# Physical Biology of Cellular Information Processing

Thesis by
Manuel Razo-Mejia

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

2022
Defended August 17, 2021

# ACKNOWLEDGEMENTS

The opportunity to reflect on having had the experience to do a Ph.D. at an institution of Caltech's caliber brought with it a deep reflection of how unlikely my twenties were. Ten years ago, when my adventure into science began, I was just a naive kid with a romantic view of what being a scientist meant and with no real hands-on experience to back up any of such idealized views. On paper, I was not supposed to be here. Not only I graduated from an institution that nobody ever heard about, but I was explicitly told by my professors at this institution that I was "just a kid from Mexico" and that I was not going to amount to much. My big dreams of making it out of my situation to study abroad and dedicate my life to understanding the natural world were almost shattered beyond repair. During this absolute low point of my life, a series of fortunate events got me here. This journey over the last decade has only been possible because I have had the immense fortune of being surrounded by incredible people. I want to take this opportunity to recall the story of how I got into Caltech, hopefully thanking every person that has helped me along the way to reach this moment.

As I just mentioned, before coming to Caltech for the first time, my scientific credentials were next to null. Circumstances I did not choose had me stuck in an unfavorable situation where almost every professor at my college thought of me as "a stubborn and hard-headed guy, too ambitious for his good." Multiple attempts to bring my spirit down were starting to be successful, as I seriously considered dropping out of college to try my luck with something else. This is where the first big name in this story appeared. Adrian Jinich, back then a staff scientist at a nearby research center and a soon-to-be graduate student at Harvard, came to my middle-of-nowhere school with a simple idea: he was going to inspire the next generation of Mexican scientists by working with students on a hands-on project during the weekends. Those Saturday meetings with Adrian were going to change my life in unexpected ways.

Adrian happened to have a textbook that has guided my thinking and my love for science for the last decade and, with a little bit of more luck, will continue to do so for decades to come. This textbook, Physical Biology of the Cell, by Rob Phillips, Jane Kondev, Julie Theriot, and Hernan Garcia instantly became a pseudo-religious bible to me. Since I was not learning anything useful in school, I dedicated countless hours to understanding the contents of this textbook. And, on a lonely night of April 2011, when I was feeling more desperate than ever after having gotten into a heated discussion with one of my professors, I sent an innocent email to the first author of the book with zero expectations of ever receiving a reply. The email said something like this:

Dear Professor Phillips PhD,

My name is Manuel Razo, I'm a Mexican student in the National Polytechnic Institute in Mexico... The reason for writing you this letter is because I've been studying your book... I found fascinating every single page in the book, to the point that I'm looking forward to work one day in a topic related with biophysics...I know that you are a very busy and important Scientist, but it would mean a lot to me if you could give me any advice of what to do in order to achieve my goals.

Little I knew that this innocent email would set my career path for the rest of my life. A few weeks after sending this email, Mr. Phillips took the time to reply to the message, giving me inspiring words about how he sees life and the meaning of being a scientist. One conversation led to another, and sooner than I realized, Rob challenged me to come up with a project to work on during a summer internship. Talks with Adrian and a lot of reading led me to an idea. I wrote down this proposal and sent it to Rob. At this time, I did not know what to expect. I had never been involved in any research project or knew what it took to become a Caltech summer student. But, to my fortune, Rob liked my naive idea and invited me to spend not only the summer but an entire year working in his lab. So there I was, a nobody with no credentials being given the opportunity of my life. I will have more to say about Mr. Phillips later on.

That year, my horizons expanded like never before. I passed from only interacting with people with my cultural background that spoke my language to be in an incredibly diverse and stimulating environment that encouraged bold ideas. At that time, I met two people that became mentors and friends. James Boedicker, a postdoc in the lab at the time, now a tenured faculty member at USC, went beyond what he needed to do when he taught me how to do experiments. He taught me how to use a micropipette, grow bacteria, and organize experimental work. But his mentoring and advice went beyond lab work. He and his wife Amy invited me for lunch or dinner on several occasions since they knew I was on an extremely tight budget. James became my second mentor and my primary guide when it came time to apply to grad school. His kindness, patience, and caring showed me that there are multiple facets to every society. Although the news wanted me to believe that every person in this country automatically hated me because of my nationality, James' friendship shattered all dumb preconceptions I could have had.

If there is something I did not expect to come across in a place like Caltech, that would be Ernie's Mexican food truck's enormous popularity. It was during the classic Thursday visits to Ernie's truck that I got to interact with Justin Bois. At the time, Justin was doing his second postdoc at UCLA. He would come once a week during our group meeting to have the chance to hang out with other physical biologists. Rob decided to recruit Justin and me as honorary TAs for his famous Bi1x class. Every Sunday for the next few months, I helped Justin develop a new module for the class involving fluorescent microscopy on fly embryos. These interactions turned into a friendship. Ever since, Justin and his wife Ramit have treated me and my family unbelievably nicely. After Rob, Justin has been the scientist that has influenced my thinking the most. His classes, sharp mind, computational skills, unbelievably ethical behavior, and his will to share his knowledge set me on the right track from the beginning of my Ph.D., when Justin started a position as a teaching faculty for the Biology division.

As the reader can infer from the context of these acknowledgments, a year at Caltech was not enough. As soon as I could, I came back to Caltech to begin my Ph.D. working again with Rob. My twenties came and went by faster than I thought. Nevertheless, I had the immense fortune to share the lab with three of my great friends and co-authors for many of those years. Nathan Belliveau, Muir Morrison, and Griffin Chure became my lab family. They helped me think through problems and worked with me on numerous successful (and unsuccessful) projects, as we all grew as scientists under Rob's tutelage. It is not an exaggeration to say that their friendship and support have been a fundamental part of why my Ph.D. experience has been the best time of my life. I would not do justice to Griffin's influence on my thinking if I did not give a special shout-out to him. For the past seven years, Griffin and I talked almost every single day. We fed off each other's ideas and together traveled the world. From Korea to New Zealand to Sweden, Griffin became my closest collaborator and someone that every time he has something to say, I am ready to take notes on whatever comes out of his creative and rigorous mind.

Within the Phillips lab, I always felt the environment encouraged collaborations and casual science conversations, for which I will be forever thankful. Rob Brewster, now a professor at U. Mass., taught me several lab tricks that I still use up to this day. Heun Jin Lee, the microscopy master, was an incredible resource and mentor that allowed my research to run smoothly. Without his work and input, it would have been incredibly hard for my projects to take off. I am also thankful to members of the Phillips lab with whom I shared projects, ideas, lunches, and visits to Red Door cafe over the years. People like Suzy Beeler, Stephanie Barnes, Soichi Hirokawa, Niko McCarty, Tom Roeschinger, and Scott Saunders made my Caltech experience much better.

I am enormously thankful to Dianne Newman. I met Dianne for the first time when I was a visiting undergrad at Caltech. I vividly remember having finished giving my MicroMorning talk when Dianne approached me. She congratulated me and told me that I should come back to Caltech since she thought I had what it took to be part of this institution. When I came back to Caltech as a Ph.D. student, I worked in her lab for ten weeks as part of a rotation where I learned about a completely different way to tackling biological problems from what I was used to. Later on, Dianne helped my sister to visit the Caltech daycare and learned from them. She also made my mom unbelievably happy when, in perfect Spanish, she told my mom that I was a great student and that she was very proud of me. Her support and kindness are things I do not take for granted, and I am enormously thankful for them.

I am also thankful to Ginger and Kim Caldwell. Thanks to Mr. and Mrs. Caldwell's kindness, I was awarded a one–year fellowship that covered my stipend. But they did not only give away the money and forgot about it. They took the time to meet with me and learn about my research projects. This experience reinforced my vision of the kindness that this country's society can have for a person like myself, pursuing the American dream.

Of course, not everything in my life for the last few years has been strictly about my Ph.D. project and my work at the lab. I owe a significant component of the enormous happiness I experienced since coming to Caltech to my friends outside of work. In particular, I want to give a shout-out to my best friends Stephanie Threatt and Porfirio Quintero. Stephanie and I share a love for hip-hop and good food. We went to concerts, amusement parks, and restaurants, and she has always been there when I needed emotional support. Porfirio and I not only share a nationality, but we shared an apartment for five years. Living with my best friend and having deep conversations and many other adventures out in the wild made all those years an authentic delight to experience. To both of you, thank you really for making my off-the-lab time in Pasadena such an amazing experience. Let's keep the HOI tradition alive for many, many years to come.

I would also like to give a shout-out to my friend Andres Ortiz. Andres and I have shared long conversations about the meaning of life, physics, math, and everything in between. His powerful intellect and will to always chat about deep stuff make Andres a delightful person to call your friend. More recently, my friend David Larios joined my inner circle and became a really close friend. David and I share a love for learning, talking about science, and, more recently, doing outreach via a podcast where we share ideas from different books that we read. We also share an openness to be vulnerable with each other. The list of friends goes on and on and it would take forever to thank them all. But I would like to give a special mention to my good friends Ellen Yan, Scott Saunders, Enrique Amaya, Alejandro Granados, Jorge Castillo, Emmanuel Flores, Teddy, Adam Gomez, Paola. My friends became my family outside of my home, so I never felt alone during this journey.

And how could I feel alone when so many people that I love were always there for me? Even when I felt like I was feeling homesick (or actually sick or injured), the Osegueda family was there for me. Kathy, Carlos, Patricio, Nico, and Matias adopted me as an extra nephew/cousin, reminding me that family doesn't necessarily need a direct genetic relationship. I treasure every lunch and every dinner as some of the happiest and most delightful times during all these years.

Shockingly enough, with all the good luck that has gotten me to this point, there is a lot more for which to be thankful. Because of Rob, my professional relationships have not been limited to Caltech only. I spent a total of six months at the Marine Biological Laboratory (MBL) as a teaching assistant for the Physiology and the Physical Biology of the Cell courses. There, I had the fortune to interact with Wallace Marshall, Jane Kondev, Hernan Garcia, Greg Huber, Julie Theriot, and Madhav Mani, all of which have been supportive and inspiring throughout these years. At MBL, I also met my good friends Cat Triandafillou and Christina Hueschen, who I'll be joining up north for the coming stage of my career.

After this long list, I would like to dedicate specific words to Rob:

Dear Mr. Phillips, Ph.D.,

It is tough to write this section and not shed tears of happiness. I do not have words to express my feelings, but I will try to do my best. Working and sharing so many moments with you has been by far the most significant honor of my entire life. To me, it is clear that the only reason I am writing a Ph.D. thesis for arguably the best research institute in the world is because of you. When I sent that innocent email, I never imagined the incredible person that would await at the receiving end. Not only you took time out of your insanely busy schedule to talk with a stranger with no credentials whatsoever, but you allowed me to prove myself capable of pursuing my scientific dreams. Even when I almost stopped believing in myself, you believed in me. And, after a decade of having sent this email, I stand proud of what we have accomplished together. This thesis reflects everything I have learned from you as I pave my path to become who I want to be.

On multiple occasions, you have told me that you don't have a recipe for a happy and fulfilling life. But the way you operate and interact with other people makes me want to shape my professional and personal life after yours. I have learned from you that science is not only our job but, if one wants it, it can be a complete lifestyle. Hike after hike, trip after trip, calculation after calculation, I have witnessed your boundless love for the natural world and the constant everyday effort that it takes to get closer to unveiling its mysteries. It is hard to imagine myself one day knowing a fraction of what you know about science, but I will for sure try to keep up with your learning pace, working as hard as I can every single day.

Although I have technically been your employee for the last decade, our relationship has been much deeper than that. Sometimes I have to pinch myself to make sure that this is not a dream. I am actually hanging out with my hero, and he is actually welcoming me in to his family. Everything you have done for me, and now for Daniela, means the world to me. Welcoming us to the most sacred family celebration over Thanksgiving has been a privilege I never expected to live. I cannot thank you, Amy, Molly, Casey, and the rest of your beautiful family, for making us feel part of the festivity and of the extended family. I have even shared some incredibly stimulating and inspiring conversations with Bob Phillips, who literally reshaped my life priorities with his wisdom and kindness.

I know that you despise honors, prizes, and people praising you. But I would not do justice to my feelings if I could not express my deep love and admiration for you. The way you see the world has shaped my vision and understanding of what it means to be alive. Your love for numbers and their explanatory power is an incredible tool I try to apply to science and mundane activities. Your passion for teaching and your unorthodox methodology has shown me that even a random kid from Mexico can get the most complex concepts if one works hard enough and puts the time into it. Each and every lesson you have given me over the last decade will continue to shape my thinking for the rest of my life.

xi

But this is not a farewell. I might be "getting fired" for the n-th time, this time for real. But that does not mean that our scientific adventures have to be over. I hope that we both share the mutual feeling that this is just the beginning. There are many more things to discover, many more books to write (plural for you, obviously), and many more courses to teach together. Thank you for being part of my life, and more importantly, thank you for letting me be a part of yours.

A mi familia:

Papá, Mamá, Hermanini. Es increíble pensar que esta aventura comenzó hace una década. ¿Cuándo íbamos a imaginar que un simple correo que le envié a Rob cambiaría mi vida de manera tan drástica? Pero no toda la fortuna que la vida me ha dado en estos años es obra del azar. Las oportunidades en la vida vienen para las personas que están preparadas para tomarlas. Y no puedo pensar en una mejor preparación que el haber crecido en un hogar como el nuestro.

No es exageración cuando digo que no hay día que reconozca el enorme privilegio que siempre he tenido. A mi hermana y a mi nunca nos faltó nada y siempre nos sobró amor. Cada característica en mi persona de la que me siento orgulloso es el resultado de los valores que ustedes me inculcaron. Mi fortaleza para levantarme todos los días a estudiar se deriva de esas mañanas de estudio a las cinco a.m. donde me mostraron cómo ser disciplinado. Mi compasión hacia mi prójimo son las enseñanzas que los abuelos nos dejaron. El hecho de que sepa hablar inglés es la consecuencia de los sacrificios que ambos hicieron para darnos la mejor educación disponible. Y, aunque los individuos en esa escuela siempre quisieron desahogar sus frustraciones familiares en mí, siempre tuve a mi hermana para defenderme y a mi madre para decirme cómo es que podía sobrellevar esos, pensando en la frase "yo puedo, yo quiero, es muy fácil, y lo voy a lograr."

Esta tesis es clara evidencia de que esa lección de vida encapsulada en la célebre frase de mi madre siempre fue cierta. Sí pude, sí quise, y definitivamente, sí lo logré. Pero, este logro no solo es mío. Después de todo, soy el reflejo de la persona que criaron. Este logro es de la familia Razo–Mejía entera, incluídos mis abuelos, mi tía Chio, mi tía Marilú y mi tío Pepe que siempre ha estado ahí para apoyarme cuando más lo he necesitado. Desde el día que partí de la casa, sabía que tenía lo que se necesitaba para buscar mi felicidad. Todos los miembros de mi familia me dieron las herramientas para construir el camino en busca de mis sueños. Ese camino continua, y las valiosas lecciones que me dieron a través de los años guiarán el resto de mi vida.

Termino una etapa maravillosa de mi vida. Entré al doctorado como un puberto sin ninguna experiencia de haber vivido fuera de casa, y salgo de esta experiencia como un hombre, con un proyecto de vida por delante. No me queda más que agradecerles por siempre haber estado ahí para mi. Gracias por ser mis papás. Gracias por ser mi hermana. Ustedes, junto con Daniela, hacen que valga la pena vivir cada momento.

A Dani:

Los últimos dos años y medio de mi doctorado son muy probablemente la mejor etapa de mi vida (pandemia incluída). La perfecta correlación entre tu llegada a mi vida y esta enorme felicidad no son coincidencia. Bien supo mi madre que yo estaba enamorado, antes de que siquiera le contara sobre tu existencia. Respiraba un aire diferente, exudaba alegría pura, y trataba a mis seres queridos con mayor atención y compasión. Todo porque, desde el sur de este continente, había arrivado la persona que me permitiría por primera vez experimentar lo que es el verdadero amor de pareja.

El camino hasta este punto, en el que ambos nos hemos comprometido a compartir nuestras vidas de aquí en adelante, ha estado lleno tanto de enormes satisfacciones como de momentos difíciles. Sin embargo, como todo lo bueno que vale la pena en esta vida, nuestro amor ha requerido—y seguirá requiriendo—de arduo

y constante trabajo. De cierta manera, el hecho de que nuestro amor no sea un producto terminado, sino un ente en constante construcción, me llena de emoción por nuestra vida juntos. Después de todo, ¿qué sería de mis más grandes momentos de alegría si no tuviera con quién compartir ese sentimiento? Todo ese fervor simplemente aislado en la privacidad de mi experiencia consciente, sin poder ser expresado hacia otra persona; ¡que enorme desperdicio! De igual manera, ¿qué sería de mis momentos de más profunda tristeza y decepción, si no pudiera aliviar mi sufrimiento al estar contigo? Tú haces que el estar vivo, experimentando esta realidad, sea una dicha constante.

Desde el fondo de mi ser, espero que el patrón de felicidad que hemos vivido durante los últimos dos años, sea una muestra de lo que nos espera por el resto de nuestras vidas. Extrañaré nuestra vida en Pasadena, nuestro taco Thursday, y las noches de ver documentales juntos. Pero la vida debe de continuar. Una nueva etapa de auto–descubrimiento y aprendizaje hará de nuestro amor una estructura aún más sólida y compleja, mientras seguimos compartiendo juntos la dicha de estar vivos. Muchísimas gracias por amarme. Yo te amo, y te amaré todos los días de mi vida.

# ABSTRACT

The state of matter that we define as *life* is different from anything else we have encountered so far in the universe. Living systems not only perpetuate their existence out of equilibrium against the will of the second law of thermodynamics, but they do so while keeping up with an ever-changing environment. A key part of this capacity to adapt to environmental changes is the ability of organisms to gather information from their surroundings to put together an adequate response to the challenges presented to them. This thesis presents an effort to understand, from first principles, this fundamental feature of information gathering that all life on earth shares. We dig into the physics behind one of the most pervasive mechanisms through which living systems sense and respond to the environment–the ability to turn *on* and *off* genes. In doing so, we hope to uncover general principles of how organisms deal with the problem of collecting information about the world that surrounds them.

In Chapter 1, we develop the theoretical and conceptual tools to navigate the rest of the thesis. I introduce the idea of gene regulation, as well as different theoretical models of this pervasive biological phenomenon. We also delve into the realm of information theory and learn how the plastic concept of information can be mathematically defined and quantified.

The second stop in our exploration (Chapter 2) asks the following question: can we understand, from first principles, how it is that proteins allow cells to regulate their genes on-demand upon sensing environmental cues? For this, we explore the physics behind transcriptional control due to allosteric transcription factors. Using simple quasi-equilibrium models of the two processes involved in this type of regulation—the regulation of the gene by the binding and unbinding of the transcription factor, and the regulation of the activity of the transcription factor itself by the binding and unbinding of an effector molecule—we are able to predict the input-output function of a simple genetic circuit, and compare such predictions with experimental determinations of the mean response of a population of bacterial cells.

We then expand on these insights to ask questions about the inescapable cell-to-cell variability that isogenic cells encounter. For this, we have to leave behind the pure thermodynamic framework and work in the language of chemical kinetics. This allows us to make predictions beyond the mean input-output gene expression response of cells by reconstructing full gene expression distributions. With these probabilistic input-output functions, in Chapter 3 we formalize the question of the *amount of information* that cells can gather from the environment. For this, we turn to information-theoretic concepts of maximal mutual information (otherwise known as channel capacity) between the state of the environment and the gene expression response from bacterial cells. Finally, we compare our predictions of the maximum amount of information—measured in bits—that cells can gather with single-cell inferences of this quantity.

# PUBLISHED CONTENT AND CONTRIBUTIONS

Muir J. Morrison, **Manuel Razo-Mejia**, and Rob Phillips. "Reconciling Kinetic and Equilibrium Models of Bacterial Transcription." PLoS Computational Biology, 2020. http://arxiv.org/abs/2006.07772. M.R.M. aided with the concept behind the paper, created data-containing figures, and wrote the manuscript.

**Manuel Razo-Mejia**, Sarah Marzen, Griffin Chure, Rachel Taubman, Muir Morrison, and Rob Phillips. "First-Principles Prediction of the Information Processing Capacity of a Simple Genetic Circuit." Physical Review E 102, no. 2 (August 13, 2020): 022404. https://doi.org/10.1103/PhysRevE.102.022404. M.R.M designed and optimized experimental procedures, collected and analyzed experimental data, and wrote the manuscript.

Peter J. Foster, **Manuel Razo-Mejia**, and Rob Phillips. "Measuring the Energetic Costs of Embryonic Development." Developmental Cell 48, no. 5 (March 2019): 591–92. https://doi.org/10.1016/j.devcel.2019.02.016. M.R.M. aided in writing the manuscript.

Rob Phillips, Nathan M. Belliveau, Griffin Chure, Hernan G. Garcia, **Manuel Razo-Mejia**, and Clarissa Scholes. "Figure 1 Theory Meets Figure 2 Experiments in the Study of Gene Expression." Annual Review of Biophysics 48, no. 1 (May 6, 2019): 121–63. https://doi.org/10.1146/annurev-biophys-052118-115525. M.R.M aided in writing and editing of the manuscript.

Griffin Chure, **Manuel Razo-Mejia**, Nathan M. Belliveau, Tal Einav, Zofii A. Kaczmarek, Stephanie L. Barnes, Mitchell Lewis, and Rob Phillips. "Predictive Shifts in Free Energy Couple Mutations to Their Phenotypic Consequences." PNAS, August 26, 2019, 201907869. https://doi.org/10.1073/pnas.1907869116. M.R.M. performed experiments, analyzed data, created data-containing figures, and provided editing and feedback on the manuscript.

**Manuel Razo-Mejia**\*, Stephanie L. Barnes\*, Nathan M. Belliveau\*, Griffin Chure\*, Tal Einav\*, Mitchell Lewis, and Rob Phillips. "Tuning Transcriptional Regulation through Signaling: A Predictive Theory of Allosteric Induction." Cell Systems 6, no. 4 (April 2018): 456-469.e10. https://doi.org/10.1016/j.cels.2018.02.004. \* Contributed equally. M.R.M. designed and performed experiments, collected and analyzed data, and aided in writing the manuscript.

# TABLE OF CONTENTS

# LIST OF ILLUSTRATIONS

# LIST OF TABLES

*Chapter 1*

# FROM BIO TO BIT: HOW DO CELLS SENSE THE WORLD AROUND THEM?

## 1.1 Introduction

In his classic 1944 book *What is Life?*, Schrödinger brought to the attention of the scientific community what he thought were two of the biggest challenges we had ahead of us if we were to understand living systems in the same way we understand the electromagnetic field or the universal law of gravitation [1]. The idea that living organisms could be "accounted for" by physics and chemistry brought with it a new agenda on what needed to be done to transition from a qualitative and descriptive study of the phenomena of life to a quantitative and predictive science in the spirit of the physical sciences. Since the publication of the book, there has been an enormous amount of progress on our understanding of living systems from a first-principles perspective, nevertheless, 75 years later Schrödinger questions are still as relevant and as vibrant as ever before [2].

One of the defining features of living organisms at all scales is their capacity of gathering information from the environment, encode an internal representation of the state of the environment, and generate a response based on this information processing capacity. Researchers in the field of origins-of-life had gone as far as declaring that life emerged when chemical systems underwent a radical phase transition after which they were able to process and use information and free energy [3]. So, although speculative, it is highly probable that the physical theory fulfilling Schrödinger's vision of accounting for the phenomena of life will be the physics of systems capable of processing information [4].

In this context, information does not take the generic concept of possessing practical knowledge about something. In this thesis, we use a precise mathematical definition of information [5]. This formal definition makes information a metric worth quantifying and predicting in various biological contexts as theoretical studies suggest that natural selection might act on the ability of an organism to process information [6]. Thus, working out the physical details of how it is that organisms sense the environment—this is, gather information about the state of the environment, encode such information in some shape or form within their physical boundaries, and take action based on this information—is at the core of the state-of-the-art research in biophysics [7].

The present thesis is an effort towards this vision of understanding biological systems as information processing machines. Our object of study will be gene regulation in bacteria. This particular system has been the subject of study for microbiologists and molecular biologists for decades, and we have come to learn a lot about the microscopic mechanistic details of how bacteria turn on and off their transcriptional machinery [8]. In particular, we will focus on what we think of as the "hydrogen atom" of gene regulation—the so-called simple-repression motif (more on that in the next section). In physics, calling something the hydrogen atom of $X$ means that for the area of study $X$, this "something" represents a system simple enough to be amenable to analytical models that standard mathematical methods can solve, but rich enough to capture the general features of the phenomena. This simple genetic circuit will allow us to write tractable mathematical models to guide our experimental efforts with the ultimate goal of testing our understanding of such systems when predicting how much information a bacterium can gather from the environment using this genetic module.

Professional biophysicists might wish to skip the rest of this chapter as we will lay the foundations needed for the rest of our enterprise. We will introduce the basics of gene expression modeling and the mathematical concept of information and work through every single physical and mathematical prerequisite needed for

the rest of the thesis. The following chapters are structured as follows: Chapter 2 builds on a decade of understanding this hydrogen atom of gene regulation and expands our models' predictive ability by including the effect of environmental effectors. This means that we will consider how gene regulation is affected by the presence of an extracellular inducer molecule. Chapter 3 will expand even further our predictive capacities by building a model capable of making predictions about the cell-to-cell variability inherent to all signaling systems working at the molecular scale. Chapter 4 serves as a Supporting Information section for Chapter 2, detailing every calculation and every inference. Likewise, Chapter 5 expands on Chapter 3, explaining every technical detail.

## 1.2 Gene Regulation as a Physics 101 Problem

As organisms navigate the challenges presented by the environment, they must constantly fight against the will of the second law of thermodynamics to bring them back to an equilibrium state. To face such challenges, cells are equipped with a toolkit of genes written in the language of A, T, C, and G of the genome. We can think of a typical bacteria genome with $\approx 5 \times 10^3$ genes as the blueprint to produce a repertoire of tools that allow cells to thrive under a myriad of circumstances that they face throughout their lives. Given the vast number of challenges that organisms face, there is constant pressure on every living system to use the right tools for the right circumstances. From cells in the fly embryo expressing different genes that will define their identity on the animal's final body plan to a simple bacteria expressing the correct enzymes to process the available nutrients in the environment, all organisms are faced with the task of orchestrating the expression of the correct subset of genes at their disposal when trying to survive.

Our understanding of how organisms regulate their genes' expression is still not as thorough as one might expect, given the effort that has gone into this question. Take, for example, *E. coli*—arguably the most well-characterized model organism— for which we know the regulatory scheme of less than 1/3 of its genes [9]. For more complex organisms such as *Drosophila*, *C. elegans*, or even humans, we are

even more hopeless on getting a holistic view of the regulatory landscape. Nevertheless, we would not be doing justice to the field's significant advances if we were to pretend we are utterly ignorant about how gene regulation takes place in bacteria. There is a rich mechanistic understanding of how the transcriptional machinery takes the information contained in DNA and transcribes it into RNA [8]. The relative simplicity of the process has inspired generations of biophysicists to try to write down minimal models that can describe and predict features of the process of gene regulation [10–12].

These modeling efforts come in two main flavors: equilibrium statistical mechanical models and kinetic models. In the following sections, we will introduce the necessary background for both approaches relevant to the rest of the thesis.

**Minimal Model of Gene Expression**

Let us begin our introduction to gene expression modeling with the simplest example. As shown in Fig. 1.1(A), we imagine that a gene promoter (the region of the gene where transcriptional regulation takes place) produces mRNA at a constant rate $r_m$. Each mRNA can stochastically decay with a rate $\gamma_m$. Our interest is to understand how the mRNA count $m$ changes over time, given these two competing processes. For that, let us write the mRNA count at time $m(t + \Delta t)$, where $t$ is the time—which we are thinking of as being "right now"—and $\Delta t$ is a little time step into the future. The mRNA count can then be predicted by computing

$$m(t + \Delta t) = m(t) + r_m \Delta t - (\gamma_m \Delta t) m(t), \tag{1.1}$$

where we can think of $r_m \Delta t$ as the probability of observing a single mRNA being produced in the time interval $[t, t + \Delta t]$ ($\Delta t$ is so small that we neglect the possibility of seeing multiple mRNAs being produced), and $\gamma_m \Delta t$ the probability of seeing a single mRNA being degraded. But since each mRNA has the same probability of being degraded, the total number of mRNAs that we would see decay in this time window would be the probability per mRNA times the total number of mRNAs. This is in contrast with the production of mRNA, which does not depend on the

current number of mRNAs. If we send the term $m(t)$ to the left-hand side of the equation and divide both sides by $\Delta t$, we obtain

$$\frac{m(t + \Delta t) - m(t)}{\Delta t} = r_m - \gamma_m m(t). \tag{1.2}$$

Upon taking the limit when $\Delta t \to 0$, we see that the left-hand side is the definition of the derivative of the mRNA count with respect to time. We then obtain an ordinary differential equation of the form

$$\frac{dm}{dt} = r_m - \gamma_m m(t). \tag{1.3}$$

Before even attempting to solve 1.3, we can perform a qualitative analysis of the dynamics [13]. It is handy to plot the contribution of each of the components (production and degradation) to the derivative $dm/dt$ as a function of $m$. This is shown in Fig. 1.1(B), where the blue horizontal line $r_m$ shows the production rate—which does not depend on $m$, and the red line shows the degradation term $m\gamma_m$ which scales linearly with $m$. Notice that we do not include the negative sign for the degradation term, i.e., we are not plotting $-m\gamma_m$. The point $m_{ss}$ where both lines intersect represents the point where the production matches the degradation. For all values less than $m_{ss}$, the production term is larger than the degradation, which means that for any value $m < m_{ss}$, the derivative is positive ($dm/dt > 0$), so over time the system will accumulate more mRNA. The opposite is true for all values after $m_{ss}$ where the degradation term is larger than the production term, implying that $dm/dt < 0$. This means that for $m > m_{ss}$, the system will tend to lose mRNA. These opposite trends point to the idea that $m_{ss}$ must be called a stable fixed point of the dynamical system. This can schematically be seen at the bottom of Fig. 1.1(B). The arrowheads' size indicates the system's trend to move either left or right in $m$. Since all arrows point at the special value, $m_{ss}$, we can say that any small perturbation of the system will be dissipated as the system relaxes back to $m_{ss}$.

This qualitative statement can be confirmed by solving Eq. 1.3. If we define the initial condition $m(t = 0) = m_o$ by separation of variables, we will obtain a solution of the form

$$m(t) = m_o e^{-\gamma_m t} + \frac{r_m}{\gamma_m}(1 - e^{-\gamma_m t}). \tag{1.4}$$

In the limit when $t \to \infty$, we can see that the steady-state solution is given by

$$m_{ss} = \frac{r_m}{\gamma_m}. \tag{1.5}$$

Fig. 1.1(C) shows the time evolution of $m$ for different initial values $m_o$. We can see that indeed regardless of the initial mRNA count, the system relaxes exponentially to $m_{ss} = r_m/\gamma_m$.

So far, our model assumes a simple constant transcription rate $r_m$; let us expand this term a little further to include regulation by a transcriptional repressor further down the road. We know that for a transcriptional event to occur, the RNA polymerase (RNAP) must bind to the promoter region and undergo a series of irreversible steps, such as opening the double helix to initiate the DNA sequence's copying into mRNA [8]. But before these irreversible steps take place, there is a chance that the RNAP falls off the promoter. If we assume that these irreversible steps take place on a much longer timescale compared to the initial binding and unbinding of the RNAP on the promoter, we can separate the time scale and investigate them independently. In particular, we can write that mRNA production happens at a rate

$$\text{mRNA production} = r_m \cdot p_{\text{bound}}, \tag{1.6}$$

where we split the original production term into two steps: $p_{\text{bound}}$, the probability of finding an RNAP bound to the promoter, and $r_m$ which captures all of the irreversible downstream steps that take place once the RNAP is engaged in a transcriptional event. A way to think about it—relevant to what I am doing right now as I type my thesis—is to think that the speed at which I type this document has to do with two things: the probability of me being actively working on these notes times the rate at which I type these notes once I engage in the activity. The reason

**(A)** kinetic scheme for mRNA production

**(B)** phase portrait of mRNA dynamics

$$\frac{dm}{dt} = r_m - \gamma_m m(t)$$

**(C)** mRNA dynamics
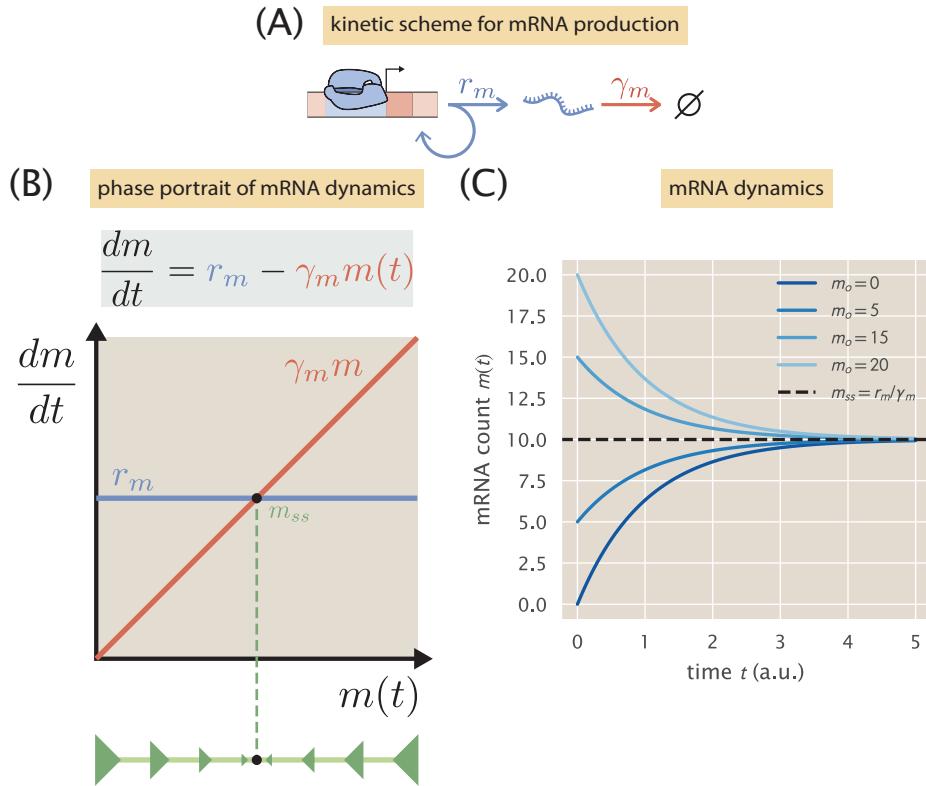
Figure 1.1: **Minimal model of gene expression.** (A) Schematic of the kinetics governing gene expression. mRNA is produced at a constant rate $r_m$ independent of the current mRNA copy number. Degradation of each mRNA occurs at a rate $\gamma_m$. (B) Example of the qualitative analysis of the mRNA dynamics via a 1D phase-portrait. The differential equation governing the dynamics contains two terms: a constant production rate given by $r_m$, and a degradation rate $\gamma_m m$, which depends on the current mRNA count. The main plot shows each of the components in the $m$ vs. $dm/dt$ plot. Since $r_m$ does not depend on the current number of mRNA, it gives a straight production rate as a function of $m$. The total degradation rate depends linearly on the mRNA copy number, giving a line with slope $\gamma_m$. When the two components are equal (bot lines crossing), we obtain the steady-state mRNA value $m_{ss}$. The bottom line shows a qualitative schematic of the flow of the system towards this steady state. The further $m$ is from $m_{ss}$, the faster it moves towards this point as schematized by the arrows' size. (C) Example of mRNA dynamics for different initial conditions. Over time, all curves converge to the steady-state mRNA value $m_{ss} = r_m/\gamma_m$. For this plot, $\gamma_m = 1$ and $r_m/\gamma_m = 10$. The Python code (`ch1_fig01C.py`) used to generate part (C) of this figure can be found on the thesis GitHub repository.

this separation makes sense is that we can include the effect of the regulation by a transcriptional repressor as a reduction of the time (the probability) that the RNAP can be bound to the promoter. Furthermore, since we are assuming that the binding and unbinding of the RNAP happen at a timescale much faster than the downstream events, we can assume that this binding reaction is in quasi-equilibrium, for which we can use the powerful theoretical framework of statistical mechanics. Let us now delve into the basics of this physical theory.

**The Unreasonable Effectiveness of Unrealistic Simplifications**

In the preface of the textbook *Molecular Driving Forces*, Dill and Bromberg introduce the idea of Statistical Mechanics as *the unreasonable effectiveness of unrealistic simplifications* [14]. Although one could make the case that all of physics follows this description, it is undoubtedly evident that statistical mechanics is a vivid example of how simple ideas can have profound consequences. Statistical mechanics can be defined as the theory that, upon assuming the atomic nature of matter, explains the phenomenology that classical thermodynamics established from the interactions of the microscopic components of a system [14]. As with any other physical theory, statistical mechanics is built from a set of *empirical* facts that define "axioms" that we take to be true. In other words, as Feynman famously described to us: if we want to come up with a new law of nature, there is a simple recipe that we must follow:

1. We guess the law. Literally. The most profound understanding of our physical reality we have comes from educated guesses made after a careful observation of nature.

2. We compute the consequences of such a guess. That is why mathematical theories allow us to sharpen our statements about how we think nature works.

3. We compare with experiments/observations. The scientific revolution came about when, after the dark ages, we finally learned it was okay to say "we don't know."

In such a simple statement, Feynman tells us, lies the key to science [15]. For our purpose of understanding the basis of statistical mechanics, we will argue that Boltzmann's law gives the main law upon which the field is founded

$$\frac{P(E_1)}{P(E_2)} = \frac{e^{-E_1/k_B T}}{e^{-E_2/k_B T}}.\qquad(1.7)$$

Let us unpack this equation. The main idea behind statistical mechanics is that macroscopic observables (temperature and pressure in classic examples) are emergent properties dictated by the dynamics of the system's microscopic components. What Boltzmann's law tells us is that the relative probability of a system in thermal equilibrium to be found in a particular microstate with energy $E_1$ compared to being in a microstate with energy $E_2$ is given by an exponential function of the negative energy of such microstate relative to the thermal energy $k_B T$. The minus sign in the exponent comes from the fact that states with negative energies are more favorable by convention in physics. Thus, having a large negative energy has a high probability when taking the exponential of minus such negative number. To provide concrete examples of what a microstate can look like, Fig. 1.2(A) shows three molecular systems relevant to biology. In the first example, we have the classic ligand-receptor binding problem; here, we imagine that a solution can be discretized in space into a series of small boxes. In each of these boxes, one and only one ligand molecule can fit in. In principle, we can list all possible spatial arrangements of ligands. We could then calculate the relative likelihood of finding the system in any configurations as long as we can assign an energy value to each of them. The second example focuses on ligand-gated ion channels. In this particular system, we care about the ion channel's state—either open or closed—and the ligands' binding configuration. If the channel responds to the ligand's concentration by changing its probability of gating, we can calculate using equilibrium statistical mechanics. Finally, the third example shows different configurations of a small patch of the cell membrane. All deformations of a membrane have ener-

getic costs associated with them. By listing all possible membrane configurations, we can calculate the most likely shape of a membrane given the forces and stresses acting on it.

The macroscopic states that we observe can then be thought of as a coarse-graining of many microstates into a single macrostate. For example, in the ligand-receptor binding case, we rarely would care about the specific position of all the ligand molecules in the solution. What we would be interested in is whether or not the ligand is bound to the receptor. We can therefore define as our "macrostate" the particular configuration of the receptor as schematically shown in Fig. 1.2(B).

If we want to know the likelihood of finding a particular system in any specific configuration, Boltzmann's law (Eq. 1.7) is then telling us a protocol we must follow:

1. Enumerate all possible microstates in which the system can be found.

2. Compute the energy of each of these microstates.

3. Define the "macrostate" we care about by grouping all microstates that belong to the same energy.

4. Compute the Boltzmann factor. This factor, sometimes called the Boltzmann weight, is defined as the exponential of the negative energy divided by the thermal energy, as indicated in Eq. 1.7.

To see this protocol in action, let us apply it to the calculation of $p_{\text{bound}}$, the probability of finding an RNAP bound to the promoter. We will go through each of the protocol steps and build up the "unrealistic simplifications" that will allow us to make this calculation.
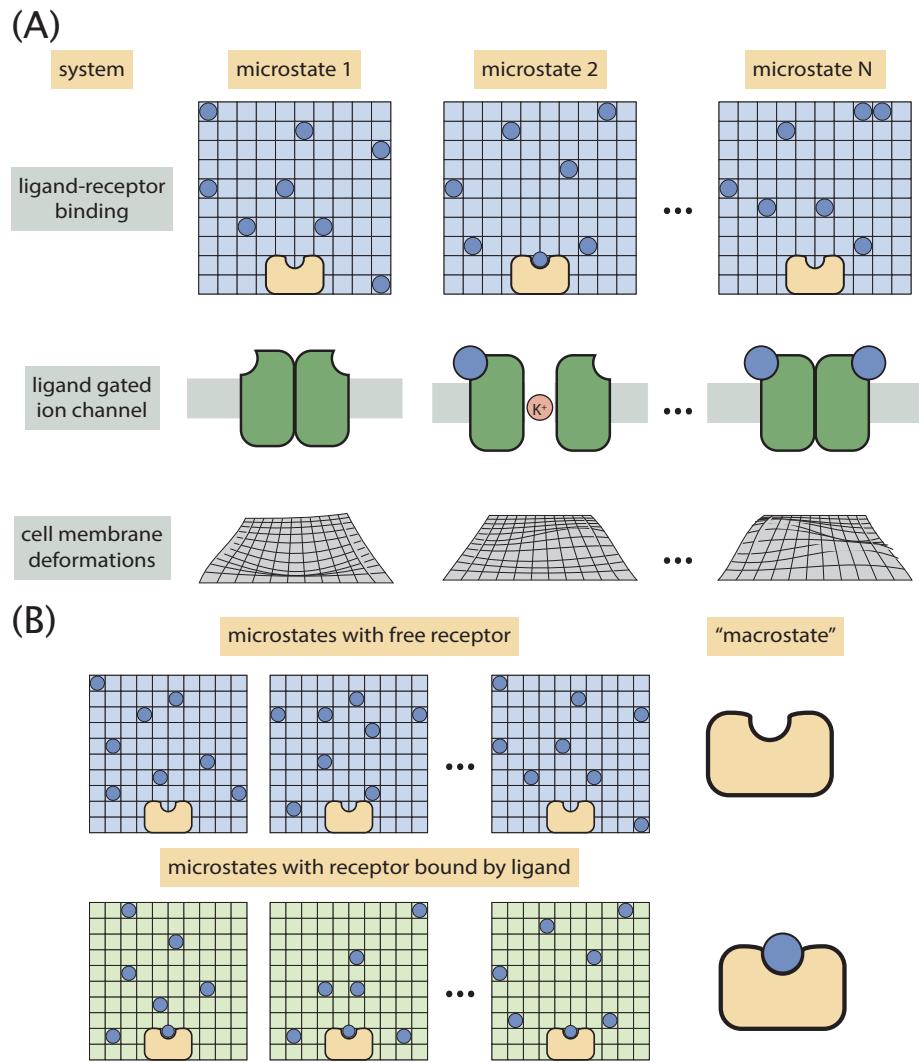
Figure 1.2: **Boltzmann's law and the definition of a micro- and macrostate.** (A) Top panel: ligand-receptor binding microstates. Middle panel: ligand-gated ion channel microstates. Bottom panel: membrane patch deformations. (B) Schematic of the definition of a "macrostate." In the ligand-receptor binding problem, we ignore all ligand molecules' spatial configuration and focus on the receptor's binding state.

**1. Enumerate possible microstates.** We begin by making a drastic coarse-graining of the bacterial genome. For us, a genome is simply made out of boxes where the RNAP can bind. We imagine that there is a single site where RNAP can bind specifically—the promoter of interest. There are also $N_{NS} \approx 5 \times 10^6$ non-specific binding sites, one per basepair (bp) in the genome. This means that because of the sequence-dependent interactions between the RNAP molecule, and the DNA, the energy associated with specific binding to the gene promoter is more favorable than the rest of the genome. We ignore the fact that the RNAP footprint where it binds to the genome is roughly 30 bp. This assumption is valid if the number of available RNAP molecules is much smaller than the number of non-specific binding sites since it is improbable that two RNAPs would fall next to each other by pure chance. A useful analogy for this point is to think about sitting $\sim$ few $\times$ 10 people on a large stadium with $\sim 10^4$ seats. If the seats are chosen randomly, we do not need to worry about doing the sampling "without replacement" because the chances of two people ending up with the same seat number are negligible. We also ignore the possibility of RNAP not being bound to the genome. This assumption is supported by experimental evidence on a particular type of *E. coli* mutant that sheds lipid vesicles without segregating DNA into such vesicles. Mass spectrometry analysis on these "min-cells" has shown that there are no RNAP molecules to be found, implying that RNAPs are bound to DNA most if not all of the time [11]. The exercise then consists of randomly choosing one box for each of the $P$ polymerases available to bind. Fig. 1.3 shows in the first column two possible configurations of our coarse-grained genome.

**2. Compute the energy for each microstate.** Let us analyze the case where all $P$ RNAP molecules are bound non-specifically to the genome. For simplicity, we assume that RNAP binds to all $N_{NS}$ non-specific binding sites with the same affinity. We assign this energy to be $\varepsilon_P^{(NS)}$. This assumption could be relaxed and we could assign instead a distribution of non-specific binding energies, as explored in [16]. But for now, we do not have to worry about this complication. For the statistical mechanics' protocol, the assignment of binding energies does not come

from some quantum first-principled calculation or anything similar. We label the interaction of the RNAP and the rest of the genome with a single value, $\varepsilon_P^{(NS)}$, that coarse-grains all of the hydrogen bonds and other effects that go into this physical process and gives an average energy. The calculation continues with this "labeled energy," and, as we will see at the end, a very clean functional form emerges. Since we have $P$ such polymerases bound non-specifically, the energy of any state with a similar configuration is then $P\varepsilon_P^{(NS)}$ as shown in Fig. 1.3 second column, top row.

**3. Define the "macrostate" we care about.** In a sense, when we speak about macrostate, it does not necessarily mean something that we can macroscopically observe. What it means is that we group a bunch of states that we take to be functionally equivalent, as shown in Fig. 1.2(B). In our case, we only care about whether or not the RNAP is bound to our promoter of interest. The configuration of the rest of the background sites is irrelevant to our question. What this means in practice is that we must compute the degeneracy or multiplicity of our state. In other words, for the *specific* state shown in the first column/top row of Fig. 1.3, we know its Boltzmann weight. Eq. 1.7 tells us that the probability of this particular configuration takes the form

$$P_{\text{state}} \propto e^{-\beta P\varepsilon_P^{(NS)}}, \tag{1.8}$$

where we define $\beta \equiv (k_B T)^{-1}$. The probability of this binding configuration takes this form since the $P$ RNAP molecules are bound non-specifically. But every single arrangement in which all RNAPs are bound non-specifically has the same Boltzmann weight. The question then becomes: in how many of such microstates can the system exist? This is a combinatorics question of the form: in how many different ways can I arrange $P$ molecules into $N_{NS}$ boxes? Which of course, the answer is

$$\text{\# states with all RNAPs bound non-specifically} = \frac{N_{NS}!}{P!(N_{NS} - P)!}, \tag{1.9}$$

as shown in the third column of Fig. 1.3. This multiplicity can be simplified if we consider that $N_{NS} \gg P$. To more easily visualize how to simplify this, let us for a second assume $N_{NS} = 100$ and $P = 3$. Given the definition of factorials, this

means that

$$\frac{N_{NS}!}{(N_{NS} - P)!} = \frac{100 \cdot 99 \cdot 98 \cdots 97 \cdots 2 \cdot 1}{97 \cdots 2 \cdot 1} = 100 \cdot 99 \cdot 98. \tag{1.10}$$

Given this result, we can simply state that $100 \cdot 99 \cdot 98 \approx 100^3$, only making a three percent error ($100 \cdot 99 \cdot 98/100^3 \approx 0.97$). Imagine that $N_{NS}$ is in the order of $10^6$. Then the error would become negligible. That is why, as shown in the third column of Fig. 1.3, we can approximate

$$\frac{N_{NS}!}{P!(N_{NS} - P)!} \approx \frac{N_{NS}^P}{P!}, \text{ for } N_{NS} \gg P. \tag{1.11}$$

For our other "macrostate," we have the case where only one out of the $P$ RNAPs is bound specifically for the promoter. We define the energy of this single RNAP specifically binding to the promoter as $\varepsilon_P^{(S)}$. We assume that the other $P - 1$ RNAPs are bound non-specifically with the usual energy $\varepsilon_P^{(NS)}$. The way to realize this state is then given by

$$\text{\# states with one RNAP bound specifically} = \frac{N_{NS}!}{(P-1)!(N_{NS} - (P-1))!} \approx \frac{N_{NS}^{P-1}}{(P-1)!}. \tag{1.12}$$

What these Boltzmann weights mean is that for us, *any* state on which a single RNAP is bound to the promoter while the rest are bound non-specifically is equivalent. Therefore the probability of finding the promoter occupied by an RNAP would be of the form

$$p_{\text{bound}} \propto e^{-\beta\epsilon_1} + e^{-\beta\epsilon_2} + e^{-\beta\epsilon_3} + \cdots \tag{1.13}$$

where $\epsilon_i$ is the energy of the $i^{\text{th}}$ state that has a single RNAP bound to the promoter. But we established that all of the $\epsilon_i$ energies are the same. So instead of writing this long sum, we multiply the Boltzmann weight of a single state by the number of states with equivalent energy, i.e., we multiply it by the state's multiplicity or degeneracy. The same logic applies for the states where none of the RNAPs are specifically bound to the promoter.

**4. Compute the Boltzmann factor.** The last step in the protocol is to follow the recipe indicated by Eq. 1.7. We exponentiate the energy, with the caveat we mentioned on the last point that this time we multiply by the multiplicity that we just computed. This is because we are lumping together all microstates into a single functional macrostate. So the Boltzmann weight for the unbound $\rho_{\text{unbound}}$ macrostate is given by

$$\rho_{\text{unbound}} = \frac{N_{NS}^{P}}{P!} e^{-\beta P \varepsilon_P^{(NS)}}. \tag{1.14}$$

For the bound state, we have

$$\rho_{\text{bound}} = \frac{N_{NS}^{P-1}}{(P-1)!} e^{-\beta \left( \varepsilon_P^{(S)} + (P-1)\varepsilon_P^{(NS)} \right)}. \tag{1.15}$$

For reasons that will become clear later in this chapter once we work with the entropy and derive the Boltzmann distribution, we know that to compute the probability of a specific microstate (or a macrostate), we simply take the Boltzmann weight of the microstate and divide by the *sum* of all of the other Boltzmann weights of the states available to the system. This sum of Boltzmann weights plays a very special role in statistical mechanics, and it is known as the *partition function* of the system. Therefore, to calculate $p_{\text{bound}}$, we compute

$$p_{\text{bound}} = \frac{\rho_{\text{bound}}}{\rho_{\text{unbound}} + \rho_{\text{bound}}}. \tag{1.16}$$

Substituting the Boltzmann weights we derived, we find

$$p_{\text{bound}} = \frac{\frac{N_{NS}^{P-1}}{(P-1)!} e^{-\beta \left( \varepsilon_P^{(S)} + (P-1)\varepsilon_P^{(NS)} \right)}}{\frac{N_{NS}^{P-1}}{(P-1)!} e^{-\beta \left( \varepsilon_P^{(S)} + (P-1)\varepsilon_P^{(NS)} \right)} + \frac{N_{NS}^{P}}{P!} e^{-\beta P \varepsilon_P^{(NS)}}}, \tag{1.17}$$

an algebraic nightmare. We can simplify this expression enormously by multiplying the numerator and denominator by $\rho_{\text{unbound}}^{-1}$. Upon simplification, we find the neat expression

$$p_{\text{bound}} = \frac{\frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_P}}{1 + \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_P}}, \tag{1.18}$$
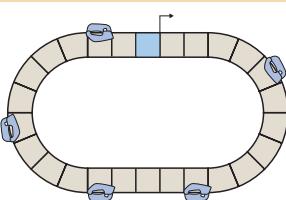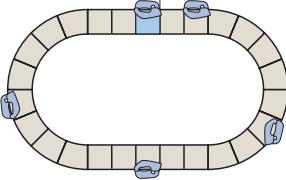
| state | energy | multiplicity | Boltzmann weight | Normalized weight |
|---|---|---|---|---|
| | $P\varepsilon_P^{(NS)}$ | $\dfrac{N_{NS}!}{P!(N_{NS}-P)!} \approx \dfrac{N_{NS}^P}{P!}$ | $\dfrac{N_{NS}^P}{P!}e^{-\beta P \varepsilon_P^{(NS)}}$ | $1$ |
| | $\varepsilon_P^{(S)} + (P-1)\varepsilon_P^{(NS)}$ | $\dfrac{N_{NS}!}{(P-1)!(N_{NS}-(P-1))!} \approx \dfrac{N_{NS}^{P-1}}{(P-1)!}$ | $\dfrac{N_{NS}^{P-1}}{(P-1)!}e^{-\beta\left[\varepsilon_P^{(S)}(P-1)\varepsilon_P^{(NS)}\right]}$ | $\dfrac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_P}$ |

Figure 1.3: **Statistical mechanics protocol for RNAP binding.** On a discretized genome, we follow the statistical mechanics' protocol to compute the Boltzmann weight of each of the relevant microstates. The $P$ available RNAPs are assumed to have two binding configurations: one specific binding to the promoter of interest (with energy $\varepsilon_P^{(S)}$) and another non-specific to any of the $N_{NS}$ non-specific binding sites (with energy $\varepsilon_P^{(NS)}$).

where $\Delta\varepsilon_P \equiv \varepsilon_P^{(S)} - \varepsilon_P^{(NS)}$. This simple expression, known as the Langmuir isothermal binding curve, tells us that the more RNAPs available (larger $P$), or the stronger the promoter is (more negative $\Delta\varepsilon_P$), the more likely it is to find the promoter bound by an RNAP, and according to Eq. 1.6, the higher the mRNA production. In the next section, we connect this model to experimental measurements.

**Figure 1 Theory in Gene Regulation**

We began this section with a simple model for the dynamics of mRNA production and degradation. We then expanded our model to deconvolve the production term into the rate at which mRNA is produced by RNAP, and the probability of finding such RNAP bound to the promoter. To calculate this probability, we used the statistical mechanics' protocol, which culminated in Eq. 1.18. So far, we are missing two important steps in our logical construction that will lead us to specific quantitative predictions that we can test experimentally:

1. The inclusion of a regulatory scheme via a transcriptional repressor.

2. The connection of the model with experimentally accessible quantities.

As hinted at earlier, for a transcriptional repressor, we imagine that the repressor's effect on the regulation of the gene acts only through changes in $p_{\text{bound}}$. To include the regulation, we add a series of microstates. Rather than having only $P$ RNAP molecules to bind the genome, we also have $R$ repressors that can bind specifically and non-specifically. Through the same statistical mechanics' protocol as for the previous case, we can arrive at the Boltzmann weights shown for the three "macrostates" in Fig. 1.4(A). For the regulated case, we have that the probability of the promoter being bound by an RNAP takes the form

$$p_{\text{bound}}(R > 0) = \frac{\frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_P}}{1 + \frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_P} + \frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_R}}, \tag{1.19}$$

where $\Delta\varepsilon_R$ is the binding energy difference between the repressor binding to a specific binding site and a non-specific one. Although exciting and insightful, the quantities we have derived so far do not have an immediate **quantitative** prediction we can connect with experimental measurements. For example, for the regulated case, the steady-state mRNA count takes the form

$$m_{ss}(R > 0) = \frac{r_m}{\gamma_m}p_{\text{bound}}(R > 0) = \frac{r_m}{\gamma_m}\frac{\frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_P}}{1 + \frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_P} + \frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_R}}. \tag{1.20}$$

Determining $r_m$ or $\gamma_m$ directly from experiments, although possible, represents an enormous technical challenge. A convenient metric we can use instead is what we call the fold-change in gene expression. Fig. 1.4(B) shows a schematic representation of what we mean by the fold-change. This ratiometric quantity normalizes the expression level of a gene with regulation given by a transcriptional repressor by the expression level of the same gene in the absence of the regulation—via a knock-out of the repressor gene, for example. Mathematically this is defined as

$$\text{fold-change} \equiv \frac{m_{ss}(R > 0)}{m_{ss}(R = 0)}. \tag{1.21}$$

This expression is convenient because upon taking the ratio of these steady-state mRNA counts, the ratio $r_m/\gamma_m$ drops out of the equation. All we are left with is then the ratio of the $p_{\text{bound}}$s

$$\text{fold-change} = \frac{p_{\text{bound}}(R > 0)}{p_{\text{bound}}(R = 0)}. \tag{1.22}$$

Substituting Eqs. 1.18 and 1.19 results in

$$\text{fold-change} = \frac{1 + \frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_P}}{1 + \frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_P} + \frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_R}}. \tag{1.23}$$

We appeal to some experimental understanding of the bacterial proteome composition [17–19]. RNAP copy number in *E. coli* is of the order $P \sim 10^3 - 10^4$ [19]. The binding affinity of these promoters is of the order $\Delta\varepsilon_P \sim -2 \pm 1\ k_B T$ [11]. Along with the value of $N_{NS} \sim 10^6$, this results in

$$\frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_P} \approx \frac{10^3}{10^6}e^{2.3} \approx \frac{10^3 \cdot 10}{10^6} \approx 10^{-2} \ll 1, \tag{1.24}$$

the so-called weak-promoter approximation. For the repressor, we have that most repressors in *E. coli* are in the order of $R \sim 10$ [18]. Their binding affinities take values between $\Delta\varepsilon_R \sim -15 \pm 5\ k_B T$ [11]. These numerical values then give

$$\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_R} \approx \frac{10}{10^6}e^{15} \approx \frac{10 \cdot 10^6}{10^6} \approx 10. \tag{1.25}$$

If we implement these approximations, we can justify simplifying the fold-change equation to take the form

$$\text{fold-change} \approx \left(1 + \frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_R}\right)^{-1}. \tag{1.26}$$

As shown in Fig. 1.4(C), this expression points directly at two experimental knobs that we can tune using molecular biology. We can modify the number of repressors by changing the ribosomal binding site sequence (RBS) of the repressor gene [20]. What that means is that with a sequence-dependent manner, the ribosome translates mRNAs according to a specific region of the gene known as the RBS [21]. Furthermore, we can change the repressor's affinity for its binding site by mutating the binding site itself [20]. Fig. 1.4(D) shows predictions of Eq. 1.26 for different binding energies.
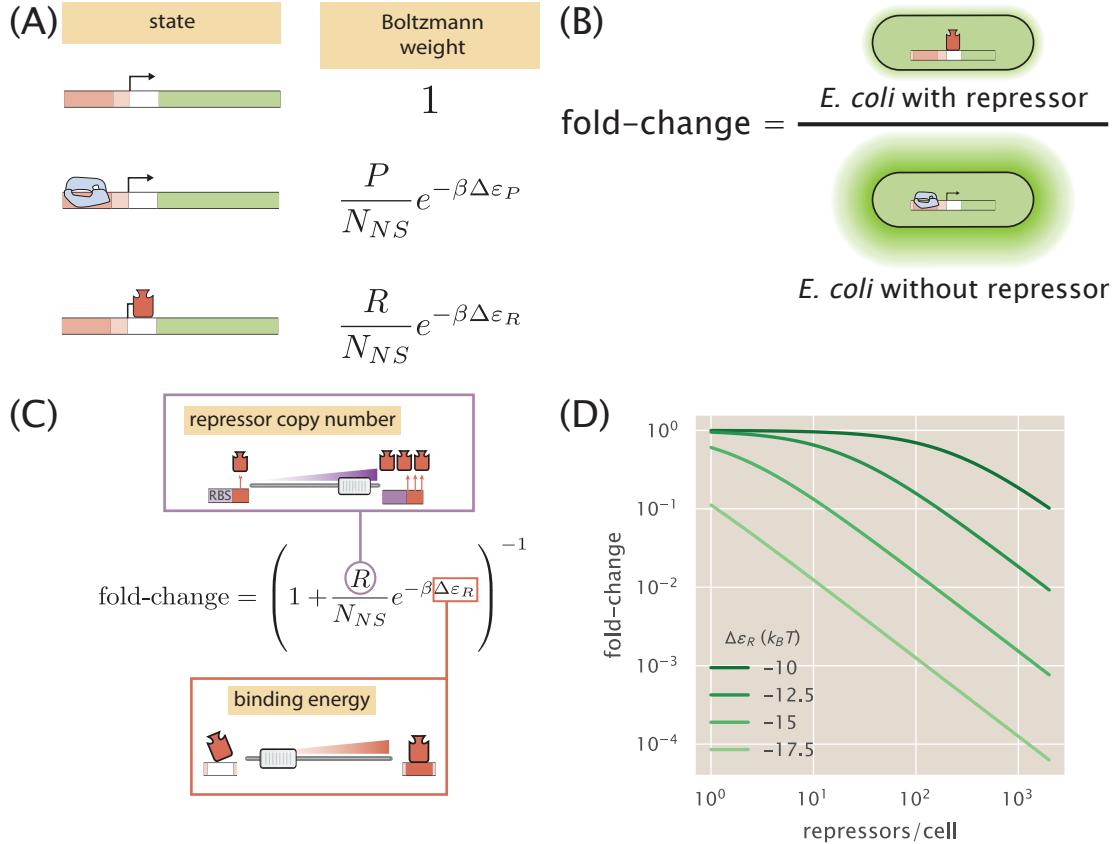
Figure 1.4: **Figure 1 theory in transcriptional regulation.** (A) States and (normalized) weights for the simple repression motif. The promoter can be found in three states: 1) empty, 2) bound by an RNAP, 3) bound by a repressor. The same statistical mechanics' protocol as in Fig. 1.3 can be used to derive the weights. (B) Schematic of the experimental determination of the fold-change in gene expression. The expression level of a regulated strain is normalized by the expression level of a strain with a repressor's knock-out. (C) Experimentally accessible knobs predicted from the theoretical model. The number of transcription factors can be tuned by changing the amount of protein produced per mRNA. The binding energy of the repressor can be tuned by mutating the basepairs in the binding site. (D) Fold-change as a function of the repressor copy number for different binding energies. The Python code (ch1_fig04D.py) used to generate part (C) of this figure can be found on the thesis GitHub repository.

