. (2013b). Theoretical and Experimental Dissection of DNA Loop-Mediated Repression. *Physical Review Letters* 110, 018101.
- Booth, I.R. (2014). Bacterial mechanosensitive channels: Progress towards an understanding of their roles in cell physiology. *Current Opinion in Microbiology* 18,

16–22.

Booth, I.R., Edwards, M.D., Murray, E., and Miller, S. (2005). The role of bacterial ion channels in cell physiology. In *Bacterial Ion Channels and Their Eukaryotic Homologs*, A. Kubalsi, and B. Martinac, eds. (Washington DC: American Society for Microbiology), pp. 291–312.

Boulton, S., and Melacini, G. (2016). Advances in NMR Methods To Map Allosteric Sites: From Models to Translation. *Chemical Reviews* *116*, 6267–6304.

Brewster, R.C., Jones, D.L., and Phillips, R. (2012). Tuning Promoter Strength through RNA Polymerase Binding Site Design in *Escherichia coli*. *PLoS Computational Biology* *8*, e1002811.

Brewster, R.C., Weinert, F.M., Garcia, H.G., Song, D., Rydenfelt, M., and Phillips, R. (2014). The Transcription Factor Titration Effect Dictates Level of Gene Expression. *Cell* *156*, 1312–1323.

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

Buchler, N.E., Gerland, U., and Hwa, T. (2003). On schemes of combinatorial transcription logic. *Proceedings of the National Academy of Sciences* *100*, 5136–5141.

Canals, M., Lane, J.R., Wen, A., Scammells, P.J., Sexton, P.M., and Christopoulos, A. (2012). A Monod-Wyman-Changeux Mechanism Can Explain G Protein-coupled Receptor (GPCR) Allosteric Modulation. *Journal of Biological Chemistry* *287*, 650–659.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software* *76*, 1–32.

Cass, J.A., Stylianidou, S., Kuwada, N.J., Traxler, B., and Wiggins, P.A. (2017). Probing bacterial cell biology using image cytometry. *Molecular Microbiology* *103*, 818–828.

- Cheng, W., and Hüllermeier, E. (2009). Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning* 76, 211–225.
- Chure, G., Lee, H.J., and Phillips, R. (2018). MCMC chains generated in "Connecting the Dots between Mechanosensitive Channel Abundance, Osmotic Shock, and Survival at Single-Cell Resolution" accessible through DOI 10.22002/D1.942.
- Chure, G., Razo-Mejia, M., Belliveau, N.M., Einav, T., Kaczmarek, Z.A., Barnes, S.L., and Phillips, R. (2019). Predictive Shifts in Free Energy Couple Mutations to Their Phenotypic Consequences. *Proceedings of the National Academy of Sciences* 116.
- Colin, R., and Sourjik, V. (2017). Emergent properties of bacterial chemotaxis pathway. *Current Opinion in Microbiology* 39, 24–33.
- Cruickshank, C.C., Minchin, R.F., Le Dain, A.C., and Martinac, B. (1997). Estimation of the pore size of the large-conductance mechanosensitive ion channel of *Escherichia Coli*. *Biophysical Journal* 73, 1925–1931.
- Czaran, T.L., Hoekstra, R.F., and Pagie, L. (2002). Chemical warfare between microbes promotes biodiversity. *Proceedings of the National Academy of Sciences* 99, 786–790.
- Daber, R., Sharp, K., and Lewis, M. (2009). One Is Not Enough. *Journal of Molecular Biology* 392, 1133–1144.
- Daber, R., Sochor, M.A., and Lewis, M. (2011). Thermodynamic Analysis of Mutant lac Repressors. *Journal of Molecular Biology* 409, 76–87.
- Diénert, F. (1900). Sur la fermentation du galactose et sur l'accoutumance des levures à ce sucre, (Sceaux).
- Dill, K.A., and Bromberg, S. (2010). *Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology* (New York: Garland Science).
- Downey, A. (2014). Probably Overthinking It: Bayes's theorem and logistic regression. *Probably Overthinking It*.