The model and the predictions presented here were worked out by Garcia and Phillips in a classic publication in 2011 [20]. In the next chapter, we build upon this theoretical scaffold to expand the predictive power of the model by including the allosteric nature of the transcription factor that allows the cells to change their genetic program upon the presence of an external molecule as a response to the environment.

**All Cells Are Equal, But Some are More Equal than Others**

One of the great discoveries that came from the single-cell biology revolution where we began to measure individual cellular behavior rather than bulk observations, was the discovery of the intrinsic cell-to-cell variability in many aspects of biology, gene expression being the canonical example [22]. This means that two cells with the same genome exposed to the same conditions will not express the same number of mRNAs and proteins of any specific gene. From a statistical physics perspective, this is not entirely "surprising" since we know that a system can be found in many different microstates as described in Fig. 1.2(A). What is different here is that a cell does not have an Avogadro number of mRNA (or, for that matter of anything) in it, making these fluctuations more relevant. If we think of fluctuations scaling as $\sqrt{N}$, that means that for an $N$ of $\approx$ ten molecules or so, these variations can be significant in terms of the downstream cellular behavior. Cells have to cope with these physical limitations on precision, many times generating systems to actively buffer as much of the "noise" as possible [23], other times using this intrinsic variability to their advantage [24].

The central assumption behind the thermodynamic models of gene regulation that we studied in the last section is that the gene expression is proportional to the probability of finding an RNAP bound to the promoter [11,25]. A consequence of this construction is that the probability space—the set of all possible events captured by the distribution—only looks at the state of the promoter itself, not at the state of the mRNA copy number. That is why thermodynamic models of this kind do not speak to the intrinsic cell-to-cell variability. For this, we need to use the so-called chemical master equation framework [26]. There are two ways of thinking about the chemical master equation:

1. The "particle" point of view.

2. The occupation number point of view.
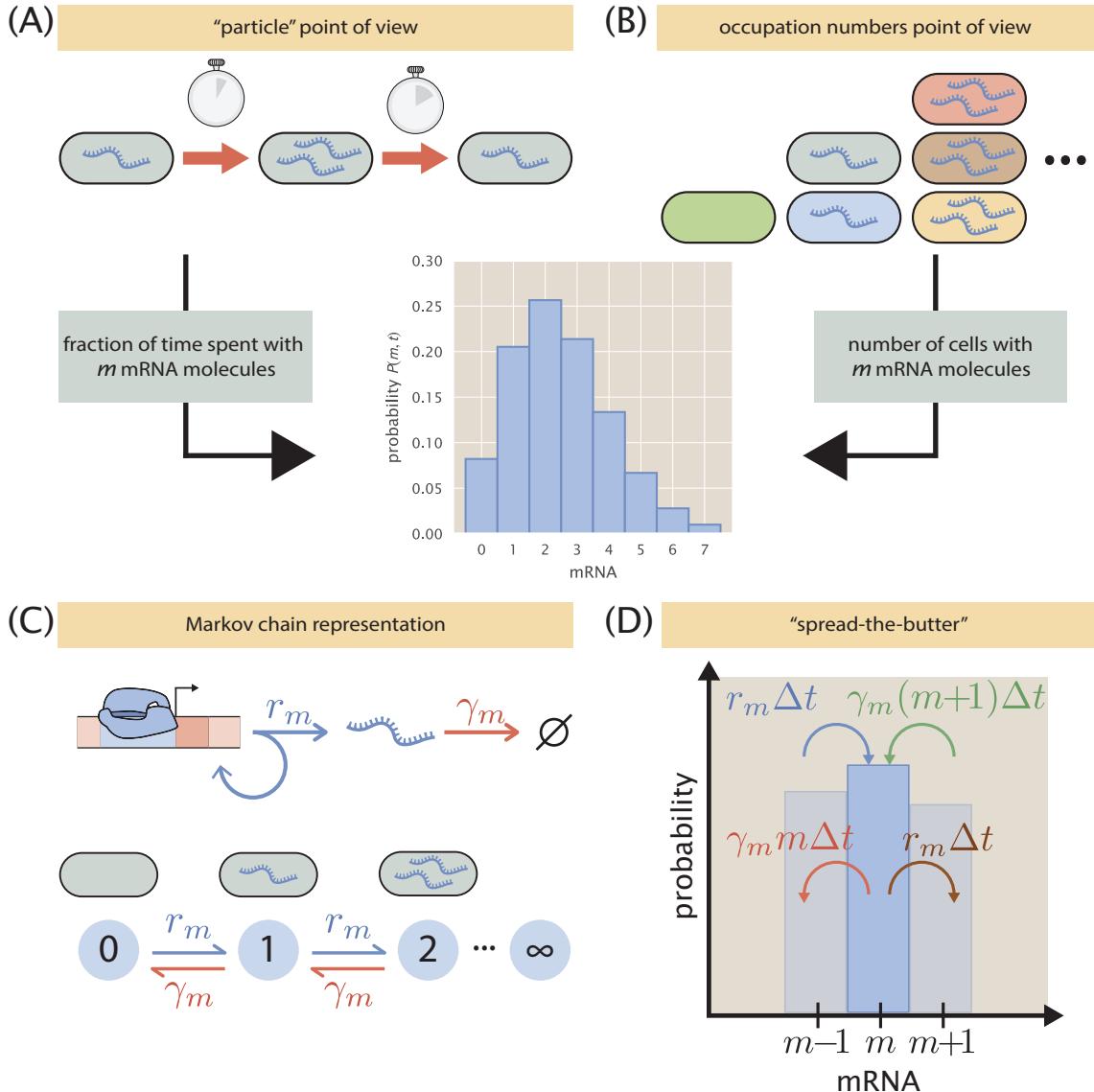
Figure 1.5: **Chemical master equation in gene regulation.** (A-B) Different points of view to understand the chemical master equation. (A) From the "particle" point of view, we imagine following the time trajectory of *a single cell*. The probability $P(m, t)$ of finding a cell with $m$ mRNAs at time $t$ is then proportional to the time this cell spent with this number of molecules. (B) On the occupation number point of view, we imagine observing a large number of isogenic cells (different colors represent the individuality of each cell). The probability $P(m, t)$ is then interpreted as the fraction of the cells representing such copy number exactly at time $t$. (C) Chemical master equations mathematize the idea of Markov processes. For the case of the unregulated promoter, the Markov process consists of a connection of an infinite number of discrete states that cells can transition between by producing or degrading mRNAs. (D) Spread-the-butter idea. Since probability is conserved, the central bar's height changes slightly by having in- and outflow of probability mass from the contiguous bins. The Python code (`ch1_fig05A.py`) used to generate the plot in part (A) of this figure can be found on the thesis GitHub repository.

Depending on the context, we might want to use either of these approaches to write down the master equation for our problem of interest. Let us look into these two different ways of interpreting a master equation using our example of a cell producing mRNA. For the particle point of view, schematized in Fig. 1.5(A), we imagine following the mRNA copy number $m$ of a single cell. The number of mRNAs in the cell change stochastically from time point to time point. On the one hand, there can be a transcriptional event that increases the number of mRNAs, and on the other hand, an mRNA can be degraded, decreasing the number of mRNAs. Suppose we imagine tracking this cell for a very long time. In that case, we can quantify the fraction of the time that the cell spent with zero mRNAs, one, two, and so on and from that, build the probability distribution $P(m, t)$ of having $m$ mRNA at time $t$ (there is a subtle point here of the process being memoryless, but we will ignore this detail). The occupation number point of view, schematized in Fig. 1.5(B), takes a different perspective. For this, we imagine tracking not one but many cells simultaneously. Each cell can either produce or degrade an mRNA on a short time window, changing its total individual count. The probability $P(m, t)$ is then built from counting how many cells out of the total have $m$ mRNAs.

Regardless of how we think about the chemical master equation, both of these perspectives describe a Markov process. These are stochastic processes in which a system transitions between different states, but the transitions between such states are only governed by the transition rates between the states and the current state of the system. In other words, a Markov process keeps no track of the states it previously visited; the only factor that determines where is the system going to head is its current state, and the transition rates out of such state—that is why these are considered memoryless processes. Fig. 1.5(C) shows a schematic of what a Markov process looks like. The schematic of the unregulated promoter indicates that there are two possible reactions: an mRNA production with rate $r_m$ and degradation with rate $\gamma_m$. The Markov process for this simple model can then be represented

as a series of nodes (representing the mRNA counts) connected with bi-directional arrows (representing the transition rates between states) indicating that the transitions can only take place between contiguous states.

In practice, the way we write down a chemical master equation is by a process christened by Professor Jane Kondev as "spread-the-butter." The idea of spread the butter is that some probability mass (the analogous of the butter) is to be spread over the range of possible values (the equivalent of the toast) where probability mass migrates in and out of a particular bin keeping the total amount of probability to add up to one. The best way to explain this concept is by following the schematic in Fig. 1.5(D) and going through the math. Let us imagine we are keeping track of a particular mRNA value $m$—the chemical master equations are in reality, a system of many coupled equations, one for each mRNA count. We want to write down an equation that describes what is the probability of finding a cell with this particular count a small time window into the future $P(m, t + \Delta t)$, where $t$ represents the time "right now," and $\Delta t$ is a tiny time increment. The master equation is nothing more than a checks and balances notebook to keep track of all the flow of probability mass in and out of the bin we are interested in, as shown in Fig. 1.5(D). Informally we would write the equation as

$$P(m, t + \Delta t) = P(m, t) + \sum \left( \begin{array}{c} \text{transitions from} \\ m' \text{ to } m \end{array} \right) - \sum \left( \begin{array}{c} \text{transitions from} \\ m \text{ to } m' \end{array} \right),$$

(1.27)

where we are describing the three main components that go into the equation for $P(m, t + \Delta t)$:

1. The probability of having $m$ mRNA right now,

2. the inflow of probability from other copy numbers $m'$ via production and degradation,

3. the outflow of probability from $m$ to other copy numbers via production and degradation.

Taking our time window $\Delta t$ to be sufficiently small, we can focus only on the two contiguous mRNA counts $m-1$ and $m+1$, and ignore the rest since jumps from further counts become increasingly improbable as the time step gets smaller. Fig. 1.5(D) shows the four in- and outflows that can happen. Let us rewrite Eq. 1.27 following this schematic. If a cell has $m-1$ mRNA and during the time window $\Delta t$ produces one molecule, then it passes from state $m-1$ to state $m$. This transition contributes to the inflow of probability mass by a factor $(r_m \Delta t) P(m-1, t)$, where we can think of $r_m \Delta t$ as the probability of the transcription event taking place during the time window, and this multiplies the probability of having $m-1$ mRNA to begin with. A similar argument can be made for all transitions in and out of $m$ depicted in Fig. 1.5(D), with the only difference that as in Eq. 1.1, the degradation of an mRNA molecule is proportional to the total number of molecules. The resulting equation for $P(m, t + \Delta t)$ then takes the form

$$
\begin{aligned}
P(m, t+\Delta t) = & P(m,t) + \overbrace{(r_m\Delta t)P(m-1,t)}^{m-1\to m} + \overbrace{(\gamma_m\Delta t)(m+1)P(m+1,t)}^{m+1\to m} \\
& - \underbrace{(r_m\Delta t)P(m,t)}_{m\to m+1} - \underbrace{(\gamma_m\Delta t)mP(m,t)}_{m\to m-1}
\end{aligned} \tag{1.28}
$$

We send the first term on the right-hand side to the left, divide both sides by $\Delta t$, and take the limit when $\Delta t \to 0$. This gives us the master equation we were searching for

$$
\frac{dP(m,t)}{dt} = r_m P(m-1,t) + \gamma_m(m+1)P(m+1,t) - r_m P(m,t) + \gamma_m m P(m,t). \tag{1.29}
$$

Eq. 1.29 is not isolated. It represents an infinite-dimensional system of coupled ordinary differential equations (one for each mRNA copy number $m$). It can therefore be tricky to work directly with these types of equations. Instead, let us take Eq. 1.28 for a ride. With modern computational power, we can explicitly use this equation as a recipe on how to update an mRNA distribution numerically. Fig. 1.6 shows such numerical integration for a system with initially no mRNAs present. This could be achieved experimentally by having an inducible system, adding the inducer, and tracking the time evolution of the single-molecule mRNA counts in-
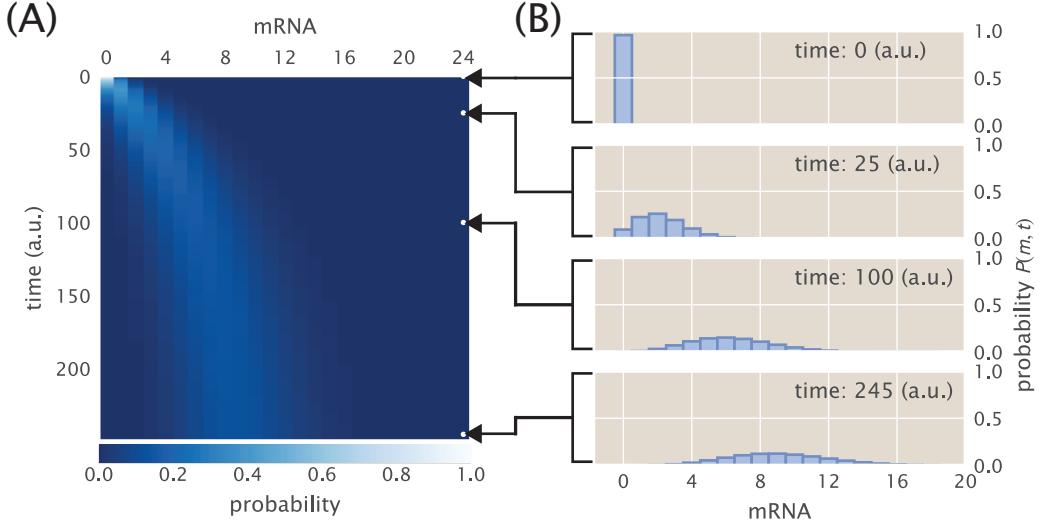
Figure 1.6: **Time evolution of mRNA distribution.** (A) Heat map of the time evolution of the mRNA distribution (Eq. 1.29) with $P(m = 0, t = 0) = 1$, i.e., a delta function at zero mRNAs at time zero. (B) Snapshots of the same time-evolving distribution at different time points. The Python code (`ch1_fig06.py`) used to generate the plot in part (A) of this figure can be found on the thesis GitHub repository.

side cells. Fig. 1.6(A) presents a heatmap of such time evolution with time running on the vertical axis, while Fig. 1.6(B) presents specific snapshots. We can see that the distribution begins as a single peak (a delta function in the physics jargon) centered at zero mRNAs. The distribution then relaxes to a broader shape and remains the same after that. This suggests that the distribution converges to a steady-state. Let us compute this steady state distribution.

In this system, where we have a series of state transitions as represented in Fig. 1.5(C), steady-state is reached when the flux of probability from two contiguous states is zero. In other words, when the probability distribution does not change over time anymore, the flow of probability from state $m = 0$ to state $m = 1$ should be the same as the reverse. The same condition applies to all other pairs of states. Mathematically this is expressed as

$$\overbrace{r_m P(0)}^{0 \to 1} = \overbrace{\gamma_m \cdot 1 \cdot P(1)}^{1 \to 0}, \tag{1.30}$$

where we removed the time dependency from $P(m, t)$ since we are at steady-state. Solving for $P(1)$ results in

$$P(1) = \left( \frac{r_m}{\gamma_m} \right) P(0). \tag{1.31}$$

The same condition applies between state $m = 1$ and $m = 2$, resulting in

$$\overbrace{r_m P(1)}^{1 \to 2} = \overbrace{\gamma_m \cdot 2 \cdot P(2)}^{2 \to 1}. \tag{1.32}$$

Again, we can solve for $P(2)$ and obtain

$$P(2) = \frac{1}{2} \left( \frac{r_m}{\gamma_m} \right) P(1). \tag{1.33}$$

Substituting the solution for $P(1)$ gives

$$P(2) = \frac{1}{2} \left( \frac{r_m}{\gamma_m} \right)^2 P(0). \tag{1.34}$$

Let us do one more example to see the general pattern. Between $m = 2$ and $m = 3$, we have

$$\overbrace{r_m P(2)}^{2 \to 3} = \overbrace{\gamma_m \cdot 3 \cdot P(3)}^{3 \to 2}. \tag{1.35}$$

Following the same procedure and substitutions results in

$$P(3) = \frac{1}{2 \cdot 3} \left( \frac{r_m}{\gamma_m} \right)^3 P(0). \tag{1.36}$$

Deducing the pattern from these examples, we can see that for any $m$, we have

$$P(m) = P(0) \frac{\left( \frac{r_m}{\gamma_m} \right)^m}{m!}. \tag{1.37}$$

All we have left is the unknown value $P(0)$. To get at it, we use the fact that the distribution must be normalized, giving

$$\sum_{m=0}^{\infty} P(m) = 1 \Rightarrow P(0) \sum_{m=0}^{\infty} \frac{\left( \frac{r_m}{\gamma_m} \right)^m}{m!} = 1. \tag{1.38}$$

We recognize the sum as the Taylor series for $e^x$. This means that our constant $P(0)$ is given by

$$P(0) = \frac{1}{\sum_{m=0}^{\infty} \frac{\left( \frac{r_m}{\gamma_m} \right)^m}{m!}} = e^{-r_m/\gamma_m}. \tag{1.39}$$

Substituting this result, we find that the mRNA steady-state distribution is a Poission distribution with mean $r_m/\gamma_m$, i.e.,

$$P(m) = \frac{e^{-r_m/\gamma_m}\left(\frac{r_m}{\gamma_m}\right)^m}{m!}. \tag{1.40}$$

## 1.3 Entropy, Information, and the Math Behind the Bit

Central to the endeavor undertaken in this thesis is the idea that cells can process information from the environment to up or down-regulate their genes to generate an appropriate response to these external signals. Information as a concept is a very plastic term that we commonly use to explain having helpful knowledge to use to our advantage. Phrases such as *"That person carries so much information in her brain. She truly knows everything!"* point at this somewhat imprecise concept of what we mean by information.

In 1948, while working at Bell Labs, Claude Shannon shocked the world with his seminal work that would go to define the field of information theory [27]. In his paper, Shannon gave us a precise mathematical definition of information. To understand Shannon's logic better, we need to put it in the context that he was thinking about: communication systems such as the telephone or the telegraph. Although seemingly unrelated to our problem of cells sensing the environment, these systems are incredibly powerful in their conceptual and explanatory reach. For Shannon, the main problem of communication consisted of reproducing a message emitted at one point in space and time with fidelity at a different point. Usually, these messages carry with them *meaning* (otherwise, why would we even want to send such messages) by which we typically mean that the message "refers to or is correlated according to some system with certain physical or conceptual entities" [27]. But for the task of engineering a reliable communication system, this meaning is irrelevant—in the same way that whatever the cell decides to do with the meaning of the signals obtained from the environment can be thought as irrelevant for the biophysics of how the signal is sensed.

As shown schematically in Fig. 1.7(A) from Shannon's original work, a communication system essentially consists of five components:

1. An **information source** which produces a message (or sequence of messages) to be communicated to the receiving terminal.

2. A **transmitter** which takes the message, converts it into a suitable signal compatible with the communication channel.

3. The **channel** that is the medium used to transmit the signal from the transmitter to the receiver.

4. The **receiver** in charge of inverting the operation done by the transmitter, reconstructing the original message.

5. The **destination** for whom the message is intended.

Fig. 1.7(B) shows an analogous schematic to Fig. 1.7(A) with the relevant components involved in the gene expression context that we focus on in this thesis. In our bacterial gene regulation model, the information source role is played by a small molecule's environmental concentration. It is this signal that the cells are trying to measure and respond to by up-regulating the expression of a gene. This signal transmitter is the allosteric transcription factor whose conformation depends on the concentration of the small molecule. The receiver of the signal is the DNA promoter that orchestrates the protein expression, which plays the receiver's role.

Having this setup in mind, the question becomes: how do we mathematically define what information is? This brings a somewhat subtle difference between two related terms that many time are incorrectly used interchangeably: *Entropy* and *Information*. Information allows the entity that possesses it to make predictions with accuracy better than random, while entropy is a quantification of how much we do not know [5]. From these definitions, we see that having information, therefore, reduces our uncertainty, i.e., reduces the entropy. This means that for Shannon, the amount of information we have from a source is related to that source's statistical

(A)



(B)
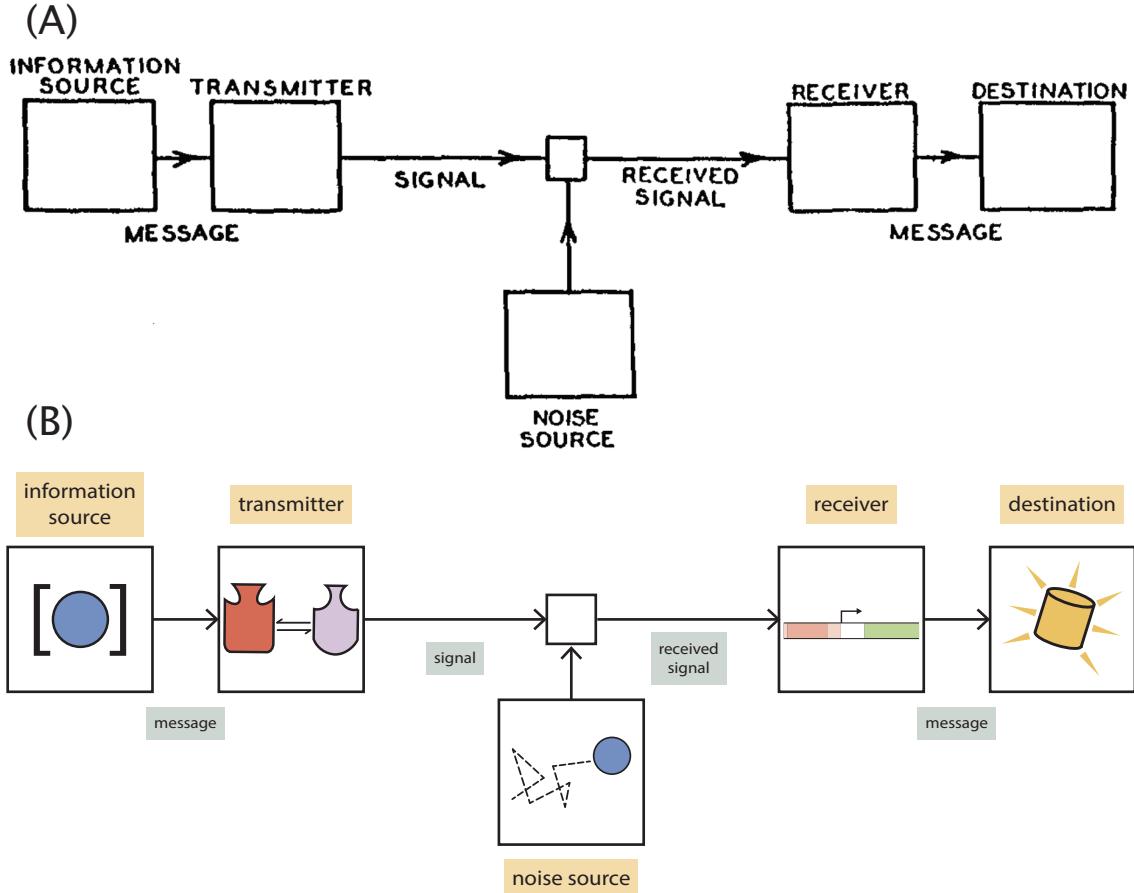


Figure 1.7: **Abstract communication system.** (A) Reproduced from Shannon's original seminal work [27]. The schematic shows an abstract communication system with all the components. (B) Adaptation of the Shannon communication system to the context of bacterial gene expression regulated by an allosteric transcription factor.

structure and how much we can predict the source's message given our knowledge of this statistical structure. Let us look at a concrete example: English text. We know that written and spoken language is not completely random. For a message to be meaningful, the choice of words has to come from a statistical structure that obeys the language's grammar rules. The choice of letters within a word also follows a certain statistical structure. Let us look at the text shown in Fig. 1.8(A). This is arguably one of the most important and most beautiful pieces of prose ever put together by a human mind as it is the last paragraph of *On the Origin of Species* by Darwin. If we ignore the paragraph's message and just quantify how often we find each of the 26 letters in the English alphabet, we obtain a distribution like the one shown in Fig. 1.8(B). This paragraph shows that the most common vowel is *e*,

exactly as in English writ-large. This distribution $P(x)$ is therefore not maximally random. In other words, if we were to put all letters in the paragraph in a hat and pick one letter at random, we could bet more money on the outcome being a letter $e$ and make money over time given this knowledge of the structure of the distribution.

A maximally random distribution would be if all letters appeared equally frequent in the paragraph, such that betting on any letter coming out of the hat would give us equal chances of guessing right. If instead of looking at the distribution of individual letters, we look at pairs of letters, the distribution $P(x, y)$ over the paragraph is shown in Fig. 1.8(C). Here we can see that, just as the letters were not completely random, the pairs of letters are also not random. For example, if we take the first letter of the pair to be $t$, we see that it is more commonly followed by the letter $h$. This implies that knowing that the first letter of the pair was $t$ reduced our uncertainty of what character could come next. We would then say that knowing the first letter gave us *information* about the possible outcomes of the second letter. In the next section, we will follow Shannon's original derivation to define both entropy and information mathematically.

**Choice, Uncertainty, and Entropy**

So far, our discussion about what entropy and information mean has been vague and not rigorous. To derive a formula to quantify these concepts, we need to get more mathematical. Let us assume that an information source (see Fig. 1.7(A)) produces elements of a message following a distribution $\mathbf{p} = \{p_1, p_2, \ldots, p_n\}$, where each $p_i$ is the probability of the $i^{\text{th}}$ element. These elements could be letters, words, sentences, basepairs, concentrations of a small molecule, etc., of which we have $n$ possibilities. What we are looking for is a metric $H(\mathbf{p})$ that quantifies how much "choice" is involved in the selection of each element of the message. In other words, how uncertain we are about the message that the information source will produce at random? We demand our desired quantity $H(\mathbf{p})$ to satisfy three reasonable conditions [27]:

**(A)** "Thus, from the war of nature, from famine and death, the most exalted object of which we are capable of conceiving, namely, the production of the higher animals, directly follows. There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved."

– Charles Darwin, 1859

**(B)** distribution *P(x)*  **(C)** joint distribution *P(x, y)*
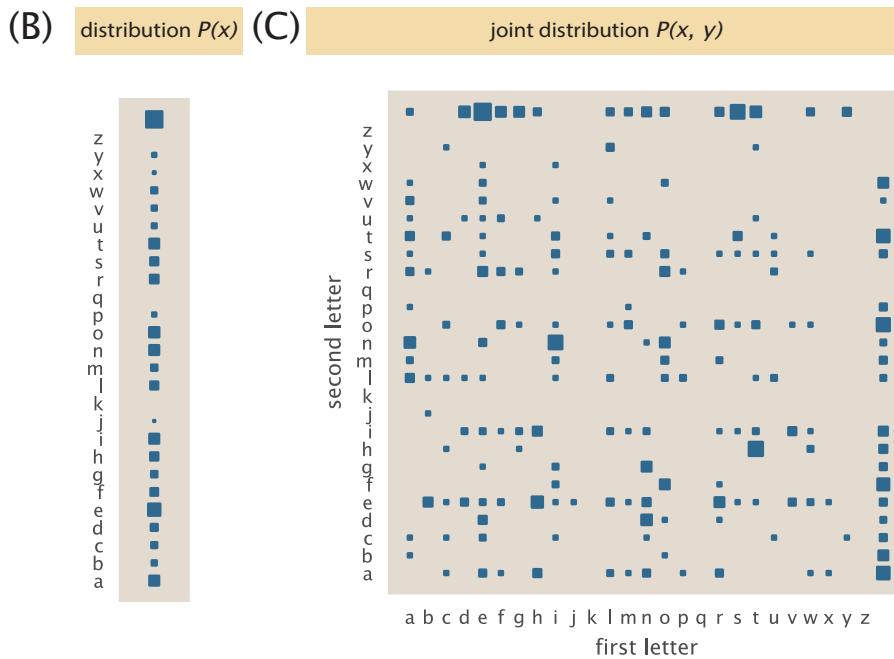
Figure 1.8: **The statistical structure of the English language.** (A) Last paragraph of *On the Origin of Species* by Charles Darwin. This serves as a rather nice not-random text example. (B) Marginal distribution $P(x)$ of all 26 letters and space. The size of the squares is proportional to how often each letter appears in the paragraph. (C) Joint distribution of pairs of characters $P(x,y)$. All pairs of characters in (A) were counted to build this histogram. The x-axis shows the first letter while the y-axis shows the second. For simplicity in (B) and (C) all punctuation was ignored. The Python code (`ch1_fig08.py`) used to generate this figure can be found on the thesis GitHub repository.

1. $H$ should be continuous in the $p_i$s. Different information sources might have slightly different distributions $\mathbf{p}$, nevertheless $H$ should still apply to all possible information sources.

2. If all of the elements of the distribution are equally likely, i.e., $p_i = 1/n$, then $H$ should be a monotonic increasing function of $n$. This means that the more options to choose from, the more uncertain we are about the possible outcome. For example, we are more uncertain about the outcome of a fair 6-sided die than of a fair coin just because of the number of possible outcomes from each of these "information sources."

3. If the act of choosing one of the possible $n$ elements of our information source can be broken down into two successive choices, the original $H$ should be the weighted sum of the individual $H$s. What this means is illustrated in Fig. 1.9(A) where we imagine having an information source with $n = 3$ choices, each with probabilities $\mathbf{p} = \{1/2, 1/3, 1/6\}$, which gives $H(1/2, 1/3, 1/6)$ for the left case. For the right case, we imagine first choosing between the upper and the lower path, and then, if the lower path is chosen, a second choice is made. This property then demands that

$$\overbrace{H(1/2, 1/3, 1/6)}^{\text{single choice}} = \overbrace{H(1/2, 1/2)}^{\text{first choice}} + \frac{1}{2}\overbrace{H(1/3, 1/6)}^{\text{second choice}}. \tag{1.41}$$

Another way to think about this property is that we want our metric of uncertainty $H$ to be *additive*.

We will now prove that the only functional form that satisfies all these three properties is given by

$$H(\mathbf{p}) = -K \sum_{i=1}^{n} p_i \log p_i, \tag{1.42}$$
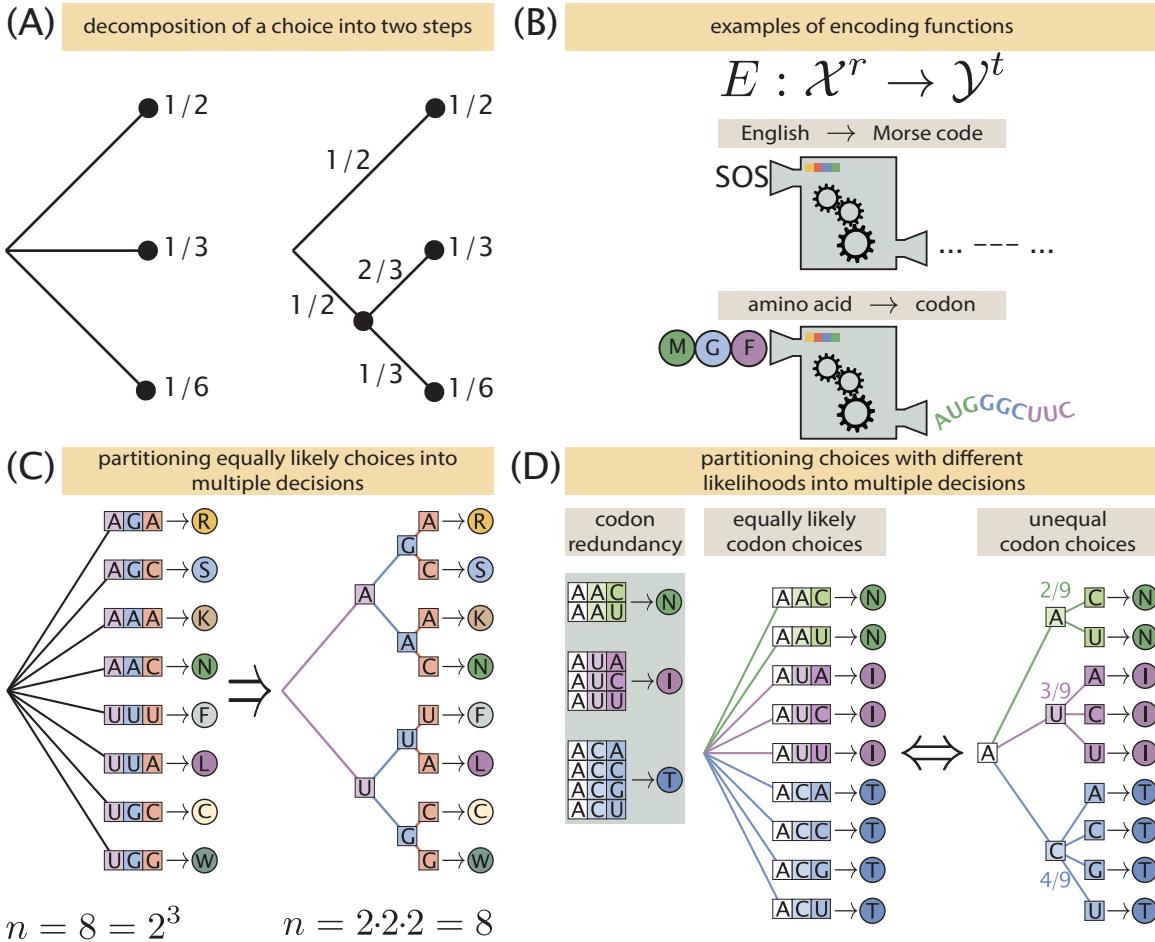
Figure 1.9: **Shannon's theorem.** (A) One of the properties of a reasonable metric for uncertainty is that we can partition choices into multiple steps, and the resulting uncertainty should remain the same. (B) Example of coding functions $E$. The English alphabet can be converted into Morse code. Amino acids can be encoded in codons. (C) Partitioning of $2^3$ equally likely choices into three decision steps, each with two choices. Eight different amino acids can be selected using two schemes: 1) each of the eight codons is chosen at random with equally likely chances, or 2) the codon is built by choosing one basepair at the time. (D) Partitioning of unequal choices. Given the redundancy of the genetic code, for equally likely codons, the resulting amino acid has different probabilities being chosen.

where $K$ is a constant having to do with the units (choice of the logarithm base). To prove this, we will follow Shannon's original work. We imagine the problem of encoding a message. For example, imagine encoding a message from the English alphabet into Morse code, or a protein sequence into the corresponding mRNA sequence, as schematically depicted in Fig. 1.9(B). In there, we take letters in the English alphabet (*SOS* for the English alphabet, *MGF* for the protein), run it through an encoding function $E$ and obtain the message ($\ldots$- - -$\ldots$ for the Morse code, *AUGGGCUUC* for the mRNA). This process of encoding can be thought of as taking a message $m_x$ written in an alphabet $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, (where $n$ is 26 for the English alphabet, and 20 for the number of amino acids) and converting it into a message $m_y$ written in a different alphabet $\mathcal{Y} = \{y_1, y_2, \ldots, y_m\}$ (where $m = 2$ for Morse code since we only have dots and dashes, and $m = 4$ for the mRNA with 4 possible nucleotides). The encoding function $E : \mathcal{X}^r \rightarrow \mathcal{Y}^t$ takes a message of length $r$ (for our exmaples $r = 3$) and translates it into a message of size $t$ (in our examples $t = 9$) such that we then have

$$m_y = E(m_x). \tag{1.43}$$

Obviously, the larger the message $m_x$ we want to encode, the larger the corresponding message $m_y$ will be. Therefore we have that

$$L(m_y) \propto L(m_x), \tag{1.44}$$

where $L(\cdot)$ is a function that counts the number of characters in a message. An essential difference between both of the examples in Fig. 1.9(B) is that, for the English to Morse code case, the number of dots and dashes for different letters is different ($e \rightarrow .$, $x \rightarrow -..-$). Meanwhile, for the amino acid to codon case, every single codon has the same length. Let us focus for now on this second coding scheme where every character from alphabet $\mathcal{X}$ is encoded with the same number of characters from alphabet $\mathcal{Y}$. We have then $L(m_x) = r$ and $L(m_y) = t$. Let us call $k$ the proportionality constant from Eq. 1.44 such that

$$L(m_y) = kL(m_x). \tag{1.45}$$

The number of messages of size $r$ that can be encoded with the alphabet $\mathcal{X}$ is given by $n^r$—because we have $n$ possible options to chose from for each of the $r$ characters, resulting in $n \cdot n \cdot n \cdots = n^r$. Likewise, the number of messages of size $t$ encoded with alphabet $\mathcal{Y}$ is $m^t$. We then demand from our coding scheme that the number of messages we can encode is at least the number of messages we could potentially send. In other words, for our coding scheme to be able to take *any* message of size $r$, it must be true that the number of possible encoded messages is at least as large as the number of possible messages to encode. This demand is expressed as

$$n^r \le m^t. \tag{1.46}$$

If our encoding did not satisfy this, we would have to increase $t$, i.e., the number of characters we use to encode our message. For example, if codons were made out of only two basepairs, the genetic code would not be able to code for all 20 amino acids plus the stop codons. On the other extreme, we could develop a ridiculously long encoding scheme (imagine a version of the genetic code where 1000 basepair represented a single amino acid). To avoid this absurd scheme, we bound the encoded message's size to be as long as necessary to encode all potential messages, but not any longer. This bound is expressed as

$$m^{t-1} < n^r \le m^t. \tag{1.47}$$

Let us now take the logarithm on our previous inequality—this preserves the inequalities since log is a monotonically increasing function—finding

$$(t-1)\log(m) < r\log(n) \le t\log(m). \tag{1.48}$$

We are free to choose the logarithm base we find convenient; therefore, let us use base $m$ for this, obtaining

$$t - 1 < r\log_m(n) \le t. \tag{1.49}$$

Dividing Eq. 1.49 by $r$ gives

$$\frac{t-1}{r} < \log_m(n) \le \frac{t}{r}. \tag{1.50}$$

Let us stare at Eq. 1.50. In Eq. 1.47, we established $t$ as the minimum number of characters from alphabet $\mathcal{Y}$ needed to encode a message of length $r$ written with alphabet $\mathcal{X}$ characters (such as *MGF* turned into *AUGGGCUUC* as in Fig. 1.9(B)). This means that, for the case where all symbols use the same number of characters when encoded, $t/r$ is the number of characters from alphabet $\mathcal{Y}$ per character from alphabet $\mathcal{X}$, i.e., the proportionality constant $k$ from Eq. 1.45. This means that Eq. 1.50 implies

$$\log_m(n) \leq k. \tag{1.51}$$

In other words, a lower bound for the number of characters from alphabet $\mathcal{Y}$ needed to encode a character from alphabet $\mathcal{X}$ is given by $\log_m(n)$. For the amino acid to codon case, the minimum number of letters in a codon would be $\log_4(20) \approx 2.16 > 2$. This shows why we could not encode all 20 amino acids with two base-pair long codons. Furthermore, Eq. 1.50 implies that

$$\frac{t}{r} - \log_m(n) < \frac{t}{r} - \frac{(t-1)}{r}, \tag{1.52}$$

given that $(t-1)/r < \log_m(n)$. Simplifying Eq. 1.52 results in

$$\frac{t}{r} - \log_m(n) < \frac{1}{r} \Rightarrow k - \log_m(n) < \frac{1}{r}. \tag{1.53}$$

Therefore, we can make $k$, the number of encoding characters, as arbitrarily close to $\log_m(n)$ as we want by increasing the length of the message being encoded, i.e., making $r \to \infty$. This would imply a genetic code, not for individual amino acids but entire polypeptides. This scheme would not work biologically; nevertheless, this mathematical limit will help us find the functional form of our desired function $H(\mathbf{p})$.

Coming back to the function $H$, let us define

$$A(n) \equiv H\left(\frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \ldots\right), \tag{1.54}$$

as the maximum possible value of $H$ when all outcomes are equally likely. Property 2 tells us that $A(n)$ increases monotonically with the length of the message. This means that if we apply the function $A(\cdot)$ to the terms in Eq. 1.47, we conserve the inequality, i.e.,

$$A(m^{t-1}) < A(n^r) \leq A(m^t). \tag{1.55}$$

Using Property 3, we can divide the $n^r$ possible choices into $r$ independent decisions, each with $n$ options to chose from. This property is depicted in Fig. 1.9(C). On the left, it shows we can choose from eight different codons that code for $2^3 = 8$ different amino acids. On the right, we can choose base by base, building up the codon in three consecutive decisions, each with two equally likely choices, for a total of $2 \cdot 2 \cdot 2 = 8$ possible outcomes. This division of choices allows us to rewrite Eq. 1.55 as

$$(t-1)A(m) < rA(n) \leq tA(m), \tag{1.56}$$

because of our requirement of the uncertainty $H$ being an additive property. For the example in Fig. 1.9(C), at each of the three decision steps, the uncertainty is given by $A(2)$. Given that the uncertainty is additive, for each of the routes, our total uncertainty is given by

$$A(2) + A(2) + A(2) = 3A(2), \tag{1.57}$$

therefore $A(2^3) = 3A(2)$. Dividing Eq. 1.56 by $r$ results in

$$\frac{(t-1)}{r}A(m) < A(n) \leq \frac{t}{r}A(m). \tag{1.58}$$

Since $\frac{(t-1)}{r}A(m) < A(n)$, it is also true that

$$\frac{t}{r}A(m) - A(n) < A(m)\left(\frac{t}{r} - \frac{(t-1)}{r}\right). \tag{1.59}$$

Simplifying terms, we are left with

$$\frac{t}{r}A(m) - A(n) < \frac{1}{r}A(m). \tag{1.60}$$

Dividing both sides by $A(m)$, we find

$$k - \frac{A(n)}{A(m)} < \frac{1}{r}. \tag{1.61}$$

We can make the ratio $A(n)/A(m)$ as close to $k$ as we want by making $r$ larger. This equation looks shockingly similar to Eq. 1.53, but what is the connection? On the one, hand Eq. 1.53 is the result of imposing the condition that our coding scheme must be able to encode any possible message from one alphabet $\mathcal{X}$ to another alphabet $\mathcal{Y}$. This condition leads us to the conclusion that the number of characters from alphabet $\mathcal{Y}$ needed to encode the characters from alphabet $\mathcal{X}$ (the constant $k$) can be made as arbitrarily close to $\log_m(n)$ as we want by writing a code, not for individual characters (individual amino acids), but for sequences of characters (polypeptides). On the other hand, Eq. 1.61 is a direct consequence of the three logical properties we imposed on our uncertainty metric $H$. These properties led us to conclude that, whatever our uncertainty function for the equally likely choices $A(\cdot)$ is, the ratio of the uncertainties for each of our two alphabets $A(n)/A(m)$ approaches the same constant $k$ as we make the encoded message longer. Since both $\log_m(n)$ and $A(n)/A(m)$ approach $k$ as $r$ grows, we can conclude that

$$\frac{A(n)}{A(m)} \to \frac{\log_m(n)}{\log_m(m)} \quad \text{as } r \to \infty. \tag{1.62}$$

We wrote the ratio $\log_m(n)/\log_m(m)$ because our choice of the logarithm base was arbitrary. Therefore, more generally, we have

$$\frac{A(n)}{A(m)} \to \frac{\log(n)}{\log(m)} \quad \text{as } r \to \infty, \tag{1.63}$$

for any base. This convergence only takes place if and only if

$$A(n) = K \log(n), \tag{1.64}$$

where $K$ is some constant. This is quite beautiful. What we just demonstrated is that the functional form for the uncertainty metric we are after scales as the logarithm of the number of possible characters in our alphabet. We know that our uncertainty function $H(1/n, 1/n, \dots)$ is a function of $1/n$ rather than of $n$. This is easily fixed by using the properties of logarithms, writing

$$H\left(\frac{1}{n}, \frac{1}{n}, \dots\right) = -K \log\left(\frac{1}{n}\right). \tag{1.65}$$

The general form of Shannon's entropy is starting to show up. After all, for the case where all choices are equally likely, we have $p_i = 1/n$. We can therefore write

$$H\left(\frac{1}{n}, \frac{1}{n}, \ldots\right) = -K \sum_{i=1}^{n} \frac{1}{n} \log\left(\frac{1}{n}\right). \tag{1.66}$$

Let us generalize the proof for cases where choices are not equally likely. To continue with the amino acid to codon encoding example, we now consider the genetic code's redundancy. Given that there are $4^3 = 64$ possible codons, multiple codons map to the same amino acid. An example of three amino acids that share the first letter is depicted on Fig. 1.9(D). The diagram on the left shows a total of nine different codons; two of such codons code for asparagine ($N$), three for isoleucine ($I$), and four for threonine ($T$). A way to express the asymmetry between the choices is to have each codon as an independent and equally likely choice, as depicted on the middle diagram of Fig. 1.9(D). Let us define the total number of codons

$$N = \sum_{i=1}^{n} n_i, \tag{1.67}$$

where $n_i$ counts the number of codons for amino acid $i$, and $n$ is the total number of amino acid choices. Let us call $H_1$ the uncertainty of this set of equal choices. From Eq. 1.65, we know that the resulting uncertainty function $H_1$ is of the form

$$H_1 = K \log\left(\sum_{i=1}^{n} n_i\right) = K \log(N), \tag{1.68}$$

since all codons are equally likely.

Although each codon is equally likely, the resulting amino acid is not. The probability of amino acid $I$ in this case is the number of codons encoding it (two) divided by the total number of codons in the example (nine). In general, we assume that each of the $n$ choices has a probability

$$p_i = \frac{\text{\# codons for amino acid } i}{\text{total \# of codons}} = \frac{n_i}{N}. \tag{1.69}$$

By Property 3 of our function $H$, we can partition the codon's choice into two consecutive decisions (not three since the first codon is the same for all amino acids in this example). This partitioning is shown on the right diagram of Fig. 1.9(D). The uncertainty $H_2$ for this case has two contributions, one for each of the decisions

$$
H_2 = \overbrace{H(p_1, p_2, \ldots, p_n)}^{\text{first choice}} + \overbrace{K \sum_{i=1}^{n} p_i \log n_i}^{\text{second choice}}. \tag{1.70}
$$

The first decision has an unknown functional form we are trying to figure out. The second choice consists of choosing between $n_i$ equally likely bases for the codon's last position, each weighted by the probability of going to this particular branch (the one that defines the amino acid) as demanded by Property 3. But whether or not we choose each codon on a single decision or in two steps, the uncertainty of this event is the same. This means that $H_1 = H_2$ as Property 3 requires. This equality results in

$$
K \log(N) = H(p_1, p_2, \ldots, p_n) + K \sum_{i=1}^{n} p_i \log(n_i). \tag{1.71}
$$

Solving for $H(p_1, p_2, \ldots, p_n)$ results in

$$
H(p_1, p_2, \ldots, p_n) = K \left[ \log N - \sum_{i=1}^{n} p_i \log(n_i) \right]. \tag{1.72}
$$

Using Eq. 1.67 results in

$$
H(p_1, p_2, \ldots, p_n) = -K \left[ \sum_{i=1}^{n} p_i \log(n_i) - \log \left( \sum_{i=1}^{n} n_i \right) \right]. \tag{1.73}
$$

Since probabilities must be normalized, i.e., $\sum_{i=1}^{n} p_i = 1$, we can write

$$
H(p_1, p_2, \ldots, p_n) = -K \left[ \sum_{i=1}^{n} p_i \log(n_i) - \sum_{i=1}^{n} p_i \log \left( \sum_{i=1}^{n} n_i \right) \right]. \tag{1.74}
$$

Using the property of logarithms, we can rewrite this as

$$
H(p_1, p_2, \ldots, p_n) = -K \left[ \sum_{i=1}^{n} p_i \log \left( \frac{n_i}{\sum_{i=1}^{n} n_i} \right) \right]. \tag{1.75}
$$

Using Eq. 1.69, we find the expected result

$$H(p_1, p_2, \ldots, p_n) = -K \sum_{i=1}^{n} p_i \log p_i. \tag{1.76}$$

Let us dissect this result. We began this derivation by stating three logical properties that a metric for uncertainty should have. The properties could be summarized simply as 1) the function exists for all possible $p_i$s, 2) the uncertainty grows as the number of possible outcomes grows, and 3) the uncertainty must be additive. We thought about a coding scheme to encode a message written in an alphabet into a different one. We demanded that our coding scheme should be able to encode *any* message we want, and this led us to conclude that the average number of characters needed to encode each character on the original message can approach $\log_m(n)$, where $n$ is the number of characters in the original alphabet and $m$ is the number of characters in the encoding alphabet. We then used the properties of our desired uncertainty function and found a non-obvious connection between the number of characters needed to pass from one alphabet to another and the uncertainty on the message. When we generalized this analysis to cases where not all outcomes are equally likely, we arrived at Eq. 1.76, the so-called Shannon entropy. This is Shannon's theorem, and what it shows is that Eq. 1.76 is the only function that satisfies the three very reasonable conditions we established for an uncertainty measurement.

To gain intuition on what this equation is telling us, let us look at two examples. In our first example, we will think about the simplest random process: a coin toss. To compute how unpredictable the outcome of our simple coin toss is, we can use Eq. 1.76. For this particular case, we only have two possible outcomes—heads with probability $p$ or tails with probability $1 - p$. The resulting entropy is of the form

$$H = -p \log(p) - (1 - p) \log(1 - p). \tag{1.77}$$

Fig. 1.10(A) plots Eq. 1.77 as a function of the probability of heads $p$. Notice that the curve is concave with a minimum at $p = 0$ and $p = 1$ and a maximum at $p = 1/2$. This shape should make intuitive sense given that Eq. 1.77 quantifies

how unpredictable the outcome of tossing the coin is. If the coin toss's outcome is always heads (p=1) or always tails (p=0), there is no uncertainty about the resulting face. The more both outcomes become (the closer $p$ gets to $1/2$), the more unpredictable the random even is. One mathematical subtly here is that for $p = 1$ or $p = 0$, we have to compute $0 \times \log(0)$, which is undefined. For these values of $p$ we take $0 \times \log(0) = 0$ since the limit where $x \to 0^+$ converges to zero. Notice that the units on the $y$-axis are given in bits. These units mean that we used base two for our logarithms. An easy way to think about what a bit means is the number of *yes/no* questions one would need to ask on average to infer the random event's outcome. For a coin, all we need is a single question (therefore one bit) to know what the outcome was.

For our second example, we go back to the mRNA steady-state distribution we derived in Eq. 1.40. We found that for our simple one-state DNA promoter, the steady-state distribution resulted in Poisson with a mean $\langle m \rangle = r_m / \gamma_m$. Fig. 1.10(B) shows the entropy of this Poisson distribution as a function of the mean mRNA. We see a quick initial increase in this entropy up to $\langle m \rangle \approx 20$, after which there is a much less steep increment. Imagine we sample a random cell from one of these Poisson distributions. Using the interpretation of bits again as the number of *yes/no* questions, what Fig. 1.10(B) tells us is that if the promoter produces $\approx 10$ mRNA on average, it will take on average 3.5 of these questions to infer the number of mRNA for random cell. For an average of $\approx 20$ mRNA, it would take four questions, and for an average of $\approx 60$ mRNA, five questions. These questions would be of the form *"is it greater than the average?"* or *"is it less than or equal to 1/3 of the average?,"* and so on.

**Information Theory and Statistical Mechanics**

Our result in Eq. 1.76 is of the same functional form as the thermodynamic entropy. The story goes that Shannon was discussing this concept with his friend John von Neumann. It was von Neumann who allegedly convinced Shannon of calling his metric of randomness *entropy* under the argument that nobody understands the
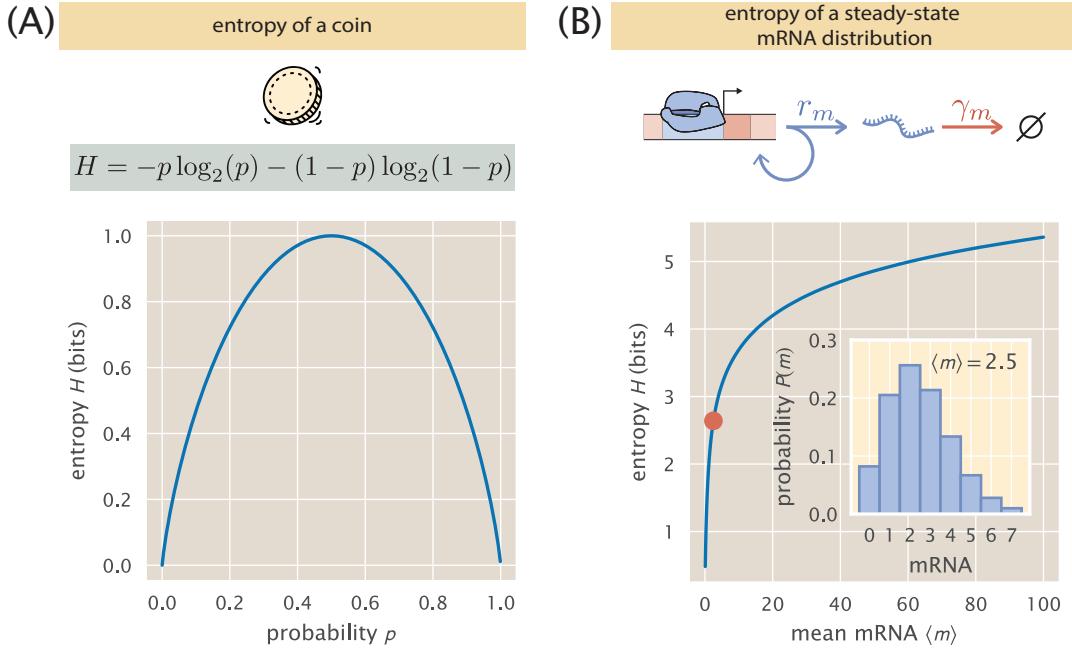
Figure 1.10: **Shannon entropy in action.** (A) The entropy of a coin as a function of the probability of heads $p$. The entropy is maximum when the coin is fair, i.e., $p = 0.5$, meaning that this is the most unpredictable coin one could have. (B) The entropy of the steady-state mRNA distribution as derived in Eq. 1.40 as a function of the mean mRNA copy number. The point shows the entropy of the distribution shown in the inset. Bot figures use base 2 for the logarithm, resulting in units of bits for the entropy. The Python code (`ch1_fig10.py`) used to generate this figure can be found on the thesis GitHub repository.

concept. But the fact that the functional forms are the same is too suggestive to dismiss a potential connection between these concepts immediately. It was until much later that E. T. Jaynes formalized ways to link both ideas [28]. Nevertheless, Jaynes himself strongly discourages people from trying to map one concept to the other explicitly. In his book "*Probability Theory: The Logic of Science*," Jaynes warns the reader about failing to distinguish information entropy, which is a property of the mathematical object we call a probability distribution, and the *experimental entropy* of thermodynamics, which is instead a property of the state of the system as defined by experimentally measurable quantities such as volume, temperature, pressure, magnetization, etc. Jaynes goes on to say: "*they should never have been called by the same name; the experimental entropy makes no reference to any probability distribution, and the information entropy makes no reference to thermodynamics*" [29].