- Dreiseitl, S., and Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics* 35, 352–359.
- Driessen, R.P.C., Sitters, G., Laurens, N., Moolenaar, G.F., Wuite, G.J.L., Goosen, N., and Dame, R.T. (2014). Effect of Temperature on the Intrinsic Flexibility of DNA and Its Interaction with Architectural Proteins. *Biochemistry* 53, 6430–6438.
- Edelstein, A.D., Tsuchida, M.A., Amodaj, N., Pinkard, H., Vale, R.D., and Stuurman, N. (2014). Advanced methods of microscope control using  $\mu$ Manager software. *Journal of Biological Methods* 1, 10.
- Edwards, M.D., Black, S., Rasmussen, T., Rasmussen, A., Stokes, N.R., Stephen, T.L., Miller, S., and Booth, I.R. (2012). Characterization of three novel mechanosensitive channel activities in *Escherichia Coli*. *Channels (Austin)* 6, 272–281.
- Einav, T., and Phillips, R. (2017). Monod-Wyman-Changeux Analysis of Ligand-Gated Ion Channel Mutants. –.
- Einav, T., Mazutis, L., and Phillips, R. (2016). Statistical Mechanics of Allosteric Enzymes. *The Journal of Physical Chemistry B* 120, 6021–6037.
- Eldar, A., and Elowitz, M.B. (2010). Functional roles for noise in genetic circuits. *Nature* 467, 167–173.
- Elf, J., Li, G.-W., and Xie, X.S. (2007). Probing Transcription Factor Dynamics at the Single-Molecule Level in a Living Cell. *Science* 316, 1191–1194.
- Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. *Science* 297, 1183–1186.
- Espah Borujeni, A., Channarasappa, A.S., and Salis, H.M. (2014). Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Research* 42, 2646–2659.
- Feeling-Taylor, A.R., Yau, S.-T., Petsev, D.N., Nagel, R.L., Hirsch, R.E., and Vekilov, P.G. (2004). Crystallization Mechanisms of Hemoglobin C in the R State. *Biophys-*

ical Journal 87, 2621–2629.

Fernández-Castané, A., Vine, C.E., Caminal, G., and López-Santín, J. (2012). Ev-  
idencing the role of lactose permease in IPTG uptake by Escherichia coli in fed-  
batch high cell density cultures. Journal of Biotechnology 157, 391–398.

Finch, J.T., Perutz, M.F., Bertles, J.F., and Dobler, J. (1973). Structure of Sickled  
Erythrocytes and of Sickle-Cell Hemoglobin Fibers. Proceedings of the National  
Academy of Sciences 70, 718–722.

Forsén, S., and Linse, S. (1995). Cooperativity: Over the Hill. Trends in Biochemical  
Sciences 20, 495–497.

Friedel, J. (1974). On the stability of the body centred cubic phase in metals at high  
temperatures. Journal de Physique Lettres 35, 59–63.

Frumkin, I., Lajoie, M.J., Gregg, C.J., Hornung, G., Church, G.M., and Pilpel, Y.  
(2018). Codon usage of highly expressed genes affects proteome-wide translation  
efficiency. Proceedings of the National Academy of Sciences 115, E4940–E4949.

Galperin, M.Y., Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2015). Expanded  
microbial genome coverage and improved protein family annotation in the COG  
database. Nucleic Acids Research 43, D261–269.

Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muñiz-  
Rascado, L., García-Sotelo, J.S., Alquicira-Hernández, K., Martínez-Flores, I., Pan-  
nier, L., Castro-Mondragón, J.A., et al. (2016). RegulonDB version 9.0: High-level  
integration of gene regulation, coexpression, motif clustering and beyond. Nucleic  
Acids Research 44, D133–D143.

Garcia, H.G., and Phillips, R. (2011). Quantitative dissection of the simple repres-  
sion input-output function. Proceedings of the National Academy of Sciences 108,  
12173–12178.

Garcia, H.G., Lee, H.J., Boedicker, J.Q., and Phillips, R. (2011a). Comparison and  
Calibration of Different Reporters for Quantitative Analysis of Gene Expression.

- Biophysical Journal 101, 535–544.
- Garcia, H.G., Lee, H.J., Boedicker, J.Q., and Phillips, R. (2011b). Comparison and Calibration of Different Reporters for Quantitative Analysis of Gene Expression. Biophysical Journal 101, 535–544.
- Garcia, H.G., Sanchez, A., Boedicker, J.Q., Osborne, M., Gelles, J., Kondev, J., and Phillips, R. (2012). Operator Sequence Alters Gene Expression Independently of Transcription Factor Occupancy in Bacteria. Cell Reports 2, 150–161.
- Gardino, A.K., Volkman, B.F., Cho, H.S., Lee, S.-Y., Wemmer, D.E., and Kern, D. (2003). The NMR Solution Structure of BeF<sub>3</sub>-Activated Spo0F Reveals the Conformational Switch in a Phosphorelay System. Journal of Molecular Biology 331, 245–254.
- Gerland, U., Moroz, J.D., and Hwa, T. (2002). Physical constraints and functional characteristics of transcription factor-DNA interaction. Proceedings of the National Academy of Sciences 99, 12015–12020.
- Gilbert, W., and Müller-Hill, B. (1966). ISOLATION OF THE LAC REPRESSOR. Proceedings of the National Academy of Sciences of the United States of America 56, 1891–1898.
- Goethe, M., Fita, I., and Rubi, J.M. (2015). Vibrational Entropy of a Protein: Large Differences between Distinct Conformations. Journal of Chemical Theory and Computation 11, 351–359.
- Good, I.J. (1950). Probability and the Weighting of Evidence (New York City).
- Harman, J.G. (2001). Allosteric regulation of the cAMP receptor protein. Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology 1547, 1–17.
- Hase, C.C., Minchin, R.F., Kloda, A., and Martinac, B. (1997). Cross-linking studies and membrane localization and assembly of radiolabelled large mechanosensitive ion channel (MscL) of *Escherichia Coli*. Biochem Biophys Res Commun 232, 777–

782.

- Haswell, E.S., Phillips, R., and Rees, D.C. (2011). Mechanosensitive Channels: What Can They Do and How Do They Do It? *Structure* *19*, 1356–1369.
- Herbig, U., Jobling, W.A., Chen, B.P.C., Chen, D.J., and Sedivy, J.M. (2004). Telomere Shortening Triggers Senescence of Human Cells through a Pathway Involving ATM, p53, and p21CIP1, but Not p16INK4a. *Molecular Cell* *14*, 501–513.
- Huang, M., Song, K., Liu, X., Lu, S., Shen, Q., Wang, R., Gao, J., Hong, Y., Li, Q., Ni, D., et al. (2018). AlloFinder: A strategy for allosteric modulator discovery and allosterome analyses. *Nucleic Acids Research* *46*, W451–W458.
- Hui, S., Silverman, J.M., Chen, S.S., Erickson, D.W., Basan, M., Wang, J., Hwa, T., and Williamson, J.R. (2015). Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria. *Molecular Systems Biology* *11*.
- Jones, D.L., Brewster, R.C., and Phillips, R. (2014). Promoter architecture dictates cell-to-cell variability in gene expression. *Science* *346*, 1533–1536.
- Jun, S., Si, F., Pugatch, R., and Scott, M. (2018). Fundamental principles in bacterial physiologyHistory, recent progress, and the future with focus on cell size control: A review. *Reports on Progress in Physics* *81*, 056601.
- Kao-Huang, Y., Revzin, A., Butler, A.P., O'Conner, P., Noble, D.W., and Hippel, P.H.V. (1977). Nonspecific DNA binding of genome-regulating proteins as a biological control mechanism: Measurement of DNA-bound *Escherichia coli* lac repressor in vivo. *Proceedings of the National Academy of Sciences* *74*, 4228–4232.
- Kepler, T.B., and Elston, T.C. (2001). Stochasticity in Transcriptional Regulation: Origins, Consequences, and Mathematical Representations. *Biophysical Journal* *81*, 3116–3136.
- Keymer, J.E., Endres, R.G., Skoge, M., Meir, Y., and Wingreen, N.S. (2006). Chemosensing in *Escherichia coli*: Two regimes of two-state receptors. *Proceedings of the National Academy of Sciences* *103*, 1786–1791.

- Kim, N.H., Lee, G., Sherer, N.A., Martini, K.M., Goldenfeld, N., and Kuhlman, T.E. (2016). Real-time transposable element activity in individual live cells. *Proceedings of the National Academy of Sciences* *113*, 7278–7283.
- Kim, P.-J., Lee, D.-Y., Kim, T.Y., Lee, K.H., Jeong, H., Lee, S.Y., and Park, S. (2007). Metabolite essentiality elucidates robustness of *Escherichia coli* metabolism. *Proceedings of the National Academy of Sciences of the United States of America* *104*, 13638–13642.
- Klumpp, S., and Hwa, T. (2008). Growth-rate-dependent partitioning of RNA polymerases in bacteria. *Proceedings of the National Academy of Sciences* *105*, 20245–20250.
- Klumpp, S., and Hwa, T. (2014). Bacterial growth: Global effects on gene expression, growth feedback and proteome partition. *Current Opinion in Biotechnology* *28*, 96–102.
- Ko, M.S.H. (1991). A stochastic model for gene induction. *Journal of Theoretical Biology* *153*, 181–194.
- Krembel, A., Colin, R., and Sourjik, V. (2015a). Importance of Multiple Methylation Sites in *Escherichia Coli* Chemotaxis. *PLoS ONE* *10*.
- Krembel, A.K., Neumann, S., and Sourjik, V. (2015b). Universal Response-Adaptation Relation in Bacterial Chemotaxis. *Journal of Bacteriology* *197*, 307–313.
- Kuhlman, T., Zhang, Z., Saier, M.H., and Hwa, T. (2007). Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proceedings of the National Academy of Sciences* *104*, 6043–6048.
- Lanfranco, M.F., Gárate, F., Engdahl, A.J., and Maillard, R.A. (2017). Asymmetric configurations in a reengineered homodimer reveal multiple subunit communication pathways in protein allostery. *Journal of Biological Chemistry* *292*, 6086–6093.
- Laxhuber, K.S., Morrison, M., Chure, G., Belliveau, N.M., Strandkvist, C., Naughton, K., and Phillips, R. (2020). Theoretical Investigation of a Genetic Switch for Metabolic