When Jaynes makes such strong remarks about the disconnection between both entropy concepts, he strictly refers to the classical thermodynamic definition. This classical definition of entropy, due to Clausius, refers to the inability of any thermal engine to convert all of the input energy into useful work. Clausius defined a new quantity $S$ as the amount of energy per unit temperature unavailable to do work. To understand this idea is to realize that from the energy liberated in gasoline combustion on a car engine, we only end up extracting $\approx 20\%$ of the energy to move the car. The other 80% is lost into heating the engine and the environment. But this is not because the engineers are using poor designs. The second law of thermodynamics on its classical definition states that nothing in the universe can convert 100% of the energy into useful work; there will always be residual energy that gets turned into heat.

At the time, the existence of atoms was not widely accepted by the scientific community. But then came Boltzmann and the statistical mechanics' conceptual revolution. The giant leap in our understanding of why the second law of thermodynamics does not allow the total conversion of energy into useful work came with Boltzmann's revolutionary entropy idea. Boltzmann hypothesized that matter was made out of atoms. Therefore, everything we can observe and measure macroscopically about any system results from the microscopic configuration of all the atoms that make up the system. Furthermore, many microscopic arrangements are indistinguishable at our macroscopic scale (recall the microstate and macrostate concept in Fig. 1.2). This line of reasoning led Boltzmann to the law we stated in Eq. 1.7. This law and all of the classic results from statistical mechanics are founded on several assumptions about the microscopic scale processes' reversibility. In other words, for Boltzmann's law to be "a legit law of nature," it must be the case that if we play a movie featuring a single atom moving around the system, the same movie played in reverse should be as equally likely to happen.

But it might be the case that the assumptions underlying statistical mechanics laws are not the most fundamental constructs of reality. As we will show next, we can derive a classic result of statistical mechanics from a completely different premise having to do more with statistical inference rather than physical laws of motion governing atoms. This becomes a circular argument where some physicists have the laws of motion as the defining foundation on which to base statistical mechanics laws is better. For others, having an information-theoretic justification for statistical mechanics independent of the underlying physical laws is more appealing. At the end of the day is a matter of taste. Having said all of this, let us delve into the connection between information-theoretic entropy and the Boltzmann distribution.

We already used the Boltzmann distribution when we computed the probability of an RNAP molecule being bound to the promoter $p_{\text{bound}}$. The Boltzmann distribution applies to systems in thermodynamic equilibrium in contact with a heat bath at a constant temperature. Think of a small Eppendorf tube ($\approx$ 2 mL) that we perfectly seal before submerging it into the ocean. The tube's temperature will equilibrate with that of the ocean, but the ocean's temperature will not be affected by the tube's presence. Submerging the tube into the reservoir allows the total energy of the tube not to be fixed. Sometimes the tube can borrow energy from the ocean; sometimes, it can give energy to it. The Boltzmann distribution precisely dictates the likelihood of such energy states. The probability of a state with energy $E_i$ is given by

$$P(E_i) = \frac{e^{-\beta E_i}}{\mathcal{Z}}, \tag{1.78}$$

where, as before, $\beta \equiv (k_B T)^{-1}$. $\mathcal{Z}$ is the partition function defined by the sum of the Boltzmann weight for all possible microstates, i.e.,

$$\mathcal{Z} \equiv \sum_{\text{states}} e^{-\beta E_i}, \tag{1.79}$$

where the sum is taken over all microstates available to the system. This equation is equivalent to Eq. 1.18 and Eq. 1.19. We can derive this functional form from the so-called maximum entropy principle. This framework is expanded more in

Chapter 5 of this thesis. But for our purposes here, the idea is that we are trying to make a "best guess" of what a distribution looks like, given limited information. For our Eppendorf tube inside the ocean, we are thinking about the distribution of all of the molecules' microstates inside the tube. Experimentally, we never get to observe any of the microstates of the system. But we know that the probability of each microstate depends on its energy, as Boltzmann told us. Let us say we can measure the average energy $\langle E \rangle$ of our little Eppendorf tube. What is then the optimal guess of the functional form of the distribution that does not use any information we do not have at hand? For example, we cannot say that there is only one microstate available to the system with energy $\langle E \rangle$, because that constrains the possibilities of the system, and measuring the average energy does not lead to such a conclusion. The next best case we can do is to maximize the Shannon entropy, subject to this constraint on the average energy. This makes sense because, as we derived in the previous section, the Shannon entropy is the only functional form that satisfies our properties for a metric of uncertainty. Maximizing the Shannon entropy leads then to a maximally uninformative distribution. Including the constraints when implementing this maximization guarantees that we use all that we know about the distribution and nothing else.

Given Property 1 of our function $H$, the Shannon entropy is continuous on the individual probabilities' values $p_i$. This means that we can maximize the Shannon entropy by taking its derivative with respect to $p_i$ and equating it to zero. This operation does not include the constraints we have on the values of the probabilities of each microstate. Let us say that each microstate available to the system with energy $E_i$ has a probability $p_i$ of happening. The constraint on the average energy is given by

$$\langle E \rangle = \sum_{\text{states}} E_i p_i, \tag{1.80}$$

where again, the sum is taken over all possible microstates. Furthermore, we know that the probability distribution must be normalized. This means that

$$\sum_{\text{states}} p_i = 1. \tag{1.81}$$

To include these constraints in our optimization, we can use the Lagrange multipliers technique. We refer the reader to any introductory text on multivariate calculus for a quick refresher of this technique. We proceed by defining a Lagrangian $\mathcal{L}$ of the form

$$\mathcal{L}(p_1, p_2, \ldots, p_N, \beta, \mu) = \overbrace{-\sum_{i=1}^{N} p_i \log(p_i)}^{\text{Shannon entropy}} - \overbrace{\beta \left( \sum_{i=1}^{N} E_i p_i - \langle E \rangle \right)}^{\text{average energy constraint}} - \overbrace{\mu \left( \sum_{i=1}^{N} p_i - 1 \right)}^{\text{normalization constraint}},$$

$$(1.82)$$

where $N$ is the total number of microstates available to the system, and $\beta$, and $\mu$ are the Lagrange multipliers associated with each of the constraints. The next step consists on computing the gradient of this Lagrangian which returns a vector of size $N$ where the $k^{\text{th}}$ entry is the derivative of the Lagrangian with respect to $p_k$. But notice that all of these derivatives will look the same. So taking one of these derivatives is enough. We then take the derivative with respect to a particular $p_k$ and equate it to zero, obtaining

$$\frac{d\mathcal{L}}{dp_k} = -\log(p_k) - 1 - \lambda - \beta E_k = 0. \tag{1.83}$$

Notice that all of the terms with $i \neq k$ disappear, leaving a simple expression. Solving for $p_k$ gives

$$p_k = \exp\left[ 1 - \lambda - E_k \right] = e^{1-\lambda} e^{-\beta E_k}. \tag{1.84}$$

Every single probability $p_k$ takes the same form. We substitute this probability $p_k$ on our normalization constraint, obtaining

$$\sum_{i=1}^{N} p_i = e^{1-\lambda} \sum_{i=1}^{N} e^{-\beta E_i} = 1. \tag{1.85}$$

This tells us that the term $e^{1-\lambda}$ is given by

$$e^{1-\lambda} = \frac{1}{\sum_{i=1}^{N} e^{-\beta E_i}}. \tag{1.86}$$

Therefore, the probability of microstate $i$ is given by

$$P(E_i) = p_i = \frac{e^{-\beta E_i}}{\sum_{i=1}^{N} e^{-\beta E_i}}, \tag{1.87}$$

exactly the Boltzmann distribution. One can show why it is the case that our Lagrange multiplier $\beta$ is exactly $1/k_B T$ as demanded by the thermodynamic version of this distribution, but that is out of the scope for our purposes. This section aims only to show the subtle and deep connection between statistical mechanics and information theory. This connection suggests that part of the unreasonable effectiveness of statistical mechanics might not come from the physical basis of its core theory; but instead from the statistical inference problem on which, given the limited information we have of any thermodynamic system's microstate, entropy maximization gives us a recipe on what the best guess for the probability distribution over the microstates is.

**Joint Uncertainty in an Uncertain World**

Part of the complexity in understanding biological systems is that their components form a network of interactions. This connectivity means that one part of the organism's state depends on many other parts' states. For example, the wild-type *lac* operon's expression depends on the conformation state of two transcription factors: CRP and LacI. The state of these transcription factors depends on the concentration of cyclic-AMP and allolactose, respectively. These concentrations rely on the state of the environment and transporters' availability to bring them into the cell. This chain of connections continues indefinitely.

The mathematical language to express the dependence between two variables is that of joint and conditional probability. Shannon's entropy (Eq. 1.76) can also be extended to account for dependence between variables. To make the notation for this extension easier to follow, let us use a different notation from now on. Let us express Shannon's entropy as

$$H(m) = -\sum_m P(m) \log P(m), \tag{1.88}$$

where instead of giving a vector of probabilities **p** to the function $H$, we now give it a random variable $m$. This notation is understood as: the entropy is calculated over the distribution of possible values that $m$ can take. If $m$ can take values $\{m_1, m_2, \ldots, m_n\}$, the probability of obtaining $m = m_k$ is given by the function $P(m = m_k)$, which for brevity we can write simply as $P(m_k)$. What Eq. 1.88 is saying is: take the random variable $m$ and all the possible values it can have; compute the Shannon entropy by summing over the probability of all those values. In this way, $H(m)$ is a shorthand for writing $H[P(m)]$.

With this notation in hand, let us think about two correlated random variables $m$ and $p$. These could be the number of mRNAs and proteins in the cells, as depicted in Fig. 1.11(A). The *joint entropy* $H(m, p)$ measures the uncertainty we have about the outcome of a pair of variables rather than a single. All it takes is to sum over both variables on Eq. 1.88 as

$$H(m, p) = -\sum_m \sum_p P(m, p) \log P(m, p). \tag{1.89}$$

Eq. 1.89 then does the same computation as Eq. 1.88, except that the sum is taken over all possible pairs of random variables $m$ and $p$. But what if we get to observe the outcome of one of the two variables (observing mRNA via RNA-seq, for example), can that tell us something about the outcome of the other one? For this, we need to understand the concept of conditional entropy.

**Thinking Conditionally, a Condition for Thinking**

In Joe Blitztein's excellent *Introduction to probability* [30], he clarifies how conditional probability is one of the most powerful concepts in probability theory. Through the concept of conditional probability, we can learn whether or not two things are somehow correlated, allowing us from there to dissect the nature of such correlation. Given the probabilistic nature of Shannon's entropy, the power of conditional entropy is extended to the so-called conditional entropy $H(p \mid m)$. Let us think of our two random variables $m$ and $p$ with a joint probability distribution $P(m, p)$. We can assume that the outcome of both random variables is correlated

**(A)** toy model for mRNA and protein copy numbers

mRNA probability distribution

$$P(m) = \frac{e^{-r_m/\gamma_m} \left(\frac{r_m}{\gamma_m}\right)^m}{m!}$$

protein probability distribution

$$P(p \mid m) = \frac{e^{-m \cdot r_p/\gamma_p} \left(m\frac{r_p}{\gamma_p}\right)^p}{p!}$$

**(B)** mRNA and protein joint distribution

**(C)** Venn diagram of additive relationships between entropies

$H(m)$    $H(p)$

$H(m|p)$    $I(m;p)$    $H(p|m)$

$H(m,p)$
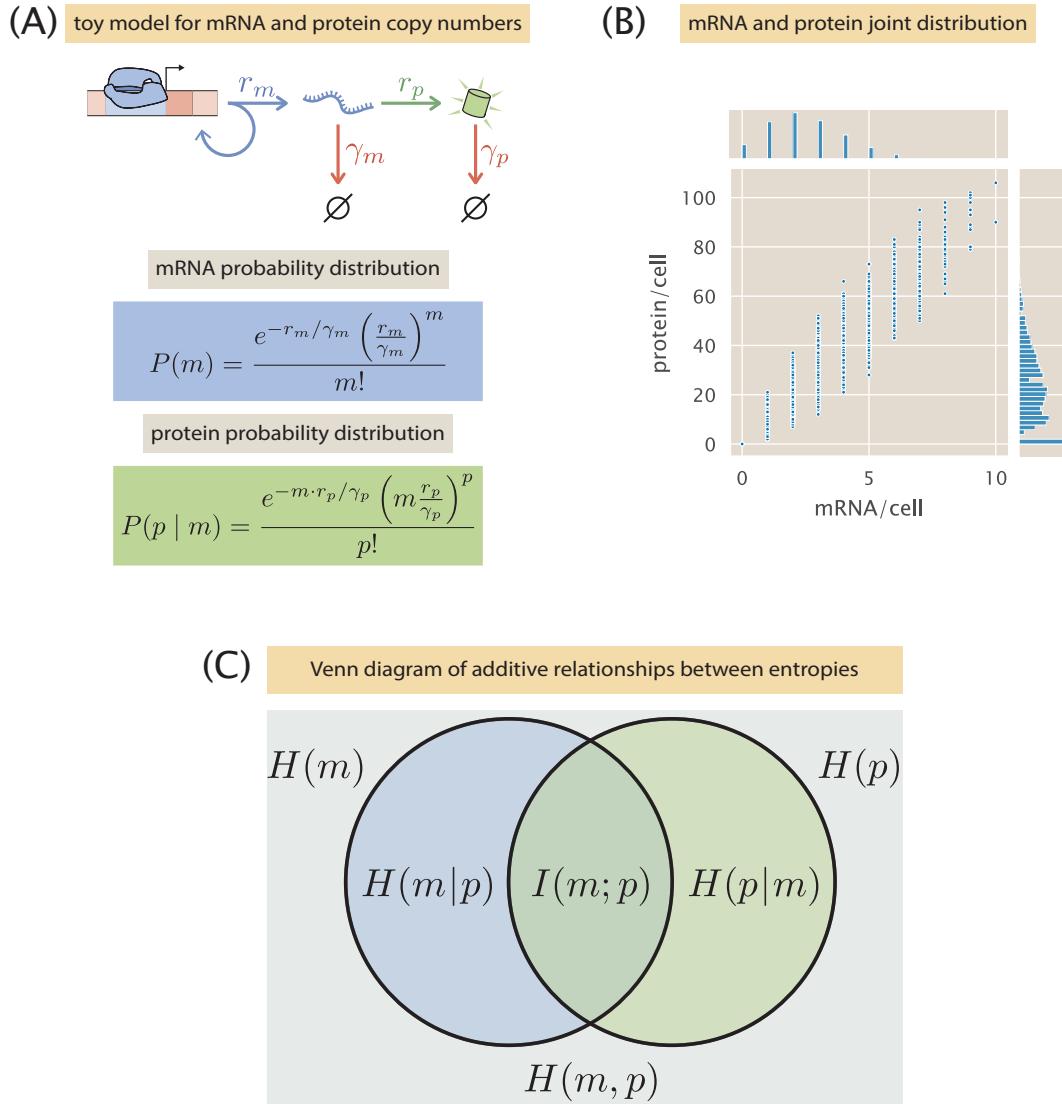
Figure 1.11: **Shannon's entropy for more than one random variable.** (A) Toy model of a random process where mRNA (random variable *m*) is stochastically produced as a Poisson process with a fixed mean. Proteins (random variable *p*) are also stochastically produced as a Poisson process, but the mean depends on the number of mRNAs. (B) Samples from the model presented in (A). The center plot shows the joint distribution $P(m, p)$, while the edge histograms show the marginal distributions $P(m)$ and $P(p)$. (C) Venn diagram of the relationship of different information metrics. The Python code (ch1_fig11.py) used to generate this figure can be found on the thesis GitHub repository.

for our mRNA-protein pair, meaning that specific pairs of values are more likely to appear. If we observed the outcome of one of the random variables and knew the correlation function between random variables, our guess for the variable's value that we did not observe would improve over a completely random choice. In our example, if we get to observe that $m$ is a small (or large) number, we would suspect that $p$ is also a small (or large) number, as shown in Fig. 1.11(B). This means that our uncertainty on the value of $p$ changed—it was reduced—upon observing the value of $m$. The new uncertainty, i.e., the entropy of $p$ having learned the value of $m$, averaged over all possible values of $m$, is computed as

$$H(p \mid m) = - \sum_{m} \sum_{p} P(m) P(p \mid m) \log P(p \mid m), \tag{1.90}$$

where $P(p \mid m)$ is read as "probability of $p$ given that we observe $m$." Finally, with all these concepts in hand, we can discuss the idea of information in the Shannon sense.

**One Person's Entropy is Another Person's Information**

So far, our discussion has focused on the concept of entropy. We first derived the Shannon entropy from three basic principles that a metric of uncertainty should satisfy. Then, we showed that one of the main statistical mechanics results, i.e., the Boltzmann distribution, could be derived from maximizing this entropy subject to certain constraints, suggesting that statistical mechanics could be nothing more than an optimal statistical inference protocol, given limited information. But no mention of information up to now. This intentional omission is because we first needed to master the idea of entropy to understand the mathematical definition of information.

Recall that $H(p)$ quantifies the uncertainty about the outcome of the random process that generates the value of the variable $p$. Furthermore, $H(p \mid m)$ quantifies the uncertainty about the outcome of the same variable, but this time observing the outcome of the random variable $m$. In the worst-case scenario, $m$ and $p$ are uncorrelated, and learning the value of $m$ does not tell us anything about $p$. In that

case, we then have that

$$H(p \mid m) = H(p) \quad \text{for } m \text{ and } p \text{ uncorrelated.} \tag{1.91}$$

If $m$ and $p$ are correlated, as depicted in Fig. 1.11(B), then the uncertainty about $p$ is reduced upon learning the value of $m$, giving us a general relationship between marginal and conditional entropy of the form

$$H(p) \geq H(p \mid m). \tag{1.92}$$

In this latter scenario, learning the value of $m$ reduced our uncertainty in the possible value of $p$. This reduction in uncertainty agrees with an informal definition of what "obtaining information" means. We can then define the mutual information $I(m; p)$ between random variable $m$ and $p$ as the reduction in uncertainty about the value of one of the random variables when we learn the value of the other random variable. For our example in which we get to observe the mRNA copy number, this would mean that the mutual information is computed as

$$I(m; p) \equiv H(p) - H(p \mid m). \tag{1.93}$$

But the mutual information is symmetric, meaning that the information about the outcome of one of the variables by observing the other variables is the same when the roles of what we get to observe are inverted. This argument means that we can mathematically show that

$$I(m; p) = H(m) - H(m \mid p). \tag{1.94}$$

This symmetry is why traditionally, the mutual information is written with a semi-colon rather than a regular comma, indicating that the order of the variables does not matter. To show the above symmetry, let us substitute the definitions of the marginal conditional entropy. This substitution for Eq. 1.93 results in

$$I(m; p) = -\sum_p P(p) \log P(p) - \left[ -\sum_m \sum_p P(m) P(p \mid m) \log P(p \mid m) \right]. \tag{1.95}$$

The trick is now to use the definition of conditional probability in the right way. We know that the conditional probability is defined as

$$P(p \mid m) \equiv \frac{P(m, p)}{P(m)}. \tag{1.96}$$

Furthermore, we know that we can obtain the probability $P(p)$ by marginalizing the joint distribution $P(m, p)$ over all values of $m$. Mathematically this is written as

$$P(p) = \sum_m P(m, p). \tag{1.97}$$

What Eq. 1.97 is stating is that to compute the probability of observing value $p$ of our random variable, we can add the probability of all pairs $m, p$ with the desired that have the desired value of $p$. For Eq. 1.95, we substitute Eq. 1.97 on the first term (outside of the log) of the right-hand side and Eq.~1.96 on the second term (in and outside of the log), obtaining

$$I(m; p) = -\sum_p \left[ \sum_m P(m, p) \right] \log P(p) + \sum_m \sum_p P(m, p) \log \frac{P(m, p)}{P(m)}. \tag{1.98}$$

Since the order of the sums do not matter, we can factorize the common terms on the left-hand side and use the properties of logarithms to write

$$I(m; p) = \sum_m \sum_p P(m, p) \log \frac{P(m, p)}{P(m)P(p)}. \tag{1.99}$$

It is now easier to see that we would arrive at the same result if we started with the opposite conditional entropy $P(m \mid p)$. These series of manipulations where we write either joint or conditional entropies will become handy in this thesis as we explore biophysical models of how to compute gene expression input-output functions (more on that in Chapter 3). Fig. 1.11(C) shows a schematic representation of the relationship of all the entropy-based quantities that we explored in this chapter. Although it is impossible to cover an entire field in a short introduction, I hope this intuitive explanation will suffice to understand the rest of the thesis.

*Chapter 2*

# TUNING TRANSCRIPTIONAL REGULATION THROUGH SIGNALING: A PREDICTIVE THEORY OF ALLOSTERIC INDUCTION

## 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, which 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 [31]. One of the most important examples of such signaling is offered by transcriptional regulation, where a transcription factor's propensity to bind to DNA will be altered upon binding to an allosteric effector.

Despite allostery's ubiquity, we 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 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 [32,33]. 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.

The ability to predict circuit behavior robustly—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 gene's 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 extrapolate

a system's behavior in other conditions accurately. 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 describes 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 [34–38].

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 [20,39,40] and the work of Monod, Wyman, and Changeux [41], we present a statistical mechanical rendering of allostery 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 [41]. 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 promoter 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 an *E. coli* genetic construct 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 [20,39,42–45]. Thus, rather than viewing the behavior of each circuit as giving rise to its 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   Results

**Characterizing Transcription Factor Induction using the Monod-Wyman-Changeux (MWC) Model**

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 [10,48]. When an effector (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 [41]. This causes the repressor to bind more weakly to the operator, which increases 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 [10,20,39,48–50] (we note that non-equilibrium models of simple repression have been shown to have the same functional form that we derive below [51]). 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
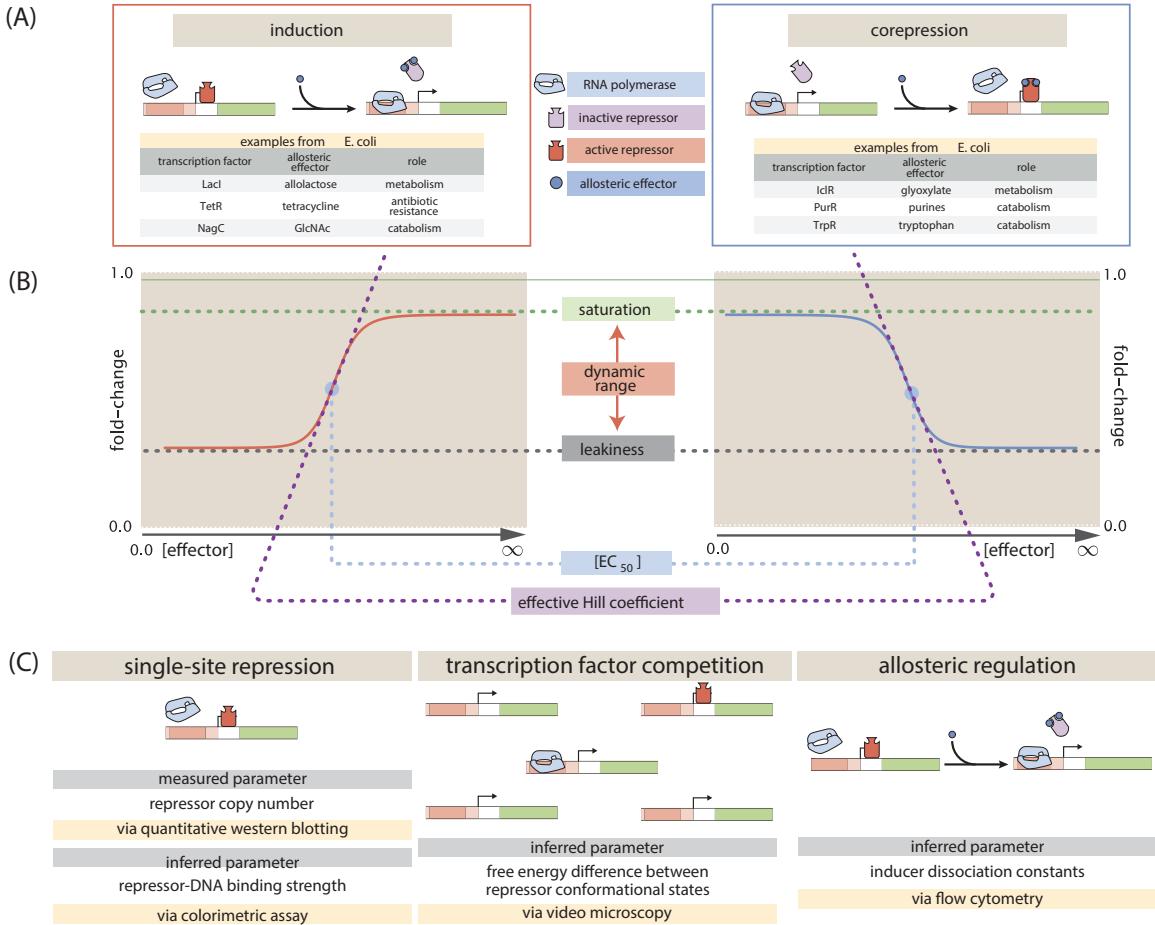
Figure 2.1: **Transcription regulation architectures involving an allosteric repressor.** We consider a promoter regulated solely by an allosteric repressor. When bound, the repressor prevents RNAP from binding and initiating transcription. Induction is characterized by the addition of an effector which binds to the repressor and stabilizes the inactive state (defined as the state which has a low affinity for DNA), thereby increasing gene expression. In corepression, the effector stabilizes the repressor's active state and thus further reduces gene expression. We list several characterized examples of induction and corepression that support different physiological roles in *E. coli* [46,47]. A schematic regulatory response of the two architectures shown in Panel plotting the fold-change in gene expression as a function of effector concentration, where fold-change is defined as the ratio of gene expression in the presence versus the absence of repressor. We consider the following key phenotypic properties that describe each response curve: the minimum response (leakiness), the maximum response (saturation), the difference between the maximum and minimum response (dynamic range), the concentration of ligand which generates a fold-change halfway between the minimal and maximal response ($[EC_{50}]$), and the log-log slope at the midpoint of the response (effective Hill coefficient). (C) Over time, we have refined our understanding of simple repression architectures. A first round of experiments 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 [20]. A second round of experiments used video microscopy to probe how the copy number of the promoter and presence of competing repressor binding sites affect gene expression and we use this data set to determine the free energy difference between the repressor's inactive and active conformations [40]. Here we used 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.

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 remain constant for a given inducer concentration [41]. 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, $\Delta\varepsilon_P$ represents the energy difference between the specific and non-specific binding for RNAP to the DNA. Likewise, $\Delta\varepsilon_{RA}$ and $\Delta\varepsilon_{RI}$ represent the difference in specific and non-specific binding energies for repressor in the active or inactive state, respectively.

Thermodynamic models of transcription [10–12,20,39,40,48–50,52] posit that gene expression is proportional to the probability that the RNAP is bound to the promoter $p_{\text{bound}}$, which is given by

$$p_{\text{bound}} = \frac{\frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_P}}{1 + \frac{R_A}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}} + \frac{R_I}{N_{NS}}e^{-\beta\Delta\varepsilon_{RI}} + \frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_P}}, \tag{2.1}$$

with $\beta = \frac{1}{k_B T}$ where $k_B$ is the Boltzmann constant and $T$ is the temperature 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,

$$\text{fold-change} \equiv \frac{p_{\text{bound}}(R > 0)}{p_{\text{bound}}(R = 0)}. \tag{2.2}$$
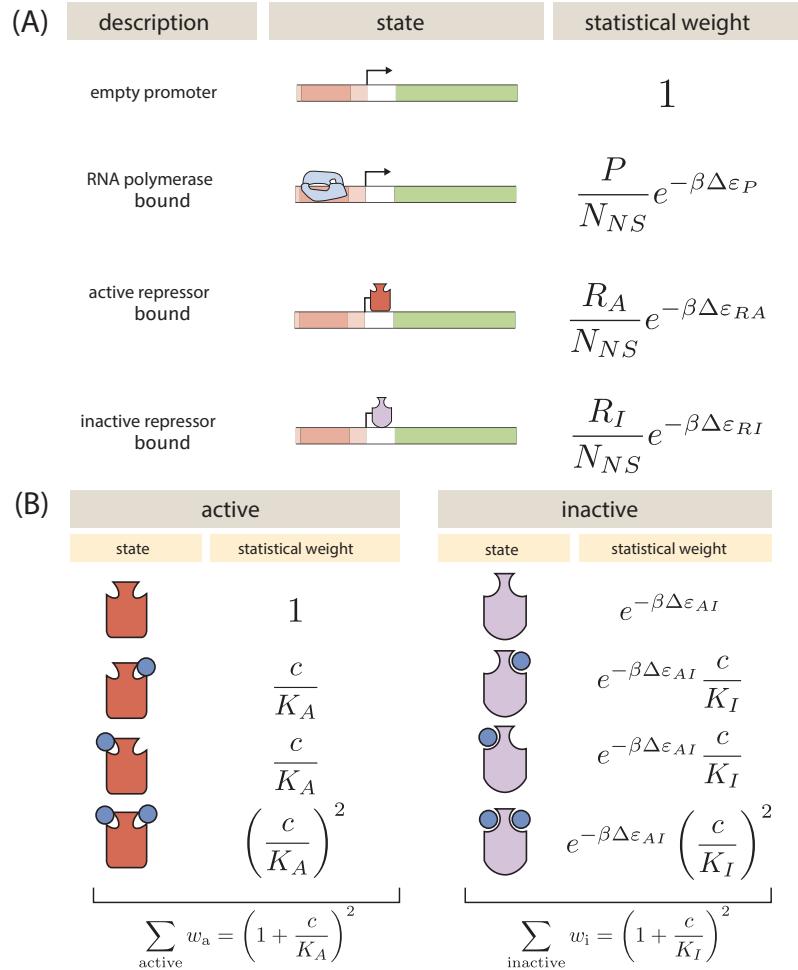
Figure 2.2: **States and weights for the simple repression motif.** RNAP (light blue) and a repressor compete for binding to a promoter of interest. There are $R_A$ repressors in the active state (red) and $R_I$ repressors in the inactive state (purple). The difference in energy between a repressor bound 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. 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 (blue circle) at concentration $c$ can bind 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.

We can simplify this expression using two well-justified approximations: (1) the RNAP binds weakly to the promoter, implying that $\frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_P} \ll 1$ ($N_{NS} = 4.6 \times 10^6$, $P \approx 10^3$ [53], $\Delta\varepsilon_P \approx -2$ to $-5\ k_BT$ [43], so that $\frac{P}{N_{NS}}e^{-\beta\Delta\varepsilon_P} \approx 0.01$) and (2) $\frac{R_I}{N_{NS}}e^{-\beta\Delta\varepsilon_{RI}} \ll 1 + \frac{R_A}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}$ 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

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

where in the last step, we have introduced the fraction $p_A(c)$ of repressors in the active state given a concentration $c$ of inducer, such that $R_A(c) = p_A(c)R$. Since inducer binding shifts the repressors from the active to the inactive state, $p_A(c)$ grows smaller as $c$ increases [54].

We use the MWC model to compute the probability $p_A(c)$ that a repressor with $n$ inducer binding sites will be active. The value of $p_A(c)$ 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,

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

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 $e^{-\Delta\beta\varepsilon_{AI}}$ is sometimes denoted by $L$ [41,54] or $K_{RR*}$ [52]). In this equation, $\frac{c}{K_A}$ and $\frac{c}{K_I}$ 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 the inactive state in the absence of inducer. Thus, a repressor that 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 $p_A(c)$ from Eq. 2.4 into Eq. 2.3 yields the general formula for induction of a simple repression regulatory architecture [51], 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(B)). 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 [11,39,40].

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 from our lab 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 $L = e^{-\beta \Delta \varepsilon_{AI}}$ [52,55], they were either based upon 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 Sec. 4.2 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 exper-

imental 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.

**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 [52,56–59]. 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 [20]. [60] extended the theory used in that work to the case of multiple promoters competing for a given transcription factor, which was validated experimentally by [39], 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 [20] and [39]. Although the current work focuses on systems with a single site of repression, in Sec. 2.5, we utilize data from [39], 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. As explained in that section, 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 Sec. 4.2). After fixing $\Delta\varepsilon_{AI}$ as described in the Sec. 2.5, 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 [61]. 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 [20] 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 the standard deviation of at least three replicates as measured by [20]. We note that $R$ refers to the number of repressor dimers in the cell, which is twice the number of repressor tetramers reported by [20]; since both heads of the repressor are always assumed to 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 [62] placed at the YFP transcription start site, thereby generating eighteen unique strains. The repressor-operator binding energies (O1 $\Delta\varepsilon_{RA} = -15.3 \pm 0.2\ k_BT$, O2 $\Delta\varepsilon_{RA} = -13.9\ k_BT \pm 0.2$, and O3

$\Delta\varepsilon_{RA} = -9.7 \pm 0.1\ k_BT$) were previously inferred by measuring the fold-change of the *lac* system at different repressor copy numbers, where the error arises from model fitting [20]. Additionally, we were able to obtain the value $\Delta\varepsilon_{AI} = 4.5\ k_BT$ by fitting to previous data as discussed in Sec. 2.5. We measure fold-change over a range of known IPTG concentrations $c$, using $n = 2$ inducer binding sites per LacI dimer and approximating the number 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 [63], under concentrations of the inducer IPTG ranging between 0 and 5 mM. We measure YFP fluorescence using flow cytometry and automatically gate the data to include only single-cell measurements (see Sec. 2.5). 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 Sec. 4.5). 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 [39] to infer that $\Delta\varepsilon_{AI} = 4.5\ k_BT$ (see Sec. 4.2). 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 ($\Delta\varepsilon_{RA} = -13.9\ k_BT$ [20]) shown in Fig. 2.4(D) (white circles). Using Markov Chain Monte Carlo, we determine the most likely parameter values to be $K_A = 139^{+29}_{-22} \times 10^{-6}\ \text{M}$ and $K_I = 0.53^{+0.04}_{-0.04} \times 10^{-6}\ \text{M}$, which are the modes of their respective distributions, where the superscripts and subscripts represent the upper and lower bounds of the 95$^{\text{th}}$ percentile of the parameter value distributions (see Fig. 2.4(B)). Unfortunately, we cannot make a meaningful value-
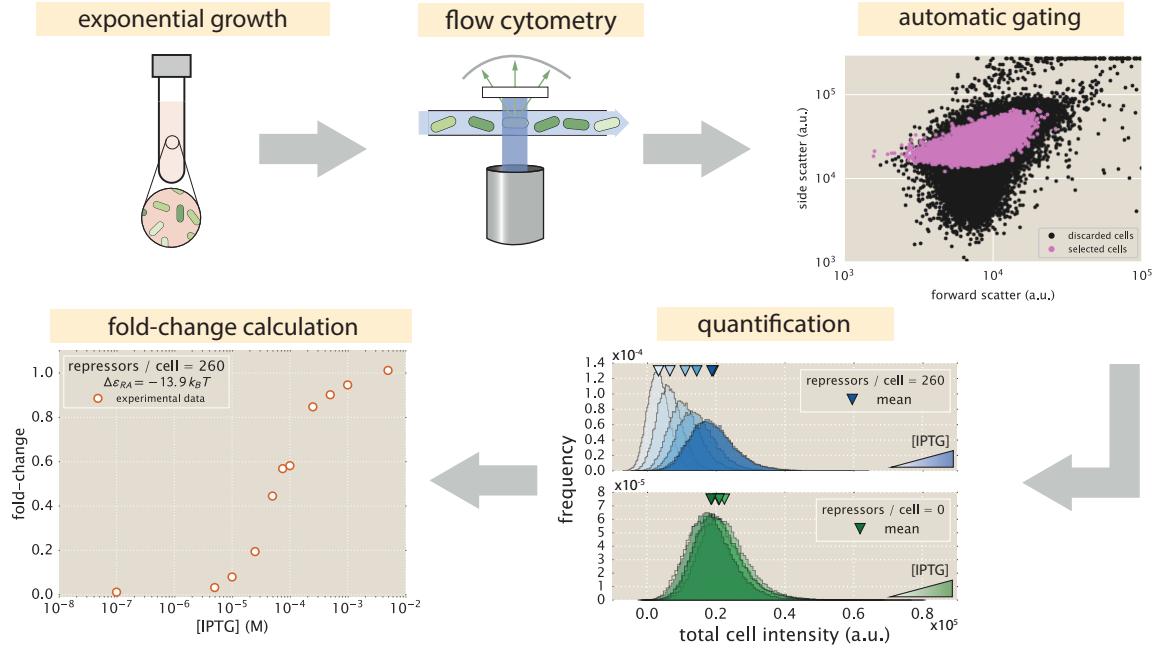
Figure 2.3: **An experimental pipeline for high-throughput fold-change measurements.** Cells are grown to an exponential steady-state, and their fluorescence is measured using flow cytometry. Automatic gating methods using forward- and side-scattering are used to ensure that all measurements come from single cells (see Sec. 2.5). 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.

for-value comparison of our parameters to those of earlier studies [52,57] because of uncertainties in gene copy number and transcription factor copy numbers in these studies. We then 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 and discussed in detail below (see Fig. 2.4(F)-(J)). The shaded regions denote the 95% credible regions. Factors determining the width of the credible regions are explored in Sec. 4.6.

We stress that the entire suite of predictions is based upon a single strain's induction profile. 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(D) could also be fit using a Hill function, such an analysis would be unable to predict any of the other curves in the figure (see Sec. 4.7). 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 [12,56].

**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 [20]. 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 $\Delta\varepsilon_{RA} = -9.7\,k_BT$ strains (Fig. 2.5(C)) 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 a better agreement for these strains. However, a discrepancy in the steepness of the response for all O3 strains remains (see Sec. 4.8). 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 $\Delta\varepsilon_{RA} = -17\,k_BT$ [20]. The global fit agrees well with the Oid microscopy data, though it asserts a stronger Oid binding energy of $\Delta\varepsilon_{RA} = -17.7\,k_BT$ (see Sec. 4.8).
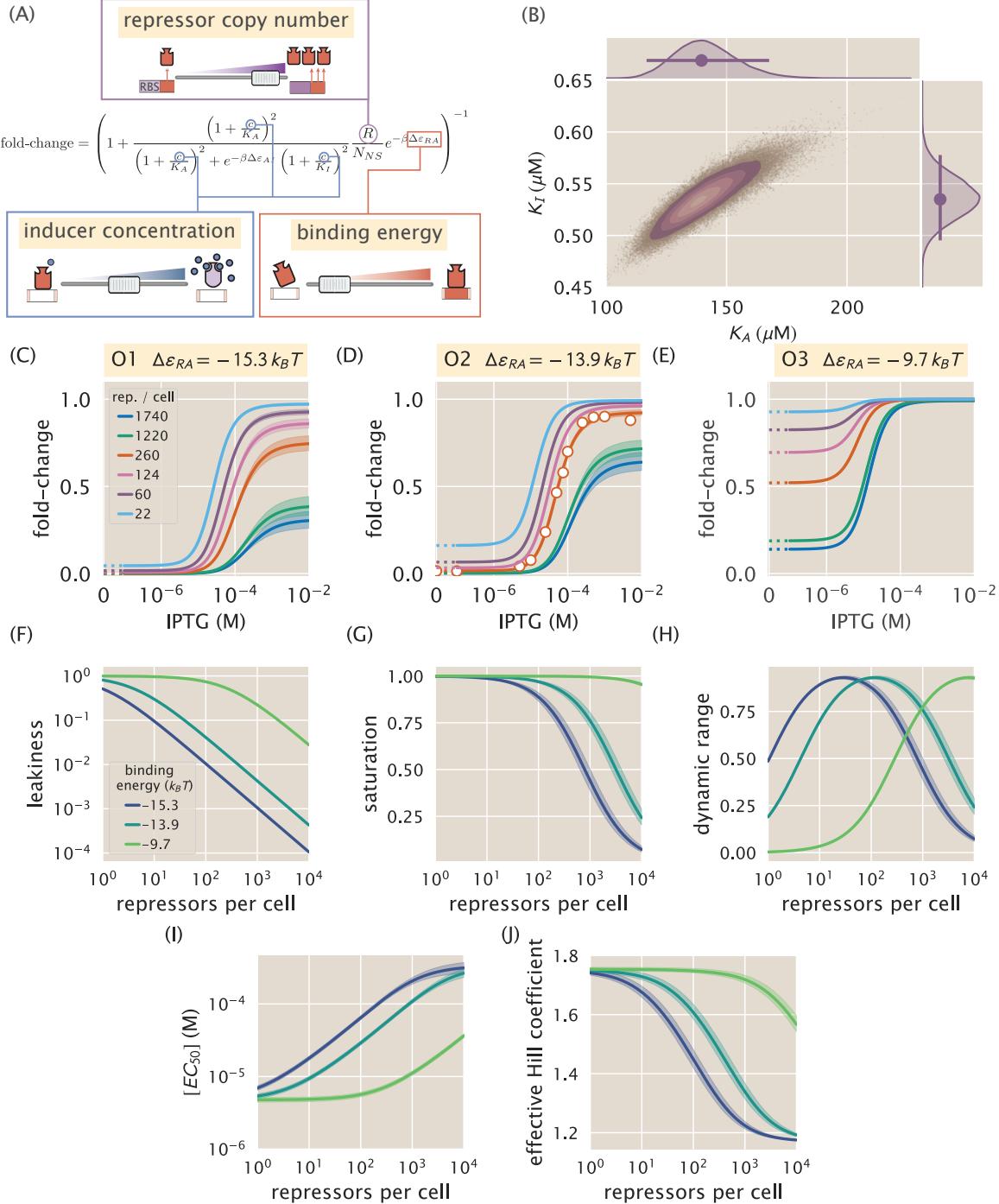
Figure 2.4: **Predicting induction profiles for different biological control parameters.** (A) Schematic representation of experimentally accessible variables. Repressor copy number $R$ is tuned by changing the sequence of the ribosomal binding site (RBS), DNA binding energy $\Delta\varepsilon_{RA}$ 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)-(E) 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. (F)-(J) 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 $10^{-2}$ $\mu$M is scaled linearly with log scaling elsewhere. The Python code (`ch2_fig04.py`) used to generate this figure can be found on the original paper's GitHub repository.

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 shown in Sec. 4.10 and Fig. 2.5(D), the inferred values of $K_A$ and $K_I$ depend minimally upon which strain is chosen, indicating 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 $\Delta\varepsilon_{RA}$ (see Sec. 4.8). The resulting parameter values were nearly identical to those fitted from any single strain. We continue using parameters fitted from the strain with $R = 260$ repressors and an O2 operator for the remainder of the text.

**Predicting the Phenotypic Traits of the Induction Response**

A subset of 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 $[EC_{50}]$) [64,65]. While these properties are all individually informative, they capture the essential features of the induction response when taken together. We reiterate that a Hill function approach cannot predict these features *a priori* and 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, Eq. 2.5. These results build upon extensive work in [66], where many such properties were computed 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

$$\text{leakiness} = \text{fold-change}(c = 0)$$

$$= \left(1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1}, \tag{2.6}$$
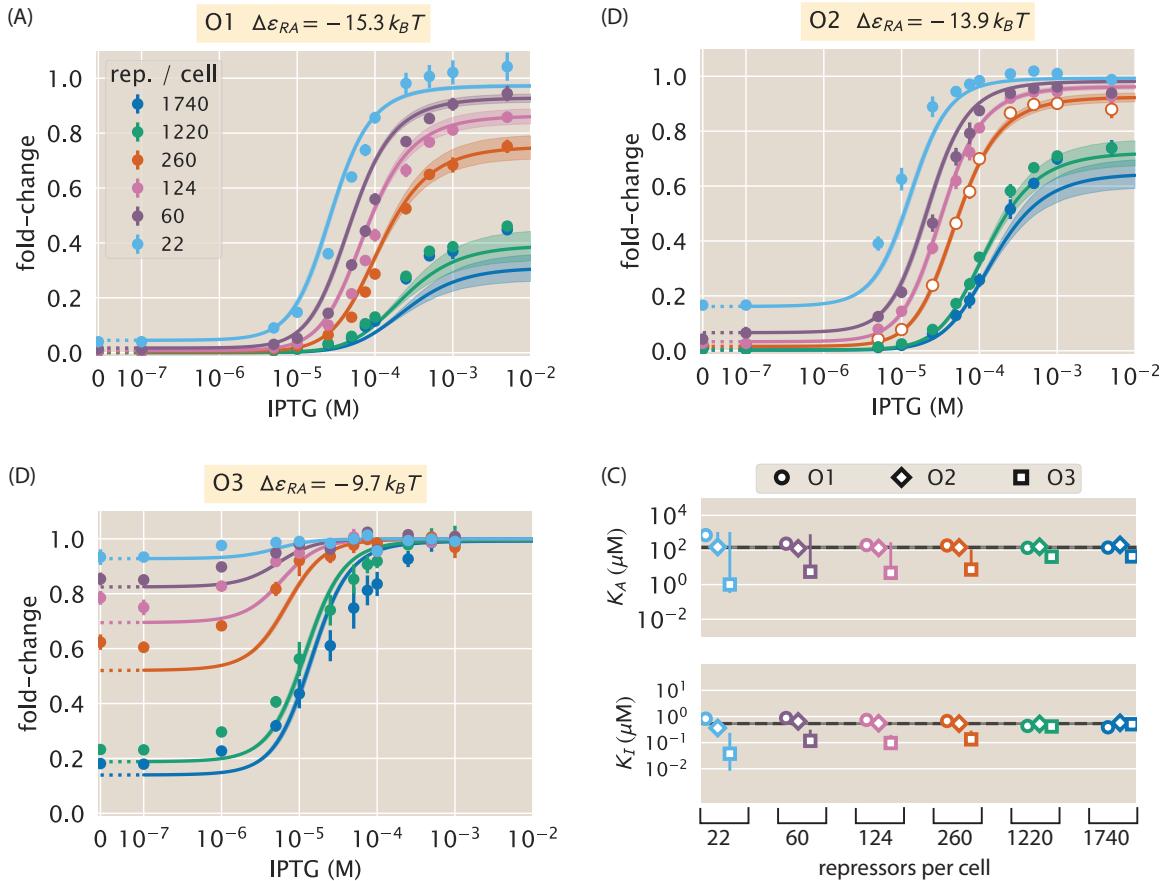
Figure 2.5: **Comparison of predictions against measured and inferred data.** (A-C) Flow cytometry measurements of fold-change over a range of IPTG concentrations for O1, O2, and O3 strains at varying repressor copy numbers overlaid on the predicted responses. Error bars for the experimental data show the standard error of the mean (eight or more replicates). As discussed in Fig. 2.4, all 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. (D) 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_fig05.py) used to generate this figure can be found on the original paper's GitHub repository.

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

$$\text{saturation} = \text{fold-change}(c \to \infty)$$

$$= \left(1 + \frac{1}{1 + e^{-\beta \Delta \varepsilon_{AI}} \left(\frac{K_A}{K_I}\right)^n} \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_{RA}}\right)^{-1}. \tag{2.7}$$

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

$$\text{dynamic range} = \text{saturation} - \text{leakiness}. \tag{2.8}$$

These three properties are shown in Fig. 2.4(F-H). We discuss these properties in greater detail in Sec. 4.11. Fig. 2.6(A-C) 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,

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

The effective Hill coefficient $h$, which quantifies the steepness of the curve at the $[EC_{50}]$ [54], is given by

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

Fig. 2.4(I),(J) shows how the $[EC_{50}]$ and effective Hill coefficient depend on the repressor copy number. Sec. 4.11 discusses the analytic forms of these two properties and their dependence on the repressor-DNA binding energy.
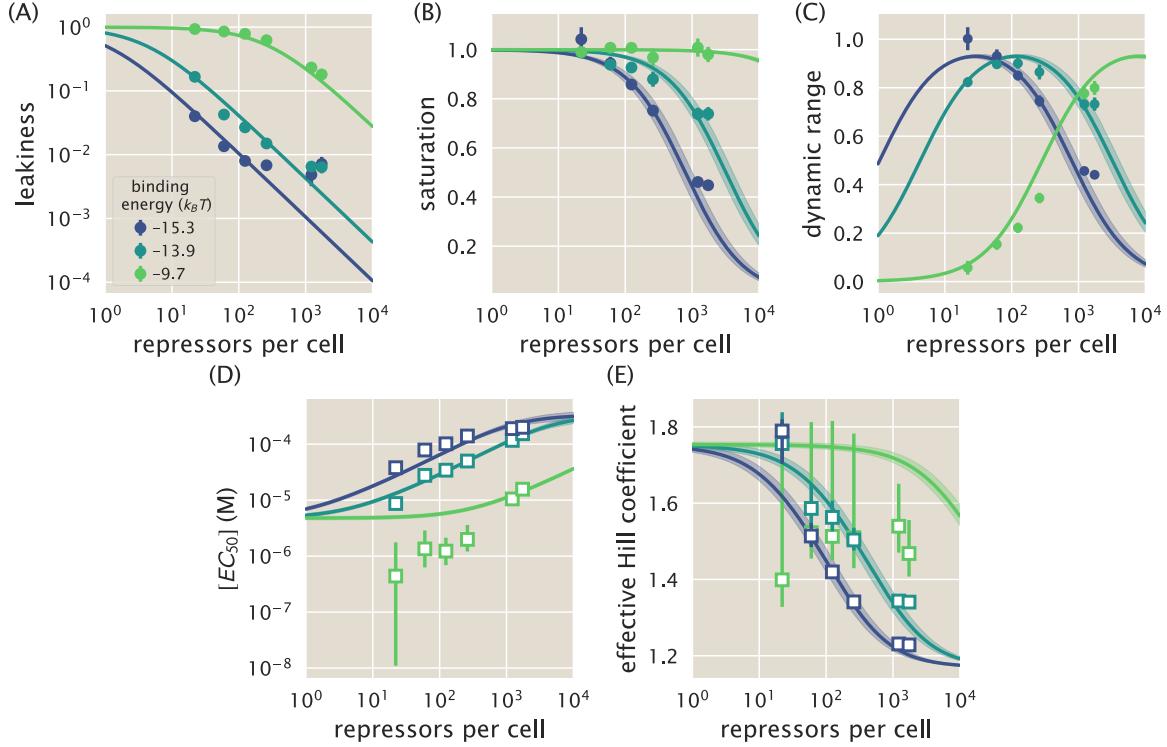
Figure 2.6: **Predictions and experimental measurements of key properties of induction profiles.**
Data for the leakiness, saturation, and dynamic range are obtained from fold-change measurements
in Fig. 2.5 in the absence of IPTG and at saturating concentrations of IPTG. The three repressor-
operator binding energies in the legend correspond to the O1 operator ($-15.3\ k_BT$), O2 operator
($-13.9\ k_BT$), and O3 operator ($-9.7\ k_BT$). Both the $[EC_{50}]$ and effective Hill coefficient are inferred
by individually fitting each operator-repressor pairing in Fig. 2.5(A-C) separately to Eq. 2.5 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. 2.9 and
Eq. 2.10. The Python code (`ch2_fig06.py`) used to generate this figure can be found on the original
paper's GitHub repository.

Fig. 2.6(D) and Fig. 2.6(E) shows the estimated values of the $[EC_{50}]$ and the effec-
tive Hill coefficient overlaid on the theoretical predictions. Both properties were
obtained by fitting Eq. 2.5 to each individual titration curve and computing the
$[EC_{50}]$ and effective Hill coefficient using Eq. 2.9 and Eq. 2.10, respectively. We
find that the predictions made with the single strain fit closely match those made
for each of the strains with O1 and O2 operators, but the predictions for the O3 op-
erator are markedly off. Sec. 4.10 shows 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 con-
stants $K_A$ 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 Sec. 4.8). It remains an open question on how to account for discrepancies in O3, particularly 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 for understanding how the system is governed and for engineering specific responses in a synthetic biology context. To address these questions, we follow the data collapse strategy used in a number of previous studies [67–69], and rewrite Eq. 2.5 as a Fermi function,

$$\text{fold-change} = \frac{1}{1 + e^{-F(c)}}, \tag{2.11}$$

where $F(c)$ is the free energy of the repressor binding to the operator of interest relative to the unbound operator state in $k_B T$ units [51,68,69], which is given by

$$F(c) = \frac{\Delta \varepsilon_{RA}}{k_B T} - \log \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{-\beta \Delta \varepsilon_{AI}} \left(1 + \frac{c}{K_I}\right)^n} - \log \frac{R}{N_{NS}}. \tag{2.12}$$

The first term in $F(c)$ 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 pHs known as the Bohr effect [51,70,71].

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. 2.7(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 energy. Fig. 2.7(B) shows this 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.

Many different combinations of parameter values can result in the same free energy as defined in Eq. 2.12. For example, suppose a system initially has a fold-change of 0.2 at a specific inducer concentration, and then operator mutations increase the $\Delta\varepsilon_{RA}$ binding energy [72]. While this initially increases 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.

## 2.4 Discussion

Since the early work by Monod, Wyman, and Changeux [41,73], 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 [74], enzymatic reactions [71,75], chemotaxis [68], quorum sensing [69], G-protein coupled receptors [76], physiologically important proteins [77,78], and beyond. One of the most ubiquitous examples of allostery 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 [46,47]. A number of studies have focused on developing a quantitative understanding of allosteric regulatory systems. [54,66] analytically derived fundamental properties of the MWC model, including the leakiness and
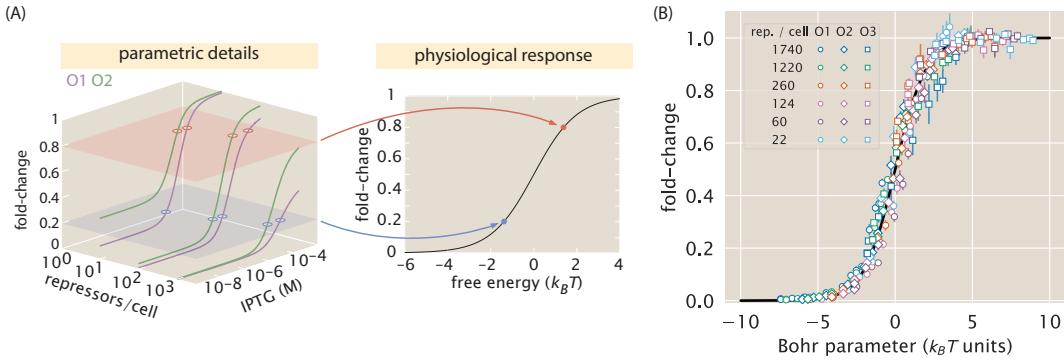
Figure 2.7: **Fold-change data from a broad collection of different strains collapse onto a single master curve.** (A) Any combination of parameters can be mapped to a single physiological response (i.e., fold-change) via the free energy, which encompasses the parametric details of the model. (B) Experimental data from collapse onto a single master curve as a function of the free energy Eq. 2.12. The free energy for each strain was calculated from Eq. 2.12 using $n = 2$, $\Delta \varepsilon_{AI} = 4.5\ k_B T$, $K_A = 139 \times 10^{-6}$ M, $K_I = 0.53 \times 10^{-6}$ M, and the strain-specific $R$ and $\Delta \varepsilon_{RA}$. 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_fig07.py`) used to generate this figure can be found on the original paper's GitHub repository.

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 [37,38,79,80]. Recently, Daber *et al.* theoretically explored the induction of simple repression within the MWC model [57] and experimentally measured how mutations alter the induction profiles of transcription factors [52]. Vilar and Saiz analyzed a variety of interactions in inducible *lac*-based systems, including the effects of oligomerization and DNA folding on transcription factor induction [36,81]. Other work has attempted to use the *lac* system to reconcile *in vitro* and *in vivo* measurements [59,82].

Although this body of work has done much to improve our understanding of allosteric transcription factors, there have been few attempts to connect quantitative models to experiments explicitly. 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 [11,20], allowing us to compose the model from a mini-

mal 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 inducer concentration. 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 generate such a wide array of predictions by using a Hill function, which abstracts away the biophysical meaning of the parameters into phenomenological parameters [83].

More precisely, we tested our model in the context of a *lac*-based simple repression system by first determining the allosteric dissociation constants $K_A$ and $K_I$ from a single induction data set (O2 operator with binding energy $\Delta\varepsilon_{RA} = -13.9\, k_B T$ and repressor copy number $R = 260$), and then using these values to make parameter-free predictions of the induction profiles for seventeen other strains where $\Delta\varepsilon_{RA}$ and $R$ were varied significantly. We next measured the induction profiles of these seventeen strains using flow cytometry and found that our predictions consistently and accurately captured the primary features for each induction data set, as shown in Fig. 2.5. Importantly, we find that fitting $K_A$ and $K_I$ to data from any other strain would have resulted in nearly identical predictions (see Sec. 4.10 for further details). This suggests that a few carefully chosen measurements can lead to a deep quantitative understanding of how simple regulatory systems work without requiring an extensive sampling of strains that span the parameter space. Moreover, the fact that we could consistently achieve reliable predictions after fitting only two free parameters stand in contrast to the common practice of fitting several free parameters simultaneously, which can nearly guarantee an acceptable fit provided that the model roughly resembles the system response, regardless of whether the details of the model are tied to any underlying molecular mechanism.

Beyond observing changes in fold-change as a function of effector concentration, our application of the MWC model allows us to predict the values of explicitly the induction curves' key parameters, namely, the leakiness, saturation, dynamic range, $[EC_{50}]$, and the effective Hill coefficient. We are consistently able to accurately predict the leakiness, saturation, and dynamic range for each of the strains. For both the O1 and O2 data sets, our model also accurately predicts the effective Hill coefficient and $[EC_{50}]$, though these predictions for O3 are noticeably less accurate. While performing a global fit for all model parameters marginally improves the prediction for O3 (see Sec. 4.8), we are still unable to predict the effective Hill coefficient or accurately 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 [84], and refitting the original binding energies from [42], but we were still 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.

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 [37,38].

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 Eq. 2.12, 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 [67–69]. 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, suppose one can find the "natural variable" that enables the output to collapse onto a single curve. In that case, 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 each construct's respective free energies, they collapse cleanly onto a single curve. This enables us to analyze how parameters can compensate for 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 $F(c) \leq -5k_BT$, 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 [11]. We can further explore different regulatory configurations such as corepression, activation, and coactivation, each of which are found in *E. coli* (see Sec. 4.12). 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 [85]. 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

$\Delta\varepsilon_{AI}$ between the allosteric states can be tuned by making single amino acid substitutions in the transcription factor [51,52]. Finally, we expect that the rigorous quantitative description of the allosteric phenomenon provided here will make it possible to construct biophysical 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 [25,86,87].