Adaptation. PLOS ONE.

Lässig, M., Mustonen, V., and Walczak, A.M. (2017). Predicting evolution. *Nature Ecology & Evolution* 1, 0077.

Levantino, M., Spilotros, A., Cammarata, M., Schirò, G., Ardiccioni, C., Vallone, B., Brunori, M., and Cupane, A. (2012). The Monod-Wyman-Changeux allosteric model accounts for the quaternary transition dynamics in wild type and a recombinant mutant human hemoglobin. *Proceedings of the National Academy of Sciences* 109, 14894–14899.

Levina, N., Totemeyer, S., Stokes, N.R., Louis, P., Jones, M.A., and Booth, I.R. (1999). Protection of *Escherichia Coli* cells against extreme turgor by activation of MscS and MscL mechanosensitive channels: Identification of genes required for MscS activity. *EMBO J* 18, 1730–1737.

Lewis, M., Chang, G., Horton, N.C., Kercher, M.A., Pace, H.C., Schumacher, M.A., Brennan, R.G., and Lu, P. (1996). Crystal Structure of the Lactose Operon Repressor and Its Complexes with DNA and Inducer. *Science* 271, 1247–1254.

Li, G.-W., Burkhardt, D., Gross, C., and Weissman, J.S. (2014). Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. *Cell* 157, 624–635.

Li, X.-t., Thomason, L.C., Sawitzke, J.A., Costantino, N., and Court, D.L. (2013). Positive and negative selection using the tetA-sacB cassette: Recombineering and P1 transduction in *Escherichia Coli*. *Nucleic Acids Research* 41, e204–e204.

Lindsley, J.E., and Rutter, J. (2006). Whence cometh the allosterome? *Proceedings of the National Academy of Sciences* 103, 10533–10535.

Liu, F., Morrison, A.H., and Gregor, T. (2013). Dynamic interpretation of maternal inputs by the *Drosophila* segmentation gene network. *Proceedings of the National Academy of Sciences of the United States of America* 110, 6724–6729.

Lo, K., Brinkman, R.R., and Gottardo, R. (2008). Automated gating of flow cytometry

- etry data via robust model-based clustering. *Cytometry Part A* 73A, 321–332.
- Loison, L. (2013). Monod before Monod: Enzymatic Adaptation, Lwoff, and the Legacy of General Biology. *History and Philosophy of the Life Sciences* 35, 167–192.
- Louhivuori, M., Risselada, H.J., van der Giessen, E., and Marrink, S.J. (2010). Release of content through mechano-sensitive gates in pressurized liposomes. *Proc Natl Acad Sci U S A* 107, 19856–19860.
- Lovely, G.A., Brewster, R.C., Schatz, D.G., Baltimore, D., and Phillips, R. (2015). Single-molecule analysis of RAG-Mediated V(D)J DNA cleavage. *Proceedings of the National Academy of Sciences* 112, E1715–E1723.
- Lutz, R., and Bujard, H. (1997). Independent and Tight Regulation of Transcriptional Units in *Escherichia Coli* Via the LacR/O, the TetR/O and AraC/I1-I2 Regulatory Elements. *Nucleic Acids Research* 25, 1203–1210.
- Martinac, B., Buechner, M., Delcour, A.H., Adler, J., and Kung, C. (1987). Pressure-sensitive ion channel in *Escherichia Coli*. *Proc Natl Acad Sci U S A* 84, 2297–2301.
- Martins, B.M.C., and Swain, P.S. (2011). Trade-Offs and Constraints in Allosteric Sensing. *PLOS Computational Biology* 7, e1002261.
- Martínez-Gómez, K., Flores, N., Castañeda, H.M., Martínez-Batallar, G., Hernández-Chávez, G., Ramírez, O.T., Gosset, G., Encarnación, S., and Bolivar, F. (2012). New insights into *Escherichia coli* metabolism: Carbon scavenging, acetate metabolism and carbon recycling responses during growth on glycerol. *Microbial Cell Factories* 11, 46.
- Marzen, S., Garcia, H.G., and Phillips, R. (2013). Statistical Mechanics of Monod-Wyman-Changeux (MWC) Models. *Journal of Molecular Biology* 425, 1433–1460.
- McLaughlin Jr, R.N., Poelwijk, F.J., Raman, A., Gosal, W.S., and Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. *Nature* 491, 138–142.