To conclude, we find that our application of the MWC model provides an accurate, predictive framework for understanding simple repression by allosteric transcription factors. To reach this conclusion, we analyzed the model in the context of a well-characterized system, in which each parameter had a clear biophysical meaning. As many of these parameters had been measured or inferred in previous studies, this gave us a minimal model with only two free parameters, which we inferred from a single data set. We then accurately predicted the behavior of seventeen other data sets in which repressor copy number and repressor-DNA binding energy were systematically varied. In addition, our model allowed us to understand how key properties such as the leakiness, saturation, dynamic range, $[EC_{50}]$, and effective Hill coefficient depended upon the small set of parameters governing this system. Finally, we show that by framing inducible simple repression in terms of free energy, the data from all of our experimental strains collapse cleanly onto a single curve, illustrating the many ways in which a particular output can be targeted. In total, these results show that a thermodynamic formulation of the MWC model supersedes phenomenological fitting functions for understanding transcriptional regulation by allosteric proteins.

## 2.5   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 [20,42]. 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 [88]. The *lacI* gene was constitutively expressed via a $P_{LtetO-1}$ promoter [79], with ribosomal binding site mutations made to vary the LacI copy number as described in [89] using site-directed mutagenesis (Quickchange II; Stratagene), with further details in [20]. 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 [90] 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 Sec. 4.15.

It is important to note that the rest of the *lac* operon (*lacZYA*) was never expressed. The LacY protein is a transmembrane protein that 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 approximately equal due to diffusion across the membrane into the cell, as suggested by previous work [91].

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 [20], the copy number is defined as the average number of repressor tetramers in each cell. To motivate this decision, we consider that the LacI repressor molecule exists as a tetramer in *E. coli* [92] 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) [57], a single LacI tetramer can be treated as two functional repressors. Therefore, we have multiplied the

number of repressors reported in [20] by a factor of two. This factor is included as a keyword argument in the numerous Python functions used to perform this analysis, as discussed in the code documentation.

A subset of strains in these experiments was measured using fluorescence microscopy to validate the flow cytometry data and results. To aid in the high-fidelity segmentation of individual cells, the strains were modified to express an mCherry fluorophore constitutively. This reporter was cloned into a pZS4*1 backbone [79] 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 concentra-

tions. 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**

Unless explicitly mentioned, 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 Sec. 4.4. 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 two to three hours using automated sampling from a 96-well plate kept at approximately $4^\circ$ - $10^\circ$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 determined to be "real" is commonly referred to as gating. These gates are typically drawn manually [93] 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 [94,95]. 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 the calculation to those that reside within the 40th percentile. This procedure is described in more detail in the supplementary information as well as in a Jupyter notebook located in this paper's Github repository.

**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

$$\text{fold-change} = \frac{\langle I_{R>0} \rangle - \langle I_{\text{auto}} \rangle}{\langle I_{R=0} \rangle - \langle I_{\text{auto}} \rangle}, \tag{2.13}$$

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

$$P(K_A, K_I \mid D) = \frac{P(D \mid K_A, K_I) P(K_A, K_I)}{P(D)}, \tag{2.14}$$

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). $P(D \mid K_A, K_I)$ 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. Our model assumes a deterministic relationship between the parameters and the data, so to construct a probabilistic relationship as required by Eq. 2.14, we assume that the experimental fold-change for the $i^{\text{th}}$ datum given the parameters is of the form

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

$$(2.15)$$

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 $P(D \mid K_A, K_I, \sigma)$ is of the form

$$P(D|K_A,K_I,\sigma) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \prod_{i=1}^{n} \exp\left[ -\frac{(\text{fold-change}_{\text{exp}}^{(i)} - \text{fold-change}(K_A,K_I,R^{(i)},\Delta\varepsilon_{RA}^{(i)},c^{(i)}))^2}{2\sigma^2} \right], \qquad (2.16)$$

where $\text{fold-change}_{\text{exp}}^{(i)}$ is the experimental fold-change and $\text{fold-change}(\cdots)$ 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. 2.16, a choice of $K_A$ and $K_I$ that provides a 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 $\tilde{k}_A = -\log\frac{K_A}{1\,\text{M}}$ and $\tilde{k}_I = -\log\frac{K_I}{1\,\text{M}}$ and fit for these parameters on a log scale. Dissociation constants are scale invariant, so that a change from $10\,\mu$M to $1\,\mu$M leads to an equivalent increase in affinity as a change from $1\,\mu$M to $0.1\,\mu$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 [61] for the scale parameter $\sigma$. This yields the complete prior

$$P(\tilde{k}_A, \tilde{k}_I, \sigma) \equiv \frac{1}{(\tilde{k}_A^{\max} - \tilde{k}_A^{\min})} \frac{1}{(\tilde{k}_I^{\max} - \tilde{k}_I^{\min})} \frac{1}{\sigma}. \tag{2.17}$$

These priors are maximally uninformative, meaning that they imply no prior knowledge of the parameter values. We defined the $\tilde{k}_A$ and $\tilde{k}_A$ 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 $P(\tilde{k}_A, \tilde{k}_I, \sigma \mid D)$ using Markov chain Monte Carlo (see GitHub repository) 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 and all relevant code can be found at this dedicated website. Data were collected, stored, and preserved using the Git version control software combined with off-site storage and hosting website GitHub. Code is used to generate all figures and complete all processing steps, 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*

# FIRST-PRINCIPLES PREDICTION OF THE INFORMATION PROCESSING CAPACITY OF A SIMPLE GENETIC CIRCUIT

## 3.1 Abstract

Given the stochastic nature of gene expression, genetically identical cells exposed to the same environmental inputs will produce different outputs. This heterogeneity has been hypothesized to have consequences for how cells can survive in changing environments. Recent work has explored the use of information theory as a framework to understand the accuracy with which cells can ascertain the state of their surroundings. Yet, the predictive power of these approaches is limited and has not been rigorously tested using precision measurements. To that end, we generate a minimal model for a simple genetic circuit in which all parameter values for the model come from independently published data sets. We then predict the information processing capacity of the genetic circuit for a suite of biophysical parameters such as protein copy number and protein-DNA affinity. Finally, we compare these parameter-free predictions with an experimental determination of protein expression distributions and the resulting information processing capacity of *E. coli* cells. We find that our minimal model captures the scaling of the cell-to-cell variability in the data and the inferred information processing capacity of our simple genetic circuit up to a systematic deviation.

## 3.2    Introduction

As living organisms thrive in a given environment, they are faced with constant changes in their surroundings. From abiotic conditions such as temperature fluctuations or changes in osmotic pressure, to biological interactions such as cell-to-cell communication in a tissue or a bacterial biofilm, living organisms of all types sense and respond to external signals. Fig. 3.1(A) shows a schematic of this process for a bacterial cell sensing a concentration of an extracellular chemical. At the molecular level, where signal transduction unfolds mechanistically, there are physical constraints on the accuracy and precision of these responses given by intrinsic stochastic fluctuations [96]. This means that two genetically identical cells exposed to the same stimulus will not have identical responses [22].

One implication of this noise in biological systems is that cells do not have an infinite resolution to distinguish signals. Consequently, there is a one-to-many mapping between inputs and outputs. Furthermore, given the limited number of possible outputs, there are overlapping responses between different inputs. This scenario can be mapped to a Bayesian inference problem where cells try to infer the state of the environment from their phenotypic response, as schematized in Fig. 3.1(B). The question then becomes this: how can one analyze this probabilistic, rather than deterministic, relationship between inputs and outputs? The abstract answer to this question was worked out in 1948 by Claude Shannon who, in his seminal work, founded the field of information theory [27]. Shannon developed a general framework for how to analyze information transmission through noisy communication channels. In his work, Shannon showed that the only quantity that satisfies three reasonable axioms for a measure of uncertainty was of the same functional form as the thermodynamic entropy–thereby christening his metric the information entropy [97]. Based on this information entropy, he also defined the relationship between inputs and outputs known as the mutual information. The mutual information $I$ between input $c$ and output $p$, given by

$$I = \sum_c P(c) \sum_p P(p \mid c) \log_2 \frac{P(p \mid c)}{P(p)}, \qquad (3.1)$$

quantifies how much we learn about the state of the input $c$ given that we get to observe the output $p$. In other words, the mutual information can be thought of as a generalized correlation coefficient that quantifies the degree to which the uncertainty about a random event decreases given the knowledge of the average outcome of another random event [98].

It is natural to conceive of scenarios in which living organisms that better resolve signals might have an evolutionary benefit, making it more likely that their offspring will have a fitness advantage [6]. In recent years, there has been a growing interest in understanding the theoretical limits on cellular information processing [99,100], and in quantifying how close evolution has pushed cellular signaling pathways to these theoretical limits [101–103]. While these studies have treated the signaling pathway as a "black box," explicitly ignoring all the molecular interactions taking place in them, other studies have explored the role that molecular players and regulatory architectures have on these information processing tasks [23,104–109]. Despite the great advances in our understanding of the information processing capabilities of molecular mechanisms, the field still lacks a rigorous experimental test of these detailed models with precision measurements on a simple system in which physical parameters can be perturbed. In this work, we approach this task with a system that is both theoretically and experimentally tractable in which molecular parameters can be varied in a controlled manner.

Over the last decade, the dialogue between theory and experiments in gene regulation has led to the predictive power of models not only over the mean level of gene expression but the noise as a function of relevant parameters such as regulatory protein copy numbers, the affinity of these proteins to the DNA promoter, as well as the extracellular concentrations of inducer molecules [20,36,110,111]. These models based on equilibrium and non-equilibrium statistical physics have reached a predictive accuracy level such that, for simple cases, it is now possible to design input-output functions [43,112]. This opens the opportunity to exploit these predictive models to tackle how much information genetic circuits can pro-

cess. This question lies at the heart of understanding the precision of the cellular response to environmental signals. Fig. 3.1(C) schematizes a scenario in which two bacterial strains respond with different levels of precision to three possible environmental states, i.e., inducer concentrations. The overlap between the three different responses precisely determines the resolution with which cells can distinguish different inputs. This is analogous to how the point spread function limits the ability to resolve two light-emitting point sources.

In this work, we follow the same philosophy of theory-experiment dialogue used to determine model parameters to predict from first principles the effect that biophysical parameters such as transcription factor copy number and protein-DNA affinity have on the information processing capacity of a simple genetic circuit. Specifically, to predict the mutual information between an extracellular chemical signal (input $c$, isopropyl $\beta$-D-1-thiogalactopyranoside or IPTG in our experimental system) and the corresponding cellular response in the form of protein expression (output $p$), we must compute the input-output function $P(p \mid c)$. To do so, we use a master-equation-based model to construct the protein copy number distribution as a function of an extracellular inducer concentration for different combinations of transcription factor copy numbers and binding sites. Having these input-output distributions allow us to compute the mutual information $I$ between inputs and outputs for any arbitrary input distribution $P(c)$. We opt to compute the channel capacity, i.e., the maximum information that can be processed by this gene regulatory architecture, defined as Eq. 3.1 maximized over all possible input distributions $P(c)$. By doing so we examine the physical limits of what cells can do in terms of information processing by harboring these genetic circuits. Nevertheless, given the generality of the input-output function $P(p \mid c)$ we derive, the model presented here can be used to compute the mutual information for any arbitrary input distribution $P(c)$. All parameters used for our model were inferred from a series of studies that span several experimental techniques [20,39,84,113], allowing us to make parameter-free predictions of this information processing capacity [16].
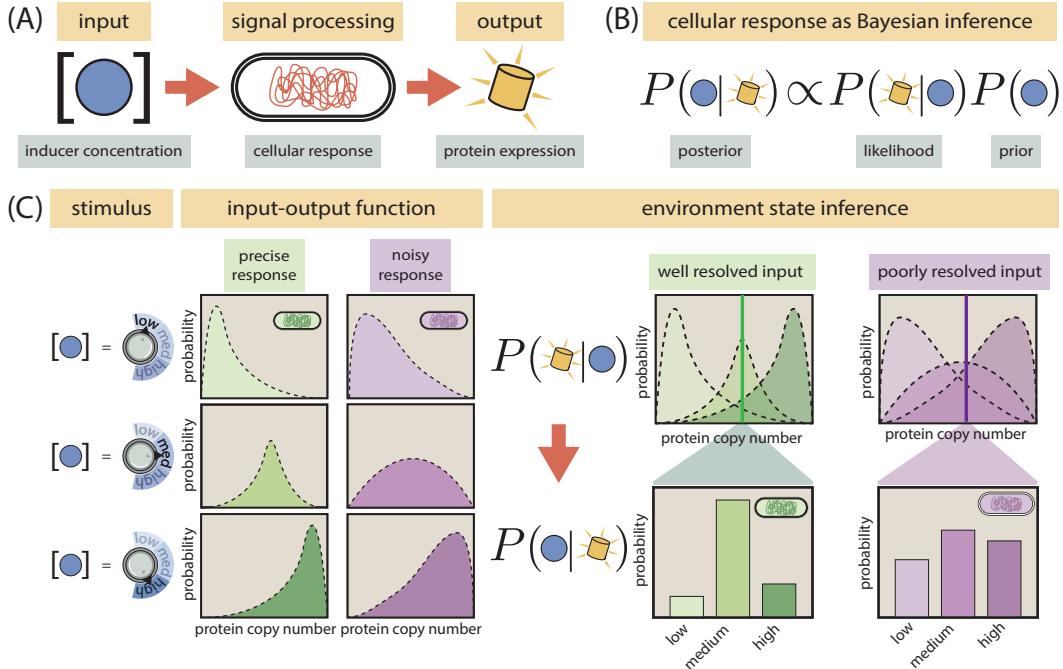
Figure 3.1: **Cellular signaling systems sense the environment with different degrees of precision.** (A) Schematic representation of a cell as a noisy communication channel. From an environmental input (inducer molecule concentration) to a phenotypic output (protein expression level), cellular signaling systems can be modeled as noisy communication channels. (B) We treat cellular response to an external stimulus as a Bayesian inference of the state of the environment. As the phenotype (protein level) serves as the internal representation of the environmental state (inducer concentration), the probability of a cell being in a specific environment given this internal representation $P(c \mid p)$ is a function of the probability of the response given that environmental state $P(p \mid c)$. (C) The precision of the inference of the environmental state depends on how well cells can resolve different inputs. For three different input levels (left panel), the green strain responds more precisely than the purple strain since the output distributions overlap less (middle panel). This allows the green strain to make a more precise inference of the environmental state given a phenotypic response (right panel).

These predictions are then contrasted with experimental data, where the channel capacity is inferred from single-cell fluorescence distributions taken at different inducer concentrations for cells with previously characterized biophysical parameters [20,113]. We find that our parameter-free predictions quantitatively track the experimental data up to a systematic deviation. The lack of numerical agreement between our model and the experimental data poses new challenges towards having a foundational, first-principles understanding of the physics of cellular decision-making.

### 3.3 Results

**Minimal Model of Transcriptional Regulation**

As a tractable circuit for which we have control over the parameters both theoretically and experimentally, we chose the so-called simple repression motif, a common regulatory scheme among prokaryotes [114]. This circuit consists of a single promoter with an RNA-polymerase (RNAP) binding site and a single binding site for a transcriptional repressor [20]. The regulation due to the repressor occurs via exclusion of the RNAP from its binding site when the repressor is bound, decreasing the likelihood of having a transcription event. As with many important macromolecules, we consider the repressor to be allosteric, meaning that it can exist in two conformations, one in which the repressor can bind to the specific binding site (active state) and one in which it cannot bind the specific binding site (inactive state). The environmental signaling occurs via the passive import of an extracellular inducer that binds the repressor, shifting the equilibrium between the two conformations of the repressor [113]. In previous work, we have extensively characterized the mean response of this circuit under different conditions using equilibrium-based models [16]. Here we build upon these models to characterize the entire distribution of gene expression with parameters such as repressor copy number and its affinity for systematically varied DNA.

As the copy number of molecular species is a discrete quantity, chemical master equations have emerged as a useful tool to model their inherent probability distribution [26]. In Fig. 3.2(A), we show the minimal model and the necessary set of parameters needed to compute the full distribution of mRNA and its protein gene product. Specifically, we assume a three-state model where the promoter can be found in a 1) transcriptionally active state (*A* state), 2) a transcriptionally inactive state without the repressor bound (*I* state), and 3) a transcriptionally inactive state with the repressor bound (*R* state). We do not assume that the transition between the active state *A* and the inactive state *I* occurs due to RNAP binding to the promoter as the transcription initiation kinetics involve several more steps than sim-

ple binding [8]. We coarse-grain all these steps into effective *"on"* and *"off"* states for the promoter, consistent with experiments demonstrating the bursty nature of gene expression in *E. coli* [110]. These three states generate a system of coupled differential equations for each of the three state distributions $P_A(m, p; t)$, $P_I(m, p; t)$ and $P_R(m, p; t)$, where $m$ and $p$ are the mRNA and protein count per cell, respectively, and $t$ is time. Given the rates depicted in Fig. 3.2(A), we define the system of ODEs for a specific $m$ and $p$. For the transcriptionally active state, we have

$$\frac{dP_A(m, p)}{dt} = -\overbrace{k_{\text{off}}^{(p)} P_A(m, p)}^{A \to I} + \overbrace{k_{\text{on}}^{(p)} P_I(m, p)}^{I \to A}$$

$$+ \overbrace{r_m P_A(m-1, p)}^{m-1 \to m} - \overbrace{r_m P_A(m, p)}^{m \to m+1} + \overbrace{\gamma_m(m+1) P_A(m+1, p)}^{m+1 \to m} - \overbrace{\gamma_m m P_A(m, p)}^{m \to m-1}$$

$$+ \overbrace{r_p m P_A(m, p-1)}^{p-1 \to p} - \overbrace{r_p m P_A(m, p)}^{p \to p+1} + \overbrace{\gamma_p(p+1) P_A(m, p+1)}^{p+1 \to p} - \overbrace{\gamma_p p P_A(m, p)}^{p \to p-1},$$

(3.2)

where overbraces label the state transitions for each term. For the transcriptionally inactive state $I$, we have

$$\frac{dP_I(m, p)}{dt} = \overbrace{k_{\text{off}}^{(p)} P_A(m, p)}^{A \to I} - \overbrace{k_{\text{on}}^{(p)} P_I(m, p)}^{I \to A} + \overbrace{k_{\text{off}}^{(r)} P_R(m, p)}^{R \to I} - \overbrace{k_{\text{on}}^{(r)} P_I(m, p)}^{I \to R}$$

$$+ \overbrace{\gamma_m(m+1) P_I(m+1, p)}^{m+1 \to m} - \overbrace{\gamma_m m P_I(m, p)}^{m \to m-1}$$

$$+ \overbrace{r_p m P_I(m, p-1)}^{p-1 \to p} - \overbrace{r_p m P_I(m, p)}^{p \to p+1} + \overbrace{\gamma_p(p+1) P_I(m, p+1)}^{p+1 \to p} - \overbrace{\gamma_p p P_I(m, p)}^{p \to p-1}.$$

(3.3)

And finally, for the repressor bound state $R$,

$$\frac{dP_R(m, p)}{dt} = -\overbrace{k_{\text{off}}^{(r)} P_R(m, p)}^{R \to I} + \overbrace{k_{\text{on}}^{(r)} P_I(m, p)}^{I \to R}$$

$$+ \overbrace{\gamma_m(m+1) P_R(m+1, p)}^{m+1 \to m} - \overbrace{\gamma_m m P_R(m, p)}^{m \to m-1}$$

$$+ \overbrace{r_p m P_R(m, p-1)}^{p-1 \to p} - \overbrace{r_p m P_R(m, p)}^{p \to p+1} + \overbrace{\gamma_p(p+1) P_R(m, p+1)}^{p+1 \to p} - \overbrace{\gamma_p p P_R(m, p)}^{p \to p-1}.$$

(3.4)

As we will discuss later, the protein degradation term $\gamma_p$ is set to zero since active protein degradation is slow compared to the cell cycle of exponentially growing bacteria, but instead, we explicitly implement binomial partitioning of the proteins into daughter cells upon division [115].

It is convenient to rewrite these equations in a compact matrix notation [26]. For this, we define the vector $\mathbf{P}(m, p)$ as

$$\mathbf{P}(m, p) = (P_A(m, p), P_I(m, p), P_R(m, p))^T, \tag{3.5}$$

where $^T$ is the transpose. By defining the matrices $\mathbf{K}$ to contain the promoter state transitions, $\mathbf{R}_m$ and $\boldsymbol{\Gamma}_m$ to contain the mRNA production and degradation terms, respectively, and $\mathbf{R}_p$ and $\boldsymbol{\Gamma}_p$ to contain the protein production and degradation terms, respectively, the system of ODEs can then be written as (see Sec. 5.2 for the full definition of these matrices)

$$
\begin{aligned}
\frac{d\mathbf{P}(m, p)}{dt} =\ & \left(\mathbf{K} - \mathbf{R}_m - m\boldsymbol{\Gamma}_m - m\mathbf{R}_p - p\boldsymbol{\Gamma}_p\right) \mathbf{P}(m, p) \\
& + \mathbf{R}_m \mathbf{P}(m - 1, p) + (m + 1)\boldsymbol{\Gamma}_m \mathbf{P}(m + 1, p) \\
& + m\mathbf{R}_p \mathbf{P}(m, p - 1) + (p + 1)\boldsymbol{\Gamma}_p \mathbf{P}(m, p + 1).
\end{aligned}
\tag{3.6}
$$

Having defined the gene expression dynamics, we now proceed to determine all rate parameters in Eq. 3.6.

**Inferring Parameters from Published Data Sets**

A decade of research in our group has characterized the simple repression motif with an ever-expanding array of predictions and corresponding experiments to uncover the physics of this genetic circuit [16]. In doing so, we have come to understand the mean response of a single promoter in the presence of varying levels of repressor copy numbers and repressor-DNA affinities [20], due to the effect that competing binding sites and multiple promoter copies impose [39], and in recent work, assisted by the Monod-Wyman-Changeux (MWC) model, we expanded the scope to the allosteric nature of the repressor [113]. All of these studies have exploited the simplicity and predictive power of equilibrium approximations to these non-equilibrium systems [48]. We have also used a similar kinetic model to that

depicted in Fig. 3.2(A) to study the noise in mRNA copy number [84]. Although these studies focus on the same experimental system described by different theoretical frameworks, in earlier work in our laboratory, an attempt to unite parametric knowledge across studies based on equilibrium and non-equilibrium models has not been performed previously. As a test case of the depth of our theoretical understanding of this simple transcriptional regulation system, we combine all of the studies mentioned above to inform the parameter values of the model presented in Fig. 3.2(A). Fig. 3.2(B) schematizes the data sets and experimental techniques used to measure gene expression along with the parameters that can be inferred from them.

Sec. 5.2 expands on the details of how the inference was performed for each of the parameters. Briefly, the promoter activation and inactivation rates $k_{\text{on}}^{(p)}$ and $k_{\text{off}}^{(p)}$, as well as the transcription rate $r_m$ were obtained in units of the mRNA degradation rate $\gamma_m$ by fitting a two-state promoter model (no state $R$ from Fig. 3.2(A)) [116] to mRNA FISH data of an unregulated promoter (no repressor present in the cell) [84]. The repressor on rate is assumed to be of the form $k_{\text{on}}^{(r)} = k_o[R]$ where $k_o$ is a diffusion-limited on rate and $[R]$ is the concentration of active repressor in the cell [84]. This concentration of active repressor is at the same time determined by the repressor copy number in the cell and the fraction of these repressors that are in the active state, i.e., able to bind DNA. Existing estimates of the transition rates between conformations of allosteric molecules set them at the microsecond scale [117]. By considering this to be representative for our repressor of interest, the separation of time-scales between the rapid conformational changes of the repressor and the slower downstream processes such as the open-complex formation processes allow us to model the probability of the repressor being in the active state as an equilibrium MWC process. The parameters of the MWC model $K_A$, $K_I$ and $\Delta\varepsilon_{AI}$ were previously characterized from video-microscopy and flow-cytometry data [113]. For the repressor off rate, $k_{\text{off}}^{(r)}$, we take advantage of the fact that the mean mRNA copy number as derived from the model in Fig. 3.2(A) cast in the language of rates is of the same functional form as the equilibrium model cast in
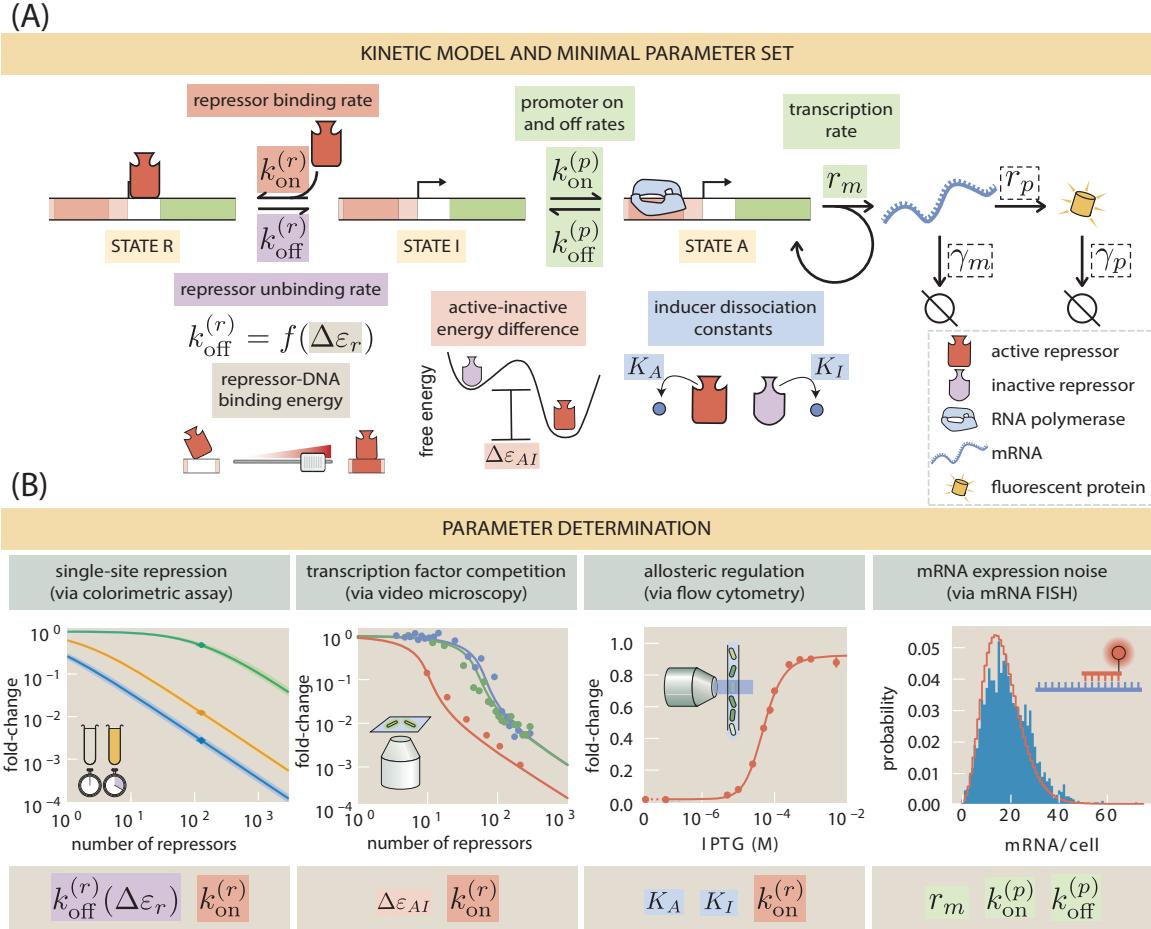
Figure 3.2: **Minimal kinetic model of transcriptional regulation for a simple repression architecture.** (A) Three-state promoter stochastic model of transcriptional regulation by a repressor. The regulation by the repressor occurs via exclusion of the transcription initiation machinery, not allowing the promoter to transition to the transcriptionally active state. All parameters highlighted with colored boxes were determined from published datasets based on the same genetic circuit. Parameters in dashed boxes were taken directly from values reported in the literature or adjusted to satisfy known biological restrictions. (B) Data sets used to infer the parameter values. From left to right, Garcia & Phillips [20] is used to determine $k_{\text{off}}^{(r)}$ and $k_{\text{on}}^{(r)}$, Brewster *et al.* [39] is used to determine $\Delta\varepsilon_{AI}$ and $k_{\text{on}}^{(r)}$, Razo-Mejia *et al.* [113] is used to determine $K_A$, $K_I$, and $k_{\text{on}}^{(r)}$, and Jones *et al.* [84] is used to determine $r_m$, $k_{\text{on}}^{(p)}$, and $k_{\text{off}}^{(p)}$.

the language of binding energies [51]. Therefore the value of the repressor-DNA binding energy $\Delta\varepsilon_r$ constrains the value of the repressor off rate $k_{\text{off}}^{(r)}$. These constraints on the rates allow us to make self-consistent predictions under both the equilibrium and the kinetic framework. Having all parameters in hand, we can now proceed to solve the gene expression dynamics.

**Computing the Moments of the mRNA and Protein Distributions**

Finding analytical solutions to chemical master equations is often fraught with difficulty. An alternative approach is to approximate the distribution. One such scheme of approximation, the maximum entropy principle, uses the distribution moments to approximate the entire distribution. In this section, we will demonstrate an iterative algorithm to compute the mRNA and protein distribution moments.

The kinetic model for the simple repression motif depicted in Fig. 3.2(A) consists of an infinite system of ODEs for each possible pair of mRNA and protein copy number, $(m, p)$. To compute any moment of the distribution, we define a vector

$$\langle \mathbf{m^x p^y} \rangle \equiv (\langle m^x p^y \rangle_A, \langle m^x p^y \rangle_I, \langle m^x p^y \rangle_R)^T, \tag{3.7}$$

where $\langle m^x p^y \rangle_S$ is the expected value of $m^x p^y$ in state $S \in \{A, I, R\}$ for $x, y \in \mathbb{N}$. In other words, just as we defined the vector $\mathbf{P}(m, p)$, here we define a vector to collect the expected value of each of the promoter states. By definition, any of these moments $\langle m^x p^y \rangle_S$ can be computed as

$$\langle m^x p^y \rangle_S \equiv \sum_{m=0}^{\infty} \sum_{p=0}^{\infty} m^x p^y P_S(m, p). \tag{3.8}$$

Summing over all possible values for $m$ and $p$ in Eq. 3.6 results in an ODE for any moment of the distribution of the form (see Sec. 5.3 for full derivation)

$$\frac{d\langle \mathbf{m^x p^y} \rangle}{dt} = \mathbf{K}\langle \mathbf{m^x p^y} \rangle$$
$$+ \mathbf{R}_m \langle \mathbf{p^y} \left[ (\mathbf{m} + 1)^{\mathbf{x}} - \mathbf{m^x} \right] \rangle + \mathbf{\Gamma}_m \langle \mathbf{m p^y} \left[ (\mathbf{m} - 1)^{\mathbf{x}} - \mathbf{m^x} \right] \rangle \tag{3.9}$$
$$+ \mathbf{R}_p \langle \mathbf{m^{(x+1)}} \left[ (\mathbf{p} + 1)^{\mathbf{y}} - \mathbf{p^y} \right] \rangle + \mathbf{\Gamma}_p \langle \mathbf{m^x p} \left[ (\mathbf{p} - 1)^{\mathbf{y}} - \mathbf{p^y} \right] \rangle.$$

Given that all transitions in our stochastic model are first-order reactions, Eq. 3.9 has no moment-closure problem [23]. This means that the dynamical equation for a given moment only depends on lower moments (see Sec. 5.3 for full proof). This feature of our model implies, for example, that the second moment of the protein distribution $\langle p^2 \rangle$ depends only on the first two moments of the mRNA distribution $\langle m \rangle$ and $\langle m^2 \rangle$, the first protein moment $\langle p \rangle$, and the cross-correlation term $\langle mp \rangle$. We can therefore define $\boldsymbol{\mu}^{(\mathbf{x}, \mathbf{y})}$ to be a vector containing all moments up to $\langle \mathbf{m^x p^y} \rangle$

for all promoter states,

$$\boldsymbol{\mu}^{(\mathbf{x,y})} = \left[ \langle \mathbf{m^0 p^0} \rangle, \langle \mathbf{m^1 p^0} \rangle, \ldots, \langle \mathbf{m^x p^y} \rangle \right]^T. \tag{3.10}$$

Explicitly for the three-state promoter model depicted in Fig. 3.2(A), this vector takes the form

$$\boldsymbol{\mu}^{(\mathbf{x,y})} = \left[ \langle m^0 p^0 \rangle_A, \langle m^0 p^0 \rangle_I, \langle m^0 p^0 \rangle_R, \ldots, \langle m^x p^y \rangle_A, \langle m^x p^y \rangle_I, \langle m^x p^y \rangle_R \right]^T. \tag{3.11}$$

Given this definition, we can compute the general moment dynamics as

$$\frac{d\boldsymbol{\mu}^{(\mathbf{x,y})}}{dt} = \mathbf{A}\boldsymbol{\mu}^{(\mathbf{x,y})}, \tag{3.12}$$

where $\mathbf{A}$ is a square matrix that contains all the numerical coefficients that relate each of the moments. We can then use Eq. 3.9 to build matrix $\mathbf{A}$ by iteratively substituting values for the exponents $x$ and $y$ up to a specified value. In the next section, we will use Eq. 3.12 to numerically integrate the dynamical equations for our moments of interest as cells progress through the cell cycle. We will then use the value of the distribution moments to approximate the full gene expression distribution. This method is computationally more efficient than trying to numerically integrate the infinite set of equations describing the full probability distribution $\mathbf{P}(m, p)$, or using a stochastic algorithm to sample from the distribution.

**Accounting for Cell Cycle Dependent Variability in Gene Dosage**

As cells progress through the cell cycle, the genome has to be replicated to guarantee that each daughter cell receives a copy of the genetic material. As replication of the genome can take longer than the total cell cycle, this implies that cells spend part of the cell cycle with multiple copies of each gene depending on the cellular growth rate and the relative position of the gene with respect to the replication origin [17]. Genes closer to the replication origin spend a larger fraction of the cell cycle with multiple copies compared to genes closer to the replication termination site [17]. Fig. 3.3(A) depicts a schematic of this process where the replication origin (*oriC*) and the relevant locus for our experimental measurements (*galK*) are highlighted.

Since this change in gene copy number has been shown to have an effect on cell-to-cell variability in gene expression [84,118], we now extend our minimal model to account for these changes in gene copy number during the cell cycle. We reason that the only difference between the single-copy state and the two-copy state of the promoter is a doubling of the mRNA production rate $r_m$. In particular, the promoter activation and inactivation rates $k_{on}^{(p)}$ and $k_{off}^{(p)}$ and the mRNA production rate $r_m$ assume that cells spend a fraction $f$ of the cell cycle with one copy of the promoter (mRNA production rate $r_m$) and a fraction $(1 - f)$ of the cell cycle with two copies of the promoter (mRNA production rate $2r_m$). This inference was performed considering that at each cell state, the mRNA level immediately reaches the steady-state value for the corresponding mRNA production rate. This assumption is justified since the timescale to reach this steady-state depends only on the degradation rate $\gamma_m$, which for the mRNA is much shorter ($\approx 3$ min) than the length of the cell cycle ($\approx 60$ min for our experimental conditions) [119]. Sec. 5.2 shows that a model accounting for this gene copy number variability can capture data from single-molecule mRNA counts of an unregulated (constitutively expressed) promoter.

Given that the protein degradation rate $\gamma_p$ in our model is set by the cell division time, we do not expect that the protein count will reach the corresponding steady-state value for each stage in the cell cycle. In other words, cells do not spend long enough with two copies of the promoter for the protein level to reach the steady-state value corresponding to a transcription rate of $2r_m$. Therefore, we use the dynamical equations developed to numerically integrate the time trajectory of the moments of the distribution with the corresponding parameters for each phase of the cell cycle. Fig. 3.3(B) shows an example corresponding to the mean mRNA level (upper panel) and the mean protein level (lower panel) for the case of the unregulated promoter. Given that we inferred the promoter rate parameters considering that mRNA reaches steady-state in each stage, we see that the numerical integration of the equations is consistent with the assumption of having the mRNA reach a stable value in each stage (see Fig. 3.3(B) upper panel). On the other hand,

the mean protein level does not reach a steady-state at either of the cellular stages. Nevertheless, it is notable that after several cell cycles, the trajectory from cycle to cycle follows a repetitive pattern (see Fig. 3.3(B) lower panel). Previously, we have experimentally observed this repetitive pattern by tracking the expression level over time with video microscopy, as observed in Fig. 18 of [16].

To test the effects of including this gene copy number variability in our model, we now compare the model's predictions with experimental data. As detailed in the Methods section, we obtained single-cell fluorescence values of different *E. coli* strains carrying a YFP gene under the control of the LacI repressor. Each strain was exposed to twelve different input inducer (IPTG) concentrations for $\approx 8$ generations for cells to adapt to the media. The strains imaged spanned three orders of magnitude in repressor copy number and three distinct repressor-DNA affinities. Since growth was asynchronous, we reason that cells were randomly sampled at all cell cycle stages. Therefore, when computing statistics from the data, such as the mean fluorescence value, we are averaging over the cell cycle. In other words, as depicted in Fig. 3.3(B), quantities such as the mean protein copy number change over time, i.e., $\langle p \rangle \equiv \langle p(t) \rangle$. This means that computing the mean of a population of unsynchronized cells is equivalent to averaging this time-dependent mean protein copy number over the span of the cell cycle. Mathematically, this is expressed as

$$\langle p \rangle_c = \int_{t_o}^{t_d} \langle p(t) \rangle P(t) dt, \tag{3.13}$$

where $\langle p(t) \rangle$ represents the first moment of the protein distribution as computed from Eq. 3.9, $\langle p \rangle_c$ represents the average protein copy number over a cell cycle, $t_o$ represents the start of the cell cycle, $t_d$ represents the time of cell division, and $P(t)$ represents the probability of any cell being at time $t \in [t_o, t_d]$ of their cell cycle. We do not consider cells uniformly distributed along the cell cycle since it is known that cell age is exponentially distributed, having more younger than older cells at any point in time [120] (see Sec. 5.10 for further details). All computations hereafter are therefore done by applying an average like that in for the span of a

cell cycle. We remind the reader that these time averages are done under a fixed environmental state. It is the trajectory of cells over cell cycles under a constant environment that we need to account for. It is through this averaging over the span of a cell cycle that we turn a periodic process like the one shown in Fig. 3.3(B) into a stationary process that we can compare with experimental data and, as we will see later, use to reconstruct the steady-state gene expression distribution.

Fig. 3.3(C) compares zero-parameter fit predictions (lines) with experimentally determined quantities (points). The upper row shows the non-dimensional quantity known as the fold-change in gene expression [20]. This fold-change is defined as the relative mean gene expression level with respect to an unregulated promoter. For protein, this is

$$\text{fold-change} = \frac{\langle p(R > 0) \rangle_c}{\langle p(R = 0) \rangle_c}, \tag{3.14}$$

where $\langle p(R > 0)i \rangle_c$ represents the mean protein count for cells with non-zero repressor copy number count $R$ over the entire cell cycle, and $\langle p(R = 0) \rangle_c$ represents the equivalent for a strain with no repressors present. The experimental points were determined from the YFP fluorescent intensities of cells with varying repressor copy numbers and a $\Delta lacI$ strain with no repressor gene present (see Methods for further details). The fold-change in gene expression has previously served as a metric to test the validity of equilibrium-based models [51]. We note that the curves shown in the upper panel of Fig. 3.3(C) are consistent with the predictions from equilibrium models [113] despite being generated from a non-equilibrium process as shown in Fig. 3.3(B). The kinetic model from Fig. 3.2(A) goes beyond the equilibrium picture to generate predictions for distribution moments other than the mean mRNA or mean protein count. To test this extended predictive power, the lower row of Fig. 3.3(C) shows the noise in gene expression defined as the standard deviation over the mean protein count, accounting for the changes in gene dosage during the cell cycle. Although our model systematically underestimates the noise in gene expression, the zero-parameter fits capture the scaling of this noise. Possible origins of this systematic discrepancy could be

the intrinsic cell-to-cell variability of rate parameters given the variability in the molecular components of the central dogma machinery [84], or noise generated by irreversible non-equilibrium reactions not explicitly taken into account in our minimal model [121]. The large errors for the highly repressed strains (lower left panel in Fig. 3.3(C)) result from having a small number in the denominator—mean fluorescence level—when computing the noise. Although the model is still highly informative about the physical nature of how cells regulate their gene expression, the lack of exact numerical agreement between theory and data opens an opportunity to gain new insights into the biophysical origin of cell-to-cell variability. In Sec. 5.8, we explore empirical ways to account for this systematic deviation. We direct the reader to Sec. 5.4 where equivalent predictions are made, ignoring the changes in gene dosage due to genome replication.

**Maximum Entropy Approximation**

Having numerically computed the moments of the mRNA and protein distributions as cells progress through the cell cycle, we now proceed to make an approximate reconstruction of the full distributions given this limited information. The maximum entropy principle, first proposed by Jaynes in 1957 [28], approximates the entire distribution by maximizing the Shannon entropy subject to constraints given by the values of the moments of the distribution [28]. This procedure leads to a probability distribution of the form (see Sec. 5.6 for full derivation)

$$P(m, p) = \frac{1}{\mathcal{Z}} \exp \left( - \sum_{(x,y)} \lambda_{(x,y)} m^x p^y \right), \tag{3.15}$$

where $\lambda_{(x,y)}$ is the Lagrange multiplier associated with the constraint set by the moment $\langle m^x p^y \rangle$, and $\mathcal{Z}$ is a normalization constant. The more moments $\langle m^x p^y \rangle$ included as constraints, the more accurate the approximation resulting from becomes.

The computational challenge then becomes an optimization routine in which the values for the Lagrange multipliers $\lambda_{(x,y)}$ that are consistent with the constraints set by the moment values $\langle m^x p^y \rangle$ need to be found. This is computationally more
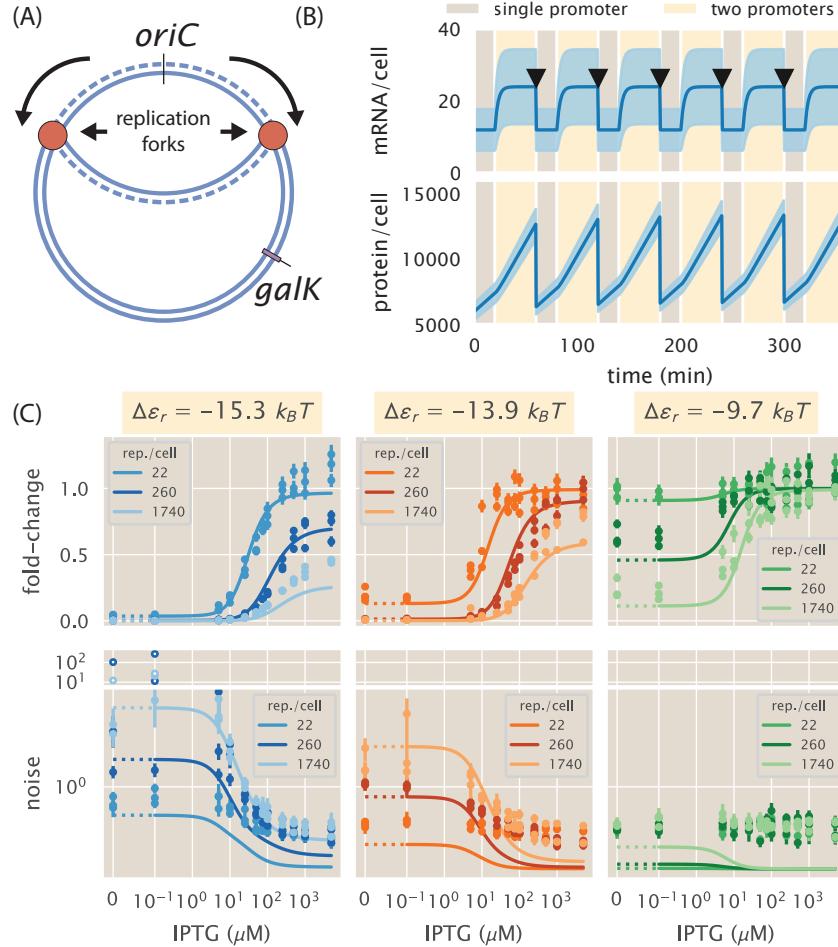
Figure 3.3: **Accounting for gene copy number variability during the cell cycle.** (A) Schematic of a replicating bacterial genome. As cells progress through the cell cycle, the genome is replicated, duplicating gene copies for a fraction of the cell cycle before the cell divides. *oriC* indicates the replication origin, and *galK* indicates the locus at which the YFP reporter construct was integrated. (B) mean (solid line) $\pm$ standard deviation (shaded region) for the mRNA (upper panel) and protein (lower panel) dynamics. Cells spend a fraction of the cell cycle with a single copy of the promoter (light brown) and the rest of the cell cycle with two copies (light yellow). Black arrows indicate the time of cell division. (C) Zero parameter-fit predictions (lines) and experimental data (circles) of the gene expression fold-change (upper row) and noise (lower row) for repressor binding sites with different affinities (different columns) and different repressor copy numbers per cell (different lines on each panel). Error bars in data represent the 95% confidence interval on the quantities as computed from 10,000 bootstrap estimates generated from $> 500$ single-cell fluorescence measurements. In the theory curves, dotted lines indicate plot in linear scale to include zero, while solid lines indicate logarithmic scale. For visual clarity, data points in the noise panel with exceptionally large values coming from highly repressed strains are plotted on a separate panel. The Python code used to generate part (B) (`ch3_fig03B.py`) and part (C) (`ch3_fig03C.py`) of this figure can be found on the original paper's GitHub repository.

efficient than sampling directly from the master equation with a stochastic algorithm (see Sec. 5.6 for further comparison between maximum entropy estimates and the Gillespie algorithm). In Sec. 5.6 we derive our implementation of a robust algorithm to find the values of the Lagrange multipliers. Fig. 3.4(A) shows an example of predicted protein distributions reconstructed using the first six moments of the protein distribution for a suite of different biophysical parameters and environmental inducer concentrations. From the predicted distributions at different inducer concentrations we can see that as repressor-DNA binding affinity (columns in Fig. 3.4(A)) and repressor copy number (rows in Fig. 3.4(A)) are varied, the responses to different signals, i.e., inducer concentrations, overlap to varying degrees. For example, the upper right corner frame with a weak binding site ($\Delta\varepsilon_r = -9.7\ k_B T$) and a low repressor copy number (22 repressors per cell) have virtually identical distributions regardless of the input inducer concentration. This means that cells with this set of parameters cannot resolve any difference in the concentration of the signal. As the number of repressors is increased, the degree of overlap between distributions decreases, allowing cells to resolve the value of the signal input better. On the opposite extreme, the lower-left panel shows a strong binding site ($\Delta\varepsilon_r = -15.3\ k_B T$) and a high repressor copy number (1740 repressors per cell). This parameter combination shows an overlap between distributions since the high degree of repression centers all distributions towards lower copy numbers, giving little ability for the cells to resolve the inputs. In Fig. 3.4(B), we show the comparison of these predicted cumulative distributions with the experimental single-cell fluorescence distributions. Given the systematic deviation of our predictions for the protein copy number noise highlighted in Fig. 3.3(C), the theoretical distributions (dashed lines) underestimate the width of the experimental data. We again direct the reader to Sec. 5.8 for an exploration of empirical changes to the moments that improve the agreement of the predictions. In the following section, we formalize how well cells can resolve different inputs from an information-theoretic perspective via channel capacity.
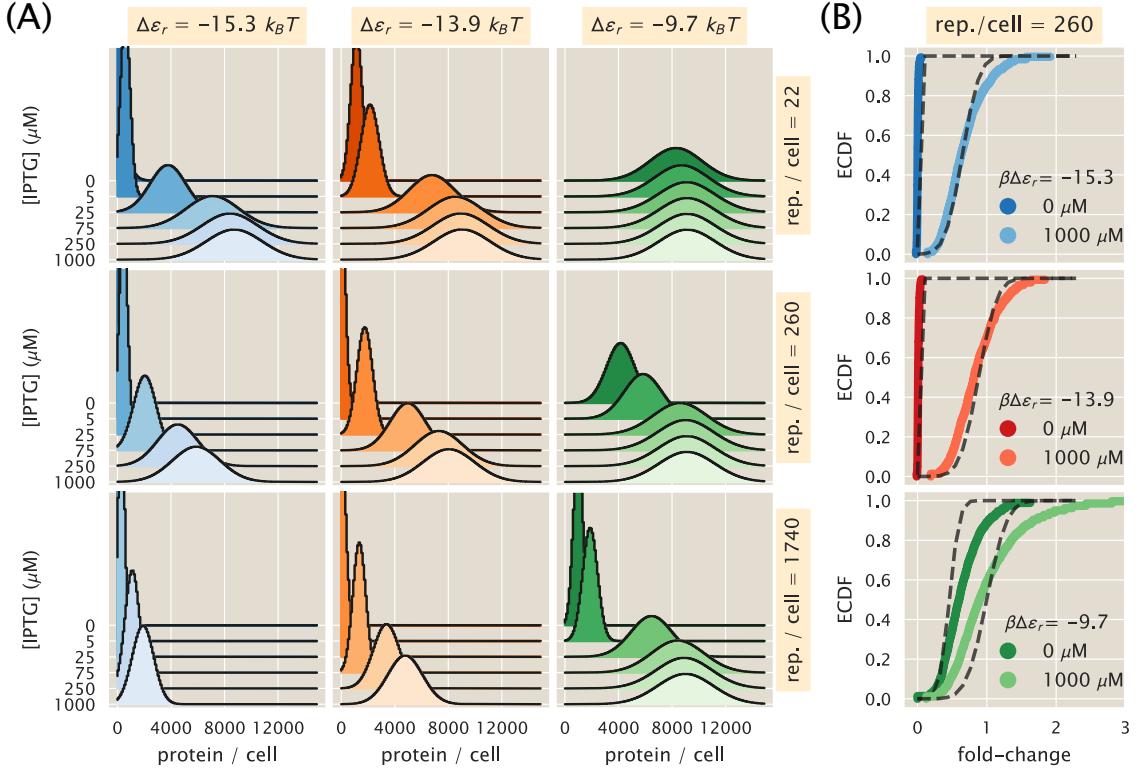
Figure 3.4: **Maximum entropy protein distributions for varying physical parameters.** (A) Predicted protein distributions under different inducer (IPTG) concentrations for different combinations of repressor-DNA affinities (columns) and repressor copy numbers (rows). The first six moments of the protein distribution used to constrain the maximum entropy approximation were computed by integrating as cells progressed through the cell cycle. (B) Theory-experiment comparison of predicted fold-change empirical cumulative distribution functions (ECDF). Each panel shows two example concentrations of inducer (colored curves) with their corresponding theoretical predictions (dashed lines). Distributions were normalized to the mean expression value of the unregulated strain to compare theoretical predictions in discrete protein counts with experimental fluorescent measurements in arbitrary units. The Python code used to generate part (A) (`ch3_fig04A.py`) and part (B) (`ch3_fig04B.py`) of this figure can be found on the original paper's GitHub repository.

## Theoretical Prediction of the Channel Capacity

We now turn our focus to the channel capacity, a metric by which we can quantify the degree to which cells can measure the environmental state (in this context, the inducer concentration). The channel capacity is defined as the mutual information $I$ between input and output (Eq. 3.1), maximized over all possible input (IPTG) distributions $P(c)$. If used as a metric of how reliably a signaling system can infer the state of the external signal, the channel capacity, when measured in bits, is commonly interpreted as the logarithm of the number of states that the signaling

system can adequately resolve. For example, a signaling system with a channel capacity of $C$ bits is interpreted as resolving $2^C$ states, though channel capacities with fractional values are allowed. We, therefore, prefer the Bayesian interpretation that the mutual information quantifies the improvement in the inference of the input when considering the output compared to just using the prior distribution of the input by itself for prediction [23,122]. Under this interpretation, a fractional bit channel capacity still quantifies an improvement in the ability of the signaling system to infer the value of the extracellular signal compared to having no sensing system at all.

Computing the channel capacity implies optimizing over an infinite space of possible distributions $P(c)$. For special cases in which the noise is small compared to the dynamic range, approximate analytical equations have been derived [108]. But given the high cell-to-cell variability that our model predicts, the so-called small noise approximation conditions are not satisfied. We, therefore, appeal to a numerical solution known as the Blahut-Arimoto algorithm [123] (see Sec. 5.7 for further details). Fig. 3.5(A) shows zero-parameter fit predictions of the channel capacity as a function of the number of repressors for different repressor-DNA affinities (solid lines). These predictions are contrasted with experimental determinations of the channel capacity as inferred from single-cell fluorescence intensity distributions taken over 12 different inducer concentrations. Briefly, we can approximate the input-output distribution $P(p \mid c)$ from single-cell fluorescence measurements. Once these conditional distributions are fixed, the task of finding the input distribution at channel capacity becomes a computational optimization routine that can be undertaken using conjugate gradient or similar algorithms. For the particular case of the channel capacity on a system with a discrete number of inputs and outputs, the Blahut-Arimoto algorithm is built to guarantee the convergence towards the optimal input distribution (see Sec. 5.7 for further details). Fig. 3.5(B) shows example input-output functions for different values of the channel capacity. This illustrates that having access to no information (zero channel capacity) is a

consequence of having overlapping input-output functions (lower panel). On the other hand, the more separated the input-output distributions are (upper panel) the higher the channel capacity can be.

All theoretical predictions in Fig. 3.5(A) are systematically above the experimental data. Although our theoretical predictions in Fig. 3.5(A) do not numerically match the experimental inference of the channel capacity, the model captures interesting qualitative features of the data worth highlighting. On one extreme, there is no information processing potential for cells with no transcription factors as this simple genetic circuit would be constitutively expressed regardless of the environmental state. As cells increase the transcription factor copy number, the channel capacity increases until it reaches a maximum before falling back down at a high repressor copy number since the promoter would be permanently repressed. The steepness of the increment in channel capacity and the height of the maximum expression are highly dependent on the repressor-DNA affinity. For strong binding sites (blue curve in Fig. 3.5(A)), there is a rapid increment in the channel capacity, but the maximum value reached is smaller compared to a weaker binding site (orange curve in Fig. 3.5(A)). In Sec 5.8, we show using the small noise approximation [101,108] that if the systematic deviation of our predictions on the cell-to-cell variability was explained with a multiplicative constant, i.e., all noise predictions could be corrected by multiplying them by a single constant, we would expect the channel capacity to be off by a constant additive factor. This factor of $\approx 0.43$ bits can recover the agreement between the model and the experimental data.
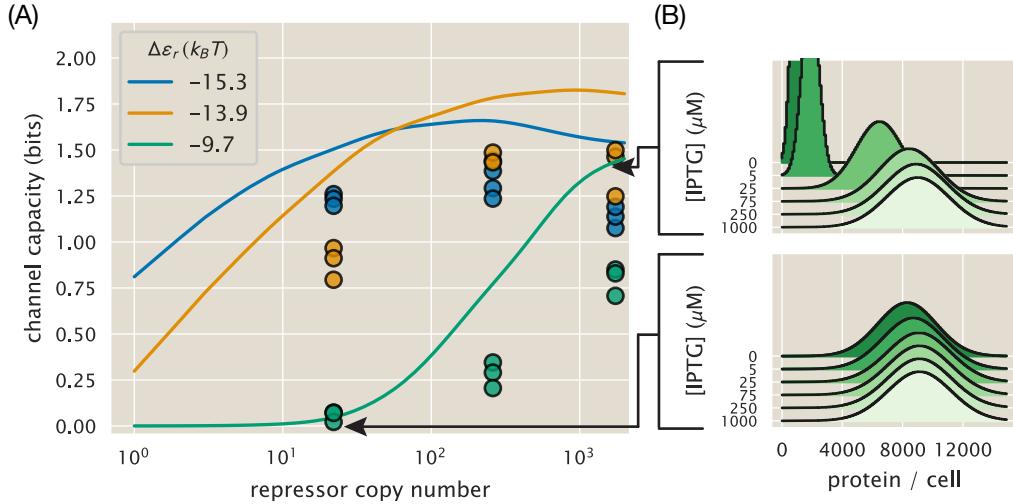
Figure 3.5: **Comparison of theoretical and experimental channel capacity.** (A) Channel capacity as inferred using the Blahut-Arimoto algorithm [123] for varying number of repressors and repressor-DNA affinities. All inferences were performed using 12 IPTG concentrations as detailed in the Methods. Curves represent zero-parameter fit predictions made with the maximum entropy distributions as shown in Fig. 3.4. Points represent inferences made from single-cell fluorescence distributions (see Sec. 5.7 for further details). Theoretical curves were smoothed using a Gaussian kernel to remove numerical precision errors. (B) Example input-output functions in opposite limits of channel capacity. The lower panel illustrates that zero channel capacity indicates that all distributions overlap. The upper panel illustrates that as the channel capacity increases, the separation between distributions increases as well. Arrows point to the corresponding channel capacity computed from the predicted distributions. The Python code used to generate part (A) (`ch3_fig05A.py`) and part (B) (`ch3_fig04B.py`) of this figure can be found on the original paper's GitHub repository.

## 3.4 Discussion

Building on Shannon's formulation of information theory, there have been significant efforts using this theoretical framework to understand the information processing capabilities of biological systems, and the evolutionary consequences for organisms harboring signal transduction systems [6,96,101,124–126]. Recently, with the mechanistic dissection of molecular signaling pathways, significant progress has been made on the question of the physical limits of cellular detection and the role that features such as feedback loops play in this task [23,99,107,127,128]. But the field still lacks a rigorous experimental test of these ideas with precision measurements on a system that is tractable both experimentally and theoretically.

In this chapter, we take advantage of the recent progress on the quantitative modeling of input-output functions of genetic circuits to build a minimal model of the simple repression motif [16]. By combining a series of studies on this circuit span-

ning diverse experimental methods for measuring gene expression under a myriad of different conditions, for the first time, we possess complete *a priori* parametric knowledge—allowing us to generate parameter-free predictions for processes related to information processing. Some of the model parameters for our kinetic formulation of the input-output function are informed by inferences made from equilibrium models. We use the fact that if both kinetic and thermodynamic languages describe the same system, the predictions must be self-consistent. In other words, if the equilibrium model can only make statements about the mean mRNA and mean protein copy number because of the way these models are constructed, those predictions must be equivalent to what the kinetic model has to say about these same quantities. This condition, therefore, constrains the values that the kinetic rates in the model can take. To test whether or not the equilibrium picture can reproduce the predictions made by the kinetic model, we compare the experimental and theoretical fold-change in protein copy number for a suite of biophysical parameters and environmental conditions (Fig. 3.3(C) upper row). The agreement between theory and experiment demonstrates that these two frameworks can indeed make consistent predictions.

The kinetic treatment of the system brings with it increasing predictive power compared to the equilibrium picture. Under the kinetic formulation, the predictions are not limited only to the mean but any of the moments of the mRNA and protein distributions. Furthermore, our formulation in terms of dynamical equations allows us to account for the time-varying nature of the moments of the mRNA and protein copy numbers. Specifically, since the protein mean lifetime is comparable with the cell cycle length, the protein copy number does not reach a steady-state over the cell cycle duration. Accounting for this effect increases the expected cell-to-cell variability when measuring non-synchronized cells. We first test these novel predictions by comparing the noise in protein copy number (standard deviation/mean) with experimental data. Our minimal model predicts the noise up to a systematic deviation. The physical or biological origins of this discrepancy remain an open question. In that way, the work presented here exposes the status

quo of our understanding of gene regulation in bacteria, posing new questions to be answered with future model refinements. We then extend our analysis to infer entire protein distributions at different input signal concentrations using the maximum entropy principle. This means that we compute moments of the protein distribution and then use these moments to build an approximation to the full distribution. These predicted distributions are then compared with experimental single-cell distributions, as shown in Fig. 3.4(B) and Sec. 5.5. Again, although our minimal model systematically underestimates the width of the distributions, it informs how changes in parameters such as protein copy number or protein-DNA binding affinity will affect the full probabilistic input-output function of the genetic circuit to a multiplicative constant. We then use our model to predict the information processing capacity.

By maximizing the mutual information between input signal concentration and output protein distribution over all possible input distributions, we predict the channel capacity of the system over a suite of biophysical parameters such as varying repressor protein copy number and repressor-DNA binding affinity. Although there is no reason to assume the simplified synthetic circuit we used as an experimental model operates optimally given the distribution of inputs, the relevance of the channel capacity comes from its interpretation as a metric of the physical limit of how precise of an inference cells can make about what the state of the environment is. Our model, despite the systematic deviations, makes non-trivial predictions such as the existence of an optimal repressor copy number for a given repressor-DNA binding energy, predicting the channel capacity up to an additive constant (see Fig. 3.5). The origin of this optimal combination of repressor copy number and binding energy differs from previous publications in which an extra term associated with the cost of producing protein was included in the model [107]. This optimal parameter combination is a direct consequence of the fact that the LacI repressor cannot be fully deactivated [113]. This implies that as the number of repressors increases, a significant number of them are still able to bind to the promoter even at saturating concentrations of inducer. This causes all of the

input-output functions to be shifted towards low expression levels, regardless of the inducer concentration, decreasing the amount of information that the circuit can process. Interestingly, the number of bits predicted and measured in our system is similar to that of the gap genes in the *Drosophila* embryo [102]. Although this is a suggestive numerical correspondence that sets current experimental data on the information processing capacity of genetic circuits between 1 and 2 bits, more work is required to fully understand the effect that different regulatory architectures have on the ability to resolve different signals.

We consider it important to highlight the limitations of the work presented here. The previously discussed systematic deviation for the noise and skewness of the predicted distributions (see Sec. 5.8), and therefore of the predicted distributions and channel capacity, remains an unresolved question. Our current best hypothesis for the origin of this unaccounted noise pertains to cell-to-cell variability in the central dogma machinery. More specifically, our model does not account for changes in RNAP and sigma factor copy numbers, changes in ribosome numbers, and even the variability in the repressor copy number. This possibility deserves to be addressed in further iterations of our minimal model. Also, as first reported in [113], our model fails to capture the steepness of the fold-change induction curve for the weakest repressor binding site (see Fig. 3.3(B)). Furthermore, the minimal model in (A), despite being widely used, is an oversimplification of the physical picture of how the transcriptional machinery works. The coarse-graining of all the kinetic steps involved in transcription initiation into two effective promoter states—active and inactive—ignores potential kinetic regulatory mechanisms of intermediate states [129]. Moreover, it has been argued that even though the mRNA count distribution does not follow a Poisson distribution, this effect could be caused by unknown factors, not at the level of transcriptional regulation [130].

The findings of this work open the opportunity to accurately test intriguing ideas that connect Shannon's metric of how accurately a signaling system can infer the state of the environment with Darwinian fitness [6]. Beautiful work along these

lines has been done in the context of the developmental program of the early *Drosophila* embryo [101,103]. These studies demonstrated that the input-output function of the pair-rule genes works at channel capacity, suggesting that selection has acted on these signaling pathways, pushing them to operate at the limit of what the physics of these systems allow. Our system differs from the early embryo because we have a tunable circuit with variable amounts of information processing capabilities. Furthermore, compared with the fly embryo in which the organism tunes both the input and output distributions over evolutionary time, we have experimental control of the distribution of inputs that the cells are exposed to. Consequently, this means that instead of seeing the final result of the evolutionary process, we would be able to set different environmental challenges and track over time the evolution of the population. These experiments could shed light on the suggestive hypothesis of information bits as a trait on which natural selection acts. We see this exciting direction as part of the overall effort in quantitative biology of predicting evolution [131].

## 3.5 Materials and Methods

### *E. coli* Strains

All strains used in this study were originally made for [113]. We chose a subset of three repressor copy numbers that span two orders of magnitude. We refer the reader to [113] for details on the construction of these strains. Briefly, the strains have a construct consisting of the *lacUV5* promoter and one of three possible binding sites for the *lac* repressor (O1, O2, and O3) controlling the expression of a YFP reporter gene. This construct is integrated into the genome at the *galK* locus. The number of repressors per cell is varied by changing the ribosomal binding site controlling the translation of the *lac* repressor gene. The repressor constructs were integrated in the *ybcN* locus. Finally, all strains used in this work constitutively express an mCherry reporter from a low copy number plasmid. This serves as a volume marker that facilitates the segmentation of cells when processing microscopy images.

## Growth Conditions

For all experiments, cultures were initiated from a 50% glycerol frozen stock at $-80°$C. Three strains—autofluorescence (*auto*), $\Delta lacI$ ($\Delta$), and a strain with a known binding site and repressor copy number (*R*)—were inoculated into individual tubes with 2 mL of Lysogeny Broth (LB Miller Powder, BD Medical) with 20 $\mu$g/mL of chloramphenicol and 30 $\mu$g/mL of kanamycin. These cultures were grown overnight at 37°C with rapid agitation to reach saturation. The saturated cultures were diluted 1:1000 into 500 $\mu$L of M9 minimal media (M9 5X Salts, Sigma-Aldrich M6030; 2 mM magnesium sulfate, Mallinckrodt Chemicals 6066-04; 100 mM calcium chloride, Fisher Chemicals C79-500) supplemented with 0.5% (w/v) glucose on a 2 mL 96-deep-well plate. The *R* strain was diluted into 12 different wells with minimal media, each with a different IPTG concentration (0 $\mu$M, 0.1 $\mu$M, 5 $\mu$M, 10 $\mu$M, 25 $\mu$M, 50 $\mu$M, 75 $\mu$M, 100 $\mu$M, 250 $\mu$M, 500 $\mu M$, 1000 $\mu$M, 5000 $\mu$M) while the *auto* and $\Delta$ strains were diluted into two wells (0 $\mu$M, 5000 $\mu$M). Each of the IPTG concentrations came from a single preparation stock kept in 100-fold concentrated aliquots. The 96-well plate was then incubated at 37°C with rapid agitation for 8 hours before imaging.

## Microscopy Imaging Procedure

The microscopy pipeline used for this work exactly followed the steps from [113]. Briefly, twelve 2% agarose (Life Technologies UltraPure Agarose, Cat.No. 16500100) gels were made out of M9 media (or PBS buffer) with the corresponding IPTG concentration (see growth conditions), and placed between two glass coverslips for them to solidify after microwaving. After the 8 hour incubation in minimal media, 1 $\mu$L of a 1:10 dilution of the cultures into fresh media or PBS buffer was placed into small squares (roughly 10 mm $\times$ 10 mm) of the different agarose gels. A total of 16 agarose squares—12 concentrations of IPTG for the *R* strain, 2 concentrations for the $\Delta$ and 2 for the *auto* strain—were mounted into a single glass-bottom dish (Ted Pella Wilco Dish, Cat. No. 14027-20) that was sealed with parafilm.

All imaging was done on an inverted fluorescent microscope (Nikon Ti-Eclipse) with a custom-built laser illumination system. The YFP fluorescence (quantitative reporter) was imaged with a CrystaLaser 514 nm excitation laser coupled with a laser-optimized (Semrock Cat. No. LF514-C-000) emission filter. All strains, including the *auto* strain, included a constitutively expressed mCherry protein to aid the segmentation. Therefore, for each image, three channels (YFP, mCherry, and bright field) on average 30 images with roughly 20 cells per condition were taken. Twenty-five images of a fluorescent slide and 25 images of the camera background noise were taken every imaging session to flatten the illumination. The image processing pipeline for this work is the same as in [113].

**Data and Code Availability**

All data and custom scripts were collected and stored using Git version control. Code for raw data processing, theoretical analysis, and figure generation is available on the GitHub repository (https://github.com/RPGroup-PBoC/chann_cap). The code can also be accessed via the paper website (https://www.rpgroup.caltech.edu/chann_cap/). Raw microscopy data are stored on the CaltechDATA data repository and can be accessed via DOI https://doi.org/10.22002/d1.1184. Bootstrap estimates of experimental channel capacity are also available on the CaltechDATA data repository via https://doi.org/10.22002/D1.1185.

*Chapter 4*

# SUPPORTING INFORMATION FOR TUNING TRANSCRIPTIONAL REGULATION THROUGH SIGNALING: A PREDICTIVE THEORY OF ALLOSTERIC INDUCTION

## 4.1 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 which enables us to collapse our experimental data onto a single master curve that captures the diverse phenomenology of the induction profiles.

## 4.2 Inferring Allosteric Parameters from Previous Data

The fold-change profile described by Eq. 2.5 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 [39] 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 shall show that if we try to fit the data into 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 determine the actual physical parameter values of the system uniquely. This problem persists even when fitting multiple data sets simultaneously, as we will see later.

In Fig. 4.1(A), 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 Eq. 2.5 to account for the underlying assumptions used when fitting previous data (see Section 3.2 for a full explanation of why this modification is needed).
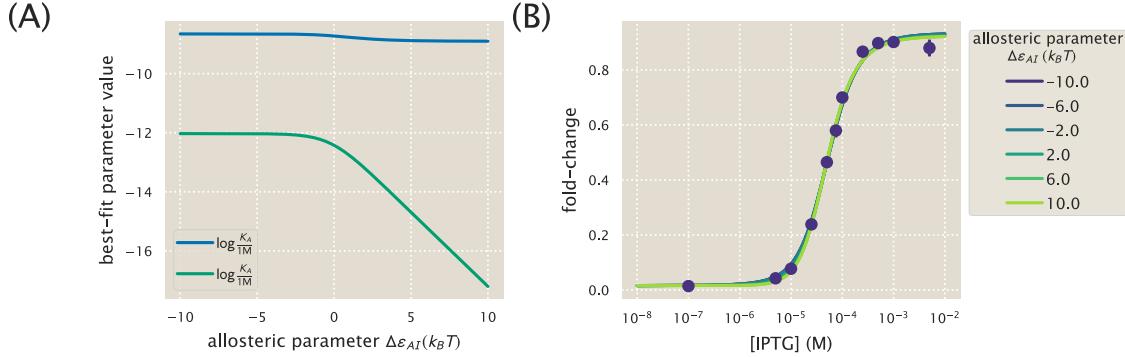
Figure 4.1: **Multiple sets of parameters yield identical fold-change responses.** (A) The data for the O2 strain ($\Delta\varepsilon_{RA} = -13.9\,k_BT$) with $R = 260$ in Fig. 2.4(D) was fit using Eq. 4.5 with $n = 2$. $\Delta\varepsilon_{AI}$ is forced to take on the value shown on the $x$-axis, while the $K_A$ and $K_I$ parameters are fit freely. (B) The resulting best-fit functions for several value of $\Delta\varepsilon_{AI}$ all yield nearly identical fold-change responses.

The best-fit curves for several different values of $\Delta\varepsilon_{AI}$ are shown in Fig. 4.1(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 Fig. 4.1(A) 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 [132,133].

**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 to determine the value of $\Delta\varepsilon_{AI}$ in Eq. 2.5 for the induction of the Lac repressor. Specifically, we consider two previous sets of work from (1) [20] and (2) [39], 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 DNA binding energy $\Delta\varepsilon_{RA}$ and the allosteric energy difference $\Delta\varepsilon_{AI}$ between the Lac repressor's active and inactive states. We refer to this convoluted energy value as $\Delta\tilde{\varepsilon}_{RA}$. We first disentangle the relationship between these parameters in Garcia and Phillips, and then use this relationship to extract the value of $\Delta\varepsilon_{AI}$ from the Brewster *et al.* dataset.

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 $\Delta\tilde{\varepsilon}_{RA}$ for each operator,

$$\text{fold-change}(c=0) = \left(1 + \frac{R}{N_{NS}} e^{-\beta\Delta\tilde{\varepsilon}_{RA}}\right)^{-1}. \tag{4.1}$$

Note that this functional form does not exactly match our fold-change in the limit $c=0$,

$$\text{fold-change}(c=0) = \left(1 + \frac{1}{1+e^{-\beta\Delta\varepsilon_{AI}}} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1}, \tag{4.2}$$

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,

$$\frac{1}{1+e^{-\beta\Delta\varepsilon_{AI}}} = p_A(0). \tag{4.3}$$

In other words, Garcia and Phillips 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, $\Delta\tilde{\varepsilon}_{RA} = -15.3\ k_BT$ for O1 and $\Delta\tilde{\varepsilon}_{RA} = -17.0\ k_BT$ for Oid) represent

$$\beta\Delta\tilde{\varepsilon}_{RA} = \beta\Delta\varepsilon_{RA} - \log\left(\frac{1}{1+e^{-\beta\Delta\varepsilon_{AI}}}\right). \tag{4.4}$$

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. 4.2 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 [20] 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 $\Delta\varepsilon_{AI}$. 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 a 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 [40], writing the fold-change in terms of the grand canonical ensemble as

$$\text{fold-change} = \frac{1}{1 + \lambda_r e^{-\beta\Delta\varepsilon_{RA}}}, \tag{4.5}$$

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

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

$$R_{\text{tot}} = R_S + R_{NS}, \tag{4.6}$$

where $R_S$ represents the number of repressors specifically bound to the promoter and $R_{NS}$ represents the number of repressors non-specifically bound throughout

the genome. The value of $R_S$ is given by

$$R_S = N \frac{\lambda_r e^{-\beta \Delta \varepsilon_{RA}}}{1 + \lambda_r e^{-\beta \Delta \varepsilon_{RA}}}, \tag{4.7}$$

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 [40]. The value of $R_{NS}$ is similarly give by

$$R_{NS} = N_{NS} \frac{\lambda_r}{1 + \lambda_r}, \tag{4.8}$$

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 Eq. 4.7 and 4.8 in Eq. 4.6 into the modified $R_A$ yields the form

$$p_A(0) R_{\text{tot}} = \frac{1}{1 + e^{-\beta \Delta \varepsilon_{AI}}} \left( N \frac{\lambda_r e^{-\beta \Delta \varepsilon_{RA}}}{1 + \lambda_r e^{-\beta \Delta \varepsilon_{RA}}} + N_{NS} \frac{\lambda_r}{1 + \lambda_r} \right), \tag{4.9}$$

where we recall from Eq. 4.4 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 Eq. 4.5 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 $\Delta \varepsilon_{AI}$ parameter. Fig. 4.5(A) shows how different values of $\Delta \varepsilon_{AI}$ 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 $\Delta \varepsilon_{AI}$. Interestingly, the inflection point of Eq. 4.9 occurs near $p_A(0) R_{\text{tot}} = N$ (as shown by the triangles in Fig. 4.5(A)), so that merely knowing where the fold-change response transitions from concave down to concave up is sufficient to obtain a rough value for $\Delta \varepsilon_{AI}$. We note, however, that for $\Delta \varepsilon_{AI} \geq 5 \, k_B T$, increasing $\Delta \varepsilon_{AI}$ further does not affect the fold-change because essentially every repressors will be in the active state in this regime. Thus, if the $\Delta \varepsilon_{AI}$ is in this regime, we can only bound it from below.
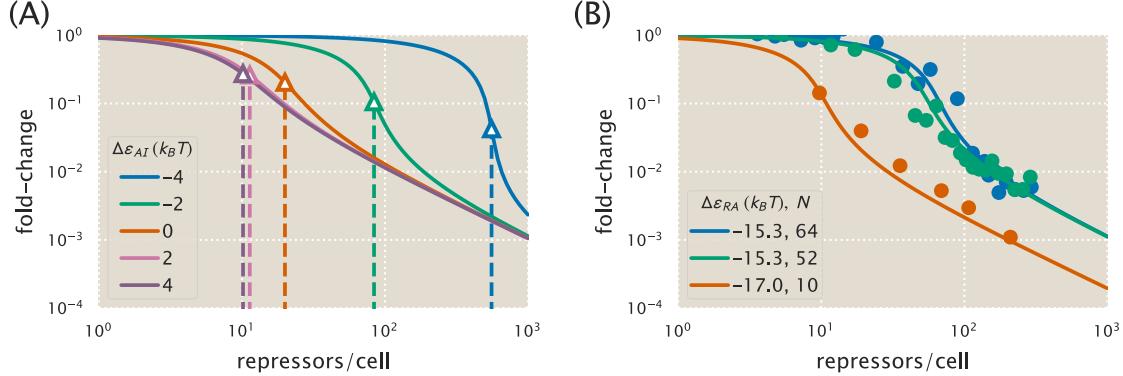
Figure 4.2: **Fold-change of multiple identical genes.** (A) In the presence of $N = 10$ identical promoters, the fold-change Eq. 4.6 depends strongly on the allosteric energy difference $\Delta\varepsilon_{AI}$ between the Lac repressor's active and inactive states. The vertical dotted lines represent the number of repressors at which $R_A = N$ for each value of $\Delta\varepsilon_{AI}$. (B) Using fold-change measurements from [39] for the operators and gene copy numbers shown, we can determine the most likely value $\Delta\varepsilon_{AI} = 4.5\,k_BT$ for LacI.

We now analyze experimental induction data for different strains with known plasmid copy numbers to determine $\Delta\varepsilon_{AI}$. Fig. 4.5(B) 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 [39]. By fitting these data to Eq. 4.5, we extracted the parameter value $\Delta\varepsilon_{AI} = 4.5\,k_BT$. Substituting this value into Eq. 4.3 shows that 99% of the repressors are in the active state in the absence of inducer and $\Delta\tilde{\varepsilon}_{RA} \approx \Delta\varepsilon_{RA}$ so that all of the previous energies and calculations made by [20,39] were accurate.

## 4.3 Induction 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 [34,36,52,57,59,62,65,134], in which reporter constructs are placed on a 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. Indeed, we can account for systems in which multiple promoters compete for the repressor of interest. Additionally, we demonstrate the importance of precise control over these parameters, as they can significantly affect the induction profile.

**Chemical Potential Formulation to Calculate Fold-Change**

This section discusses a simple repression construct that we generalize in two ways from the scenario discussed in the text. First, we will allow the repressor to bind to $N_S$ identical specific promoters whose fold-change we are interested in measuring. Each promoter contains 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. We can write the fold-change in the grand canonical ensemble as

$$\text{fold-change} = \frac{1}{1 + \lambda_r e^{-\beta \Delta \varepsilon_{RA}}}, \tag{4.10}$$

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, $R_{tot}$, so that it is given by

$$R_{tot} = R_S + R_{NS} + R_C, \tag{4.11}$$

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

$$R_S = N_S \frac{\lambda_r e^{-\beta \Delta \varepsilon_{RA}}}{1 + \lambda_r e^{-\beta \Delta \varepsilon_{RA}}}, \tag{4.12}$$

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

$$R_{NS} = N_{NS} \frac{\lambda_r}{1 + \lambda_r}, \tag{4.13}$$

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

$$R_C = N_C \frac{\lambda_r e^{-\beta \Delta \varepsilon_C}}{1 + \lambda_r e^{-\beta \Delta \varepsilon_C}}, \tag{4.14}$$

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 $R_{tot}$ in Eq. 4.11 by the number of active repressors in the cell, $p_A(c) R_{tot}$. Here, $p_A$ denotes the probability that the repressor is in the active state (Eq. 4.13),

$$p_A(c) = \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{-\beta \Delta \varepsilon_{AI}} \left(1 + \frac{c}{K_I}\right)^n}. \tag{4.15}$$

Substituting Eq. 4.15 in Eqs. 4.12-4.14 into the modified Eq. 4.11 yields the form

$$p_A(c) R_{tot} = N_S \frac{\lambda_r e^{-\beta \Delta \varepsilon_{RA}}}{1 + \lambda_r e^{-\beta \Delta \varepsilon_{RA}}} + N_{NS} \frac{\lambda_r}{1 + \lambda_r} + N_C \frac{\lambda_r e^{-\beta \Delta \varepsilon_C}}{1 + \lambda_r e^{-\beta \Delta \varepsilon_C}}. \tag{4.16}$$

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 4.10 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. 4.16 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.

**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 $N_S = 1$ (see Fig. 2.5). Here we predict the induction profiles for similar strains in which $R$ is varied, but $N_S > 1$, as shown in Fig. 4.3. 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 $N_S$.
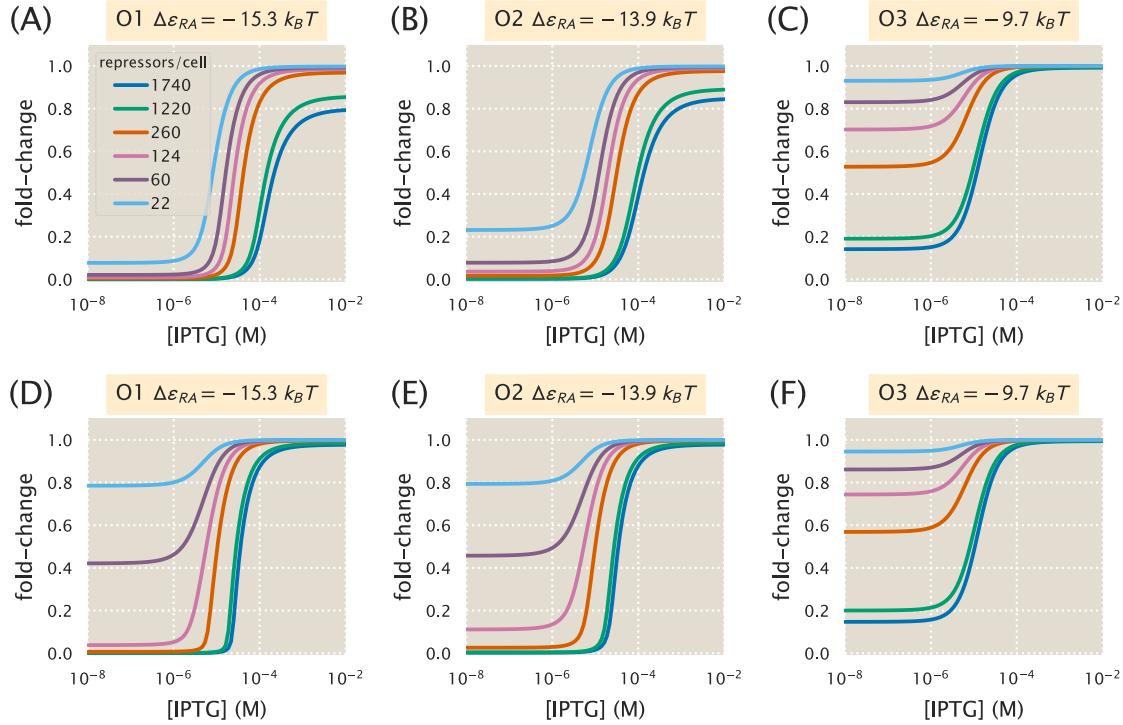
Figure 4.3: **Induction with variable $R$ and multiple specific binding sites.** Induction profiles are shown for strains with variable R and $\Delta\varepsilon_{RA} = -15.3, -13.9$, or $-9.7\ k_B T$. The number of specific sites, $N_S$, is held constant at ten as $R$ and $\Delta\varepsilon_{RA}$ are varied. $N_S$ is held constant at 100 as $R$ and $\Delta\varepsilon_{RA}$ are varied. These situations mimic the common scenario in which a promoter construct is placed on either a low or high copy number plasmid.

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

The second set of scenarios we consider is when the repressor copy number $R = 260$ is held constant while the number of specific promoters $N_S$ is varied (see Fig. 4.4). 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.
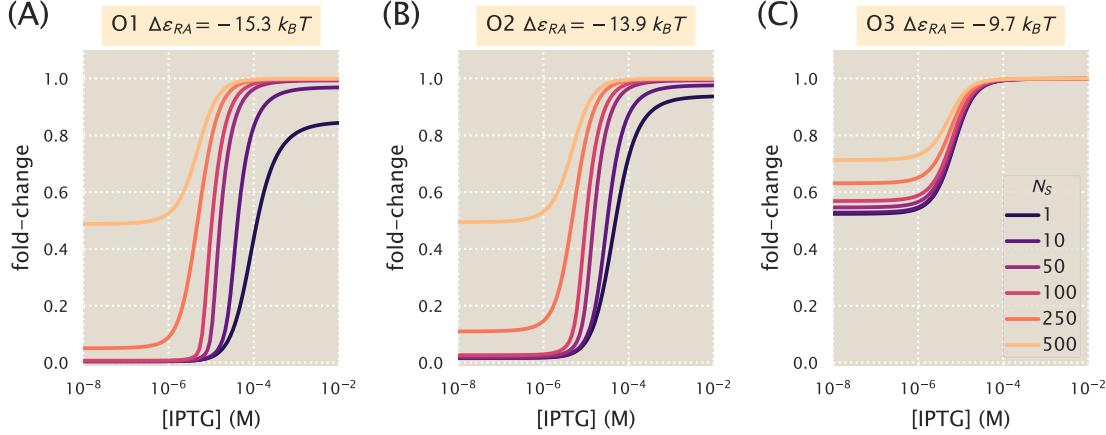
Figure 4.4: **Induction with variable specific sites and fixed R.** Induction profiles are shown for strains with $R = 260$ and $\Delta\varepsilon_{RA} = -15.3\ k_BT$, $\Delta\varepsilon_{RA} = -13.9\ k_BT$, or $\Delta\varepsilon_{RA} = -9.7\ k_BT$. The number of specific sites $N_S$ is varied from 1 to 500.

## 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 [114,135]. If the number of competitor sites and their average binding energy is known, 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 a variable number of competitor sites $N_C$ with strong binding energy $\Delta\varepsilon_C = -17.0\ k_BT$. 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 $\Delta\varepsilon_{RA} = -9.7\ k_BT$, 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.

## 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
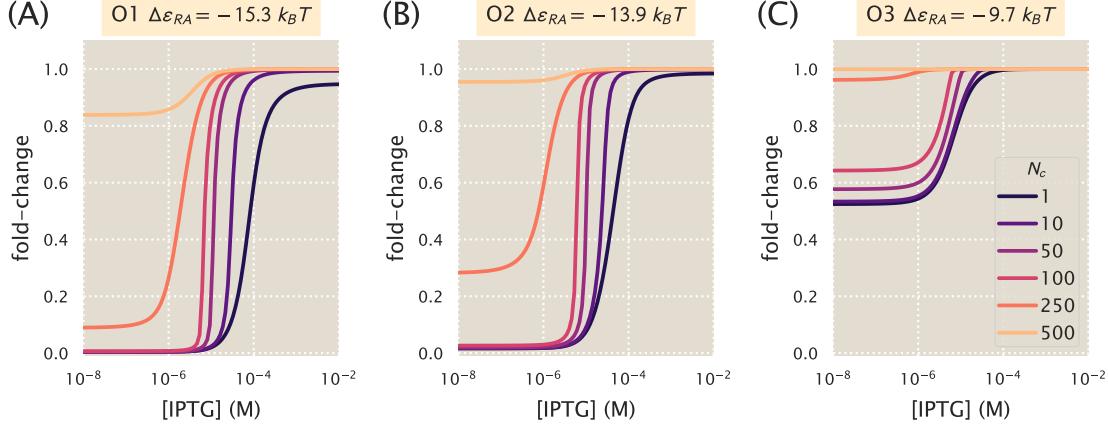
Figure 4.5: **Induction with variable competitor sites, a single specific site, and fixed $R$.** Induction profiles are shown for strains with $R = 260$, $N_s = 1$, and $\Delta\varepsilon_{RA} = -15.3\ k_B T$ for the O1 operator, $\Delta\varepsilon_{RA} = -13.9\ k_B T$ for the O2 operator, or $\Delta\varepsilon_{RA} = -9.7\ k_B T$ 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.

leakiness, saturation, and dynamic range (see Fig. 2.1) by numerically solving Eq. 4.16 in the absence of inducer, $c = 0$, and in the presence of saturating inducer $c \to \infty$. Using Eq. 4.15, the former case is given by

$$R_{\text{tot}} \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} = N_S \frac{\lambda_r e^{-\beta\Delta\varepsilon_{RA}}}{1 + \lambda_r e^{-\beta\Delta\varepsilon_{RA}}} + N_{NS} \frac{\lambda_r}{1 + \lambda_r} + N_C \frac{\lambda_r e^{-\beta\Delta\varepsilon_C}}{1 + \lambda_r e^{-\beta\Delta\varepsilon_C}}, \qquad (4.17)$$

whereupon substituting in the value of $\lambda_r$ into Eq. 4.10 will yield the leakiness. Similarly, the limit of saturating inducer is found by determining $\lambda_r$ from the form

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

$$(4.18)$$
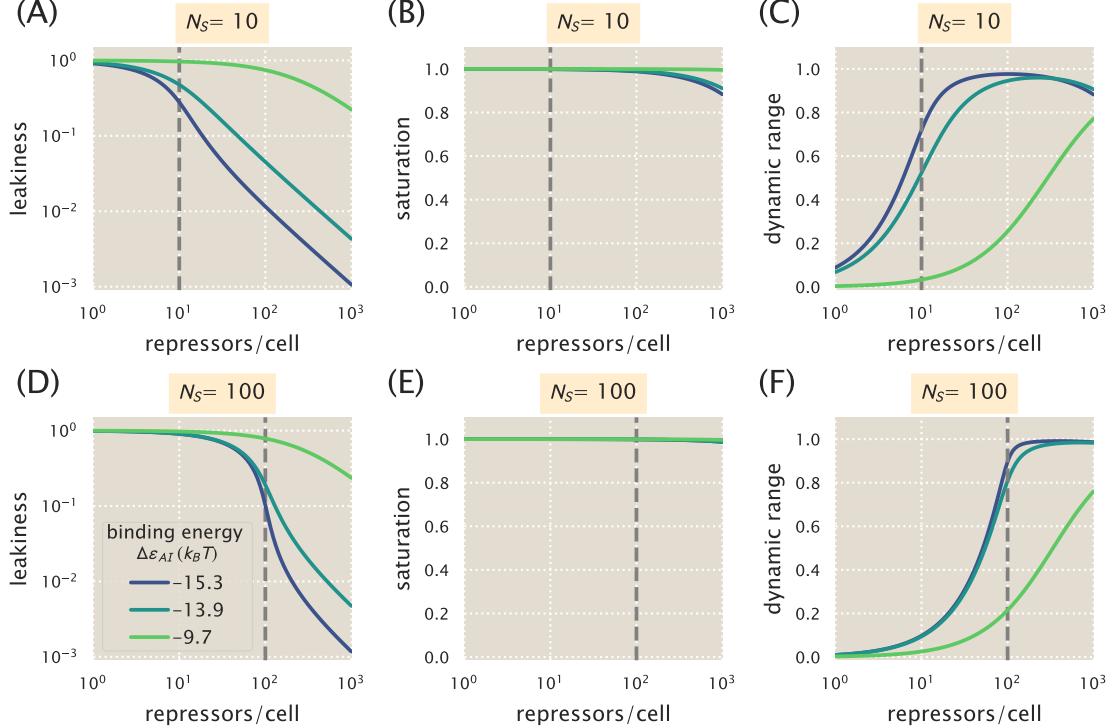
Figure 4.6: **Phenotypic properties of induction with multiple specific binding sites.** The leakiness (A, D), saturation (B, E), and dynamic range (C, F) are shown for systems with a number of specific binding sites $N_S = 10$ or $N_S = 100$. The dashed vertical line indicates the point at which $N_S = R$.

In Fig. 4.6, we show how the leakiness, saturation, and dynamic range vary with $R$ and $\Delta\varepsilon_{RA}$ 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 $\Delta\varepsilon_{RA} = -15.3$ and $-13.9\ k_B T$ 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.
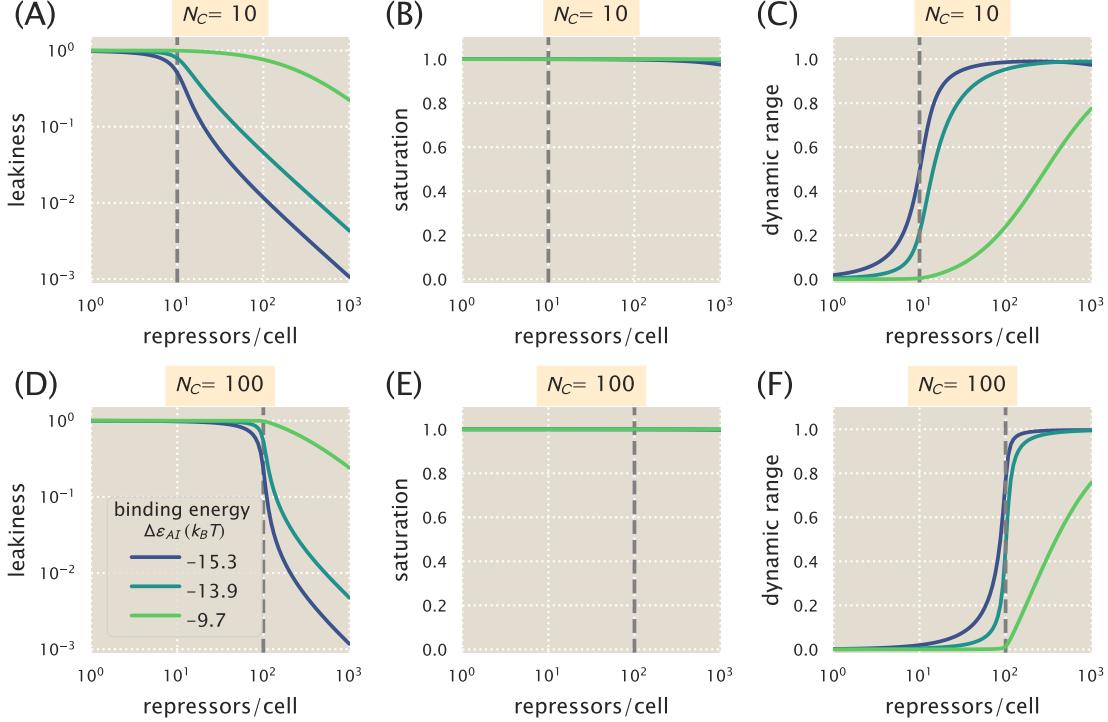
Figure 4.7: **Phenotypic properties of induction with a single specific site and multiple competitor sites.** The leakiness, saturation, and dynamic range are shown for systems with a single specific binding site $N_S = 1$ and a number of competitor sites $N_C = 10$ or $N_C = 100$. All competitor sites have a binding energy of $\Delta\varepsilon_C = -17.0\ k_BT$. The dashed vertical line indicates the point at which $N_C = R$.

In Fig. 4.7, we similarly show how the leakiness, saturation, and dynamic range vary with $R$ and $\Delta\varepsilon_{RA}$ in systems with $N_S = 1$ and multiple competitor sites $N_C = 10$ or $N_C = 100$. Each of the competitor sites has a binding energy of $\Delta\varepsilon_C = -17.0\ k_BT$. The phenotypic profiles are very similar to those for multiple specific sites shown in Fig. 4.7, 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. In either case, the influence of the repressors is dampened when there are insufficient repressors to interact with all of the specific binding sites.

This section of the appendix 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.

## 4.4  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 provided 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 [136], which was sufficient for 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. The 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 4.1.

Table 4.1: **Instrument settings for data collection using the Miltenyi Biotec MACSQuant flow cytometer.** All experimental measurements were collected using these values.

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

**Experimental Measurement**

Before 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. A single data set consisted of seven bacterial strains, all sharing the same operator, with varying repressor copy numbers ($R = 0, 22, 60, 124, 260, 1220$, and 1740), in addition to an autofluorescent strain, under twelve IPTG concentrations. Data collection took place over two to three hours. During this time, the cultures were held at approximately $4°C$ by placing the 96-well plate on a MACSQuant ice block. Because the ice block thawed over the course of the experiment, the samples measured last were approximately at room temperature. This means that samples may have grown slightly by the end of the experiment. To confirm that this continued growth did not alter the measured results, a subset of experiments were run in reverse, meaning that the fully induced cultures were measured first and the uninduced samples last. The plate arrangements and corresponding fold-change measurements are shown in Fig. 4.8(A) and (B), respectively. The measured fold-change values in the reverse ordered plate appear to be drawn from
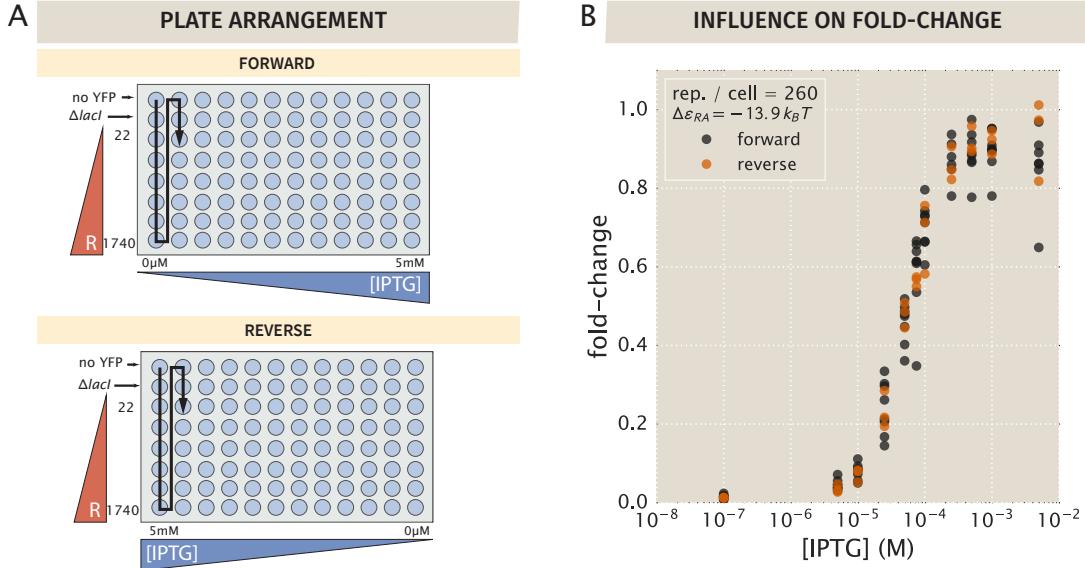
Figure 4.8: **Plate arrangements for flow cytometry.** (A) Samples were measured primarily in the forward arrangement with a subset of samples measured in reverse. The black arrow indicates the order in which samples were processed by the flow cytometer. (B) The experimentally measured fold-change values for the two sets of plate arrangements show that samples measured in the forward arrangement appear to be indistinguishable from those measured in reverse order.

the same distribution as those measured in the forward order, meaning that any growth that might have occurred during the experiment did not significantly affect the results. Both the forward and reverse data sets were used in our analysis.

**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 determined to be "real" is commonly referred to as gating. These gates are typically drawn manually [93] 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 [94,95].

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 the highest density of points in these two channels corresponds to single-cell measurements. Therefore, everything extending outside of this region was discarded to exclude sources of error such as cell clustering, particulates, or other spurious events.

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

$$(x - \mu)^T \Sigma^{-1} (x - \mu) \leq \chi_\alpha^2(p), \tag{4.19}$$

where $x$ is the $2 \times 1$ vector containing the $\log(\text{FSC})$ and $\log(\text{SSC})$, $\mu$ 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 $\Sigma$ 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. 4.9 shows an example of different gating contours that would arise from different values of $\alpha$ in Eq. 4.19. In this work, we chose $\alpha = 0.4$, which we deemed as a sufficient constraint to minimize the noise in the data. As explained in Section 4.6, we compared our high throughput flow cytometry data with single-cell microscopy, confirming that the automatic gating did not introduce systematic biases to the analysis pipeline. The specific code where this gating is implemented can be found in GitHub repository.
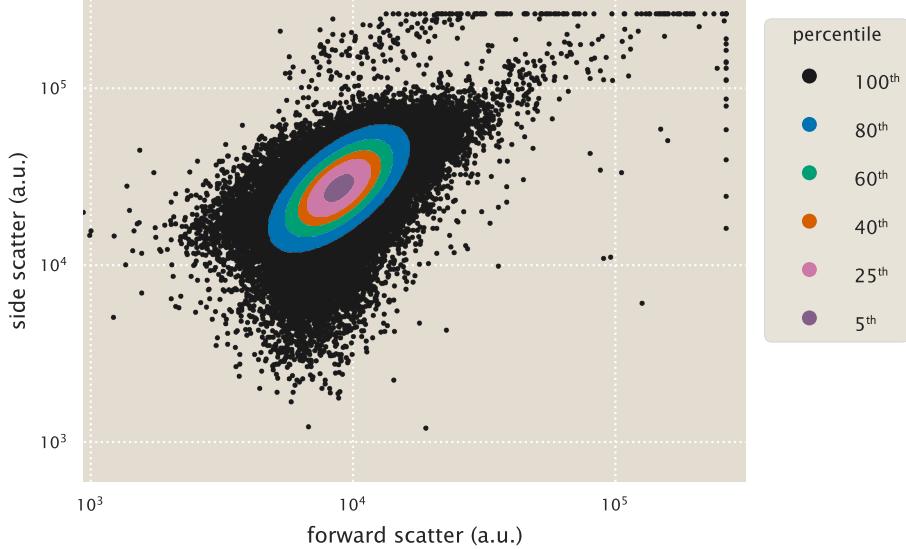
Figure 4.9: **Representative unsupervised gating contours.** Points indicate individual flow cytometry measurements of forward scatter and side scatter. Colored points indicate arbitrary gating contours ranging from 100% ($\alpha = 1.0$) to 5% ($\alpha = 0.05$). All measurements for this work were made computing the mean fluorescence from the 40th percentile ($\alpha = 0.4$), shown as orange points.

## Comparison of Flow Cytometry with Other Methods

Previous work from our lab experimentally determined fold-change for similar simple repression constructs using a variety of different measurement methods [39,42]. 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 [20]. Ref. [39] 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.
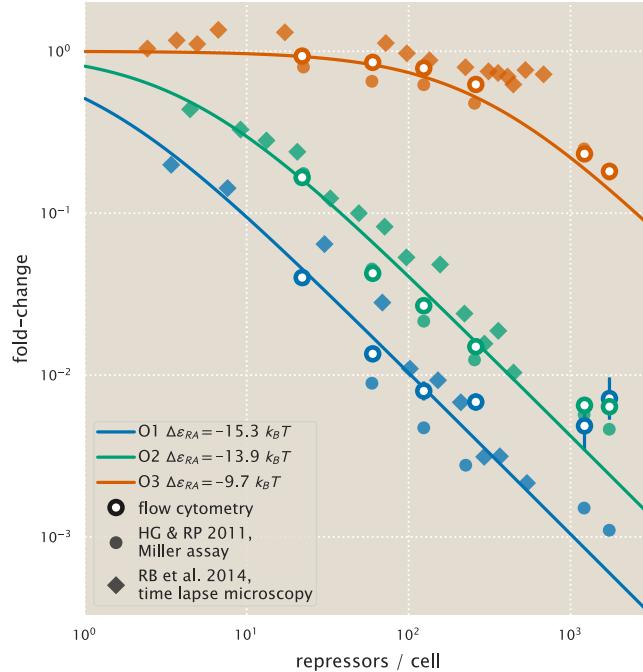
Figure 4.10: **Comparison of experimental methods to determine the fold-change.** The fold-change in gene expression for equivalent simple-repression constructs has been determined using three independent methods: flow cytometry (this work), colorimetric miller assays [20], and video microscopy [39]. All three methods give consistent results, although flow cytometry measurements lose accuracy for fold-change less than $10^{-2}$. Note that the repressor-DNA binding energies $\Delta\varepsilon_{ra}$ used for the theoretical predictions were determined in [20].

Fig. 4.10 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 crucial 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 an extensive dynamic range such as the Miller assay is needed to resolve the fold-change at such low expression levels accurately.

## 4.5 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 [20] or through single-cell microscopy [39]. 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 ($\Delta\varepsilon_{RA}$) of $-13.9\,k_BT$ and $-15.3\,k_BT$ 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. An outline of our microscopy workflow can be seen in Fig. 4.11.

**Strains and Growth Conditions**

Cells were grown identically to those used for measurement via flow cytometry (see Methods). 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°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°C with vigorous aeration for approximately 8 hours. The cultures were diluted 10-fold into M9 glucose minimal medium without IPTG before mounting for microscopy. Each construct was measured using the same range of inducer concentration values as was performed in the flow cytometry measurements (between $100\,\text{nM}$ and $5\,\text{mM}$ IPTG). Each condition was measured in triplicate in microscopy, whereas approximately ten measurements were made using flow cytometry.
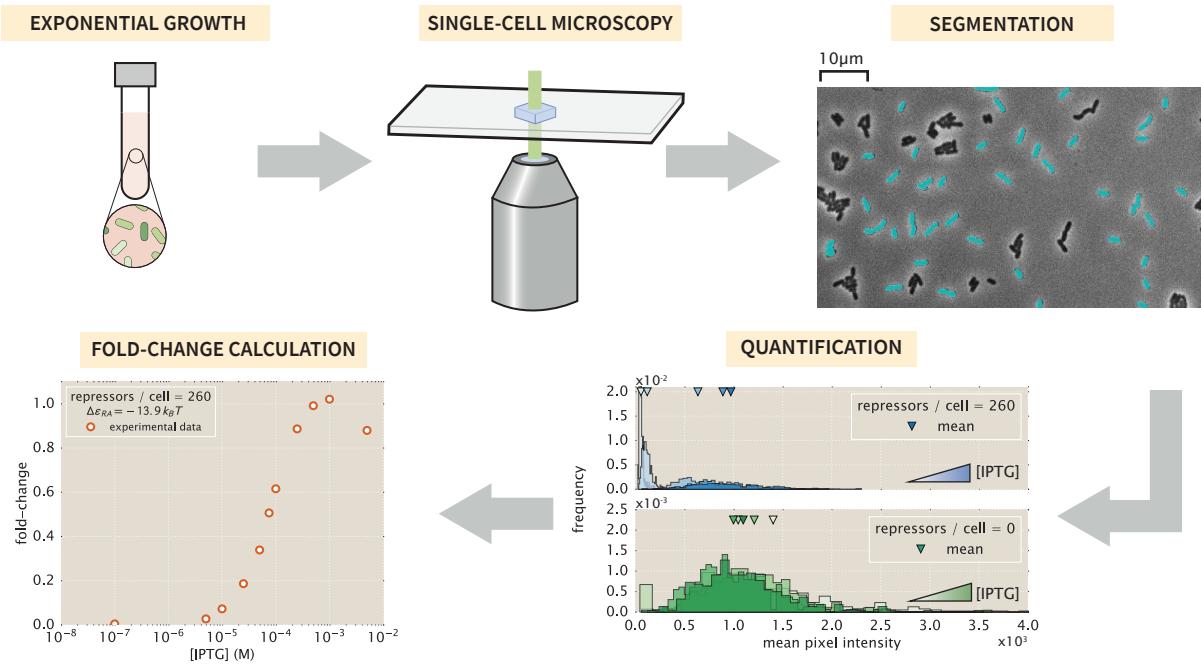
Figure 4.11: **Experimental workflow for single-cell microscopy.** For comparison with the flow cytometry results, the cells were grown in an identical manner to those described in Chapter 2. Once cells had reached mid to late exponential growth, the cultures were diluted and placed on agarose substrates and imaged under $100\times$ magnification. Regions of interest representing cellular mass were segmented, and average single-cell intensities were computed. The means of the distributions were used to compute the fold-change in gene expression.

## Imaging Procedure

During the last hour of cell growth, an agarose mounting substrate was prepared to contain 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 the addition of the IPTG to the appropriate final concentration. This solution was then thoroughly mixed, and a $500\,\mu$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 were 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 operated by the open-source MicroManager control software [137]. 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 the 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. Thus, each channel used the same exposure time 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 non-uniformity was corrected for in post-processing by capturing twenty images of a homogeneously fluorescent plastic slide (Autofluorescent Plastic Slides, Chroma Cat. No. 920001) and averag-
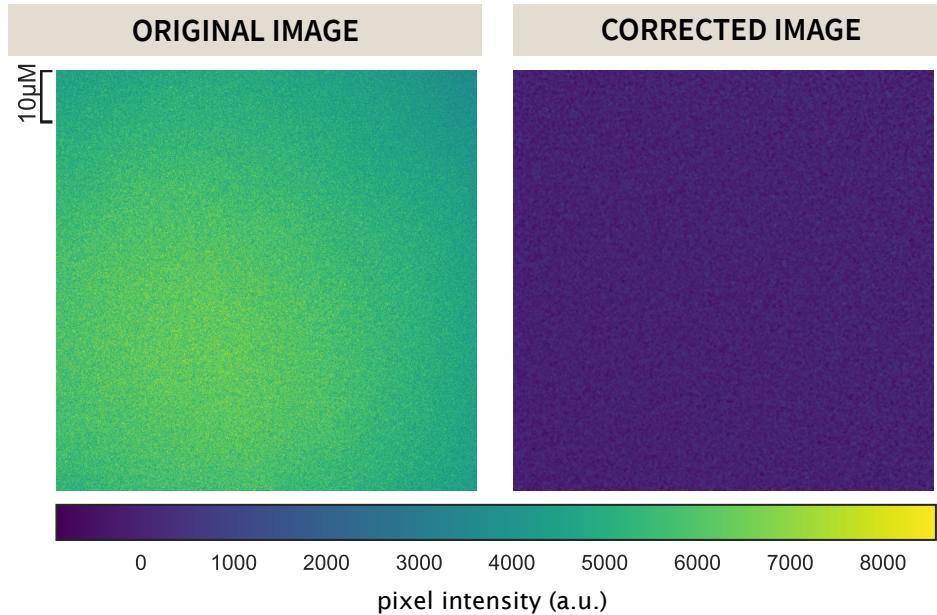
Figure 4.12: **Correction for uneven illumination.** A representative image of the illumination profile of the 512 nm excitation beam on a homogeneously fluorescent slide is shown in the left panel. This is corrected for using Eq. 4.20 and is shown in the right panel.

ing to generate a map of illumination intensity at any pixel $I_{\text{YFP}}$. 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

$$I_{\text{flat}} = \frac{I - I_{\text{dark}}}{I_{\text{YFP}} - I_{\text{dark}}} \langle I_{\text{YFP}} - I_{\text{dark}} \rangle, \tag{4.20}$$

where $I_{\text{flat}}$ is the renormalized image and $I$ is the original fluorescence image. An example of this correction can be seen in Fig. 4.12.

**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 [138], 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 $2\,\mu$m long and $1\,\mu$m wide, so that cells are likely to have an area between $0.5\,\mu$m$^2$ and $6\,\mu$m. To determine the eccentricity bounds, we assumed that a single cell could be approximated by an ellipse with semi-major ($a$) and semi-minor ($b$) axis lengths of $0.5\,\mu$m and $0.25\,\mu$m, respectively. The eccentricity of this hypothetical cell can be computed as

$$\text{eccentricity} = \sqrt{1 - \left(\frac{b}{a}\right)^2}, \tag{4.21}$$

yielding a value of approximately 0.8. Any objects with an eccentricity below these values were not considered to be single cells. After imposing both an area (Fig. 4.13(A)) and eccentricity filter (Fig. 4.13(B)), the remaining objects were considered cells of interest (Fig. 4.13(C)), and the mean fluorescence intensity of each cell was extracted.

**Calculation of Fold-Change**

Cells exhibited background fluorescence even in the absence of an expressed fluorophore. We corrected 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
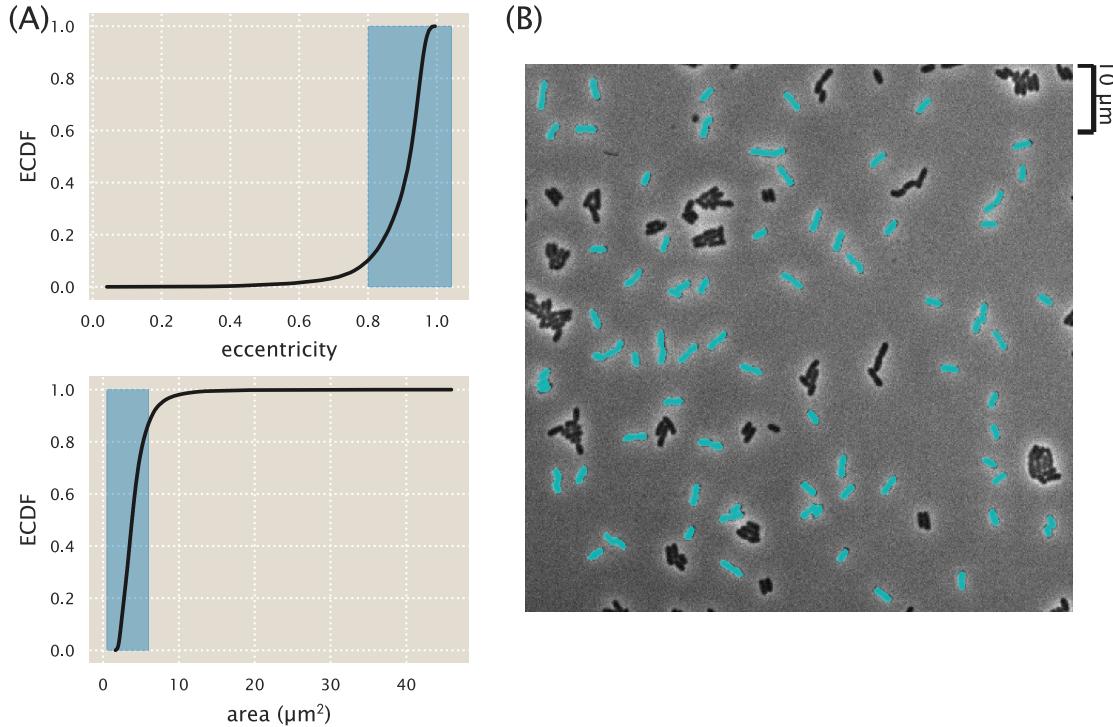
Figure 4.13: **Segmentation of single bacterial cells.** Objects were selected if they had an eccentricity greater than 0.8 and an area between $0.5\,\mu\text{m}^2$ and $6\,\mu\text{m}^2$. Highlighted in blue are the regions considered to be representative of single cells. The black lines correspond to the empirical cumulative distribution functions for the parameter of interest. A representative final segmentation mask is shown in which segmented cells are depicted in cyan over the phase contrast image.

in gene expression was, therefore, calculated as

$$\text{fold-change} = \frac{\langle I_{R>0}\rangle - \langle I_{\text{auto}}\rangle}{\langle I_{R=0}\rangle - \langle I_{\text{auto}}\rangle}, \tag{4.22}$$

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. These fold-change values were very similar to those obtained through flow cytometry and were well described using the thermodynamic parameters used in Chapter 2. With these experimentally measured fold-change values, the best-fit parameter values of the model were inferred and compared to those obtained from flow cytometry.
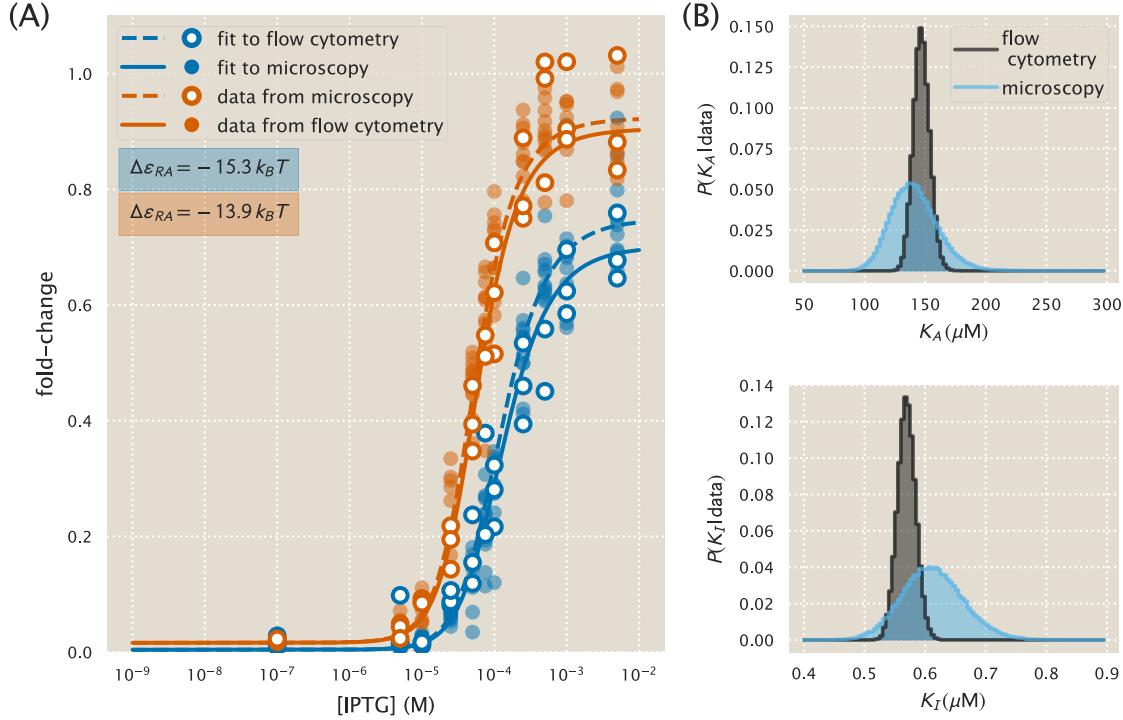
Figure 4.14: **Comparison of measured fold-change between flow cytometry and single-cell microscopy.** Experimentally measured fold-change values obtained through single-cell microscopy and flow cytometry are shown as white-filled and solid-colored circles, respectively. Solid and dashed lines indicate the predicted behavior using the most likely parameter values of $K_A$ and $K_I$ inferred from flow cytometry data and microscopy data, respectively. The red and blue plotting elements correspond to the different operators O1 and O2 with binding energies $\Delta\varepsilon_{RA}$ of $-13.9\ k_BT$ and $-15.3\ k_BT$, respectively [20]. The marginalized posterior distributions for $K_A$ and $K_I$ are shown in the top and bottom panels, respectively. The posterior distribution determined using the microscopy data is wider than that computed using the flow cytometry data due to a smaller fig collection of data sets (three for microscopy and ten for flow cytometry).