- Michel, D. (2010). How transcription factors can adjust the gene expression flood-gates. *Progress in Biophysics and Molecular Biology* 102, 16–37.
- Milo, R., Hou, J.H., Springer, M., Brenner, M.P., and Kirschner, M.W. (2007). The relationship between evolutionary and physiological variation in hemoglobin. *Proceedings of the National Academy of Sciences* 104, 16998–17003.
- Milo, R., Jorgensen, P., Moran, U., Weber, G., and Springer, M. (2010). BioNumbers—the database of key numbers in molecular and cell biology. *Nucleic Acids Research* 38, D750–D753.
- Mirny, L.A. (2010). Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences* 107, 22534–22539.
- Mishra, V., Skotak, M., Schuetz, H., Heller, A., Haorah, J., and Chandra, N. (2016). Primary blast causes mild, moderate, severe and lethal TBI with increasing blast overpressures: Experimental rat injury model. *Scientific Reports* 6, 26992.
- Mondal, J., Bratton, B.P., Li, Y., Yethiraj, A., and Weisshaar, J.C. (2011). Entropy-Based Mechanism of Ribosome-Nucleoid Segregation in *E. coli* Cells. *Biophysical Journal* 100, 2605–2613.
- Monod, J. (1941). Sur un phénomène nouveau de croissance complexe dans les culures bactériennes. *Comptes Rendus Des Séances de L'Académie Des Sciences* 212, 934–939.
- Monod, J. (1947). The Phenomenon of Enzymatic Adaptation And Its bearings on Problems of Genetics and Cellular Differentiation. *Growth Symposium* 9, 223–289.
- Monod, J. (1966). From Enzymatic Adaptation to Allosteric Transitions. *Science* 154, 475–483.
- Monod, J., Changeux, J.-P., and Jacob, F. (1963). Allosteric proteins and cellular control systems. *Journal of Molecular Biology* 6, 306–329.
- Monod, J., Wyman, J., and Changeux, J.-P. (1965). On the nature of allosteric transitions: A plausible model. *Journal of Molecular Biology* 12, 88–118.

- Moon, T.S., Lou, C., Tamsir, A., Stanton, B.C., and Voigt, C.A. (2012). Genetic programs constructed from layered logic gates in single cells. *Nature* *491*, 249–253.
- Motlagh, H.N., Wrabl, J.O., Li, J., and Hilser, V.J. (2014). The ensemble nature of allostery. *Nature* *508*, 331–339.
- Murphy, K.F., Balázsi, G., and Collins, J.J. (2007). Combinatorial promoter design for engineering noisy gene expression. *Proceedings of the National Academy of Sciences* *104*, 12726–12731.
- Nagai, T., Ibata, K., Park, E.S., Kubota, M., Mikoshiba, K., and Miyawaki, A. (2002). A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. *Nature Biotechnology* *20*, 87–90.
- Naismith, J.H., and Booth, I.R. (2012). Bacterial mechanosensitive channelsMscS: Evolution's solution to creating sensitivity in function. *Annu Rev Biophys* *41*, 157–177.
- Nayak, C.R., and Rutenberg, A.D. (2011). Quantification of Fluorophore Copy Number from Intrinsic Fluctuations during Fluorescence Photobleaching. *Bioophysical Journal* *101*, 2284–2293.
- Nordström, K., and Dasgupta, S. (2006). Copy-number control of the Escherichia coli chromosome: A plasmidologist's view. *EMBO Reports* *7*, 484–489.
- Oehler, S., Amouyal, M., Kolkhof, P., von Wilcken-Bergmann, B., and Müller-Hill, B. (1994). Quality and position of the three lac operators of E. Coli define efficiency of repression. *The EMBO Journal* *13*, 3348–3355.
- O'Gorman, R.B., Rosenberg, J.M., Kallai, O.B., Dickerson, R.E., Itakura, K., Riggs, A.D., and Matthews, K.S. (1980). Equilibrium binding of inducer to lac repressor operator DNA complex. *Journal of Biological Chemistry* *255*, 10107–10114.
- Peebo, K., Valgepea, K., Maser, A., Nahku, R., Adamberg, K., and Vilu, R. (2015). Proteome reallocation in Escherichia coli with increasing specific growth rate. *Molecular BioSystems* *11*, 1184–1193.

- Perutz, M.F., and Mitchison, J.M. (1950). State of Hæmoglobin in Sickle-Cell Anæmia. *Nature* *166*, 677–679.
- Phillips, R. (2001). Crystals, Defects and Microstructures by Rob Phillips ([/core/books/crystals-defects-and-microstructures/85B8CE3A333532ABC0C21CC7E2096B50](http://core/books/crystals-defects-and-microstructures/85B8CE3A333532ABC0C21CC7E2096B50)).
- Phillips, R. (2015). Napoleon Is in Equilibrium. *Annual Review of Condensed Matter Physics* *6*, 85–111.
- Phillips, R., Belliveau, N.M., Chure, G., Garcia, H.G., Razo-Mejia, M., and Scholes, C. (2019). Figure 1 Theory Meets Figure 2 Experiments in the Study of Gene Expression. *Annual Review of Biophysics* *48*, 121–163.
- Pilizota, T., and Shaevitz, J.W. (2012). Fast, Multiphase Volume Adaptation to Hyperosmotic Shock by *Escherichia coli*. *PLOS ONE* *7*, e35205.
- Pilizota, T., and Shaevitz, J.W. (2014). Origins of *Escherichia coli* Growth Rate and Cell Shape Changes at High External Osmolality. *Biophysical Journal* *107*, 1962–1969.
- Poelwijk, F.J., Heyning, P.D., de Vos, M.G., Kiviet, D.J., and Tans, S.J. (2011). Optimality and evolution of transcriptionally regulated gene expression. *BMC Systems Biology* *5*, 128.
- Raman, A.S., White, K.I., and Ranganathan, R. (2016). Origins of Allostery and Evolvability in Proteins: A Case Study. *Cell* *166*, 468–480.
- Razo-Mejia, M., Boedicker, J.Q., Jones, D., DeLuna, A., Kinney, J.B., and Phillips, R. (2014). Comparison of the theoretical and real-world evolutionary potential of a genetic circuit. *Physical Biology* *11*, 026005.
- Razo-Mejia, M., Barnes, S.L., Belliveau, N.M., Chure, G., Einav, T., Lewis, M., and Phillips, R. (2018). Tuning Transcriptional Regulation through Signaling: A Predictive Theory of Allosteric Induction. *Cell Systems* *6*, 456–469.e10.
- Razo-Mejia, M., Marzen, S., Chure, G., Morrison, M., Taubman, R., and Phillips, R. (2020). First-principles prediction of the information processing capacity of a

simple genetic circuit. In Preparation.

Reynolds, K.A., McLaughlin, R.N., and Ranganathan, R. (2011). Hot Spots for Allosteric Regulation on Protein Surfaces. *Cell* 147, 1564–1575.

Rogers, J.K., Guzman, C.D., Taylor, N.D., Raman, S., Anderson, K., and Church, G.M. (2015). Synthetic biosensors for precise gene control and real-time monitoring of metabolites. *Nucleic Acids Research* 43, 7648–7660.

Rohlhill, J., Sandoval, N.R., and Papoutsakis, E.T. (2017). Sort-Seq Approach to Engineering a Formaldehyde-Inducible Promoter for Dynamically Regulated *Escherichia coli* Growth on Methanol. *ACS Synthetic Biology* 6, 1584–1595.

Rosenfeld, N., Elowitz, M.B., and Alon, U. (2002). Negative autoregulation speeds the response times of transcription networks. *J Mol Biol* 323, 785–793.

Rosenfeld, N., Young, J.W., Alon, U., Swain, P.S., and Elowitz, M.B. (2005). Gene Regulation at the Single-Cell Level. *Science* 307, 1962–1965.

Rosenfeld, N., Perkins, T.J., Alon, U., Elowitz, M.B., and Swain, P.S. (2006). A Fluctuation Method to Quantify In Vivo Fluorescence Data. *Biophysical Journal* 91, 759–766.

Rydenfelt, M., Cox, R.S., Garcia, H., and Phillips, R. (2014a). Statistical mechanical model of coupled transcription from multiple promoters due to transcription factor titration. *Physical Review E* 89, 012702.