## Parameter Estimation and Comparison

To confirm quantitative consistency between flow cytometry and microscopy, the parameter values of $K_A$ and $K_I$ were also estimated from three biological replicates of IPTG titration curves obtained by microscopy for strains with $R = 260$ and operators O1 and O2. Fig. 4.14(A) shows the data from these measurements (orange circles) and the ten biological replicates from our flow cytometry measurements (blue circles), along with the fold-change predictions from each inference. In comparison with the values obtained by flow cytometry, each parameter estimate overlapped with the 95% credible region of our flow cytometry estimates, as shown in Fig. 4.14(B). Specifically, these values were $K_A = 142^{+40}_{-34}\ \mu$M and $K_I = 0.6^{+0.1}_{-0.1}\ \mu$M

from microscopy and $K_A = 149^{+14}_{-12}\,\mu\text{M}$ and $K_I = 0.57^{+0.03}_{-0.02}\,\mu\text{M}$ from the flow cytometry data. We note that the credible regions from the microscopy data shown in Fig. 4.14(B) are much broader than those from flow cytometry due to the fewer number of replicates performed.

## 4.6 Fold-Change Sensitivity Analysis

In Fig. 2.5, we found that the width of the credible regions varied widely depending on the repressor copy number $R$ and repressor operator binding energy $\Delta\varepsilon_{RA}$. More precisely, the credible regions were much narrower for low repressor copy numbers $R$ and weak binding energy $\Delta\varepsilon_{RA}$. 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 given by Eq. 2.7. While it is straightforward to consider the width of the credible regions at any other inducer concentration, it shows that the credible region is 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

$$\Delta\text{fold-change}_{K_A} \equiv \text{fold-change}(K_A, K_I^{\text{fit}}) - \text{fold-change}(K_A^{\text{fit}}, K_I^{\text{fit}}), \qquad (4.23)$$

the difference between the fold-change at a particular $K_A$ value relative to the best-fit dissociation constant $K_A^{\text{fit}} = 139 \times 10^{-6}\,\text{M}$. For simplicity, we keep the inactive state dissociation constant fixed at its best-fit value $K_I^{\text{fit}} = 0.53 \times 10^{-6}\,\text{M}$. A larger difference $\Delta\text{fold-change}_{K_A}$ implies a wider credible region. Similarly, we define the analogous quantity

$$\Delta\text{fold-change}_{K_I} = \text{fold-change}(K_A^{\text{fit}}, K_I) - \text{fold-change}(K_A^{\text{fit}}, K_I^{\text{fit}}) \qquad (4.24)$$

to measure the sensitivity of the fold-change to $K_I$ at a fixed $K_A^{\text{fit}}$. Fig. 4.15 shows both of these quantities in the limit $c \to \infty$ for different repressor-DNA binding energies $\Delta\varepsilon_{RA}$ and repressor copy numbers $R$. See our GitHub repository for the code that reproduces these plots.

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, $\Delta\text{fold-change}_{K_A} \approx \frac{\partial\text{fold-change}}{\partial K_A}\left(K_A - K_A^{\text{fit}}\right)$, where the partial derivative has the form

$$\frac{\partial\text{fold-change}}{\partial K_A} = \frac{e^{-\beta\Delta\varepsilon_{AI}}\frac{n}{K_I}\left(\frac{K_A}{K_I}\right)^{n-1}}{\left(1+e^{-\beta\Delta\varepsilon_{AI}}\left(\frac{K_A}{K_I}\right)^n\right)^2}\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\left(1+\frac{1}{1+e^{-\beta\Delta\varepsilon_{AI}}\left(\frac{K_A}{K_I}\right)^n}\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)^{-2}. \qquad (4.25)$$

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

$$\frac{\partial\text{fold-change}}{\partial K_I} = -\frac{e^{-\beta\Delta\varepsilon_{AI}}\frac{n}{K_I}\left(\frac{K_A}{K_I}\right)^{n}}{\left(1+e^{-\beta\Delta\varepsilon_{AI}}\left(\frac{K_A}{K_I}\right)^n\right)^2}\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\left(1+\frac{1}{1+e^{-\beta\Delta\varepsilon_{AI}}\left(\frac{K_A}{K_I}\right)^n}\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)^{-2}. \qquad (4.26)$$

From Eqs. 4.25 and 4.26, 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. 4.15.

Lastly, we note that Eqs. 4.25 and 4.26 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 equal the fold-change squared is roughly one. 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.
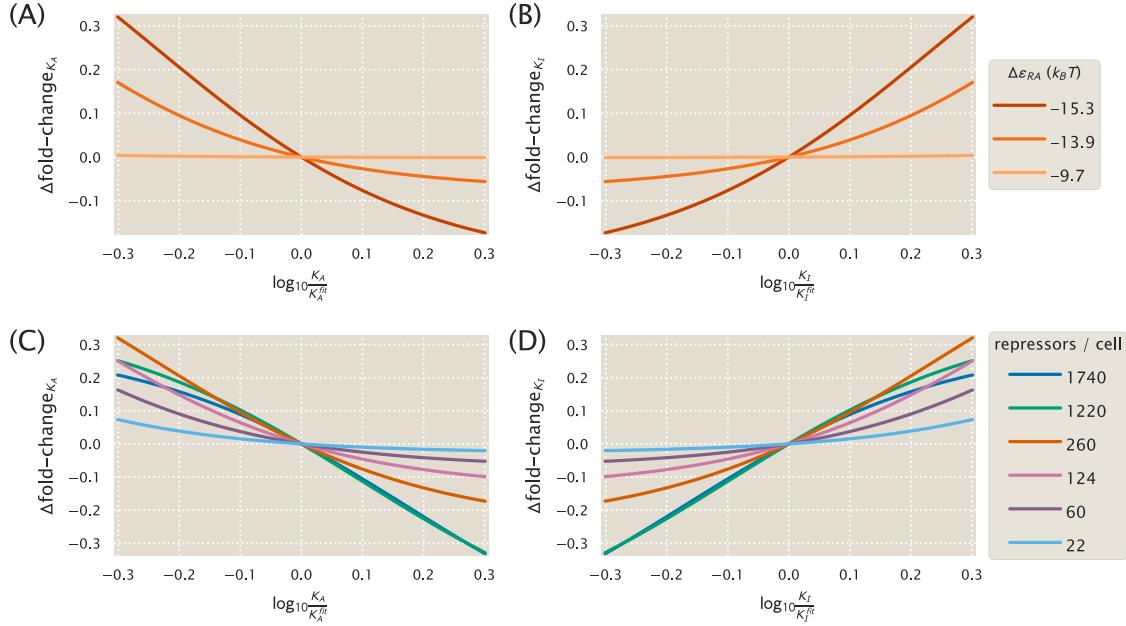
Figure 4.15: **Determining how sensitive the fold-change values are to the fit values of the dissociation constants.** (A) The difference $\Delta$fold-change$_{K_A}$ in fold change when the dissociation constant $K_A$ is slightly offset from its best-fit value $K_A = 139^{+29}_{-22} \times 10^{-6}$ M, as given by Eq. 4.23. Fold-change is computed in the limit of saturating inducer concentration ($c \to \infty$, see Eq. 2.7) where the credible regions in Fig. 2.4 are the widest. The O3 strain ($\Delta\varepsilon_{RA} = -9.7\ k_BT$) 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 $R = 260$. (B) As in Panel (A), but plotting the sensitivity of fold-change to the $K_I$ parameter relative to the best-fit value $K_I = 0.53^{+0.04}_{-0.04} \times 10^{-6}$ 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. (C) As in Panel (A), but showing how the fold-change sensitivity for different repressor copy numbers. The strains with lower repressor copy numbers are less sensitive to changes in the dissociation constants, and hence their corresponding curves in Fig. 2.4 have tighter credible regions. All curves were made using $\Delta\varepsilon_{RA} = -13.9\ k_BT$. (D) As in Panel (C), the sensitivity of fold-change with respect to $K_I$ is again smallest (in magnitude) for the low repressor copy number strains.

## 4.7    Alternate Characterizations of Induction

In this section, we discuss a different way to describe the induction data, namely, through using the conventional Hill approach. We first demonstrate how using a Hill function to characterize a single induction curve enables us to extract features (such as the midpoint and sharpness) of that single response, but precludes any predictions of the other seventeen strains. We then discuss how a thermodynamic model of simple repression coupled with a Hill approach to the induction response can both characterize an induction profile and predict the response of all eighteen strains, although we argue that such a description provides no insight into the allosteric nature of the protein and how mutations to the repressor would affect induction. We conclude the section by discussing the differences between such a model and the statistical mechanical model used in Chapter 2.

**Fitting Induction Curves using a Hill Function Approach**

The Hill equation is a phenomenological function commonly used to describe data with a sigmoidal profile [37,56,58]. Its simplicity and ability to estimate the cooperativity of a system (through the Hill coefficient) has led to its widespread use in many domains of biology [139]. Nevertheless, the Hill function is often criticized as a physically unrealistic model and the extracted Hill coefficient is often difficult to contextualize in the physics of a system [140]. In the present work, we note that a Hill function, even if it is only used because of its simplicity, presents no mechanism to understand how a regulatory system's behavior will change if physical parameters such as repressor copy number or operator binding energy are varied. In addition, the Hill equation provides no foundation to explore how mutating the repressor (e.g., at its inducer-binding interface) would modify its induction profile, although statistical mechanical models have proved capable of characterizing such scenarios [68,69,71].

Consider the general Hill equation for a single induction profile given by

$$\text{fold-change} = (\text{leakiness}) + (\text{dynamic range})\frac{\left(\frac{c}{K}\right)^n}{1 + \left(\frac{c}{K}\right)^n}, \qquad (4.27)$$

where, as in Chapter 2, the leakiness represents the minimum fold-change, the dynamic range represents the difference between the maximum and minimum fold-change, $K$ is the repressor-inducer dissociation constant, and $n$ denotes the Hill coefficient that characterizes the sharpness of the curve ($n > 1$ signifies positive cooperativity, $n = 1$ denotes no cooperativity, and $n < 1$ represents negative cooperativity). Fig. 4.16 shows how the individual induction profiles can be fit (using the same Bayesian methods as described in Sec. 4.8 to this Hill response, yielding a similar response to that shown in Fig. 2.5. However, characterizing the induction response in this manner is unsatisfactory because each curve must be fit independently, thus removing our predictive power for other repressor copy numbers and binding sites.

The fitted parameters obtained from this approach are shown in Fig. 4.17. These are rather unsatisfactory because they do not reflect the properties of the physical system under consideration. For example, the dissociation constant $K$ between LacI and inducer should not be affected by either the copy number of the repressor or the DNA binding energy. Yet, we see upward trends as $R$ is increased or the binding energy is decreased. Here, the $K$ parameter ultimately describes the midpoint of the induction curve and, therefore, cannot strictly be considered a dissociation constant. Similarly, the Hill coefficient $n$ does not directly represent the cooperativity between the repressor and the inducer. The molecular details of the copy number and DNA binding strength are subsumed in this parameter. While the leakiness and dynamic range describe important phenotypic properties of the induction response, this Hill approach leaves us with no means to predict them for other strains. In summary, the Hill equation (Eq. 4.27) cannot predict how an induction profile varies with repressor copy number, operator binding energy, or how mutations alter the induction profile. To that end, we turn to a more sophisticated approach where we use the Hill function to describe the available fraction of repressor as a function of inducer concentration.
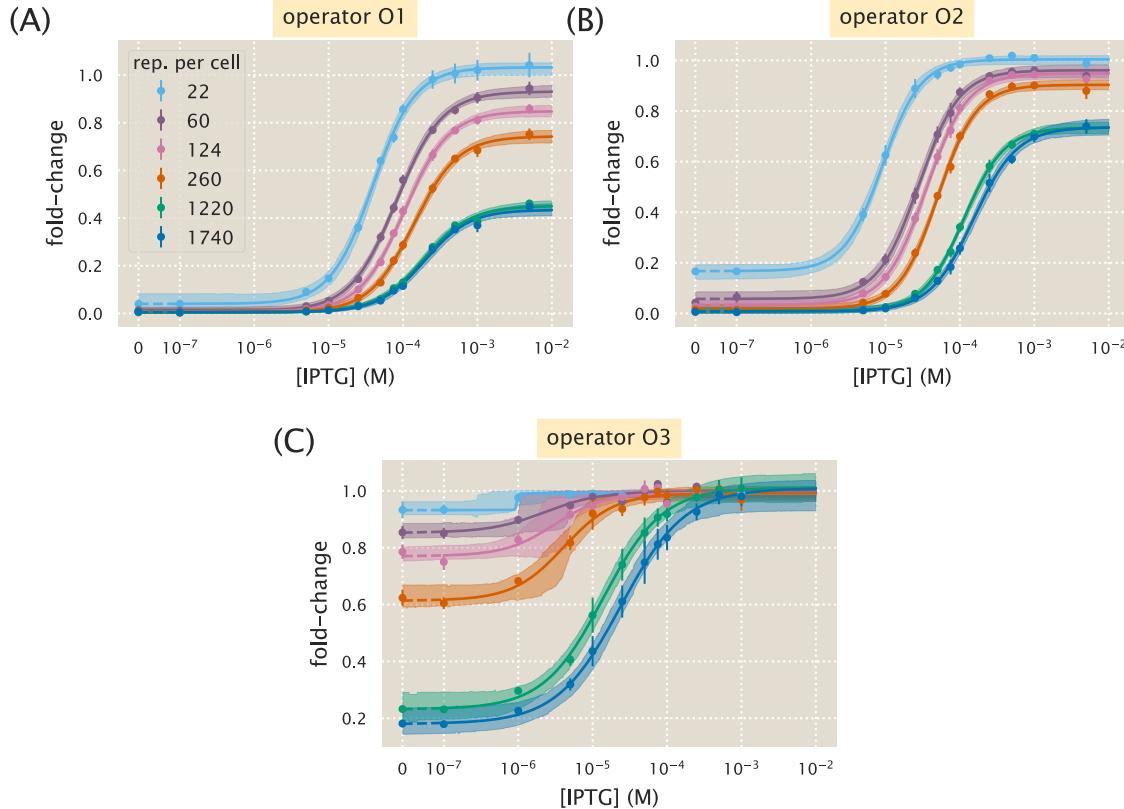
Figure 4.16: **Hill function and MWC analysis of each induction profile.** Data for each individual strain was fit to the general Hill function in Fig. 2.5. (A) strains with O1 binding site, (B) strains with O2 binding site, and (C) strains with O3 binding site. Shaded regions indicate the bounds of the 95% credible region.

**Fitting Induction Curves using a Combination Thermodynamic Model and Hill Function Approach**

Motivated by the inability in the previous section to characterize all eighteen strains using the Hill function with a single set of parameters, here we combine the Hill approach with a thermodynamic model of simple repression to garner predictive power. More specifically, we will use the thermodynamic model in Fig. 2.2(A), but substitute the statistical model in Fig. 2.2(B) with the phenomenological Hill function (Eq. 4.27).

Following Eqs. 2.1, 2.2, 2.3 and fold-change is given by

$$\text{fold-change} = \left(1 + p_A(c)\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1} \tag{4.28}$$
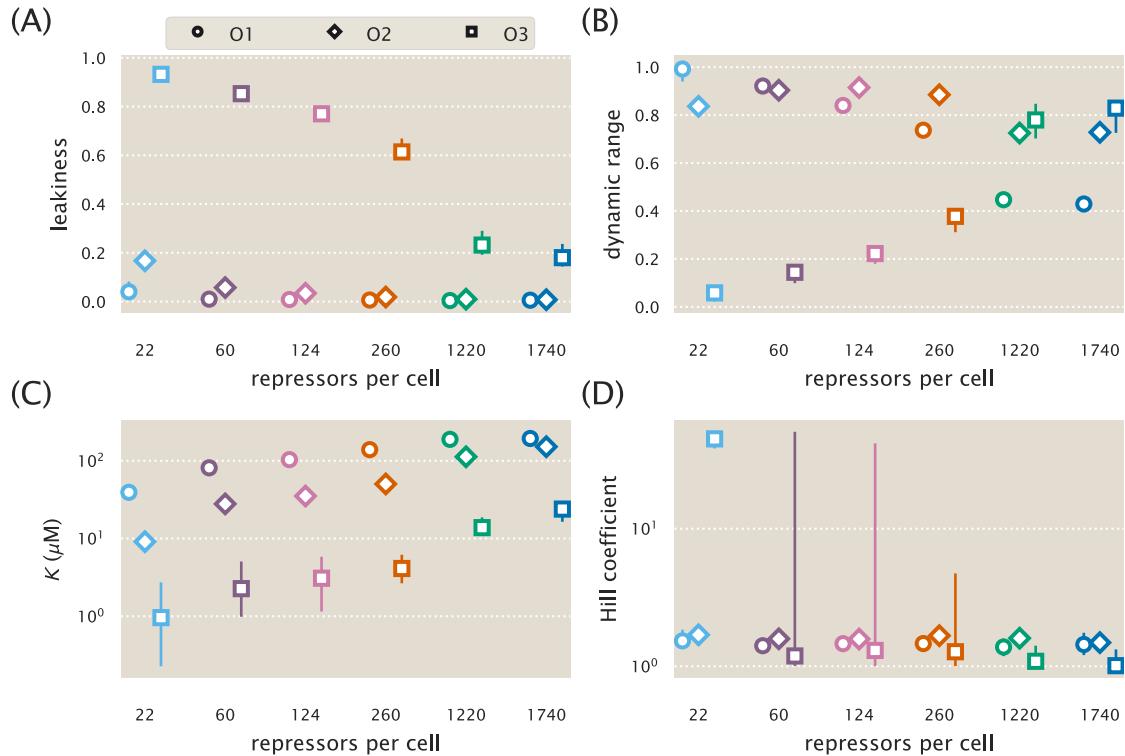
Figure 4.17: **Parameter values for the Hill equation fit to each individual titration.** The resulting fit parameters from the Hill function fits of Fig. 4.16 are summarized. The large parameter intervals for many of the O3 strains are due to the flatter induction profile (as seen by its smaller dynamic range) and the ability for a large range of $K$ and $n$ values to describe the data.

where the Hill function

$$p_A(c) = p_A^{\max} - p_A^{\text{range}} \frac{\left(\frac{c}{K_D}\right)^n}{1 + \left(\frac{c}{K_D}\right)^n} \tag{4.29}$$

represents the fraction of repressors in the allosterically active state, with $p_A^{\max}$ denoting the fraction of active repressors in the absence of inducer and $p_A^{\max} - p_A^{\text{range}}$ the minimum fraction of active repressors in the presence of saturating inducer. The Hill function characterizes the inducer-repressor binding while the thermodynamic model with the known constants $R$, $N_{NS}$, and $\Delta\varepsilon_{RA}$ describes how the induction profile changes with repressor copy number and repressor-operator binding energy.
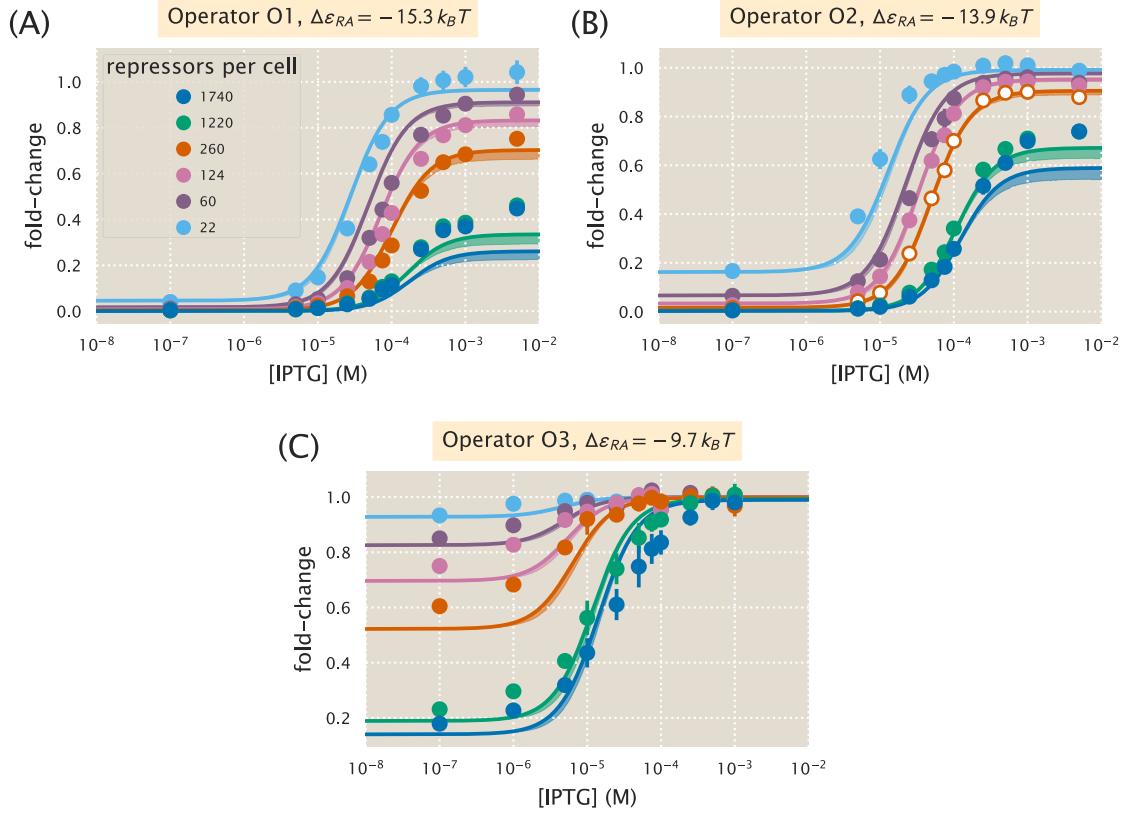
Figure 4.18: **A thermodynamic model coupled with a Hill analysis can characterize induction.** Combining a thermodynamic model of simple repression with the Hill function to characterize the repressor-inducer binding successfully characterizes the induction profiles of all eighteen strains. As in the main text, data was only fit for the O2 $R = 260$ strain using Eq. 4.27 and the parameters $p_A^{\text{max}} = 0.90^{+0.03}_{-0.01}$, $p_A^{\text{range}} = -0.90^{+0.02}_{-0.03}$, $n = 1.6^{+0.2}_{-0.1}$, and $K_D = 4^{+2}_{-1} \times 10^{-6}\,\text{M}$. Shaded regions indicate bounds of the 95% credible region.

As in Chapter 2, we can fit the four Hill parameters—the vertical shift and stretch parameters $p_A^{\text{max}}$ and $p_A^{\text{range}}$, the Hill coefficient $n$, and the inducer-repressor dissociation constant $K_D$—for a single induction curve and then use the fully characterized Eq. 4.27 to describe the response of each of the eighteen strains. Fig. 4.18 shows this process carried out by fitting the O2 $R = 260$ strain (white circles in Panel (B)) and predicting the behavior of the remaining seventeen strains.

Although the curves in Fig. 4.18 are nearly identical to those in Fig. 2.5 (which were made using the MWC model), we stress that the Hill function approach is more complex than the MWC model (containing four parameters instead of three) and obscures the relationships to the physical parameters of the system. For example, it is not clear whether the fit parameter $K_D = 4^{+2}_{-1} \times 10^{-6}$ M relays the dissociation constant between the inducer and active-state repressor, between the inducer and the inactive-state repressor, or some mix of the two quantities.

In addition, the MWC model naturally suggests further quantitative tests for the fold-change relationship. For example, mutating the repressor's inducer binding site would likely alter the repressor-inducer dissociation constants $K_A$ and $K_I$, and it would be interesting to find out if such mutations also modify the allosteric energy difference $\Delta \varepsilon_{AI}$ between the repressor's active and inactive conformations. For our purposes, the Hill function falls short of the connection to the physics of the system and provides no intuition about how transcription depends upon such mutations. For these reasons, we present the thermodynamic model coupled with the statistical mechanical MWC model approach in the paper.

## 4.8 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 [20]. However, any error in these previous measurements of $R$ and $\Delta \varepsilon_{RA}$ will necessarily propagate into our own fold-change predictions. This section takes an alternative approach to fitting the system's physical parameters to that used in Chapter 2. First, rather than fitting only a single strain, we fit the entire data set in Fig. 2.5 along with microscopy data for the synthetic operator Oid (see Sec. 4.9). 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. As a point of reference, we state all of the parameters of the MWC model derived in the text in Table 4.2.

To fit all of the parameters simultaneously, we follow a similar approach to the one detailed in the Methods section. 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 Chapter 2, we fit the logarithms $\tilde{k}_A = -\log\frac{K_A}{1\mathrm{M}}$ and $\tilde{k}_I = -\log\frac{K_I}{1\mathrm{M}}$ of the dissociation constants, which grants better numerical stability.

As in Eqs. 4.28 and 4.29, 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,

$$P(\tilde{k}_A, \tilde{k}_I, \mathbf{R}, \Delta\boldsymbol{\varepsilon}_{\mathbf{RA}}, \sigma \mid D) = \frac{P(D \mid \tilde{k}_A, \tilde{k}_I, \mathbf{R}, \Delta\boldsymbol{\varepsilon}_{\mathbf{RA}}, \sigma)P(\tilde{k}_A, \tilde{k}_I, \mathbf{R}, \Delta\boldsymbol{\varepsilon}_{\mathbf{RA}}, \sigma)}{P(D)}, \quad (4.30)$$

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

Given $n$ independent measurements of the fold-change, the first term in Eq. 4.30 can be written as

$$P(D|\tilde{k}_A, \tilde{k}_I, \mathbf{R}, \Delta\boldsymbol{\varepsilon}_{\mathbf{RA}}, \sigma) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \prod_{i=1}^{n} \exp\left[-\frac{(\mathrm{fc}_{\mathrm{exp}}^{(i)} - \mathrm{fc}(\tilde{k}_A, \tilde{k}_I, R^{(i)}, \Delta\varepsilon_{RA}^{(i)}, c^{(i)}))^2}{2\sigma^2}\right], \quad (4.31)$$

where $\mathrm{fc}_{\mathrm{exp}}^{(i)}$ is the $i^{\mathrm{th}}$ experimental fold-change and $\mathrm{fc}(\cdots)$ 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 Eq. 4.30 represents the prior information of the parameter values. We assume that all parameters are independent of each other so that

$$P(\tilde{k}_A, \tilde{k}_I, \mathbf{R}, \mathbf{\Delta \varepsilon_{RA}}, \sigma) = P(\tilde{k}_A) \cdot P(\tilde{k}_I) \cdot \prod_i P(R^{(i)}) \cdot \prod_j P(\Delta \varepsilon_{RA}^{(j)}) \cdot P(\sigma), \qquad (4.32)$$

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 knew 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 [61]. We do, however, have prior information for the repressor copy numbers and the repressor-DNA binding energies from [20]. 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

$$P(R^{(i)}) = \frac{1}{\sqrt{2\pi\sigma_{R_i}^2}} \exp\left(-\frac{(R^{(i)} - \bar{R}^{(i)})^2}{2\sigma_{R_i}^2}\right), \qquad (4.33)$$

where $\bar{R}^{(i)}$ is the mean repressor copy number and $\sigma_{R_i}$ is the variability associated with this parameter as reported in [20]. 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

$$P(\Delta\varepsilon_{RA}^{(j)}) = \frac{1}{\sqrt{2\pi\sigma_{\varepsilon_j}^2}} \exp\left(-\frac{(\Delta\varepsilon_{RA}^{(j)} - \Delta\bar{\varepsilon}_{RA}^{(j)})^2}{2\sigma_{\varepsilon_j}^2}\right), \qquad (4.34)$$

where $\Delta\bar{\varepsilon}_{RA}^{(j)}$ is the binding energy and $\sigma_{\varepsilon_j}$ is the variability associated with that parameter around the mean value as reported in [20].

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 strong confidence in the previously reported value. Mathematically, the exponential in Eq. 4.33 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 explore parameter space in freely the vicinity of the best estimates. Putting all these terms together, we use Markov chain Monte Carlo to sample the posterior distribution $P(\tilde{k}_A, \tilde{k}_I, \mathbf{R}, \Delta\varepsilon_{\mathbf{RA}}, \sigma \mid D)$, enabling us to determine both the most likely value for each physical parameter as well as its associated credible region (see the GitHub repository for the implementation).

Fig. 4.19 shows the result of this global fit. When compared with Fig. 2.5, we can see that fitting for the binding energies and the repressor copy numbers improve the agreement between the theory and the data. Table 4.3 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 previously reported values provided in [20]. 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 $k_B T$, representing 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 Sec. 4.9.

Fig. 4.20 shows the same key properties as in Fig. 2.6, 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 due to only minor improvements between the theoretical curves and data. For the O3 operator data, again, the agreement between the predicted $[EC_{50}]$ and the effective Hill coefficient remains poor due to the theory being unable to capture the steepness of the response curves.

Table 4.2: **Key model parameters for induction of an allosteric repressor.**

| Parameter | Description |
| --- | --- |
| $c$ | Concentration of the inducer |
| $K_A, K_I$ | Dissociation constant between an inducer and the repressor in the active/inactive state |
| $\Delta\varepsilon_{AI}$ | The difference between the free energy of repressor in the inactive and active states |
| $\Delta\varepsilon_P$ | Binding energy between the RNAP and its specific binding site |
| $\Delta\varepsilon_{RA}, \Delta\varepsilon_{RI}$ | Binding energy between the operator and the active/inactive repressor |
| $n$ | Number of inducer binding sites per repressor |
| $P$ | Number of RNAP |
| $R_A, R_I, R$ | Number of active/inactive/total repressors |
| $p_A = \frac{R_A}{R}$ | Probability that a repressor will be in the active state |
| $p_{\text{bound}}$ | Probability that an RNAP is bound to the promoter of interest, assumed to be proportional to gene expression |

| Parameter | Description |
|---|---|
| fold-change | Ratio of gene expression in the presence of repressor to that in the absence of repressor |
| $F$ | Free energy of the system |
| $N_{NS}$ | The number of non-specific binding sites for the repressor in the genome |
| $\beta = \frac{1}{k_B T}$ | The inverse product of the Boltzmann constant $k_B$ and the temperature $T$ of the system |

Table 4.3: **Global fit of all parameter values using the entire data set in Fig. 2.5.** In addition to fitting the repressor inducer dissociation constants $K_A$ and $K_I$ as was done in the text, we also fit the repressor DNA binding energy $\Delta\varepsilon_{RA}$ as well as the repressor copy numbers $R$ for each strain. The middle columns show the previously reported values for all $\Delta\varepsilon_{RA}$ and $R$ values, with $\pm$ representing the standard deviation of three replicates. The right column shows the global fits from this work, with the subscript and superscript notation denoting the 95% credible region. Note that there is overlap between all of the repressor copy numbers and that the net difference in the repressor-DNA binding energies is less than 1 $k_B T$. The logarithms $\tilde{k}_A = -\log\frac{K_A}{1\mathrm{M}}$ and $\tilde{k}_I = -\log\frac{K_I}{1\mathrm{M}}$ of the dissociation constants were fit for numerical stability.

| | Reported Values [20] | Global Fit |
|---|---|---|
| $\tilde{k}_A$ | – | $-5.33^{+0.06}_{-0.05}$ |
| $\tilde{k}_I$ | – | $0.31^{+0.05}_{-0.06}$ |
| $K_A$ | – | $205^{+11}_{-12}\,\mu\mathrm{M}$ |
| $K_I$ | – | $0.73^{+0.04}_{-0.04}\,\mu\mathrm{M}$ |
| $R_{22}$ | $22 \pm 4$ | $20^{+1}_{-1}$ |
| $R_{60}$ | $60 \pm 20$ | $74^{+4}_{-3}$ |
| $R_{124}$ | $124 \pm 30$ | $130^{+6}_{-6}$ |

|  | **Reported Values [20]** | **Global Fit** |
|---|---|---|
| $R_{260}$ | $260 \pm 40$ | $257^{+9}_{-11}$ |
| $R_{1220}$ | $1220 \pm 160$ | $1191^{+32}_{-55}$ |
| $R_{1740}$ | $1740 \pm 340$ | $1599^{+75}_{-87}$ |
| O1 $\Delta\varepsilon_{RA}$ | $-15.3 \pm 0.2\ k_B T$ | $-15.2^{+0.1}_{-0.1}\ k_B T$ |
| O2 $\Delta\varepsilon_{RA}$ | $-13.9 \pm 0.2\ k_B T$ | $-13.6^{+0.1}_{-0.1}\ k_B T$ |
| O3 $\Delta\varepsilon_{RA}$ | $-9.7 \pm 0.1\ k_B T$ | $-9.4^{+0.1}_{-0.1}\ k_B T$ |
| Oid $\Delta\varepsilon_{RA}$ | $-17.0 \pm 0.2\ k_B T$ | $-17.7^{+0.2}_{-0.1}\ k_B T$ |

## 4.9 Applicability of Theory to the Oid Operator Sequence

In addition to the native operator sequences (O1, O2, and O3) considered in Chapter 2, 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 ($\Delta\varepsilon_{RA} = -17.0\ k_B T$) [42]. Accordingly, fluorescence data for Oid were obtained using microscopy, which is more sensitive than flow cytometry. Sec. 4.6 gives a detailed explanation of how microscopy measurements were used to obtain induction curves.

We follow the approach of Chapter 2 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. 4.21(A), 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 Sec. 4.8) to see whether we might find
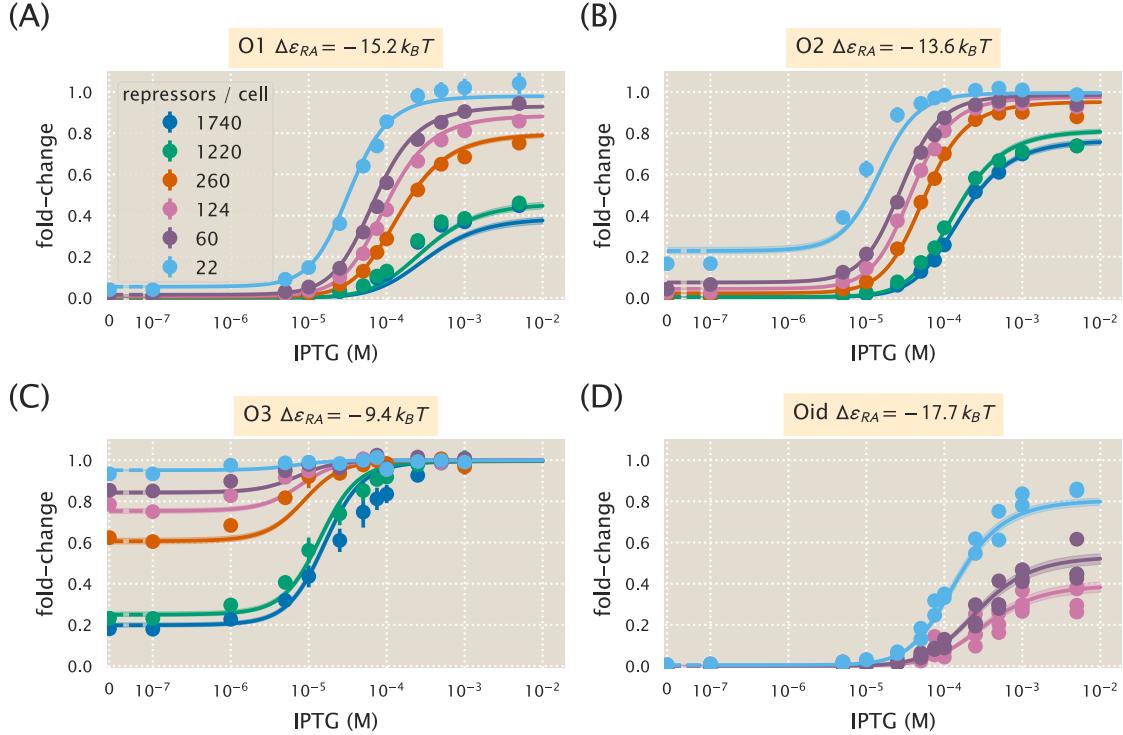
Figure 4.19: **Global fit of dissociation constants, repressor copy numbers, and binding energies.** Theoretical predictions resulting from simultaneously fitting the dissociation constants $K_A$ and $K_I$, the six repressor copy numbers $R$, and the four repressor-DNA binding energies $\Delta\varepsilon_{RA}$ using the entire data set from Fig. 2.5 as well as the microscopy data for the Oid operator. Error bars of experimental data show the standard error of the mean (eight or more replicates), and shaded regions denote the 95% credible region. Where error bars are not visible, they are smaller than the point itself. All of the data points are shown for the Oid operator since a smaller number of replicates were taken. The shaded regions are significantly smaller than in Fig. 2.5 because this fit was based on all data points, and hence the fit parameters are much more tightly constrained. The dashed lines at 0 IPTG indicate a linear scale, whereas solid lines represent a log scale.

better agreement with the observed data. Interestingly, we find that the majority of the parameters remain essentially unchanged, but our estimate for the Oid binding energy $\Delta\varepsilon_{RA}$ is shifted to $-17.7\,k_BT$ instead of the value $-17.0\,k_BT$ found by [20]. In Fig. 4.21(B), we again plot the Oid fold-change data with theoretical predictions using the new estimate for the Oid binding energy from our global fit. This parameter modification gives substantially better agreement between theory and data.
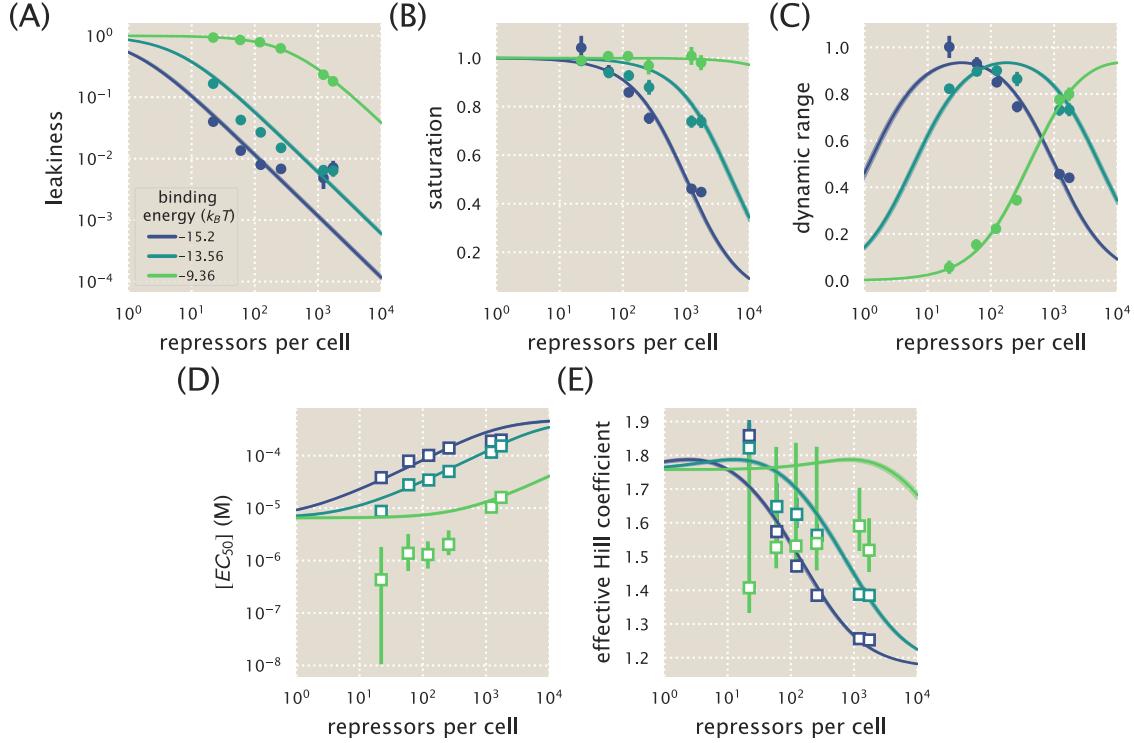
Figure 4.20: **Key properties of induction profiles as predicted with a global fit using all available data.** Data for the (A) leakiness, (B) saturation, and (C) dynamic range are obtained from fold-change measurements in Fig. 2.5 in the absence and presence of IPTG. All prediction curves were generated using the parameters listed in Table 4.3. Both the (D) $[EC_{50}]$ and (E) effective Hill coefficient are inferred by individually fitting all parameters–$K_A$, $K_I$, $R$, $\Delta\varepsilon_{RA}$–to each operator-repressor pairing in Fig. 2.4(A)-(C) separately to Eq. 2.5 to smoothly interpolate between the data points. Note that the error bars are smaller than some of the points.

Fig. 4.22 shows the cumulative data from [20] and [39], as well as our data with $c = 0\,\mu$M, which all measured fold-change for the same simple repression architecture utilizing different reporters and measurement techniques. We find that the binding energies from the global fit, including $\Delta\varepsilon_{RA} = -17.7\,k_BT$, compare reasonably well with all previous measurements.

## 4.10   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 ($R = 260$, $\Delta\varepsilon_{RA} = -13.9\,k_BT$, $n = 2$, and $\Delta\varepsilon_{AI} = 4.5\,k_BT$). After determining these parameters, we were able to predict the fold-change of the remaining strains without
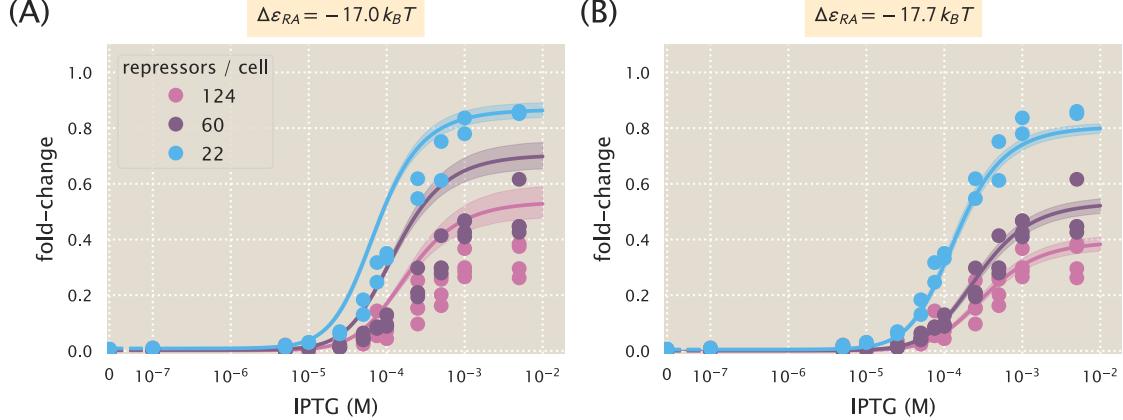
Figure 4.21: **Predictions of fold-change for strains with an Oid binding sequence versus experimental measurements with different repressor copy numbers.** Experimental data is plotted against the parameter-free predictions that are based on our fit to the O2 strain with $R = 260$. Here we use the previously measured binding energy $\Delta\varepsilon_{RA} = -17.0\,k_BT$ [20]. 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 (see Sec. 4.8). Here the major difference in the inferred parameters is a shift in the binding energy for Oid from $\Delta\varepsilon_{RA} = -17.0\,k_BT$ to $\Delta\varepsilon_{RA} = -17.7\,k_BT$, which now shows agreement between the theoretical predictions and experimental data. Shaded regions from the theoretical curves denote the 95% credible region. These are narrower in Panel because the inference of parameters was performed with much more data, and hence the best-fit values are more tightly constrained. Individual data points are shown due to the small number of replicates. The dashed lines at 0 IPTG indicate a linear scale, whereas solid lines represent a log scale.

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. 2.4(D), 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.
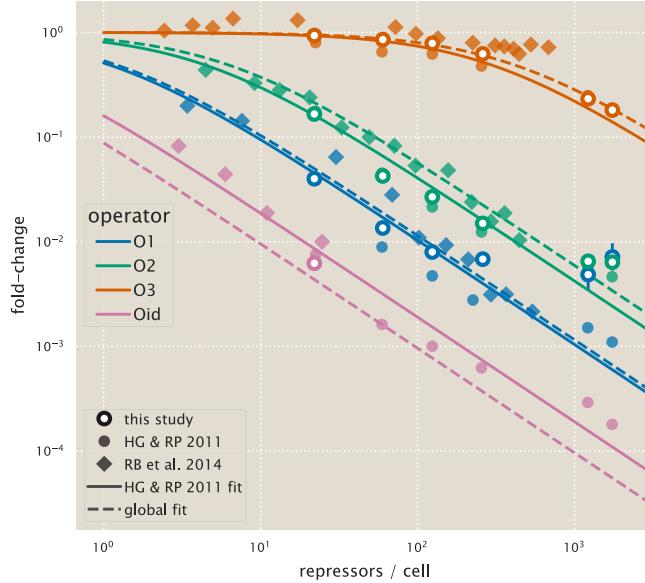
Figure 4.22: **Comparison of fold-change predictions based on binding energies from Garcia and Phillips and those inferred from this work.** Fold-change curves for the different repressor-DNA binding energies $\Delta\varepsilon_{RA}$ are plotted as a function of repressor copy number when IPTG concentration $c = 0$. Solid curves use the binding energies determined from [20], while the dashed curves use the inferred binding energies we obtained when performing a global fit of $K_A$, $K_I$, repressor copy numbers, and the binding energies using all available data from our work. Fold-change measurements from our experiments (outlined circles) [20] (solid circles), and [39] (diamonds) show that the small shifts in binding energy that we infer are still in agreement with prior data. Note that only a single flow cytometry data point is shown for Oid from this study, since the $R = 60$ and $R = 124$ curves from Fig. 4.21 had extremely low fold-change in the absence of inducer ($c = 0$) to be indistinguishable from autofluorescence, and their fold-change values in this limit were negative and hence do not appear on this plot.

We follow the approach taken in Chapter 2 and use Eq. 2.5 to 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. 4.23, we plot these fold-change predictions along with experimental data for each of our strains that contain an O1 operator. To make sense of this plot, 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, shaded in gray). 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. 4.24 and Fig. 4.25,

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. 2.4(D).

### 4.11 Properties of Induction Titration Curves

In this section, we expand on the phenotypic properties of the induction response that were explored in Chapter 2 (see Fig. 2.1). 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 Chapter 2, 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. Using Eqs. 2.6, 2.7, and 2.8, the dynamic range is given by

$$\text{dynamic range} = \left(1 + \frac{1}{1 + e^{-\beta \Delta \varepsilon_{AI}} \left(\frac{K_A}{K_I}\right)^n} \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_{RA}}\right)^{-1} - \left(1 + \frac{1}{1 + e^{-\beta \Delta \varepsilon_{AI}}} \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_{RA}}\right)^{-1}. \quad (4.35)$$

The dynamic range, saturation, and leakiness were plotted with our experimental data in Fig. 2.6(A)-(C) as a function of repressor copy number. Fig. 4.26 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. 2.6(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, $Re^{-\beta \Delta \varepsilon_{RA}}$. Hence, increasing $R$ on a logarithmic scale (as in Fig. 2.6(A)-(C)) is equivalent to decreasing $\Delta \varepsilon_{RA}$ on a linear scale (as in Fig. 4.26).

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 bind-
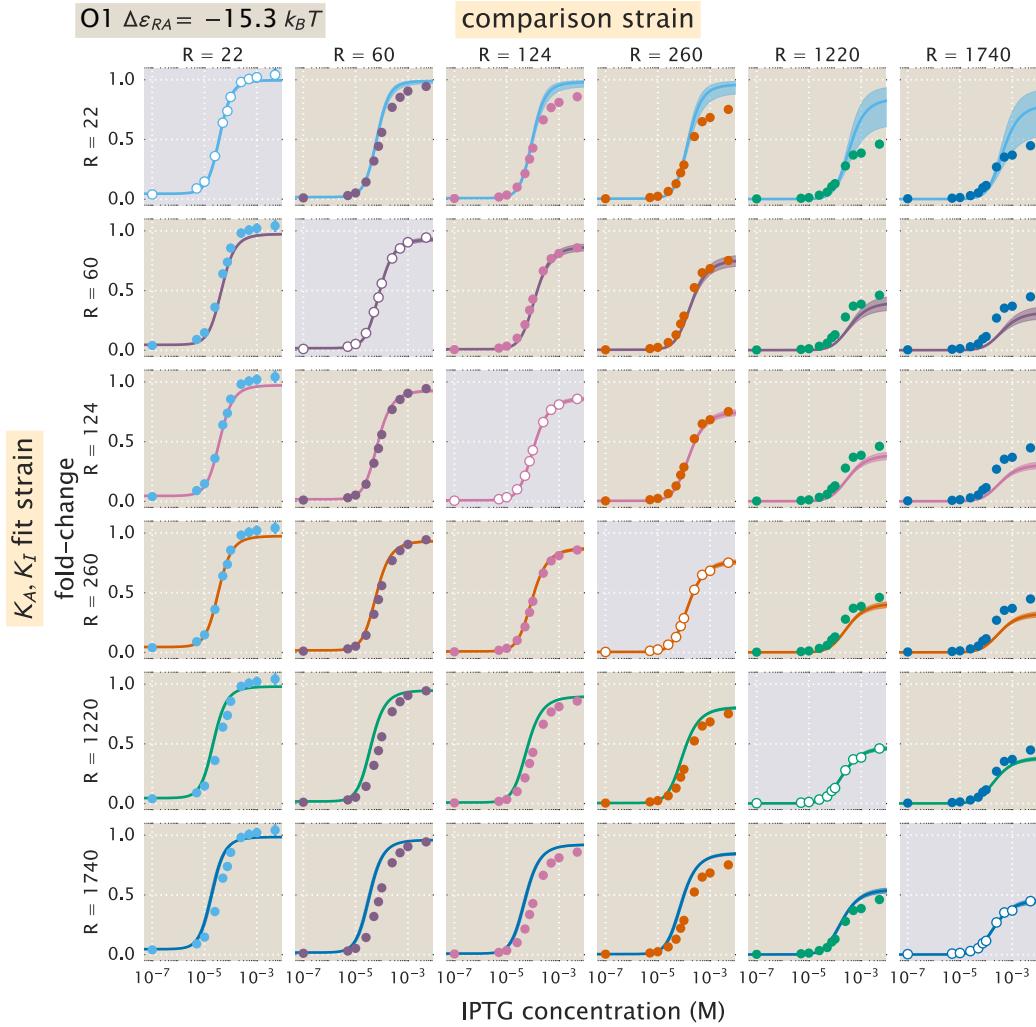
Figure 4.23: **O1 strain fold-change predictions based on strain-specific parameter estimation of $K_A$ and $K_I$.** Fold-change in expression is plotted as a function of IPTG concentration for all strains containing an O1 operator. The solid points correspond to the mean experimental value. The solid lines correspond to Eq. 2.5 using the parameter estimates of $K_A$ and $K_I$. Each row uses a single set of parameter values based on the strain noted on the left axis. The shaded plots along the diagonal are those where the parameter estimates are plotted along with the data used to infer them. Values for repressor copy number and operator binding energy are from [20]. The shaded region on the curve represents the uncertainty from our parameter estimates and reflects the 95% highest probability density region of the parameter predictions.
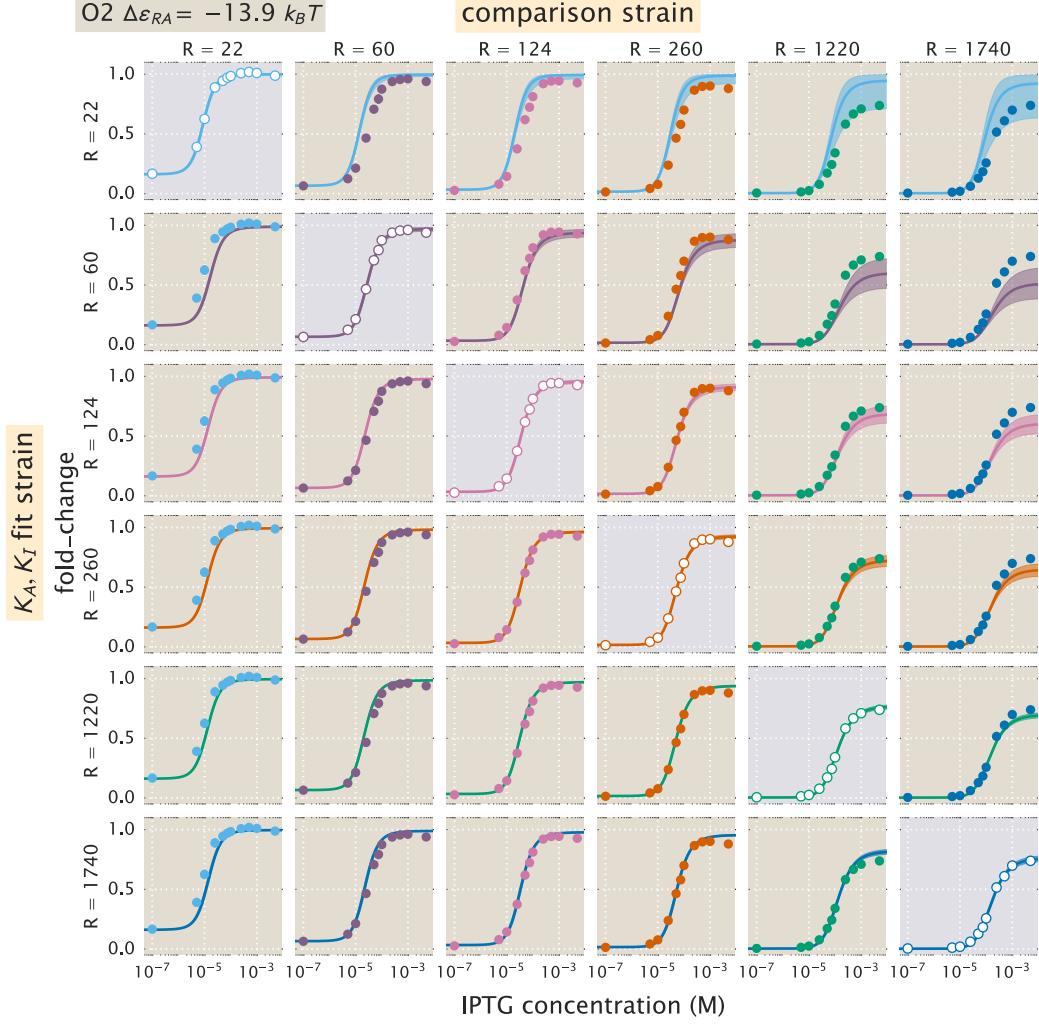
Figure 4.24: **O2 strain fold-change predictions based on strain-specific parameter estimation of $K_A$ and $K_I$.** Fold-change in expression is plotted as a function of IPTG concentration for all strains containing an O2 operator. The plots and data shown are analogous to Fig. 4.23, but for the O2 operator.

ing energy). Differentiating the dynamic range Eq. 4.35 and setting it equal to zero, we find that this peak occurs at

$$\frac{R^*}{N_{NS}} = e^{-\beta(\Delta\varepsilon_{AI} - \Delta\varepsilon_{RA})} \sqrt{e^{\Delta\varepsilon_{AI}} + 1} \sqrt{e^{\Delta\varepsilon_{AI}} + \left(\frac{K_A}{K_I}\right)^n}. \qquad (4.36)$$

The magnitude of the peak is given by

$$\text{max dynamic range} = \frac{\left(\sqrt{e^{\Delta\varepsilon_{AI}} + 1} - \sqrt{e^{\Delta\varepsilon_{AI}} + \left(\frac{K_A}{K_I}\right)^n}\right)^2}{\left(\frac{K_A}{K_I}\right)^n - 1}, \qquad (4.37)$$
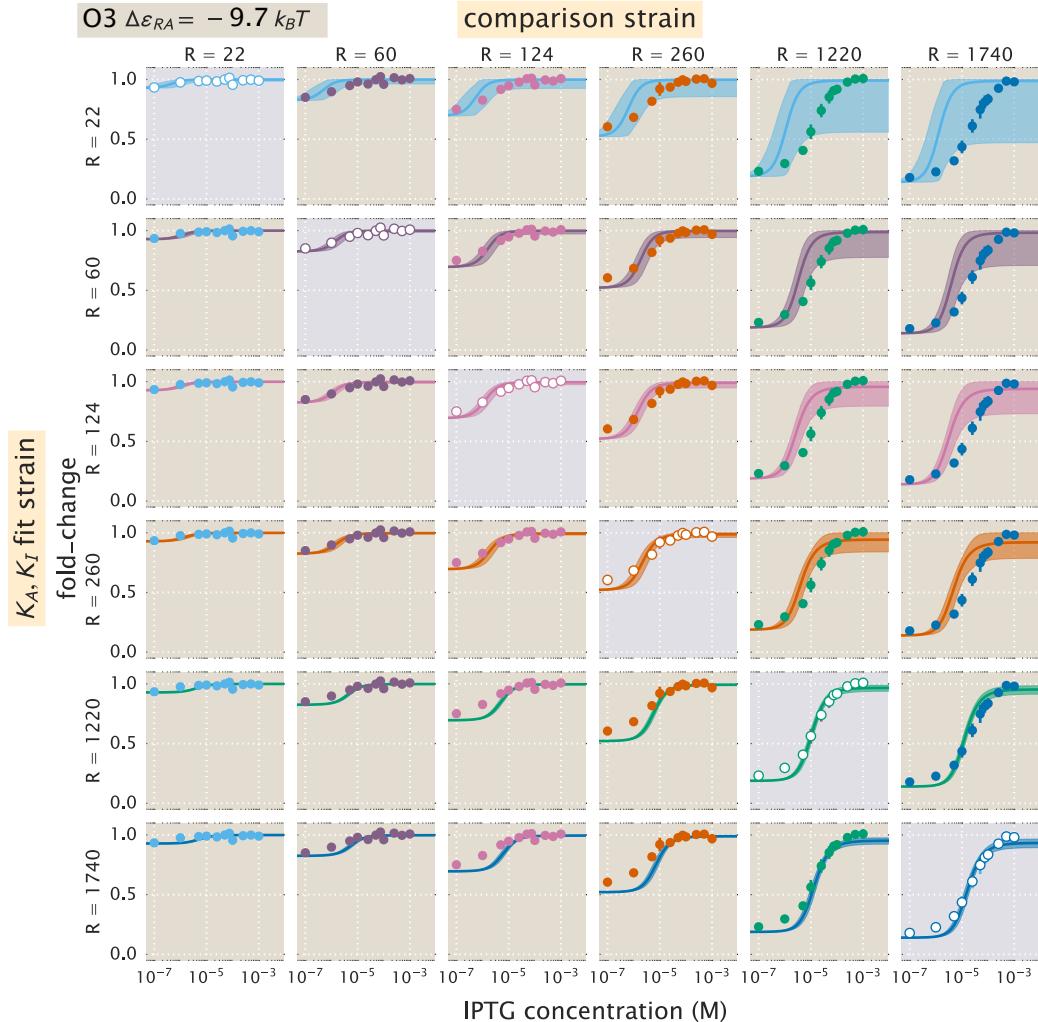
Figure 4.25: **O3 strain fold-change predictions based on strain-specific parameter estimation of $K_A$ and $K_I$.** Fold-change in expression is plotted as a function of IPTG concentration for all strains containing an O3 operator. The plots and data shown are analogous to Fig. 4.23, but for the O3 operator. We note that when using the $R = 22$ O3 strain to predict $K_A$ and $K_I$, the large uncertainty in the estimates of these parameters (see Fig. 2.4(D)) leads to correspondingly wider credible regions.
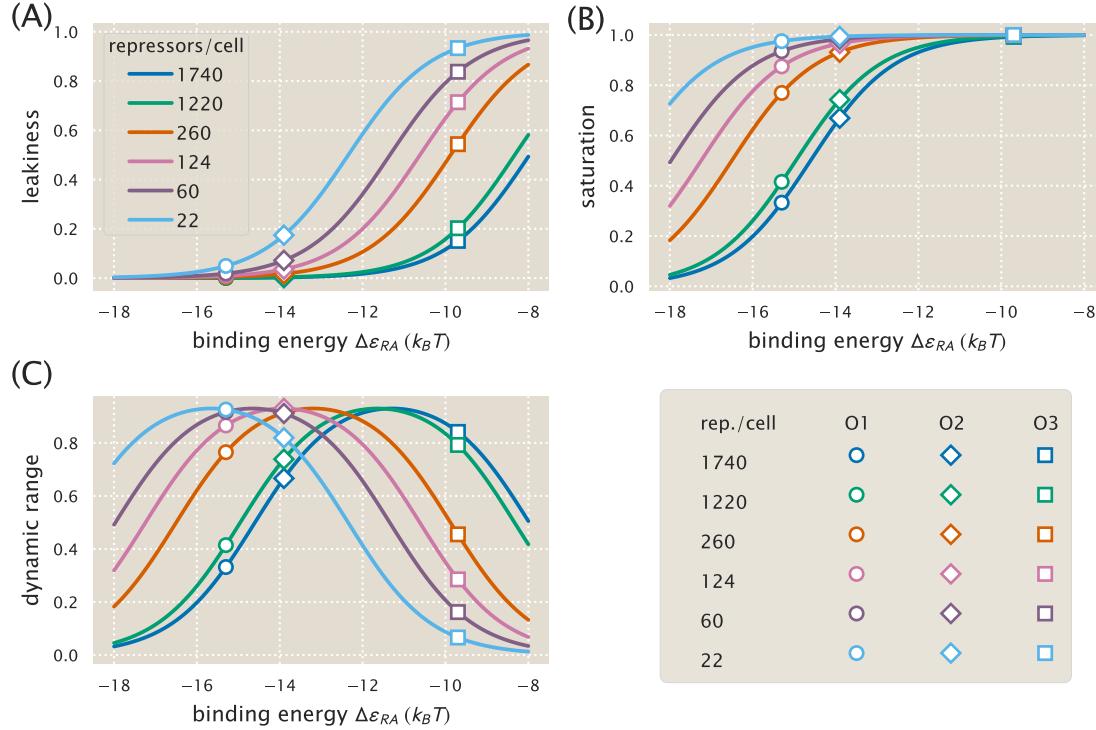
Figure 4.26: **Dependence of leakiness, saturation, and dynamic range on the operator binding energy and repressor copy number.** Increasing the repressor copy number or decreasing the repressor-operator binding energy suppresses gene expression and decreases the leakiness and saturation. The dynamic 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 $\Delta\varepsilon_{RA}$: for small repressor-operator binding energies, the leakiness is small, but the saturation increases with $\Delta\varepsilon_{RA}$; for large repressor-operator binding energies, the saturation is near unity, and the leakiness increases with $\Delta\varepsilon_{RA}$, thereby decreasing the dynamic range. Repressor copy number does not affect the maximum dynamic range. Circles, diamonds, and squares represent $\Delta\varepsilon_{RA}$ values for the O1, O2, and O3 operators, respectively, demonstrating the expected values of the properties using those strains.

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
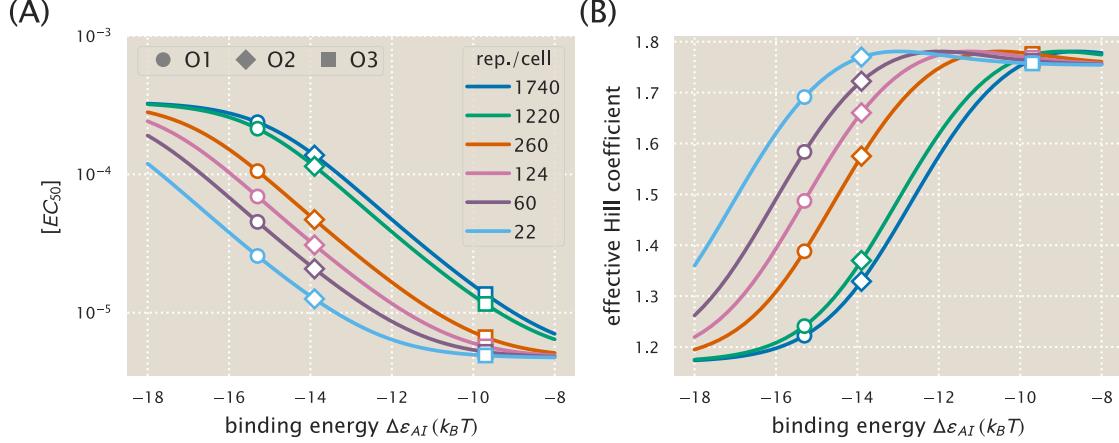
Figure 4.27: $[EC_{50}]$ **and effective Hill coefficient depend strongly on repressor copy number and operator binding energy.** $[EC_{50}]$ values range from very small and tightly clustered at weak operator binding energies (e.g. O3) to relatively large and spread out for stronger operator binding energies (O1 and O2). 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. Circles, diamonds, and squares represent $\Delta\varepsilon_{RA}$ values for the O1, O2, and O3 operators, respectively.

$[EC_{50}]$ is given by

$$\frac{[EC_{50}]}{K_A} = \frac{\frac{K_A}{K_I} - 1}{\frac{K_A}{K_I} - \left(\frac{\left(1+\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)+\left(\frac{K_A}{K_I}\right)^n\left(2e^{-\beta\Delta\varepsilon_{AI}}+\left(1+\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)\right)}{2\left(1+\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)+e^{-\beta\Delta\varepsilon_{AI}}+\left(\frac{K_A}{K_I}\right)^n e^{-\beta\Delta\varepsilon_{AI}}}\right)^{\frac{1}{n}}} - 1. \quad (4.38)$$

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 $c = [EC_{50}]$ (see Eq. 2.10). In Fig. 2.6(D)-(E), we show how these two properties vary with repressor copy number, and in Fig. 4.27 we demonstrate how they depend on the repressor-operator binding energy. Both the $[EC_{50}]$ and $h$ vary significantly with repressor copy numbers 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 precisely 2.

## 4.12 Applications to Other Regulatory Architectures

This section discusses how the theoretical framework presented in this work is sufficiently general to include various regulatory architectures outside of simple repression by LacI. We begin by noting that the same formula for fold-change given in Eq. 2.5 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 [11,54].

**Corepression**

Consider a regulatory architecture where binding of a transcriptional repressor impedes 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 Chapter 2, 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 Fig. 2.2 and yield the same fold-change equation as Eq. 2.5,

$$\text{fold-change} \approx \left(1 + \frac{\left(1 + \frac{c}{K_A}\right)^n}{\left(1 + \frac{c}{K_A}\right)^n + e^{\beta\Delta\varepsilon_{AI}}\left(1 + \frac{c}{K_I}\right)^n}\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1}, \quad (4.39)$$

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 from the induction data (see Fig. 2.4). Much as was performed in Chapter 2, we can make some predictions about the response of a corepressor. In Fig. 4.28(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 the 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 that 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. The 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

$$\text{fold-change} = \frac{1 + \dfrac{\left(1+\frac{c}{K_A}\right)^n}{\left(1+\frac{c}{K_A}\right)^n + e^{\beta\Delta\varepsilon_{AI}}\left(1+\frac{c}{K_I}\right)^n} \dfrac{A}{N_{NS}} e^{-\beta\Delta\varepsilon_{AA}} e^{-\beta\varepsilon_{AP}}}{1 + \dfrac{\left(1+\frac{c}{K_A}\right)^n}{\left(1+\frac{c}{K_A}\right)^n + e^{\beta\Delta\varepsilon_{AI}}\left(1+\frac{c}{K_I}\right)^n} \dfrac{A}{N_{NS}} e^{-\beta\Delta\varepsilon_{AA}}}, \qquad (4.40)$$

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 and when both RNAP and the activator are simultaneously bound to the DNA. Fig. 4.28(B) 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 ($< 0.5\ k_BT$) can dramatically increase fold-change.

As in the case of induction, the Eq. 4.40 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. Thus, while these cases might be viewed as separate biological phenomena, they can all be described by the same underlying formalism mathematically.

## 4.13   Definition of the non-specific background $N_{NS}$

In this section, we will explore the definition of the non-specific background $N_{NS}$. As raised by an anonymous reviewer, the nature of this parameter seems to raise some controversy on what the right value should be, or whether or not the arbitrary definition of its value should also be applied to the $\Delta\varepsilon_{AI}$ parameter.

Specifically, during the first round, a reviewer did not like the idea that the value of $N_{NS} = 4.6 \times 10^6$ assumed that the entirety of the genome was available for non-specific binding of the repressor. We will consider how reasonable this is at the end of the section. However, as we will show first, the specific value of $N_{NS}$ is analogous to the zero potential energy or the reference concentration state. Thus, it is only the free energy differences that matter at the end of the day. For the
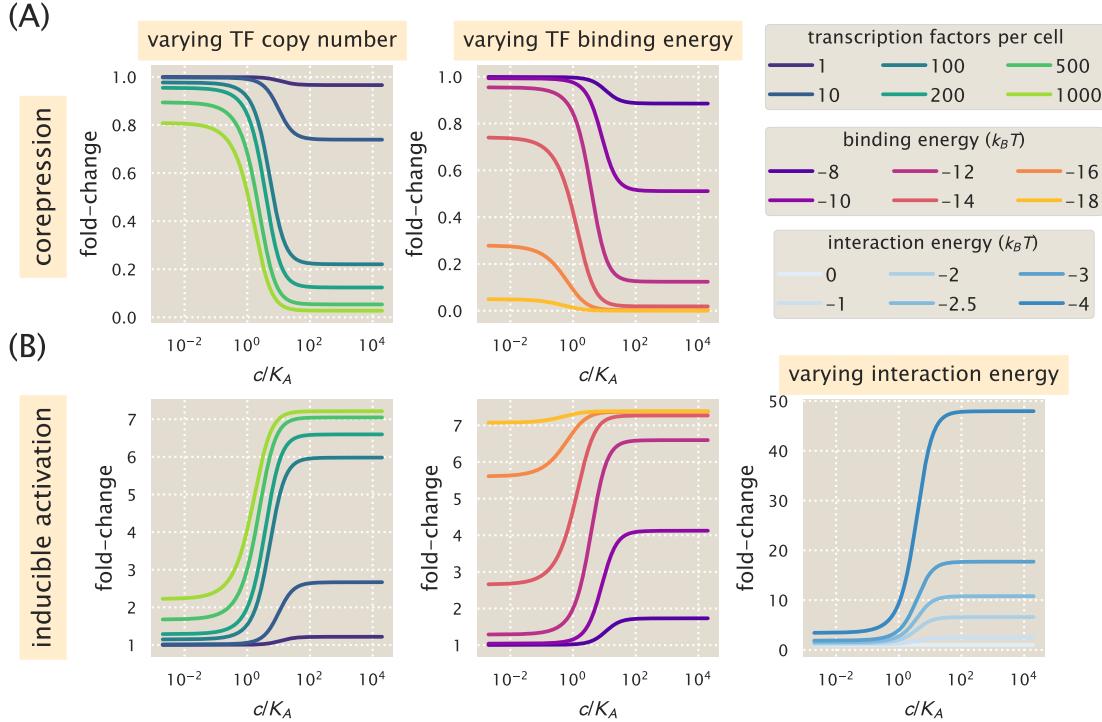
Figure 4.28: **Representative fold-change predictions for allosteric corepression and activation.** (A) Contrary to the case of induction described in Chapter 2, the addition of a corepressor decreases fold-change in gene expression. The left and right panels demonstrate how varying the values of the repressor copy number $R$ and repressor-DNA binding energy $\Delta\varepsilon_{RA}$, respectively, change the predicted response profiles. (B) In the case of inducible activation, binding of an effector molecule to an activator transcription factor increases the fold-change in gene expression. Note that for activation, the fold-change is greater than 1. The left and center panels show how changing the activator copy number $A$ and activator-DNA binding energy $\Delta\varepsilon_{AA}$ alter the response, respectively. The right panel shows how varying the polymerase-activator interaction energy $\varepsilon_{AP}$ alters the fold-change. Relatively small perturbations to this energetic parameter drastically change the level of activation and play a major role in dictating the dynamic range of the system.

second round of reviews, the same reviewer was willing to agree on our point if and only if we were to acknowledge that other parameters such as $\Delta\varepsilon_{AI}$, the free energy difference between the active and inactive state of the repressor, also had an arbitrary definition that could be set to any value. In this section, we will show that such a statement is an erroneous interpretation of the parameters. This free energy difference value cannot be re-defined to take any value if one is consistent with the experimental data.

Let us start by showing why the specific value of $N_{NS}$ is not the critical variable. Under the weak promoter approximation, the fold-change equation is equivalent

to a two-state Fermi function of having the promoter occupied by a repressor or having an empty promoter. This is

$$\text{fold-change} \rightarrow p_{bound}^r = \frac{1}{1 + \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_{RA}}}. \tag{4.41}$$

This expression can be rewritten as

$$p_{bound}^r = \frac{1}{1 + e^{-\beta \Delta E}}, \tag{4.42}$$

where $\Delta E$ is the free energy difference between the empty and occupied promoter. This definition implies that

$$\Delta E \equiv \overbrace{\Delta \varepsilon_{RA}}^{\text{enthalpic term}} \overbrace{- k_B T \ln \left( \frac{R}{N_{NS}} \right)}^{\text{entropic term}}. \tag{4.43}$$

Given that the parameter $\Delta \varepsilon_{RA}$ is inferred rather than directly measured, this puts us in the position of being able to re-define $N_{NS}$ at will as long as $\Delta E$ is in accordance with the experimental data. In other words, the parameter that matters is the free energy difference rather than its components. For example, if for a given operator and a given repressor copy number we choose a different value of $N_{NS}$, it still should hold that

$$\Delta E = \Delta \varepsilon_{RA}' - k_B T \ln \left( \frac{R}{N_{NS}'} \right), \tag{4.44}$$

where $N_{NS}'$ is the changed value of the non-specific background and $\Delta \varepsilon_{RA}'$ is a different value for the repressor binding energy that compensates for the difference in the non-specific background.

Let $N_{NS}' \equiv \alpha N_{NS}$. Since the value of $\Delta E$ has to be preserved, it should be true that

$$\Delta E = \Delta \varepsilon_{RA}' - k_B T \ln \left( \frac{R}{\alpha N_{NS}} \right) = \Delta \varepsilon_{RA} - k_B T \ln \left( \frac{R}{N_{NS}} \right). \tag{4.45}$$

Solving Eq. 4.45 for $\Delta \varepsilon_{RA}'$ gives

$$\Delta \varepsilon_{RA}' = \Delta \varepsilon_{RA} + k_B T \ln \left( \frac{N_{NS}}{\alpha N_{NS}} \right)$$
$$= \Delta \varepsilon_{RA} - k_B T \ln \alpha. \tag{4.46}$$

Eq. 4.46 implies that we can redefine $N_{NS}$ to be any value as long as $\Delta\varepsilon_{RA}$ compensates to maintain the value of $\Delta E$. This statement holds true whether we are considering a single promoter or multiple promoters. The same cannot be said about the $\Delta\varepsilon_{AI}$ parameter. The parameter $\Delta\varepsilon_{AI}$ by itself sets the fraction of inactive repressors in the absence of inducer via

$$p_{act} = \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}}, \tag{4.47}$$

where we have a Fermi function for a two-state system in which the repressor can be in an active or inactive state again.

As shown before, the reason why we could define $N_{NS}$ to be any value is that the parameter that matters is itself $\Delta E$ the free energy difference. Therefore the repressor binding energy $\Delta\varepsilon_{RA}$ could compensate for changes in the value of $N_{NS}$. For the case of $\Delta\varepsilon_{AI}$ Eq. 4.47 tells us that $\Delta\varepsilon_{AI}$ has no entropic term that can be compensated with an enthalpic term, or vice versa.

One could argue that for the case of a single promoter, the fold-change equation does allow this parameter to be re-defined arbitrarily since the full equation in the absence of inducer can be written as

$$\text{fold-change} = \frac{1}{1 + \left(\frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}}\right)\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}}. \tag{4.48}$$

So when we define the free energy $\Delta E$, we would include an extra term of the form

$$\Delta E = \Delta\varepsilon_{RA} - k_B T \left[\ln\left(\frac{R}{N_{NS}}\right) + \ln\left(\frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}}\right)\right]. \tag{4.49}$$

If we were only to use Eq. 4.49, the statement brought up by the anonymous reviewer would be true since changes in $\Delta\varepsilon_{AI}$ could be compensated by changes in $\Delta\varepsilon_{RA}$ or $N_{NS}$. But as specified in Sec. 4.2, this is not the case for cells with multiple promoters.

The case of multiple promoters can be handled using the Canonical ensemble as is or using the Grand Canonical ensemble as detailed in [40]. Our point is more clearly seen in the case of the Canonical ensemble. Under this formalism, the fold-change equation is given by [39]

$$\text{fold-change} = \frac{\sum_{m=0}^{\min(N,R)} \frac{R!}{(N_{NS})^m (R-m)!} \binom{N}{m} e^{-\beta m \Delta \varepsilon_{RA}} (N-m)}{N \sum_{m=0}^{\min(N,R)} \frac{R!}{(N_{NS})^m (R-m)!} \binom{N}{m} e^{-\beta m \Delta \varepsilon_{RA}}}, \quad (4.50)$$

where $N$ is the number of promoters. Notice that we can group the terms including $N_{NS}$ and $\Delta \varepsilon_{RA}$ as

$$\text{fold-change} = \frac{\sum_{m=0}^{\min(N,R)} \frac{R!}{(R-m)!} \binom{N}{m} \left( \frac{e^{-\beta \Delta \varepsilon_{RA}}}{N_{NS}} \right)^m (N-m)}{N \sum_{m=0}^{\min(N,R)} \frac{R!}{(R-m)!} \binom{N}{m} \left( \frac{e^{-\beta \Delta \varepsilon_{RA}}}{N_{NS}} \right)^m}, \quad (4.51)$$

to highlight that it is a combination of these two parameters that matter, rather than their individual values. For the case of the $\Delta \varepsilon_{AI}$ parameter, this is not the case. Every term containing $R$ on Eq. 4.51 is effectively multiplied by Eq. 2.4. Since these terms are included inside the factorials, it is not true that a simple compensation by the other parameters allows us to define $\Delta \varepsilon_{AI}$ to be any value. Therefore as defined in Sec. 4.3, the parameter $\Delta \varepsilon_{AI}$ can be independently inferred using multiple promoter measurements of fold change.

As a final note, we can also check whether $N_{NS} = 4.6 \times 10^6$ is at all a reasonable value to use. One potential point of concern is whether the chromosomal DNA is occupied by other transcription factors that may reduce the availability of the DNA for repressor or RNAP to bind. Here we consider data from a recent census of protein abundance across the *E. coli* genome. In that work, Schmidt *et al.* [135] measured the protein copy number across more than half the coding genes (greater than 95% by total protein mass). During exponential growth in M9 minimal media with 0.5 % glucose, they found that about 6 % of the protein mass, or 311,000 monomer copies per cell, are proteins such as transcription factors that will be bound to the DNA (about two-thirds of these are nucleoid-associated proteins such as HNS and HU).

To make a simple estimate of DNA occupancy, let us assume that all transcription factors bind DNA as dimers and occupy a DNA length of 15 bp (this appears to vary from 7 bp to 38 bp in *E. coli* on RegulonDB [141]), we find that about 2.3 kbp or about half of the genome will be occupied. In the most extreme case, we could assume that this fraction is inaccessible, which would reduce $N_N S$ by a factor of about 2. Applying this to Eq. 4.46, we see that this has a negligible effect on the actual binding energy that we would infer and only corresponds to a change in energy $\varepsilon_{RA}$ by about $0.7\,k_B T$.

## 4.14 Measurement of Steady State

All measurements have been performed with cells in an exponential growth phase, where we expect an average expression to be maintained across the cell population. Here we wanted to use one of our strains (O2 $\Delta\varepsilon_{RA} = -13.9\,k_B T$, $R = 260$) to show that gene expression is under steady-state for our experimental conditions. As a reminder, we begin by growing an overnight culture in Lysogeny Broth for each of the required strains under our standard protocol. After approximately 12 hours, the saturated cultures are diluted 1000-fold into a 2 mL 96-deep-well plate. Each well contains 500 $\mu$L of M9 minimal media supplemented with 0.5% w/v glucose and the appropriate IPTG inducer concentration.
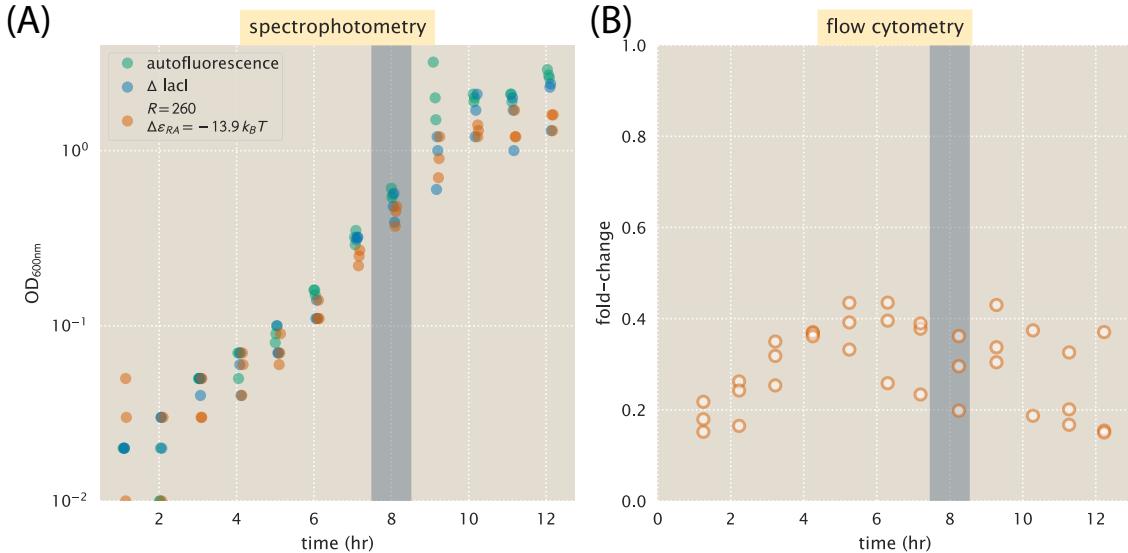
Figure 4.29: **Time course measurement of single-cell fluorescence by flow cytometry—data set 1.** Flow cytometry measurements were performed at different time points following a 1000-fold dilution of an overnight culture. Cell strains were grown in M9 minimal media supplemented with 0.5% w/v glucose and IPTG $c = 50\,\mu$M. $OD_{600nm}$ measurements are shown for the three strains. (B) The fold-change is calculated for each measurement shown in Panel (A). Note that each measurement represents a different culture grown in a 96-deep-well plate.

Here we follow the protocol as noted above but take measurements in one-hour increments after the 1000-fold dilution. We performe this in triplicate with our O2 $\Delta\varepsilon_{RA} = -13.9\,k_B T$, $R = 260$ strain (IPTG inducer concentration $c = 50\,\mu$M), and also include an autofluorescence strain and O2 $\Delta lacI$ strain. In Fig. 4.29(A) we plot the optical density ($OD_{600nm}$) as a function of time and see that growth is reasonably consistent between strains and their replicates. The shaded gray bar indicates an OD 0.3, which is the density at which we typically make our measurements. In Fig. 4.29(B), we show the associated fold-change measurements (using flow cytometry). While it does look like there is a steady increase in fold-change from 0 to 4 hours, it seems to level off past this time point. However, there was also a large degree of variation in our measurements, making it difficult to say that the fold-change is not changing over time.

In Fig. 4.31, we also plot the raw fluorescence values against the measured $OD_{600nm}$ values. The variation is rather large, but it does appear that the overall expression is relatively constant across two decades of $OD_{600nm}$.
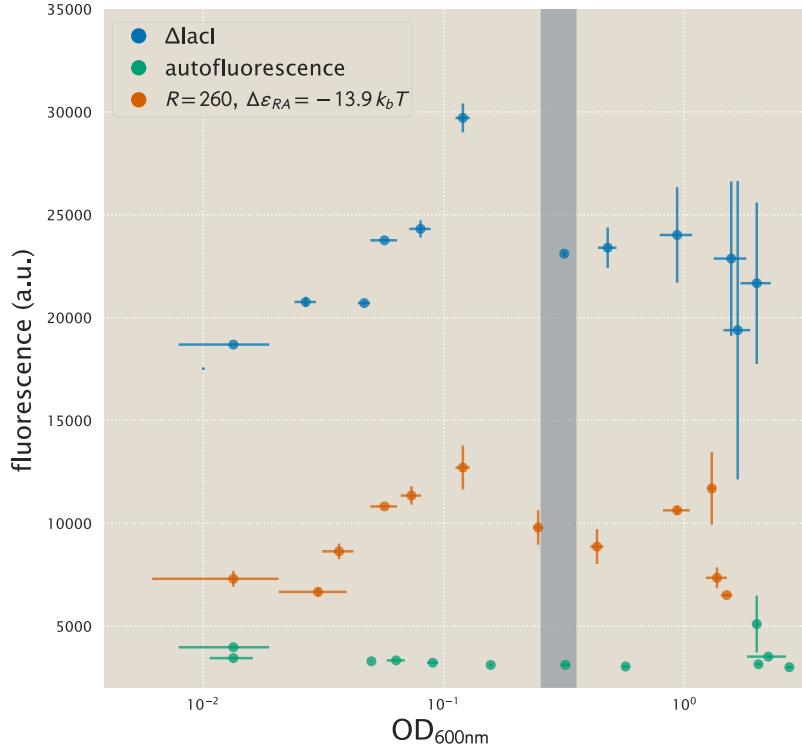
Figure 4.30: **Time course measurement of single-cell fluorescence versus OD$_{600nm}$—data set 1.** Fluorescence measurements used to calculate fold-change from Fig. 4.29(B) are plotted against their OD$_{600nm}$. Error bars represent standard deviation from the triplicate culture measurements from growth in a 96-deep-well plate.

In a separate set of replicates, we observed more consistent fold-change measurements over these later time points. Fig. 4.31(A) shows the average single-cell fluorescence from these measurements. While there does appear to be a downward trend in both the $R = 260$ strain and the $\Delta lacI$ strain, this is perhaps due to cultures leaving exponential growth (mistakenly, OD$_{600nm}$ was not measured in this attempt). In contrast to the data in Fig. 4.30(B), we found that fold-change did not appreciably change across these measurements (Fig. 4.31(B)). Given the differences across these two sets of experiments, it will be essential to perform more experiments before drawing any definite opinions about the above results.
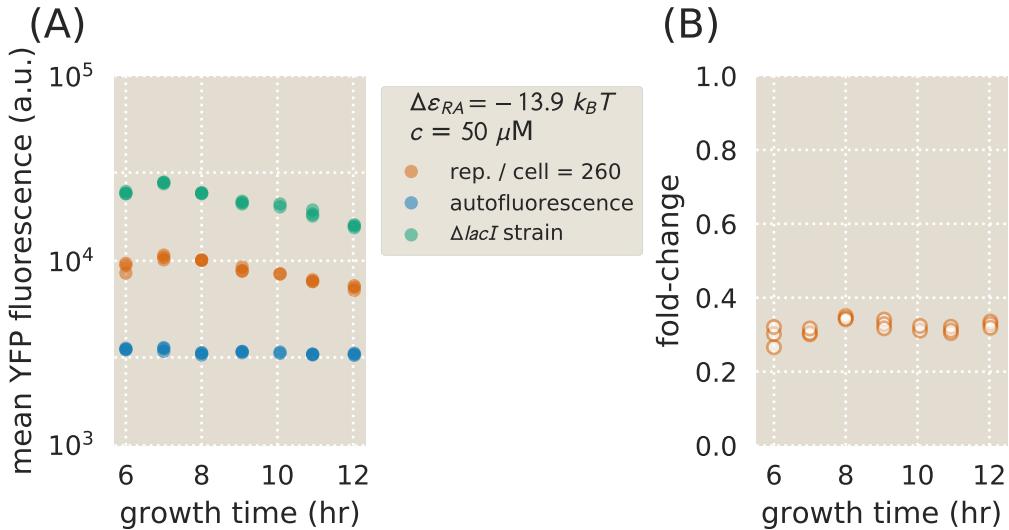
Figure 4.31: **Time course measurement of single-cell fluorescence by flow cytometry - data set 2.** Flow cytometry measurements were performed at different time points following a 1000-fold dilution of an overnight culture. Cell strains were grown in M9 minimal media supplemented with 0.5% w/v glucose and IPTG $c = 50\,\mu$M. Mean fluorescence values are shown for strain O2 $\Delta\varepsilon_{RA} = -13.9\,k_BT$ with $R = 260$, O2 $\Delta lacI$, and an autofluorescence strain. Data points represent measurements from separate $500\,\mu$L cell cultures. The fold change is calculated for each measurement shown in Panel .

### *E. coli* **Primer and Strain List**

Here we provide additional details about the strains' genotypes and 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 [88] as described in [20]. We follow the notation of Lutz and Bujard [79] 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 the expression of ei-

ther YFP or LacI (1 = $P_{LtetO-1}$, and 5 = *lacUV5*). Note that in 25x+11-yfp, x refers to the LacI operator used, which is centered at +11 (or begins at the transcription start site). For the different LacI constructs, 3*1x-lacI, x refers to the different ribosomal binding site modifications that provide different repressor copy numbers and follows from [20]. 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 4.4. 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*1x-*lacI*, mCherry is driven by a $P_{LtetO-1}$ promoter.

Table 4.4: *E. coli* **strains used in this work.** Each strain contains a unique operator-yfp construct for measurement of fluorescence, and *R* refers to the dimer copy number as measured by [20].

| Strain | Genotype |
| --- | --- |
| O1, $R = 0$ | HG105::galK <>25O1+11-yfp |
| O1, $R = 22$ | HG104::galK <>25O1+11-yfp |
| O1, $R = 60$ | HG105::galK <>25O1+11-yfp, ybcN <>3*1RBS1147-lacI |
| O1, $R = 124$ | HG105::galK <>25O1+11-yfp, ybcN <>3*1RBS1027-lacI |
| O1, $R = 260$ | HG105::galK <>25O1+11-yfp, ybcN <>3*1RBS446-lacI |
| O1, $R = 1220$ | HG105::galK <>25O1+11-yfp, ybcN <>3*1RBS1-lacI |
| O1, $R = 1740$ | HG105::galK <>25O1+11-yfp, ybcN <>3*1-lacI (RBS1L) |
| O2, $R = 0$ | HG105::galK <>25O2+11-yfp |
| O2, $R = 22$ | HG104::galK <>25O2+11-yfp |
| O2, $R = 60$ | HG105::galK <>25O2+11-yfp, ybcN <>3*1RBS1147-lacI |
| O2, $R = 124$ | HG105::galK <>25O2+11-yfp, ybcN <>3*1RBS1027-lacI |
| O2, $R = 260$ | HG105::galK <>25O2+11-yfp, ybcN <>3*1RBS446-lacI |

| Strain | Genotype |
| --- | --- |
| O2, $R = 1220$ | HG105::galK <>25O2+11-yfp, ybcN <>3*1RBS1-lacI |
| O2, $R = 1740$ | HG105::galK <>25O2+11-yfp, ybcN <>3*1-lacI (RBS1L) |
| O3, $R = 0$ | HG105::galK <>25O3+11-yfp |
| O3, $R = 22$ | HG104::galK <>25O3+11-yfp |
| O3, $R = 60$ | HG105::galK <>25O3+11-yfp, ybcN <>3*1RBS1147-lacI |
| O3, $R = 124$ | HG105::galK <>25O3+11-yfp, ybcN <>3*1RBS1027-lacI |
| O3, $R = 260$ | HG105::galK <>25O3+11-yfp, ybcN <>3*1RBS446-lacI |
| O3, $R = 1220$ | HG105::galK <>25O3+11-yfp, ybcN <>3*1RBS1-lacI |
| O3, $R = 1740$ | HG105::galK <>25O3+11-yfp, ybcN <>3*1-lacI (RBS1L) |
| Oid, $R = 0$ | HG105::galK <>25Oid+11-yfp |
| Oid, $R = 22$ | HG104::galK <>25Oid+11-yfp |
| Oid, $R = 60$ | HG105::galK <>25Oid+11-yfp, ybcN <>3*1RBS1147-lacI |
| Oid, $R = 124$ | HG105::galK <>25Oid+11-yfp, ybcN <>3*1RBS1027-lacI |
| Oid, $R = 260$ | HG105::galK <>25Oid+11-yfp, ybcN <>3*1RBS446-lacI |
| Oid, $R = 1220$ | HG105::galK <>25Oid+11-yfp, ybcN <>3*1RBS1-lacI |
| Oid, $R = 1740$ | HG105::galK <>25Oid+11-yfp, ybcN <>3*1-lacI (RBS1L) |

*Chapter 5*

# SUPPORTING INFORMATION FOR FIRST-PRINCIPLES PREDICTION OF THE INFORMATION PROCESSING CAPACITY OF A SIMPLE GENETIC CIRCUIT

## 5.1    Abstract

Given the stochastic nature of gene expression, genetically identical cells exposed to the same environmental inputs will produce different outputs. This heterogeneity has been hypothesized to have consequences for how cells are able to survive in changing environments. Recent work has explored the use of information theory as a framework to understand the accuracy with which cells can ascertain the state of their surroundings. Yet the predictive power of these approaches is limited and has not been rigorously tested using precision measurements. To that end, we generate a minimal model for a simple genetic circuit in which all parameter values for the model come from independently published data sets. We then predict the information processing capacity of the genetic circuit for a suite of biophysical parameters such as protein copy number and protein-DNA affinity. We compare these parameter-free predictions with an experimental determination of protein expression distributions and the resulting information processing capacity of *E. coli* cells. We find that our minimal model captures the scaling of the cell-to-cell variability in the data and the inferred information processing capacity of our simple genetic circuit up to a systematic deviation.

## 5.2 Three-State Promoter Model for Simple Repression

To tackle the question of how much information the simple repression motif can process, we require the joint probability distribution of mRNA and protein $P(m, p; t)$. To obtain this distribution, we use the chemical master equation formalism. Specifically, we assume a three-state model, where the promoter can be found 1) in a transcriptionally active state ($A$ state), 2) in a transcriptionally inactive state without the repressor bound ($I$ state), and 3) with the repressor bound ($R$ state). (see Fig. 3.2(A)). These three states generate a system of coupled differential equations for each of the three state distributions $P_A(m, p)$, $P_I(m, p)$, and $P_R(m, p)$. Given the rates shown in Fig. 3.2(A), let us define the system of ODEs. For the transcriptionally active state, we have

$$
\begin{aligned}
\frac{dP_A(m, p)}{dt} = & -\overbrace{k_{\text{off}}^{(p)} P_A(m, p)}^{A \to I} + \overbrace{k_{\text{on}}^{(p)} P_I(m, p)}^{I \to A} \\
& + \overbrace{r_m P_A(m-1, p)}^{m-1 \to m} - \overbrace{r_m P_A(m, p)}^{m \to m+1} + \overbrace{\gamma_m(m+1) P_A(m+1, p)}^{m+1 \to m} - \overbrace{\gamma_m m P_A(m, p)}^{m \to m-1} \\
& + \overbrace{r_p m P_A(m, p-1)}^{p-1 \to p} - \overbrace{r_p m P_A(m, p)}^{p \to p+1} + \overbrace{\gamma_p(p+1) P_A(m, p+1)}^{p+1 \to p} - \overbrace{\gamma_p p P_A(m, p)}^{p \to p-1}.
\end{aligned}
\tag{5.1}
$$

For the inactive promoter state $I$, we have

$$
\begin{aligned}
\frac{dP_I(m, p)}{dt} = & \overbrace{k_{\text{off}}^{(p)} P_A(m, p)}^{A \to I} - \overbrace{k_{\text{on}}^{(p)} P_I(m, p)}^{I \to A} + \overbrace{k_{\text{off}}^{(r)} P_R(m, p)}^{R \to I} - \overbrace{k_{\text{on}}^{(r)} P_I(m, p)}^{I \to R} \\
& + \overbrace{\gamma_m(m+1) P_I(m+1, p)}^{m+1 \to m} - \overbrace{\gamma_m m P_I(m, p)}^{m \to m-1} \\
& + \overbrace{r_p m P_I(m, p-1)}^{p-1 \to p} - \overbrace{r_p m P_I(m, p)}^{p \to p+1} + \overbrace{\gamma_p(p+1) P_I(m, p+1)}^{p+1 \to p} - \overbrace{\gamma_p p P_I(m, p)}^{p \to p-1}.
\end{aligned}
\tag{5.2}
$$

And finally, for the repressor bound state $R$, we have

$$
\frac{dP_R(m,p)}{dt} = -\overbrace{k_{\text{off}}^{(r)}P_R(m,p)}^{R\to I} + \overbrace{k_{\text{on}}^{(r)}P_I(m,p)}^{I\to R}
$$

$$
+ \overbrace{\gamma_m(m+1)P_R(m+1,p)}^{m+1\to m} - \overbrace{\gamma_m m P_R(m,p)}^{m\to m-1}
$$

$$
+ \overbrace{r_p m P_R(m,p-1)}^{p-1\to p} - \overbrace{r_p m P_R(m,p)}^{p\to p+1} + \overbrace{\gamma_p(p+1)P_R(m,p+1)}^{p+1\to p} - \overbrace{\gamma_p p P_R(m,p)}^{p\to p-1}.
$$

$$(5.3)$$

For an unregulated promoter, i.e., a promoter in a cell that has no repressors present and therefore constitutively expresses the gene, we use a two-state model in which the state $R$ is not allowed. All the terms in the system of ODEs containing $k_{\text{on}}^{(r)}$ or $k_{\text{off}}^{(r)}$ are then set to zero.

It is convenient to express this system using matrix notation [26]. For this, we define $\mathbf{P}(m,p) = (P_A(m,p), P_I(m,p), P_R(m,p))^T$. Then the system of ODEs can be expressed as

$$
\begin{aligned}
\frac{d\mathbf{P}(m,p)}{dt} = {} & \mathbf{K}\mathbf{P}(m,p) - \mathbf{R}_m\mathbf{P}(m,p) + \mathbf{R}_m\mathbf{P}(m-1,p) \\
& - m\boldsymbol{\Gamma}_m\mathbf{P}(m,p) + (m+1)\boldsymbol{\Gamma}_m\mathbf{P}(m+1,p) \\
& - m\mathbf{R}_p\mathbf{P}(m,p) + m\mathbf{R}_p\mathbf{P}(m,p-1) \\
& - p\boldsymbol{\Gamma}_p\mathbf{P}(m,p) + (p+1)\boldsymbol{\Gamma}_p\mathbf{P}(m,p+1),
\end{aligned}
$$

$$(5.4)$$

where we defined matrices representing the promoter state transition $\mathbf{K}$,

$$
\mathbf{K} \equiv \begin{bmatrix} -k_{\text{off}}^{(p)} & k_{\text{on}}^{(p)} & 0 \\ k_{\text{off}}^{(p)} & -k_{\text{on}}^{(p)} - k_{\text{on}}^{(r)} & k_{\text{off}}^{(r)} \\ 0 & k_{\text{on}}^{(r)} & -k_{\text{off}}^{(r)} \end{bmatrix},
$$

$$(5.5)$$

mRNA production, $\mathbf{R}_m$, and degradation, $\boldsymbol{\Gamma}_m$, as

$$
\mathbf{R}_m \equiv \begin{bmatrix} r_m & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},
$$

$$(5.6)$$

and

$$\mathbf{\Gamma}_m \equiv \begin{bmatrix} \gamma_m & 0 & 0 \\ 0 & \gamma_m & 0 \\ 0 & 0 & \gamma_m \end{bmatrix}. \tag{5.7}$$

For the protein, we also define production $\mathbf{R}_p$ and degradation $\mathbf{\Gamma}_p$ matrices as

$$\mathbf{R}_p \equiv \begin{bmatrix} r_p & 0 & 0 \\ 0 & r_p & 0 \\ 0 & 0 & r_p \end{bmatrix} \tag{5.8}$$

and

$$\mathbf{\Gamma}_p \equiv \begin{bmatrix} \gamma_p & 0 & 0 \\ 0 & \gamma_p & 0 \\ 0 & 0 & \gamma_p \end{bmatrix}. \tag{5.9}$$

The corresponding equation for the unregulated two-state promoter takes the same form with the definition of the matrices following the same scheme without including the third row and third column, and setting $k_{\mathrm{on}}^{(r)}$ and $k_{\mathrm{off}}^{(r)}$ to zero.

A closed-form solution for this master equation might not even exist. The approximate solution of chemical master equations of this kind is an active area of research. As we will see later in this chapter, the two-state promoter master equation has been analytically solved for the mRNA [116] and protein distributions [142]. For our purposes, we will detail how to use the Maximum Entropy principle to approximate the full distribution for the two- and three-state promoter.

## 5.3   Parameter Inference

(Note: The Python code used for the calculations presented in this section can be found in the <span style="color:orange">following link</span> as an annotated Jupyter notebook.)

To generate falsifiable predictions with meaningful parameters, we infer the kinetic rates for this three-state promoter model using different data sets generated in our lab over the last decade concerning different aspects of the regulation of the simple repression motif. For example, for the unregulated promoter transition rates $k_{\mathrm{on}}^{(p)}$ and $k_{\mathrm{off}}^{(p)}$ and the mRNA production rate $r_m$, we use single-molecule

mRNA FISH counts from an unregulated promoter [84]. Once these parameters are fixed, we use the values to constrain the repressor rates $k_{\text{on}}^{(r)}$ and $k_{\text{off}}^{(r)}$. These repressor rates are obtained using information from mean gene expression measurements from bulk LacZ colorimetric assays [20]. We also expand our model to include the allosteric nature of the repressor protein, taking advantage of video microscopy measurements done in the context of multiple promoter copies [39] and flow-cytometry measurements of the mean response of the system to different levels of induction [113]. In what follows, we detail the steps taken to infer the parameter values. At each step, the values of the parameters inferred in previous steps constrain the values of the parameters that are not yet determined, building in this way a self-consistent model informed by work that spans several experimental techniques.

**Unregulated Promoter Rates**

We begin our parameter inference problem with the promoter on and off rates $k_{\text{on}}^{(p)}$ and $k_{\text{off}}^{(p)}$, as well as the mRNA production rate $r_m$. In this case, there are only two states available to the promoter— the inactive state $I$ and the transcriptionally active state $A$. That means that the third ODE for $P_R(m, p)$ is removed from the system. The mRNA steady-state distribution for this particular two-state promoter model was solved analytically by Peccoud and Ycart [116]. This distribution $P(m) \equiv P_I(m) + P_A(m)$ is of the form

$$P(m|k_{\text{on}}^{(p)}, k_{\text{off}}^{(p)}, r_m, \gamma_m) = \frac{\Gamma\left(\frac{k_{\text{on}}^{(p)}}{\gamma_m} + m\right)}{\Gamma(m+1)\Gamma\left(\frac{k_{\text{off}}^{(p)} + k_{\text{on}}^{(p)}}{\gamma_m} + m\right)} \frac{\Gamma\left(\frac{k_{\text{off}}^{(p)} + k_{\text{on}}^{(p)}}{\gamma_m}\right)}{\Gamma\left(\frac{k_{\text{on}}^{(p)}}{\gamma_m}\right)} \left(\frac{r_m}{\gamma_m}\right)^m {}_1F_1\left(\frac{k_{\text{on}}^{(p)}}{\gamma_m} + m, \frac{k_{\text{off}}^{(p)} + k_{\text{on}}^{(p)}}{\gamma_m} + m, -\frac{r_m}{\gamma_m}\right),$$

$$(5.10)$$
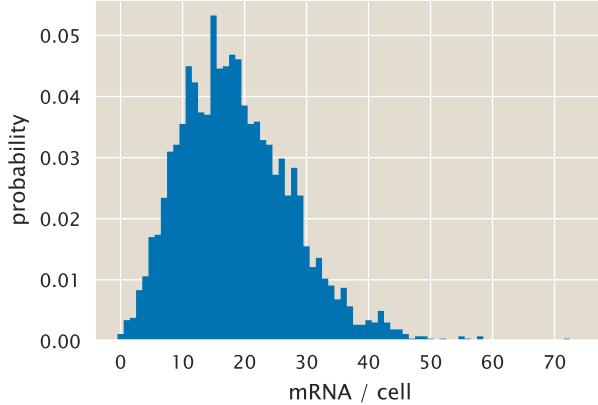
Figure 5.1: ***lacUV5* mRNA per cell distribution.** Data from [84] of the unregulated *lacUV5* promoter as inferred from single-molecule mRNA FISH. The Python code (`ch5_fig01.py`) used to generate this figure can be found on the original paper's GitHub repository.

where $\Gamma(\cdot)$ is the gamma function, and $F_1^1$ is the confluent hypergeometric function of the first kind. This rather complicated expression will aid us in finding parameter values for the rates. The inferred rates $k_{\text{on}}^{(p)}$, $k_{\text{off}}^{(p)}$, and $r_m$ will be expressed in units of the mRNA degradation rate $\gamma_m$. This is because the model in Eq. 5.10 is homogeneous in time, meaning that if we divide all rates by a constant, it would be equivalent to multiplying the characteristic time scale by the same constant. As we will discuss in the next section, Eq. 5.10 has degeneracy in the parameter values. What this means is that a change in one of the parameters, specifically $r_m$, can be compensated by a change in another parameter, specifically $k_{\text{off}}^{(p)}$, to obtain the same distribution. To work around this intrinsic limitation of the model, we will include information from what we know from equilibrium-based models in our inference prior.

**Bayesian Parameter Inference of RNAP Rates**

To make progress at inferring the unregulated promoter state transition rates, we make use of the single-molecule mRNA FISH data from Jones *et al.* [84]. Fig. 5.1 shows the distribution of mRNA per cell for the *lacUV5* promoter used for our inference. This promoter, being very strong, has a mean copy number of $\langle m \rangle \approx 18$ mRNA/cell.

Having these data in hand, we now turn to Bayesian parameter inference. Writing Bayes' theorem, we have

$$P(k_{\text{on}}^{(p)}, k_{\text{off}}^{(p)}, r_m \mid D) = \frac{P(D \mid k_{\text{on}}^{(p)}, k_{\text{off}}^{(p)}, r_m) P(k_{\text{on}}^{(p)}, k_{\text{off}}^{(p)}, r_m)}{P(D)}, \qquad (5.11)$$

where $D$ represents the data. For this case, the data consists of single-cell mRNA counts $D = \{m_1, m_2, \ldots, m_N\}$, where $N$ is the number of cells. We assume that each cell's measurement is independent of the others such that we can rewrite Eq. 5.11 as

$$P(k_{\text{on}}^{(p)}, k_{\text{off}}^{(p)}, r_m \mid \{m_i\}) \propto \left[ \prod_{i=1}^{N} P(m_i \mid k_{\text{on}}^{(p)}, k_{\text{off}}^{(p)}, r_m) \right] P(k_{\text{on}}^{(p)}, k_{\text{off}}^{(p)}, r_m), \qquad (5.12)$$

where we ignore the normalization constant $P(D)$. The likelihood term $P(m_i \mid k_{\text{on}}^{(p)}, k_{\text{off}}^{(p)}, r_m)$ is exactly given Eq. 5.10 by with $\gamma_m = 1$. Given that we have this functional form for the distribution, we can use Markov Chain Monte Carlo (MCMC) sampling to explore the 3D parameter space in order to fit 5.10 to the mRNA-FISH data.

**Constraining the Rates Given Prior Thermodynamic Knowledge**

One of the Bayesian approach's strengths is that we can include all the prior knowledge on the parameters when performing an inference [97]. Basic features such as the fact that the rates have to be strictly positive constrain these parameters' values. We know more than the simple constraint of non-negative values for the specific rates analyzed in this section. For example, the expression of an unregulated promoter has been studied from a thermodynamic perspective [43]. Given the underlying assumptions of these equilibrium models, in which the probability of finding the RNAP bound to the promoter is proportional to the transcription rate [50], they can only make statements about the mean expression level. Nevertheless, if both the thermodynamic and kinetic models describe the same process, the mean gene expression level predictions must agree. That means that we can

use what we know about the mean gene expression and how this is related to parameters such as molecule copy numbers and binding affinities to constrain the values that the rates in question can take.

In the case of this two-state promoter, it can be shown that the mean number of mRNA is given by [26] (see Sec. 5.3 for moment computation)

$$\langle m \rangle = \frac{r_m}{\gamma_m} \frac{k_{on}^{(p)}}{k_{on}^{(p)} + k_{off}^{(p)}}.$$ (5.13)

Another way of expressing this is as $\frac{r_m}{\gamma_m} \times p_{active}^{(p)}$, where $p_{active}^{(p)}$ is the probability of the promoter being in the transcriptionally active state. The thermodynamic picture has an equivalent result where the mean number of mRNA is given by [43,50]

$$\langle m \rangle = \frac{r_m}{\gamma_m} \frac{\frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_p}}{1 + \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_p}},$$ (5.14)

where $P$ is the number of RNAP per cell, $N_{NS}$ is the number of non-specific binding sites, $\Delta \varepsilon_p$ is the RNAP binding energy in $k_B T$ units and $\beta \equiv (k_B T)^{-1}$. Using Eq. 5.13 and Eq. 5.14, we can easily see that if these frameworks are to be equivalent, then it must be true that

$$\frac{k_{on}^{(p)}}{k_{off}^{(p)}} = \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_p},$$ (5.15)

or equivalently

$$\ln \left( \frac{k_{on}^{(p)}}{k_{off}^{(p)}} \right) = -\beta \Delta \varepsilon_p + \ln P - \ln N_{NS}.$$ (5.16)

To put numerical values into these variables, we can use information from the literature. The RNAP copy number is of order $P \approx 1000 - 3000$ RNAP/cell for a one-hour doubling time [53]. As for the number of non-specific binding sites and the binding energy, we have that $N_{NS} = 4.6 \times 10^6$ [50] and $-\beta \Delta \varepsilon_p \approx 5 - 7$ [43]. Given these values, we define a Gaussian prior for the log ratio of these two quantities of the form

$$P \left( \ln \left( \frac{k_{on}^{(p)}}{k_{off}^{(p)}} \right) \right) \propto \exp \left\{ -\frac{\left( \ln \left( \frac{k_{on}^{(p)}}{k_{off}^{(p)}} \right) - \left( -\beta \Delta \varepsilon_p + \ln P - \ln N_{NS} \right) \right)^2}{2\sigma^2} \right\},$$ (5.17)

where $\sigma$ is the variance that accounts for the uncertainty in these parameters. We include this prior as part of the prior term $P(k_{\text{on}}^{(p)}, k_{\text{off}}^{(p)}, r_m)$ of Eq. 5.12. We then use MCMC to sample the posterior distribution given by Eq. 5.12. Fig. 5.2 shows the MCMC samples of the posterior distribution. For the case of the $k_{\text{on}}^{(p)}$ parameter, there is a single symmetric peak. $k_{\text{off}}^{(p)}$ and $r_m$ have a rather long tail towards large values. The 2D projection of $k_{\text{off}}^{(p)}$ vs. $r_m$ shows that the model is sloppy, meaning that the two parameters are highly correlated. This feature is a common problem for many non-linear systems used in biophysics and systems biology [143]. What this implies is that we can change the value of $k_{\text{off}}^{(p)}$, and then compensate by a change in $r_m$ to maintain the shape of the mRNA distribution. Therefore, it is impossible for the data and the model to narrow down a single value for the parameters. Nevertheless, since we included the prior information on the rates as given by the analogous form between the equilibrium and non-equilibrium expressions for the mean mRNA level, we obtained a more constrained parameter value for the RNAP rates and the transcription rate we will take as the peak of this long-tailed distribution.

The inferred values $k_{\text{on}}^{(p)} = 4.3_{-0.3}^{+1}$, $k_{\text{off}}^{(p)} = 18.8_{-10}^{+120}$ and $r_m = 103.8_{-37}^{+423}$ are given in units of the mRNA degradation rate $\gamma_m$. Given the asymmetry of the parameter distributions, we report the upper and lower bound of the 95th percentile of the posterior distributions. Assuming a mean life-time for mRNA of $\approx$ 3 min (from this link), we have an mRNA degradation rate of $\gamma_m \approx 2.84 \times 10^{-3} s^{-1}$. Using this value gives the following values for the inferred rates: $k_{\text{on}}^{(p)} = 0.024_{-0.002}^{+0.005} s^{-1}$, $k_{\text{off}}^{(p)} = 0.11_{-0.05}^{+0.66} s^{-1}$, and $r_m = 0.3_{-0.2}^{+2.3} s^{-1}$.

Fig. 5.3 compares the experimental data from Fig. 5.1 with the resulting distribution obtained by substituting the most likely parameter values into Eq. 5.10. As we can see, this two-state model fits the data adequately.
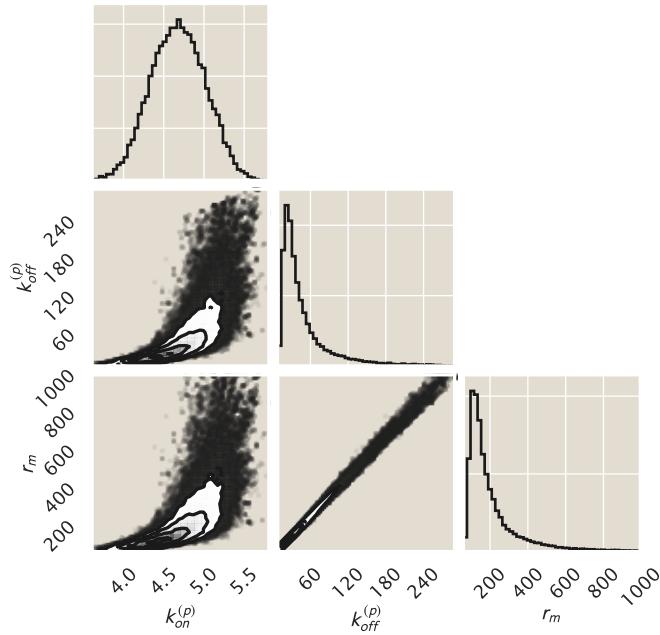
Figure 5.2: **MCMC posterior distribution.** Sampling out of Eq. 5.12, the plot shows 2D and 1D projections of the 3D parameter space. The parameter values are (in units of the mRNA degradation rate $\gamma_m$) $k_{on}^{(p)} = 4.3^{+1}_{-0.3}$, $k_{off}^{(p)} = 18.8^{+120}_{-10}$ and $r_m = 103.8^{+423}_{-37}$ which are the modes of their respective distributions, where the superscripts and subscripts represent the upper and lower bounds of the 95$^{th}$ percentile of the parameter value distributions. The Python code (`ch5_fig02.py`) used to generate this figure can be found on the original paper's GitHub repository.
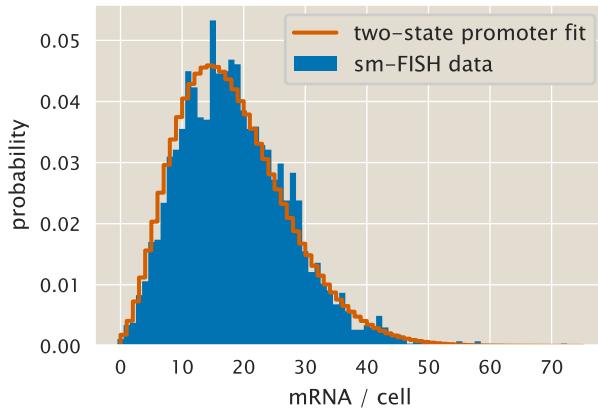


Figure 5.3: **Experimental vs. theoretical distribution of mRNA per cell using parameters from Bayesian inference.** Dotted line shows the result of using Eq. 5.10 along with the parameters inferred for the rates. Blue bars are the same data as Fig. 5.1 obtained from [84]. The Python code (`ch5_fig03.py`) used to generate this figure can be found on the original paper's GitHub repository.

**Accounting for Variability in the Number of Promoters**

As discussed in ref. [84] and further expanded in [118], an essential source of cell-to-cell variability in gene expression in bacteria is the fact that, depending on the growth rate and the position relative to the chromosome replication origin, cells can have multiple copies of any given gene. Genes closer to the replication origin have, on average, higher gene copy numbers compared to genes at the opposite end. For the locus in which our reporter construct is located (*galK*) and the doubling time of the mRNA FISH experiments, we expect to have $\approx 1.66$ copies of the gene [17,84]. This implies that the cells spend 2/3 of the cell cycle with two copies of the promoter and the rest with a single copy.

To account for this variability in gene copy, we extend the model assuming that when cells have two copies of the promoter, the mRNA production rate is $2r_m$ compared to the rate $r_m$ for a single promoter copy. The probability of observing a particular mRNA copy $m$ is therefore given by

$$P(m) = P(m|\text{one promoter}) \cdot P(\text{one promoter}) + P(m|\text{two promoters}) \cdot P(\text{two promoters}). \tag{5.18}$$

Both terms $P(m \mid \text{promoter copy})$ are given by Eq. 5.10 with the only difference being the rate $r_m$. It is important to acknowledge that Eq. 5.18 assumes that once the gene is replicated, the time scale in which the mRNA count relaxes to the new steady state is much shorter than the time that the cells spend in this two promoter copies state. This approximation should be valid for a short-lived mRNA molecule, but the assumption is not applicable for proteins whose degradation rate is comparable to the cell cycle length as explored in Sec. 5.5.