Rydenfelt, M., Garcia, H.G., Cox, R.S., and Phillips, R. (2014b). The Influence of Promoter Architectures and Regulatory Motifs on Gene Expression in *Escherichia coli*. *PLoS ONE* 9, e114347.

Salis, H.M., Mirsky, E.A., and Voigt, C.A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology* 27, 946–950.

Schaechter, M., Maaløe, O., and Kjeldgaard, N.O. (1958). Dependency on Medium and Temperature of Cell Size and Chemical Composition during Balanced Growth

- of *Salmonella typhimurium*. *Microbiology* *19*, 592–606.
- Schatz, D.G., and Baltimore, D. (2004). Uncovering the V(D)J recombinase. *Cell* *116*, S103–S108.
- Schatz, D.G., and Ji, Y. (2011). Recombination centres and the orchestration of V(D)J recombination. *Nature Reviews Immunology* *11*, 251–263.
- Schlake, T., and Bode, J. (1994). Use of Mutated FLP Recognition Target (FRT) Sites for the Exchange of Expression Cassettes at Defined Chromosomal Loci. *Biochemistry* *33*, 12746–12751.
- Schmidt, A., Kochanowski, K., Vedelaar, S., Ahrné, E., Volkmer, B., Callipo, L., Knoops, K., Bauer, M., Aebersold, R., and Heinemann, M. (2016). The quantitative and condition-dependent *Escherichia coli* proteome. *Nature Biotechnology* *34*, 104–110.
- Schumann, U., Edwards, M.D., Rasmussen, T., Bartlett, W., van West, P., and Booth, I.R. (2010). YbdG in *Escherichia Coli* is a threshold-setting mechanosensitive channel with MscM activity. *Proc Natl Acad Sci U S A* *107*, 12664–12669.
- Scott, M., Gunderson, C.W., Mateescu, E.M., Zhang, Z., and Hwa, T. (2010). Interdependence of Cell Growth and Gene Expression: Origins and Consequences. *Science* *330*, 1099–1102.
- Scott, M., Klumpp, S., Mateescu, E.M., and Hwa, T. (2014). Emergence of robust growth laws from optimal regulation of ribosome synthesis. *Molecular Systems Biology* *10*, 747–747.
- Setty, Y., Mayo, A.E., Surette, M.G., and Alon, U. (2003). Detailed map of a cis-regulatory input function. *Proceedings of the National Academy of Sciences* *100*, 7702–7707.
- Sharan, S.K., Thomason, L.C., Kuznetsov, S.G., and Court, D.L. (2009). Recombineering: A homologous recombination-based method of genetic engineering. *Nat Protoc* *4*, 206–223.

- Shehata, T.E., and Marr, A.G. (1975). Effect of temperature on the size of Escherichia coli cells. *Journal of Bacteriology* *124*, 857–862.
- Shis, D.L., Hussain, F., Meinhardt, S., Swint-Kruse, L., and Bennett, M.R. (2014). Modular, Multi-Input Transcriptional Logic Gating with Orthogonal LacI/GalR Family Chimeras. *ACS Synthetic Biology* *3*, 645–651.
- Sivia, D., and Skilling, J. (2006). *Data Analysis: A Bayesian Tutorial* (OUP Oxford).
- Sochor, M.A. (2014). In vitro transcription accurately predicts lac repressor phenotype in vivo in Escherichia coli. *PeerJ* *2*, e498.
- Soufi, B., Krug, K., Harst, A., and Macek, B. (2015). Characterization of the *E. coli* proteome and its modifications during growth and ethanol stress. *Frontiers in Microbiology* *6*.
- Sourjik, V., and Berg, H.C. (2002). Receptor sensitivity in bacterial chemotaxis. *Proceedings of the National Academy of Sciences* *99*, 123–127.
- Stahler, G.J., Mennis, J., Belenko, S., Welsh, W.N., Hiller, M.L., and Zajac, G. (2013). Predicting Recidivism For Released State Prison Offenders. *Criminal Justice and Behavior* *40*, 690–711.
- Stokes, N.R., Murray, H.D., Subramaniam, C., Gourse, R.L., Louis, P., Bartlett, W., Miller, S., and Booth, I.R. (2003). A role for mechanosensitive channels in survival of stationary phase: Regulation of channel expression by RpoS. *Proceedings of the National Academy of Sciences* *100*, 15959–15964.
- Stylianidou, S., Brennan, C., Nissen, S.B., Kuwada, N.J., and Wiggins, P.A. (2016). SuperSegger: Robust image segmentation, analysis and lineage tracking of bacterial cells. *Molecular Microbiology* *102*, 690–700.
- Süel, G.M., Lockless, S.W., Wall, M.A., and Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology* *10*, 59–69.
- Swain, P.S., Stevenson, K., Leary, A., Montano-Gutierrez, L.F., Clark, I.B.N., Vogel,

- J., and Pilizota, T. (2016). Inferring time derivatives including cell growth rates using Gaussian processes. *Nature Communications* 7, 13766.
- Swem, L.R., Swem, D.L., Wingreen, N.S., and Bassler, B.L. (2008). Deducing Receptor Signaling Parameters from In Vivo Analysis: LuxN/AI-1 Quorum Sensing in *Vibrio harveyi*. *Cell* 134, 461–473.
- Taheri-Araghi, S., Bradde, S., Sauls, J.T., Hill, N.S., Levin, P.A., Paulsson, J., Vergasola, M., and Jun, S. (2015a). Cell-size control and homeostasis in bacteria. *Current Biology* : CB 25, 385–391.
- Taheri-Araghi, S., Bradde, S., Sauls, J.T., Hill, N.S., Levin, P.A., Paulsson, J., Vergasola, M., and Jun, S. (2015a). Cell-size control and homeostasis in bacteria. *Current Biology* : CB 25, 385–391.
- Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. (2018). Validating Bayesian Inference Algorithms with Simulation-Based Calibration. *arXiv:1804.06788 [Stat]*.
- Thomason, L.C., Costantino, N., and Court, D.L. (2007). E. Coli Genome Manipulation by P1 Transduction. *Current Protocols in Molecular Biology* 79, 1.17.1–1.17.8.
- Tungtur, S., Skinner, H., Zhan, H., Swint-Kruse, L., and Beckett, D. (2011). In vivo tests of thermodynamic models of transcription repressor function. *Biophysical Chemistry* 159, 142–151.
- Ullmann, A. (2011). In Memoriam: Jacques Monod (19101976). *Genome Biology and Evolution* 3, 1025–1033.
- Ursell, T., Phillips, R., Kondev, J., Reeves, D., and Wiggins, P.A. (2008). The role of lipid bilayer mechanics in mechanosensation. In *Mechanosensitivity in Cells and Tissues 1: Mechanosensitive Ion Channels*, A. Kamkin, and I. Kiseleva, eds. (Springer-Verlag), pp. 37–70.
- Valgepea, K., Adamberg, K., Seiman, A., and Vilu, R. (2013). *Escherichia coli* achieves faster growth by increasing catalytic and translation rates of proteins.

Molecular BioSystems 9, 2344.

van den Berg, J., Galbiati, H., Rasmussen, A., Miller, S., and Poolman, B. (2016). On the mobility, membrane location and functionality of mechanosensitive channels in *Escherichia Coli*. Scientific Reports 6.

Velyvis, A., Yang, Y.R., Schachman, H.K., and Kay, L.E. (2007). A solution NMR study showing that active site ligands and nucleotides directly perturb the allosteric equilibrium in aspartate transcarbamoylase. Proceedings of the National Academy of Sciences 104, 8815–8820.

Victorelli, S., and Passos, J.F. (2017). Telomeres and Cell Senescence - Size Matters Not. EBioMedicine 21, 14–20.

Vilar, J.M.G., and Leibler, S. (2003). DNA Looping and Physical Constraints on Transcription Regulation. Journal of Molecular Biology 331, 981–989.

Vilar, J.M.G., and Saiz, L. (2013). Reliable Prediction of Complex Phenotypes from a Modular Design in Free Energy Space: An Extensive Exploration of the lac Operon. ACS Synthetic Biology 2, 576–586.

Weinert, F.M., Brewster, R.C., Rydenfelt, M., Phillips, R., and Kegel, W.K. (2014). Scaling of Gene Expression with Transcription-Factor Fugacity. Physical Review Letters 113.

Yakovchuk, P., Protozanova, E., and Frank-Kamenetskii, M.D. (2006). Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. Nucleic Acids Research 34, 564–574.

Zeldovich, K.B., and Shakhnovich, E.I. (2008). Understanding Protein Evolution: From Protein Physics to Darwinian Selection. Annual Review of Physical Chemistry 59, 105–127.

Zhu, X.-D., and Sadowski, P.D. (1995). Cleavage-dependent Ligation by the FLP Recombinase CHARACTERIZATION OF A MUTANT FLP PROTEIN WITH AN ALTERATION IN A CATALYTIC AMINO ACID. Journal of Biological Chemistry

270, 23044–23054.

## QUESTIONNAIRE

*Chapter 11*

**CONSENT FORM**