To repeat the Bayesian inference, including this variability in gene copy number, we must split the mRNA count data into two sets—cells with a single copy of the promoter and cells with two copies of the promoter. There is no labeling of the locus for the single-molecule mRNA FISH data, making it impossible to determine the promoter's number of copies for any given cell. We, therefore, follow Jones *et al.* [84] in using the cell area as a proxy for the stage in the cell cycle. They sorted cells by area in their approach, considering cells below the 33$^{\text{th}}$ percentile having
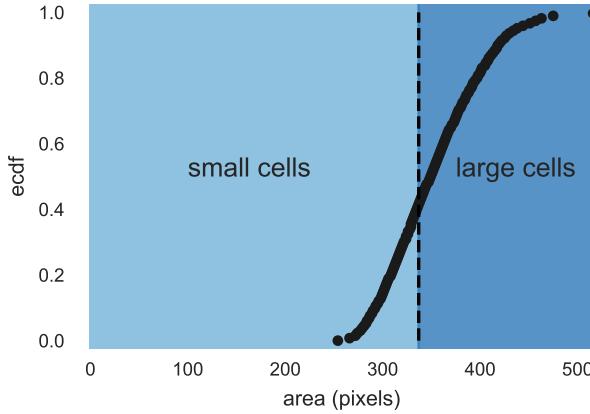
Figure 5.4: **Separation of cells based on cell size.** Using the area as a proxy for position in the cell cycle, cells can be sorted into two groups—small cells (with one promoter copy) and large cells (with two promoter copies). The vertical black line delimits the threshold that divides both groups as weighted by Eq. 5.19. The Python code (`ch5_fig04.py`) used to generate this figure can be found on the original paper's GitHub repository.

a single promoter copy and the rest as having two copies. This approach ignores that cells are not uniformly distributed along the cell cycle. As first derived in [120], populations of cells in a log-phase are exponentially distributed along the cell cycle. This distribution is of the form

$$P(a) = (\ln 2) \cdot 2^{1-a}, \tag{5.19}$$

where $a \in [0, 1]$ is the stage of the cell cycle, with $a = 0$ being the start of the cycle and $a = 1$ being the cell division (see Sec. 5.10 for a derivation of Eq. 5.19). Fig. 5.4 shows the separation of the two groups based on the area where Eq. 5.19 was used to weight the distribution along the cell cycle.

A subtle but important consequence of Eq. 5.19 is that computing any quantity for a single cell is not equivalent to computing the same quantity for a population of cells. For example, let us assume that we want to compute the mean mRNA copy number $\langle m \rangle$. For a single cell, this would be of the form

$$\langle m \rangle_{\text{cell}} = \langle m \rangle_1 \cdot f + \langle m \rangle_2 \cdot (1 - f), \tag{5.20}$$

where $\langle m \rangle_i$ is the mean mRNA copy number with $i$ promoter copies in the cell, and $f$ is the fraction of the cell cycle that cells spend with a single copy of the promoter. For a single cell, the probability of having a single promoter copy is equivalent to
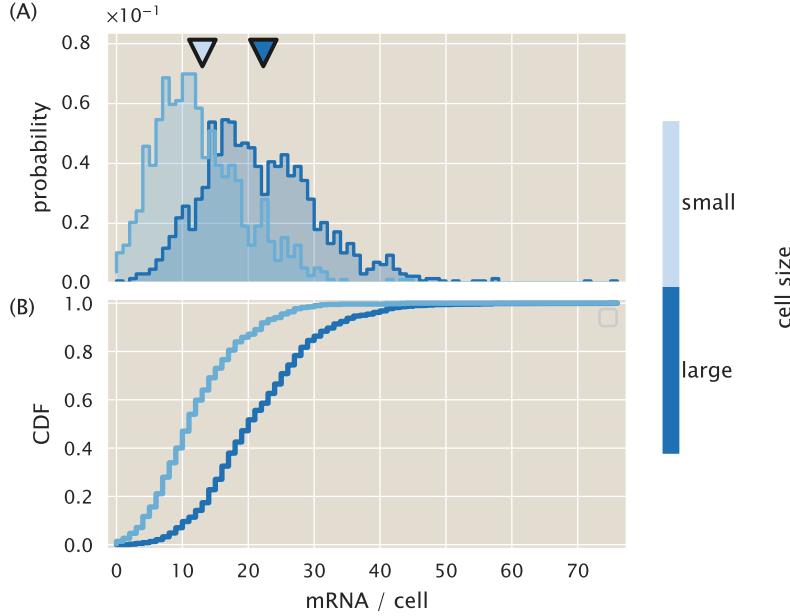
Figure 5.5: **mRNA distribution for small and large cells.** (A) histogram and (B) the cumulative distribution function of the small and large cells as determined in Fig. 5.4. The triangles above histograms in (A) indicate the mean mRNA copy number for each group. The Python code (ch5_fig05.py) used to generate this figure can be found on the original paper's GitHub repository.

this fraction $f$. But Eq. 5.19 tells us that if we sample unsynchronized cells, we are not sampling uniformly across the cell cycle. Therefore for a population of cells, the mean mRNA is given by

$$\langle m \rangle_{\text{population}} = \langle m \rangle_1 \cdot \phi + \langle m \rangle_2 \cdot (1 - \phi) \tag{5.21}$$

where the probability of sampling a cell with one promoter $\phi$ is given by

$$\phi = \int_0^f P(a)da, \tag{5.22}$$

where $P(a)$ is given by Eq. 5.19. What this equation computes is the probability of sampling a cell during a stage of the cell cycle $< f$ where the reporter gene has not been replicated yet. Fig. 5.5 shows the distribution of both groups. As expected, larger cells have a higher mRNA copy number on average.

We modify Eq. 5.12 to account for the two separate groups of cells. Let $N_s$ be the number of cells in the small size group and $N_l$ the number of cells in the large size group. Then the posterior distribution for the parameters is of the form

$$P(k_{\text{on}}^{(p)}, k_{\text{off}}^{(p)}, r_m | \{m_i\}) \propto \left[ \prod_{i=1}^{N_s} P(m_i | k_{\text{on}}^{(p)}, k_{\text{off}}^{(p)}, r_m) \right] \left[ \prod_{j=1}^{N_l} P(m_j | k_{\text{on}}^{(p)}, k_{\text{off}}^{(p)}, 2r_m) \right] P(k_{\text{on}}^{(p)}, k_{\text{off}}^{(p)}, r_m), \tag{5.23}$$

where we split the product of small and large cells.

For the two-promoter model, the prior shown in Eq. 5.17 requires a small modification. Eq. 5.21 gives the mean mRNA copy number of a population of asynchronous cells growing at a steady-state. Given that we assume that the only difference between having one vs. two promoter copies state is the change in transcription rate from $r_m$ in the single promoter case to $2r_m$ in the two-promoter case, we can write Eq. 5.21 as

$$\langle m \rangle = \phi \cdot \frac{r_m}{\gamma_m} \frac{k_{on}^{(p)}}{k_{on}^{(p)} + k_{off}^{(p)}} + (1 - \phi) \cdot \frac{2r_m}{\gamma_m} \frac{k_{on}^{(p)}}{k_{on}^{(p)} + k_{off}^{(p)}}. \tag{5.24}$$

This can be simplified to

$$\langle m \rangle = (2 - \phi) \frac{r_m}{\gamma_m} \frac{k_{on}^{(p)}}{k_{on}^{(p)} + k_{off}^{(p)}}. \tag{5.25}$$

Equating Eq. 5.25 and Eq. 5.14 to again require self-consistent predictions of the mean mRNA from the equilibrium and kinetic models gives

$$(2 - \phi) \frac{k_{on}^{(p)}}{k_{on}^{(p)} + k_{off}^{(p)}} = \frac{\frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_p}}{1 + \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_p}}. \tag{5.26}$$

Solving for $\frac{k_{on}^{(p)}}{k_{off}^{(p)}}$ results in

$$\left( \frac{k_{on}^{(p)}}{k_{off}^{(p)}} \right) = \frac{\rho}{[(1 + \rho)(2 - \phi) - \rho]}, \tag{5.27}$$

where we define $\rho \equiv \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_p}$. To simplify things further, we notice that for the specified values of $P = 1000 - 3000$ per cell, $N_{NS} = 4.6 \times 10^6$ bp, and $-\beta \Delta \varepsilon_p = 5 - 7$, we can safely assume that $\rho \ll 1$. This simplifying assumption has been previously called the weak promoter approximation [20]. Given this, we can simplify Eq. 5.27 as

$$\frac{k_{on}^{(p)}}{k_{off}^{(p)}} = \frac{1}{2 - \phi} \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_p}. \tag{5.28}$$

Taking the log of both sides gives

$$\ln \left( \frac{k_{on}^{(p)}}{k_{off}^{(p)}} \right) = -\ln(2 - \phi) + \ln P - \ln N_{NS} - \beta \Delta \varepsilon_p. \tag{5.29}$$
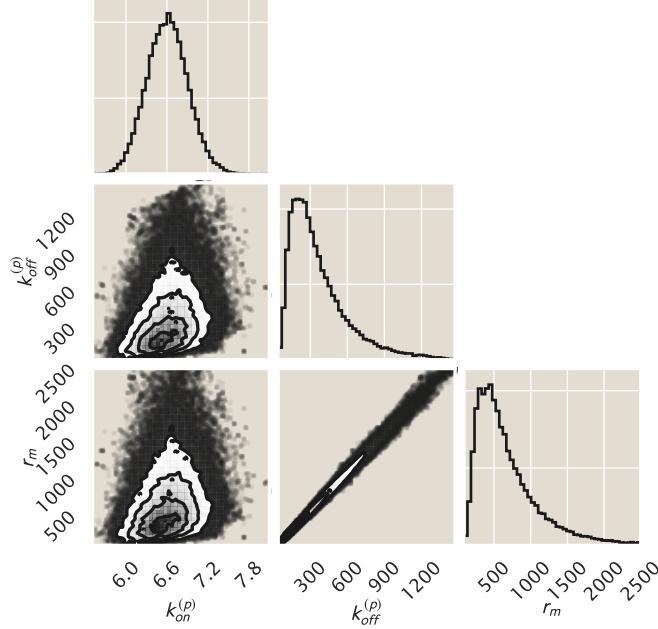
Figure 5.6: **MCMC posterior distribution for a multi-promoter model.** Sampling out of Eq. 5.23, the plot shows 2D and 1D projections of the 3D parameter space. The parameter values are (in units of the mRNA degradation rate $\gamma_m$) $k_{on}^{(p)} = 6.4^{+0.8}_{-0.4}$, $k_{off}^{(p)} = 132^{+737}_{-75}$ and $r_m = 257^{+1307}_{-132}$ which are the modes of their respective distributions, where the superscripts and subscripts represent the upper and lower bounds of the 95$^{th}$ percentile of the parameter value distributions. The sampling was bounded to values $< 1000$ for numerical stability when computing the confluent hypergeometric function. The Python code (`ch5_fig06.py`) used to generate this figure can be found on the original paper's GitHub repository.

With this, we can set as before a Gaussian prior to constrain the ratio of the RNAP rates as

$$
P\left(\ln\left(\frac{k_{on}^{(p)}}{k_{off}^{(p)}}\right)\right) \propto \exp\left\{-\frac{\left(\ln\left(\frac{k_{on}^{(p)}}{k_{off}^{(p)}}\right) - \left[-\ln(2-\phi) - \beta\Delta\varepsilon_p + \ln P - \ln N_{NS}\right)\right]^2}{2\sigma^2}\right\}.
$$
(5.30)

Fig. 5.6 shows the result of sampling out of Eq. 5.23. Again we see that the model is highly sloppy with large credible regions obtained for $k_{off}^{(p)}$ and $r_m$. Nevertheless, the prior information allows us to get parameter values consistent with the equilibrium picture.
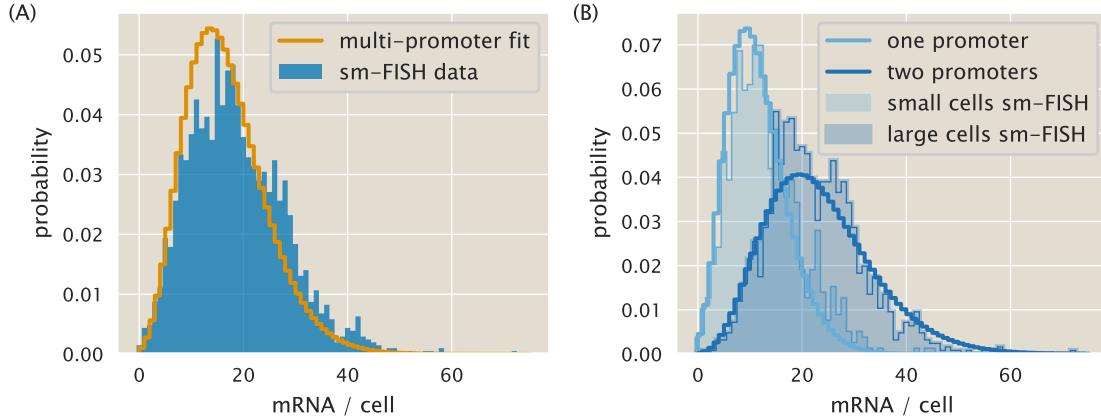
Figure 5.7: **Experimental vs. theoretical distribution of mRNA per cell using parameters for multi-promoter model.** (A) Solid line shows the result of using Eq. 5.18 with the parameters inferred by sampling Eq. 5.23. Blue bars are the same data as Fig. 5.1 from [84]. (B) Split distributions of small cells (light blue bars) and large cells (dark blue) with the corresponding theoretical predictions with transcription rate $r_m$ (light blue line) and transcription rate $2r_m$ (dark blue line). The Python code (ch5_fig07.py) used to generate this figure can be found on the original paper's GitHub repository.

Using again a mRNA mean lifetime of $\approx 3$ min gives the following values for the parameters: $k_{on}^{(p)} = 0.03^{+0.004}_{-0.002}s^{-1}$, $k_{off}^{(p)} = 0.7^{+4.1}_{-0.4}s^{-1}$, and $r_m = 1.4^{+7.3}_{-0.7}s^{-1}$. Fig. 5.7 shows the result of applying Eq. 5.18 using these parameter values. Specifically Fig. 5.7(A) shows the global distribution, including cells with one and two promoters and Fig. 5.7(B) split the distributions within the two populations. Given that the model adequately describes both populations independently and pooled together, we confirm that using the cell area as a proxy for the stage in the cell cycle and the doubling of the transcription rate once cells have two promoters are reasonable approximations.

It is hard to compare literature-reported values because these kinetic rates are effective parameters hiding a lot of the complexity of transcription initiation [8]. Also, the parameters' non-identifiability restricts our explicit comparison of the actual numerical values of the inferred rates. Nevertheless, from the model, we can see that the mean burst size for each transcription event is given by $r_m/k_{\text{off}}^{(p)}$. We obtain a mean burst size of $\approx 1.9$ transcripts per cell from our inferred values. This mean burst size is similar to the reported burst size of 1.15 on a similar system on *E. coli* [144].

**Repressor Rates from a Three-State Regulated Promoter**

Having determined the unregulated promoter transition rates, we now proceed to determine the repressor rates $k_{\text{on}}^{(r)}$ and $k_{\text{off}}^{(r)}$. These rates' values are constrained again by the correspondence between our kinetic picture and what we know from equilibrium models [51]. For this analysis, we again exploit the feature that, at the mean, both the kinetic language and the thermodynamic language should have equivalent predictions. Over the last decade, there has been a great effort in developing equilibrium models for gene expression regulation [48,50,145]. In particular, our group has extensively characterized the simple repression motif using this formalism [20,39,113].

The dialogue between theory and experiments has led to simplified expressions that capture the phenomenology of the gene expression response as a function of natural variables such as molecule count and affinities between molecular players. A particularly interesting quantity for the simple repression motif used by Garcia & Phillips [20] is the fold-change in gene expression, defined as

$$\text{fold-change} = \frac{\langle \text{gene expression}(R \neq 0) \rangle}{\langle \text{gene expression}(R = 0) \rangle}, \tag{5.31}$$

where $R$ is the number of repressors per cell and $\langle \cdot \rangle$ is the population average. The fold-change is simply the mean expression level in the presence of the repressor relative to the mean expression level in the absence of regulation. In the language

of statistical mechanics, this quantity takes the form

$$\text{fold-change} = \left(1 + \frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_r}\right)^{-1}, \tag{5.32}$$

where $\Delta\varepsilon_r$ is the repressor-DNA binding energy, and as before, $N_{NS}$ is the number of non-specific binding sites where the repressor can bind [20].

To compute the fold-change in the chemical master equation language, we compute the first moment of the steady-state mRNA distribution $\langle m \rangle$ for both the three-state promoter ($R \neq 0$) and the two-state promoter case ($R = 0$) (see Sec. 5.3 for moment derivation). The unregulated (two-state) promoter mean mRNA copy number is given by Eq. 5.25. For the regulated (three-state) promoter, we have an equivalent expression of the form

$$\langle m(R \neq 0) \rangle = (2 - \phi)\frac{r_m}{\gamma_m}\frac{k_{\text{off}}^{(r)}k_{\text{on}}^{(p)}}{k_{\text{off}}^{(p)}k_{\text{off}}^{(r)} + k_{\text{off}}^{(p)}k_{\text{on}}^{(r)} + k_{\text{off}}^{(r)}k_{\text{on}}^{(p)}}. \tag{5.33}$$

Computing the fold-change then gives

$$\text{fold-change} = \frac{\langle m(R \neq 0) \rangle}{\langle m(R = 0) \rangle} = \frac{k_{\text{off}}^{(r)}\left(k_{\text{off}}^{(p)} + k_{\text{on}}^{(p)}\right)}{k_{\text{off}}^{(p)}k_{\text{on}}^{(r)} + k_{\text{off}}^{(r)}\left(k_{\text{off}}^{(p)} + k_{\text{on}}^{(p)}\right)}, \tag{5.34}$$

where the factor $(2 - \phi)$ due to the multiple promoter copies, the transcription rate $r_m$, and the mRNA degradation rate $\gamma_m$ all cancel out.

Given that the number of repressors per cell $R$ is an experimental variable that we can control, we assume that the rate at which the promoter transitions from the transcriptionally inactive state to the repressor bound state, $k_{\text{on}}^{(r)}$, is given by the concentration of repressors $[R]$ times a diffusion-limited on rate $k_o$. For the diffusion-limited constant $k_o$, we use the value used by Jones *et al.* [84]. With this in hand, we can rewrite Eq. 5.34 as

$$\text{fold-change} = \left(1 + \frac{k_0[R]}{k_{\text{off}}^{(r)}}\frac{k_{\text{off}}^{(p)}}{k_{\text{on}}^{(p)} + k_{\text{off}}^{(p)}}\right)^{-1}. \tag{5.35}$$

We note that both Eq. 5.32 and Eq. 5.35 have the same functional form. Therefore if both languages predict the same output for the mean gene expression level, it must be true that

$$\frac{k_o[R]}{k_{\text{off}}^{(r)}} \frac{k_{\text{off}}^{(p)}}{k_{\text{on}}^{(p)} + k_{\text{off}}^{(p)}} = \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_r}. \tag{5.36}$$

Solving for $k_{\text{off}}^{(r)}$ gives

$$k_{\text{off}}^{(r)} = \frac{k_o[R] N_{NS} e^{\beta \Delta \varepsilon_r}}{R} \frac{k_{\text{off}}^{(p)}}{k_{\text{on}}^{(p)} + k_{\text{off}}^{(p)}}. \tag{5.37}$$

Since the reported value of $k_o$ is given in units of $\text{nM}^{-1}\text{s}^{-1}$ for the units to cancel properly, the repressor concentration must be given in nM rather than absolute count. If we consider that the repressor concentration is equal to

$$[R] = \frac{R}{V_{cell}} \cdot \frac{1}{N_A}, \tag{5.38}$$

where $R$ is the absolute repressor copy number per cell, $V_{cell}$ is the cell volume, and $N_A$ is Avogadro's number. The *E. coli* cell volume is 2.1 fL [146], and Avogadro's number is $6.022 \times 10^{23}$. If we further include the conversion factor to turn M into nM, we find that

$$[R] = \frac{R}{2.1 \times 10^{-15} L} \cdot \frac{1}{6.022 \times 10^{23}} \cdot \frac{10^9 \text{ nmol}}{1 \text{ mol}} \approx 0.8 \times R. \tag{5.39}$$

Using this, we simplify Eq. 5.37 as

$$k_{\text{off}}^{(r)} \approx 0.8 \cdot k_o \cdot N_{NS} e^{\beta \Delta \varepsilon_r} \cdot \frac{k_{\text{off}}^{(p)}}{k_{\text{on}}^{(p)} + k_{\text{off}}^{(p)}}. \tag{5.40}$$

What Eq. 5.40 shows is the direct relationship that must be satisfied if the equilibrium model is set to be consistent with the non-equilibrium kinetic picture. Table 5.1 summarizes the values obtained for the three operator sequences used throughout this work. To compute these numbers, the number of non-specific binding sites $N_{NS}$ was taken to be $4.6 \times 10^6$ bp, i.e., the size of the *E. coli* K12 genome.

Table 5.1: **Binding sites and corresponding parameters.**

| Operator | $\Delta\varepsilon_r$ $(k_B T)$ | $k_{\text{off}}^{(r)}$ $(s^{-1})$ |
|----------|------------|------------|
| O1 | -15.3 | 0.002 |
| O2 | -13.9 | 0.008 |
| O3 | -9.7 | 0.55 |

*In-vivo* measurements of the Lac repressor off rate have been done with single-molecule resolution [147]. The authors report a mean residence time of $5.3 \pm 0.2$ minutes for the repressor on an O1 operator. The corresponding rate is $k_{\text{off}}^{(r)} \approx 0.003$ $(s^{-1})$, very similar value to what we inferred from our model. In this same reference, the authors determined that, on average, the repressor takes $30.9 \pm 0.5$ seconds to bind to the operator [147]. Given the kinetic model presented in Fig. 3.2(A), this time can be converted to the probability of not being on the repressor bound state $P_{\text{not } R}$. This is computed as

$$P_{\text{not } R} = \frac{\tau_{\text{not } R}}{\tau_{\text{not } R} + \tau_R},\tag{5.41}$$

where $\tau_{\text{not } R}$ is the average time that the repressor does not occupy the operator, and $\tau_R$ is the average time that the repressor spends bound to the operator. Substituting the numbers from [147] gives $P_{\text{not } R} \approx 0.088$. From our model, we can compute the zeroth moment $\langle m^0 p^0 \rangle$ for each of the three promoter states. This moment is equivalent to the probability of being on each of the promoter states. Upon substitution of our inferred rate parameters, we can compute $P_{\text{not } R}$ as

$$P_{\text{not } R} = 1 - P_R \approx 0.046,\tag{5.42}$$

where $P_R$ is the probability of the promoter being bound by the repressor. The value we obtained is within a factor of two from the one reported in [147].

## 5.4  Computing Moments from the Master Equation

This section will compute the moment equations for the distribution $P(m, p)$. Without loss of generality, here, we will focus on the three-state regulated promoter. The computation of the two-state promoter's moments follows the same procedure, changing only the matrices' definition in the master equation.

**Computing Moments of a Distribution**

(Note: The Python code used for the calculations presented in this section can be found in the following link as an annotated Jupyter notebook.)

To compute any moment of our chemical master equation, let us define a vector

$$\langle \mathbf{m^x p^y} \rangle \equiv (\langle m^x p^y \rangle_A, \langle m^x p^y \rangle_I, \langle m^x p^y \rangle_R)^T, \tag{5.43}$$

where $\langle m^x p^y \rangle_S$ is the expected value of $m^x p^y$ in state $S \in \{A, I, R\}$ with $x, y \in \mathbb{N}$. In other words, just as we defined the vector $\mathbf{P}(m, p)$, here we define a vector to collect the expected value of each of the promoter states. By definition, these moments $\langle m^x p^y \rangle_S$ are computed as

$$\langle m^x p^y \rangle_S \equiv \sum_{m=0}^{\infty} \sum_{p=0}^{\infty} m^x p^y P_S(m, p). \tag{5.44}$$

To simplify the notation, let $\sum_x \equiv \sum_{x=0}^{\infty}$. Since we are working with a system of three ODEs, one for each state, let us define the following operation:

$$\langle \mathbf{m^x p^y} \rangle = \sum_m \sum_p m^x p^y \mathbf{P}(m, p) \equiv \begin{bmatrix} \sum_m \sum_p m^x p^y P_A(m, p) \\ \sum_m \sum_p m^x p^y P_I(m, p) \\ \sum_m \sum_p m^x p^y P_R(m, p) \end{bmatrix}. \tag{5.45}$$

With this in hand, we can then apply this sum over $m$ and $p$ to Eq. 3.9. For the left-hand side, we have

$$\sum_m \sum_p m^x p^y \frac{d\mathbf{P}(m, p)}{dt} = \frac{d}{dt} \left[ \sum_m \sum_p m^x p^y \mathbf{P}(m, p) \right], \tag{5.46}$$

where we made use of the linearity property of the derivative to switch the order between the sum and the derivative. Notice that the right-hand side of Eq. 5.46 contains the definition of a moment from Eq. 5.44. That means that we can rewrite it as

$$\frac{d}{dt}\left[\sum_m \sum_p m^x p^y \mathbf{P}(m,p)\right] = \frac{d\langle \mathbf{m^x p^y}\rangle}{dt}.$$

(5.47)

Distributing the sum on the right-hand side of Eq. 3.9 gives

$$\frac{d\langle \mathbf{m^x p^y}\rangle}{dt} = \mathbf{K}\sum_m \sum_p m^x p^y \mathbf{P}(m,p)$$

$$- \mathbf{R}_m \sum_m \sum_p m^x p^y \mathbf{P}(m,p) + \mathbf{R}_m \sum_m \sum_p m^x p^y \mathbf{P}(m-1,p)$$

$$- \mathbf{\Gamma}_m \sum_m \sum_p (m) m^x p^y \mathbf{P}(m,p) + \mathbf{\Gamma}_m \sum_m \sum_p (m+1) m^x p^y \mathbf{P}(m+1,p) \quad (5.48)$$

$$- \mathbf{R}_p \sum_m \sum_p (m) m^x p^y \mathbf{P}(m,p) + \mathbf{R}_p \sum_m \sum_p (m) m^x p^y \mathbf{P}(m,p-1)$$

$$- \mathbf{\Gamma}_p \sum_m \sum_p (p) m^x p^y \mathbf{P}(m,p) + \mathbf{\Gamma}_p \sum_m \sum_p (p+1) m^x p^y \mathbf{P}(m,p+1).$$

Let us look at each term on the right-hand side individually. For the terms in Eq. 5.48 involving $\mathbf{P}(m,p)$, we can again use Eq. 5.44 to rewrite them in a more compact form. This means that we can rewrite the state transition term as

$$\mathbf{K}\sum_m \sum_p m^x p^y \mathbf{P}(m,p) = \mathbf{K}\langle \mathbf{m^x p^y}\rangle.$$

(5.49)

The mRNA production term involving $\mathbf{P}(m,p)$ can be rewritten as

$$\mathbf{R}_m \sum_m \sum_p m^x p^y \mathbf{P}(m,p) = \mathbf{R}_m \langle \mathbf{m^x p^y}\rangle.$$

(5.50)

In the same way, the mRNA degradation term gives

$$\mathbf{\Gamma}_m \sum_m \sum_p (m) m^x p^y \mathbf{P}(m,p) = \mathbf{\Gamma}_m \left\langle \mathbf{m}^{(x+1)} \mathbf{p^y}\right\rangle.$$

(5.51)

For the protein production and degradation terms involving $\mathbf{P}(m,p)$, we have

$$\mathbf{R}_p \sum_m \sum_p (m) m^x p^y \mathbf{P}(m,p) = \mathbf{R}_p \left\langle \mathbf{m}^{(x+1)} \mathbf{p^y}\right\rangle,$$

(5.52)

and

$$\mathbf{\Gamma}_p \sum_m \sum_p (p) m^x p^y \mathbf{P}(m, p) = \mathbf{\Gamma}_p \left\langle \mathbf{m^x p^{(y+1)}} \right\rangle, \tag{5.53}$$

respectively.

For the terms of the sum in Eq. 5.48 involving $\mathbf{P}(m \pm 1, p)$ or $\mathbf{P}(m, p \pm 1)$, we can reindex the sum to work around this mismatch. To be more specific, let us again look at each term case by case. For the mRNA production term involving $\mathbf{P}(m - 1, p)$ we define $m' \equiv m - 1$. Using this, we write

$$\mathbf{R}_m \sum_m \sum_p m^x p^y \mathbf{P}(m - 1, p) = \mathbf{R}_m \sum_{m'=-1}^{\infty} \sum_p (m' + 1)^x p^y \mathbf{P}(m', p). \tag{5.54}$$

Since having negative numbers of mRNA or protein does not make physical sense, we have that $\mathbf{P}(-1, p) = 0$. Therefore we can rewrite the sum starting from 0 rather than from -1, obtaining

$$\mathbf{R}_m \sum_{m'=-1}^{\infty} \sum_p (m' + 1)^x p^y \mathbf{P}(m', p) = \mathbf{R}_m \sum_{m'=0}^{\infty} \sum_p (m' + 1)^x p^y \mathbf{P}(m', p). \tag{5.55}$$

Recall that our distribution $\mathbf{P}(m, p)$ takes $m$ and $p$ as numerical inputs and returns a probability associated with such a molecule count. Nevertheless, $m$ and $p$ themselves are dimensionless quantities that serve as indices of how many molecules are in the cell. The distribution is the same whether the variable is called $m$ or $m'$; for a specific number, let us say $m = 5$, or $m' = 5$, $\mathbf{P}(5, p)$ will return the same result. This means that the variable name is arbitrary, and the right-hand side of Eq. 5.55 can be written as

$$\mathbf{R}_m \sum_{m'=0}^{\infty} \sum_p (m' + 1)^x p^y \mathbf{P}(m', p) = \mathbf{R}_m \langle (\mathbf{m} + 1)^x \mathbf{p^y} \rangle, \tag{5.56}$$

since the left-hand side corresponds to the definition of a moment.

For the mRNA degradation term involving $\mathbf{P}(m + 1, p)$, we follow a similar procedure in which we define $m' = m + 1$ to obtain

$$\mathbf{\Gamma}_m \sum_m \sum_p (m + 1) m^x p^y \mathbf{P}(m + 1, p) = \mathbf{\Gamma}_m \sum_{m'=1}^{\infty} \sum_p m' (m' - 1)^x p^y \mathbf{P}(m', p). \tag{5.57}$$

Since the term on the right-hand side of the equation is multiplied by $m'$, starting the sum over $m'$ from zero rather than one will not affect the result since this factor will not contribute to the total sum. Nevertheless, this is useful since our definition of a moment from Eq. 5.44 requires the sum to start at zero. This means that we can rewrite this term as

$$\boldsymbol{\Gamma}_m \sum_{m'=1}^{\infty} m' \sum_p (m'-1)^x p^y \mathbf{P}(m',p) = \boldsymbol{\Gamma}_m \sum_{m'=0}^{\infty} m' \sum_p (m'-1)^x p^y \mathbf{P}(m',p). \qquad (5.58)$$

Here again, we can change the arbitrary label $m'$ back to $m$, obtaining

$$\boldsymbol{\Gamma}_m \sum_{m'=0}^{\infty} m' \sum_p (m'-1)^x p^y \mathbf{P}(m',p) = \boldsymbol{\Gamma}_m \langle \mathbf{m}(\mathbf{m}-\mathbf{1})^{\mathbf{x}} \mathbf{p}^{\mathbf{y}} \rangle. \qquad (5.59)$$

The protein production term involving $\mathbf{P}(m, p-1)$ can be reindexed by defining $p' \equiv p - 1$. This gives

$$\mathbf{R}_p \sum_m \sum_p (m) m^x p^y \mathbf{P}(m, p-1) = \mathbf{R}_p \sum_m \sum_{p'=-1}^{\infty} m^{(x+1)} (p+1)^y \mathbf{P}(m, p'). \qquad (5.60)$$

We again use the fact that negative molecule copy numbers are assigned with probability zero to begin the sum from 0 rather than -1 and the arbitrary nature of the label $p'$ to write

$$\mathbf{R}_p \sum_m \sum_{p'=0}^{\infty} m^{(x+1)} (p+1)^y \mathbf{P}(m, p') = \mathbf{R}_p \left\langle \mathbf{m}^{(\mathbf{x+1})} (\mathbf{p}+\mathbf{1})^{\mathbf{y}} \right\rangle. \qquad (5.61)$$

Finally, we take care of the protein degradation term involving $\mathbf{P}(m, p+1)$. As before, we define $p' = p + 1$ and substitute this to obtain

$$\boldsymbol{\Gamma}_p \sum_m \sum_p (p+1) m^x p^y \mathbf{P}(m, p+1) = \boldsymbol{\Gamma}_p \sum_m \sum_{p'=1}^{\infty} (p') m^x (p'-1)^y \mathbf{P}(m, p'). \qquad (5.62)$$

Just as with the mRNA degradation term, having a term $p'$ inside the sum allows us to start the sum over $p'$ from 0 rather than 1. We can therefore write

$$\boldsymbol{\Gamma}_p \sum_m \sum_{p'=0}^{\infty} (p') m^x (p'-1)^y \mathbf{P}(m, p') = \boldsymbol{\Gamma}_p \langle \mathbf{m}^{\mathbf{x}} \mathbf{p}(\mathbf{p}-\mathbf{1})^{\mathbf{y}} \rangle. \qquad (5.63)$$

Putting all these terms together, we can write the general moment ODE. This is of the form

$$\frac{d\langle \mathbf{m^x p^y}\rangle}{dt} = \mathbf{K}\langle \mathbf{m^x p^y}\rangle \text{ (promoter state transition)}$$

$$- \mathbf{R}_m\langle \mathbf{m^x p^y}\rangle + \mathbf{R}_m\langle(\mathbf{m+1})^{\mathbf{x}}\mathbf{p^y}\rangle \text{ (mRNA production)}$$

$$- \mathbf{\Gamma}_m\left\langle \mathbf{m}^{(\mathbf{x+1})}\mathbf{p^y}\right\rangle + \mathbf{\Gamma}_m\langle \mathbf{m(m-1)^x p^y}\rangle \text{ (mRNA degradation)} \qquad (5.64)$$

$$- \mathbf{R}_p\left\langle \mathbf{m}^{(\mathbf{x+1})}\mathbf{p^y}\right\rangle + \mathbf{R}_p\left\langle \mathbf{m}^{(\mathbf{x+1})}(\mathbf{p+1})^{\mathbf{y}}\right\rangle \text{ (protein production)}$$

$$- \mathbf{\Gamma}_p\left\langle \mathbf{m^x p}^{(\mathbf{y+1})}\right\rangle + \mathbf{\Gamma}_p\langle \mathbf{m^x p(p-1)^y}\rangle \text{ (protein degradation)}.$$

**Moment Closure of the Simple-Repression Distribution**

A very interesting and useful feature of Eq. 5.64 is that for a given value of $x$ and $y$, the moment $\langle \mathbf{m^x p^y}\rangle$ is only a function of lower moments. Specifically $\langle \mathbf{m^x p^y}\rangle$ is a function of moments $\left\langle \mathbf{m}^{\mathbf{x'}}\mathbf{p}^{\mathbf{y'}}\right\rangle$ that satisfy two conditions:

$$\begin{aligned} &1) y' \leq y, \\ &2) x' + y' \leq x + y. \end{aligned} \qquad (5.65)$$

To prove this, we rewrite Eq. 5.64 as

$$\frac{d\langle \mathbf{m^x p^y}\rangle}{dt} = \mathbf{K}\langle \mathbf{m^x p^y}\rangle$$

$$+ \mathbf{R}_m\langle \mathbf{p^y}\left[(\mathbf{m+1})^{\mathbf{x}} - \mathbf{m^x}\right]\rangle$$

$$+ \mathbf{\Gamma}_m\langle \mathbf{mp^y}\left[(\mathbf{m-1})^{\mathbf{x}} - \mathbf{m^x}\right]\rangle \qquad (5.66)$$

$$+ \mathbf{R}_p\left\langle \mathbf{m}^{(\mathbf{x+1})}\left[(\mathbf{p+1})^{\mathbf{y}} - \mathbf{p^y}\right]\right\rangle$$

$$+ \mathbf{\Gamma}_p\langle \mathbf{m^x p}\left[(\mathbf{p-1})^{\mathbf{y}} - \mathbf{p^y}\right]\rangle,$$

where the factorization is valid given the linearity of expected values. The objective is to find the highest moment for each term once the relevant binomial, such as $(m-1)^x$, is expanded. Take, for example, a simple case in which we want to find the second moment of the mRNA distribution. We then set $x = 2$ and $y = 0$.

Eq. 5.66 then becomes

$$\frac{\langle \mathbf{m^2 p^0} \rangle}{dt} = \mathbf{K} \langle \mathbf{m^2 p^0} \rangle$$

$$+ \mathbf{R}_m \langle \mathbf{p^0} \left[ (\mathbf{m}+1)^2 - \mathbf{m^2} \right] \rangle$$

$$+ \mathbf{\Gamma}_m \langle \mathbf{m p^0} \left[ (\mathbf{m}-1)^2 - \mathbf{m^2} \right] \rangle \tag{5.67}$$

$$+ \mathbf{R}_p \langle \mathbf{m^{(2+1)}} \left[ (\mathbf{p}+1)^0 - \mathbf{p^0} \right] \rangle$$

$$+ \mathbf{\Gamma}_p \langle \mathbf{m^2 p} \left[ (\mathbf{p}-1)^0 - \mathbf{p^0} \right] \rangle.$$

Simplifying this equation gives

$$\frac{d \langle \mathbf{m^2} \rangle}{dt} = \mathbf{K} \langle \mathbf{m^2} \rangle + \mathbf{R}_m \langle [2\mathbf{m}+1] \rangle + \mathbf{\Gamma}_m \langle [-2\mathbf{m^2}+\mathbf{m}] \rangle. \tag{5.68}$$

Eq. 5.68 satisfies both of our conditions. Since we set $y$ to be zero, none of the terms depend on any moment that involves the protein number. Therefore $y' \leq y$ is satisfied. Also, the highest moment in Eq. 5.68 also satisfies $x' + y' \leq x + y$ since the second moment of mRNA does not depend on any moment higher than $\langle \mathbf{m^2} \rangle$. To demonstrate that this is true for any $x$ and $y$, we now rewrite Eq. 5.66, making use of the binomial expansion

$$(z \pm 1)^n = \sum_{k=0}^{n} \binom{n}{k} (\pm 1)^k z^{n-k}. \tag{5.69}$$

Just as before, let us look at each term individually. For the mRNA production term, we have

$$\mathbf{R}_m \langle \mathbf{p^y} \left[ (\mathbf{m}+1)^{\mathbf{x}} - \mathbf{m^x} \right] \rangle = \mathbf{R}_m \left\langle \mathbf{p^y} \left[ \sum_{k=0}^{\mathbf{x}} \binom{\mathbf{x}}{k} \mathbf{m^{x-k}} - \mathbf{m^x} \right] \right\rangle. \tag{5.70}$$

When $k = 0$, the term inside the sum on the right-hand side cancels with the other $m^x$ so that we can simplify to

$$\mathbf{R}_m \langle \mathbf{p^y} \left[ (\mathbf{m}+1)^{\mathbf{x}} - \mathbf{m^x} \right] \rangle = \mathbf{R}_m \left\langle \mathbf{p^y} \left[ \sum_{k=1}^{\mathbf{x}} \binom{\mathbf{x}}{k} \mathbf{m^{x-k}} \right] \right\rangle. \tag{5.71}$$

Once the sum is expanded, we can see that the highest moment in this sum is given by $\langle \mathbf{m^{(x-1)} p^y} \rangle$ which satisfies both of the conditions on Eq. 5.65.

For the mRNA degradation term, we similarly have

$$\Gamma_m \langle \mathbf{m}\mathbf{p}^{\mathbf{y}}\left[(\mathbf{m}-\mathbf{1})^{\mathbf{x}} - \mathbf{m}^{\mathbf{x}}\right] \rangle = \Gamma_m \left\langle \mathbf{m}\mathbf{p}^{\mathbf{y}} \left[ \sum_{\mathbf{k}=\mathbf{0}}^{\mathbf{x}} \binom{\mathbf{x}}{\mathbf{k}}(-\mathbf{1})^{\mathbf{k}}\mathbf{m}^{\mathbf{x}-\mathbf{k}} - \mathbf{m}^{\mathbf{x}}\right] \right\rangle. \quad (5.72)$$

Simplifying terms, we obtain

$$\Gamma_m \left\langle \mathbf{m}\mathbf{p}^{\mathbf{y}}\left[ \sum_{\mathbf{k}=\mathbf{0}}^{\mathbf{x}} \binom{\mathbf{x}}{\mathbf{k}}(-\mathbf{1})^{\mathbf{k}}\mathbf{m}^{\mathbf{x}-\mathbf{k}} - \mathbf{m}^{\mathbf{x}}\right] \right\rangle = \Gamma_m \left\langle \mathbf{p}^{\mathbf{y}}\left[ \sum_{\mathbf{k}=\mathbf{1}}^{\mathbf{x}} \binom{\mathbf{x}}{\mathbf{k}}(-\mathbf{1})^{\mathbf{k}}\mathbf{m}^{\mathbf{x}+\mathbf{1}-\mathbf{k}}\right] \right\rangle.$$
$$(5.73)$$

The largest moment in this case is $\langle \mathbf{m}^{\mathbf{x}}\mathbf{p}^{\mathbf{y}} \rangle$, which again satisfies the conditions on Eq. 5.65.

The protein production term gives

$$R_p \left\langle \mathbf{m}^{(\mathbf{x}+\mathbf{1})}\left[(\mathbf{p}+\mathbf{1})^{\mathbf{y}} - \mathbf{p}^{\mathbf{y}}\right] \right\rangle = R_p \left\langle \mathbf{m}^{(\mathbf{x}+\mathbf{1})}\left[ \sum_{\mathbf{k}=\mathbf{0}}^{\mathbf{y}} \binom{\mathbf{y}}{\mathbf{k}}(-\mathbf{1})^{\mathbf{k}}\mathbf{p}^{\mathbf{y}-\mathbf{k}} - \mathbf{p}^{\mathbf{y}}\right] \right\rangle. \quad (5.74)$$

Upon simplification, we obtain

$$R_p \left\langle \mathbf{m}^{(\mathbf{x}+\mathbf{1})}\left[ \sum_{\mathbf{k}=\mathbf{0}}^{\mathbf{y}} \binom{\mathbf{y}}{\mathbf{k}}(-\mathbf{1})^{\mathbf{k}}\mathbf{p}^{\mathbf{y}-\mathbf{k}} - \mathbf{p}^{\mathbf{y}}\right] \right\rangle = R_p \left\langle \mathbf{m}^{(\mathbf{x}+\mathbf{1})}\left[ \sum_{\mathbf{k}=\mathbf{1}}^{\mathbf{y}} \binom{\mathbf{y}}{\mathbf{k}}(-\mathbf{1})^{\mathbf{k}}\mathbf{p}^{\mathbf{y}-\mathbf{k}}\right] \right\rangle.$$
$$(5.75)$$

Here the largest moment is given by $\langle \mathbf{m}^{\mathbf{x}+\mathbf{1}}\mathbf{p}^{\mathbf{y}-\mathbf{1}} \rangle$, that again satisfies both of our conditions. For the last term, for protein degradation, we have

$$R_p \left\langle \mathbf{m}^{(\mathbf{x}+\mathbf{1})}\left[(\mathbf{p}+\mathbf{1})^{\mathbf{y}} - \mathbf{p}^{\mathbf{y}}\right] \right\rangle = R_p \left\langle \mathbf{m}^{(\mathbf{x}+\mathbf{1})}\left[ \sum_{\mathbf{k}=\mathbf{1}}^{\mathbf{y}} \binom{\mathbf{y}}{\mathbf{k}}(-\mathbf{1}^{\mathbf{k}})\mathbf{p}^{\mathbf{y}-\mathbf{k}}\right] \right\rangle. \quad (5.76)$$

The largest moment involved in this term is therefore $\langle \mathbf{m}^{\mathbf{x}}\mathbf{p}^{\mathbf{y}-\mathbf{1}} \rangle$. With this, we show that the four terms involved in our general moment equation depend only on lower moments that satisfy Eq. 5.65.

As a reminder, we showed in this section that the kinetic model introduced in Fig. 3.2(A) has no moment-closure problem. In other words, moments of the joint mRNA and protein distribution can be computed from knowledge of lower moments. This allows us to cleanly integrate the distribution moment dynamics as cells progress through the cell cycle.

**Computing Single Promoter Steady-State Moments**

(Note: The Python code used for the calculations presented in this section can be found in the following link as an annotated Jupyter notebook.)

One of the main factors contributing to cell-to-cell variability in gene expression is the change in gene copy number during the cell cycle as cells replicate their genome before cell division. Our minimal model accounts for this variability by considering the time trajectory of the distribution moments given by Eq. 5.66. These predictions will be contrasted with the predictions from a kinetic model that does not account for gene copy numbers changes during the cell cycle in Sec. 4.4.

Suppose we do not account for the change in gene copy number during the cell cycle or the partition of proteins during division. In that case, the dynamics of the moments of the distribution described in this section will reach a steady state. To compute the kinetic model's steady-state moments with a single gene across the cell cycle, we use the moment closure property of our master equation. By equating Eq. 5.66 to zero for a given $\mathbf{x}$ and $\mathbf{y}$, we can solve the resulting linear system and obtain a solution for $\langle \mathbf{m^x p^y} \rangle$ at steady state as a function of moments $\langle \mathbf{m}^{x'} \mathbf{p}^{y'} \rangle$ that satisfy Eq. 5.65. Then, by solving for the zero$^{\text{th}}$ moment $\langle \mathbf{m^0 p^0} \rangle$ subject to the constraint that the probability of the promoter being in any state should add up to one, we can substitute back all of the solutions in terms of moments $\langle \mathbf{m}^{x'} \mathbf{p}^{y'} \rangle$ with solutions in terms of the rates shown in Fig. 3.2. In other words, through an iterative process, we can get at the value of any moment of the distribution. We start by solving for the zero$^{\text{th}}$ moment. Since all higher moments depend on lower moments, we can use the solution of the zero$^{\text{th}}$ moment to compute the first mRNA moment. This solution is then used for higher moments in a hierarchical iterative process.

## 5.5 Accounting for the Variability in Gene Copy Number During the Cell Cycle

(Note: The Python code used for the calculations presented in this section can be found in the following link as an annotated Jupyter notebook.)

When growing in rich media, bacteria can double every $\approx$ 20 minutes. With two replication forks, each traveling at $\approx$ 1000 bp per second, and a genome of $\approx$ 5 Mbp for *E. coli* [148], a cell would need $\approx$ 40 minutes to replicate its genome. The apparent paradox of growth rates faster than one division per 40 minutes is solved because cells have multiple replisomes, i.e., molecular machines that replicate the genome running in parallel. Cells can have up to 8 copies of the genome being replicated simultaneously, depending on the growth rate [17].

This observation implies that during the cell cycle, gene copy number varies. This variation depends on the growth rate and the relative position of the gene with respect to the replication origin, having genes close to the replication origin spending more time with multiple copies than genes closer to the replication termination site. This change in gene dosage directly affects cell-to-cell variability in gene expression [84,118].

**Numerical Integration of Moment Equations**

(Note: The Python code used for the calculations presented in this section can be found in the following link as an annotated Jupyter notebook)

For our specific locus (*galK*) and a doubling time of $\approx$ 60 min for our experimental conditions, cells have on average 1.66 copies of the reporter gene during the cell cycle [84]. This means that cells spend 60% of the time having one copy of the gene and 40% of the time with two copies. To account for this variability in gene copy number across the cell cycle, we numerically integrate the moment equations derived in for a time $t = [0, t_s]$ with an mRNA production rate $r_m$, where $t_s$ is the time point at which the replication fork reaches our specific locus. For the remaining time before the cell division $t = [t_s, t_d]$ that the cell spends with two promoters, we assume that the only parameter that changes is the mRNA production rate from $r_m$ to $2r_m$. This simplifying assumption ignores potential changes in protein translation rate $r_p$ or changes in the repressor copy number that would be reflected in changes on the repressor on rate $k_{on}^{(r)}$.

**Computing Distribution Moments After Cell Division**

(Note: The Python code used for the calculations presented in this section can be found in the following link as an annotated Jupyter notebook)

We have already solved a general form for the dynamics of the moments of the distribution, i.e., we wrote differential equations for the moments $\frac{d\langle m^x p^y \rangle}{dt}$. Given that we know all parameters for our model, we can numerically integrate these equations to compute how the distribution moments evolve as cells progress through their cell cycle. Once the cell reaches a time $t_d$ and divides, the mRNA and proteins that we are interested in undergo a binomial partitioning between the two daughter cells. In other words, each molecule flips a coin to decide to which daughter cell to go. The question then becomes given that we have a value for the moment $\langle m^x p^y \rangle_{t_d}$ at a time before the cell division, what would the value of this moment be after the cell division takes place $\langle m^x p^y \rangle_{t_o}$?

The probability distribution of mRNA and protein after the cell division $P_{t_o}(m, p)$ must satisfy

$$P_{t_o}(m, p) = \sum_{m'=m}^{\infty} \sum_{p'=p}^{\infty} P(m, p \mid m', p') P_{t_d}(m', p'),  \tag{5.77}$$

where we are summing over all the possibilities of having $m'$ mRNA and $p'$ proteins before cell division. Note that the sums start at $m$ and $p$; this is because for a daughter cell to have these copy numbers before cell division, it is a requirement that the mother cell had at least such a copy number since we are not assuming that there is any production at the instantaneous cell division time. Since we assume that the partition of mRNA is independent of the partition of protein, the conditional probability $P(m, p \mid m', p')$ is given by a product of two binomial distributions, one for the mRNA and one for the protein, i.e.

$$P(m, p \mid m', p') = \binom{m'}{m} \left( \frac{1}{2} \right)^{m'} \cdot \binom{p'}{p} \left( \frac{1}{2} \right)^{p'}.  \tag{5.78}$$

Because of this product of binomial probabilities, we are allowed to extend the sum from Eq. 5.77 to start at $m' = 0$ and $p' = 0$ as

$$P_{t_o}(m, p) = \sum_{m'=0}^{\infty} \sum_{p'=0}^{\infty} P(m, p \mid m', p') P_{t_d}(m', p'), \qquad (5.79)$$

since the product of the binomial distributions in Eq. 5.78 is zero for all $m' < m$ and/or $p' < 0$. Thus, to simplify notation, from now on in this section we will assume that a sum of the form $\sum_x \equiv \sum_{x=0}^{\infty}$.

We can then compute the distribution moments after the cell division $\langle m^x p^y \rangle_{t_o}$ as

$$\langle m^x p^y \rangle_{t_o} = \sum_m \sum_p m^x p^y P_{t_o}(m, p), \qquad (5.80)$$

for all $x, y \in \mathbb{N}$. Substituting Eq. 5.77 results in

$$\langle m^x p^y \rangle_{t_o} = \sum_m \sum_p m^x p^y \sum_{m'} \sum_{p'} P(m, p \mid m', p') P_{t_d}(m', p'). \qquad (5.81)$$

We can rearrange the sums to be

$$\langle m^x p^y \rangle_{t_o} = \sum_{m'} \sum_{p'} P_{t_d}(m', p') \sum_m \sum_p m^x p^y P(m, p \mid m', p'). \qquad (5.82)$$

The fact that Eq. 5.78 is the product of two independent events allows us to rewrite the joint probability $P(m, p \mid m', p')$ as

$$P(m, p \mid m', p') = P(m \mid m') \cdot P(p \mid p'). \qquad (5.83)$$

With this, we can then write the moment $\langle m^x p^y \rangle_{t_o}$ as

$$\langle m^x p^y \rangle_{t_o} = \sum_{m'} \sum_{p'} P_{t_d}(m', p') \sum_m m^x P(m \mid m') \sum_p p^y P(p \mid p'). \qquad (5.84)$$

Notice that both terms summing over $m$ and $p$ are the conditional expected values, i.e.

$$\sum_z z^x P(z \mid z') \equiv \langle z^x \mid z' \rangle, \quad \text{for } z \in \{m, p\}. \qquad (5.85)$$

These conditional expected values are the expected values of a binomial random variable $z \sim \text{Bin}(z', 1/2)$, which can be easily computed, as we will show later in this section. We then rewrite the expected values after the cell division in terms of these moments of a binomial distribution

$$\langle m^x p^y \rangle_{t_o} = \sum_{m'} \sum_{p'} \langle m^x \mid m' \rangle \langle p^y \mid p' \rangle P_{t_d}(m', p'). \qquad (5.86)$$

To see how this general formula for the moments after the cell division works, let us compute the mean protein per cell after the cell division $\langle p \rangle_{t_o}$. That means setting $x = 0$, and $y = 1$. This results in

$$\langle p \rangle_{t_o} = \sum_{m'} \sum_{p'} \left\langle m^0 \mid m' \right\rangle \langle p \mid p' \rangle \, P_{t_d}(m', p'). \tag{5.87}$$

The zeroth moment $\left\langle m^0 \mid m' \right\rangle$ by definition must be one, i.e.,

$$\left\langle m^0 \mid m' \right\rangle = \sum_m m^0 P(m \mid m') = \sum_m P(m \mid m') = 1, \tag{5.88}$$

since the probability distribution must be normalized. This leaves us then with

$$\langle p \rangle_{t_o} = \sum_{m'} \sum_{p'} P_{t_d}(m', p') \langle p \mid p' \rangle. \tag{5.89}$$

If we take the sum over $m'$, we simply compute the marginal probability distribution $\sum_{m'} P_{t_d}(m', p') = P_{t_d}(p')$; then we have

$$\langle p \rangle_{t_o} = \sum_{p'} \langle p \mid p' \rangle \, P_{t_d}(p'). \tag{5.90}$$

For the particular case of the first moment of the binomial distribution with parameters $p'$ and $1/2$, we know that

$$\langle p \mid p' \rangle = \frac{p'}{2}. \tag{5.91}$$

Therefore the moment after division is equal to

$$\langle p \rangle_{t_o} = \sum_{p'} \frac{p'}{2} P_{t_d}(p') = \frac{1}{2} \sum_{p'} p' P_{t_d}(p'). \tag{5.92}$$

Notice that this is just $1/2$ of the expected value of $p'$ averaging over the distribution before cell division, i.e.

$$\langle p \rangle_{t_o} = \frac{\langle p' \rangle_{t_d}}{2}, \tag{5.93}$$

where $\langle \cdot \rangle_{t_d}$ highlights that the moment is computed prior to the cell division. This result makes perfect sense. What this is saying is that the mean protein copy number right after the cell divides is half of the mean protein copy number just before the cell division. That is exactly what we would expect. So, in principle, to know the first moment of either the mRNA distribution $\langle m \rangle_{t_o}$ or the protein distribution $\langle m \rangle_{t_o}$ right after cell division, it suffices to multiply the moments before the cell division $\langle m \rangle_{t_d}$ or $\langle p \rangle_{t_d}$ by $1/2$. Let us now explore how this generalizes to any other moment $\langle m^x p^y \rangle_{t_o}$.

**Computing the Moments of a Binomial Distribution**

The last section's result depended on us knowing the functional form of the first moment of the binomial distribution. For higher moments, we need some systematic way to compute such moments. Luckily for us, we can do so by using the so-called moment generating function (MGF). The MGF of a random variable $X$ is defined as

$$M_X(t) = \left\langle e^{tX} \right\rangle , \tag{5.94}$$

where $t$ is a dummy variable. Once we know the MGF, we can obtain any moment of the distribution by simply computing

$$\langle X^n \rangle = \left. \frac{d^n}{dt^n} M_X(t) \right|_{t=0} , \tag{5.95}$$

i.e., taking the $n$-th derivative of the MGF returns the $n$-th moment of the distribution. For the particular case of the binomial distribution $X \sim \mathrm{Bin}(N, q)$, it can be shown that the MGF is of the form

$$M_X(t) = \left[ (1 - q) + q e^t \right]^N . \tag{5.96}$$

As an example, let us compute the first moment of this binomially distributed variable. For this, the first derivative of the MGF results in

$$\frac{dM_X(t)}{dt} = N[(1 - q) + q e^t]^{N-1} q e^t . \tag{5.97}$$

We just need to follow Eq. 5.95 and set $t = 0$ to obtain the first moment

$$\frac{dM_X(t)}{dt}\bigg|_{t=0} = Nq, \tag{5.98}$$

which is precisely the expected value of a binomially distributed random variable.

So according to Eq. 5.86, to compute any moment $\langle m^x p^y \rangle$ after cell division, we can just take the $x$-th derivative and the $y$-th derivative of the binomial MGF to obtain $\langle m^x \mid m' \rangle$ and $\langle p^y \mid p' \rangle$, respectively, and take the expected value of the result. Let us compute the specific case for the moment $\langle mp \rangle$ to illustrate the procedure. When computing the moment after cell division $\langle mp \rangle_{t_o}$ which is of the form

$$\langle mp \rangle_{t_o} = \sum_{m'} \sum p' \langle m \mid m' \rangle \langle p \mid p' \rangle P_{t_d}(m', p'), \tag{5.99}$$

the product $\langle m \mid m' \rangle \langle p \mid p' \rangle$ is then

$$\langle m \mid m' \rangle \langle p \mid p' \rangle = \frac{m'}{2} \cdot \frac{p'}{2}, \tag{5.100}$$

where we used the result in Eq. 5.98, substituting $m$ and $p$ for $X$, respectively, and $q$ for $1/2$. Substituting this result into the moment gives

$$\langle mp \rangle_{t_o} = \sum_{m'} \sum_{p'} \frac{m'p'}{4} P_{t_d}(m', p') = \frac{\langle m'p' \rangle_{t_d}}{4}. \tag{5.101}$$

Therefore to compute the moment after cell division $\langle mp \rangle_{t_o}$ we simply have to divide by 4 the corresponding equivalent moment before the cell division.

Not all moments after cell division depend only on the equivalent moment before cell division. For example, if we compute the third moment of the protein distribution $\langle p^3 \rangle_{t_o}$, we find

$$\langle p^3 \rangle_{t_o} = \frac{\langle p^3 \rangle_{t_d}}{8} + \frac{3 \langle p^2 \rangle_{t_d}}{8}. \tag{5.102}$$

For this particular case, the third moment of the protein distribution depends on the third moment and the second moment before the cell division. In general, all moments after cell division $\langle m^x p^y \rangle_{t_o}$ linearly depend on moments before cell division. Furthermore, there is "moment closure" for this specific case in the sense that all moments after cell division depend on lower moments before cell division. To generalize these results to all the moments computed in this work, let us then define a vector to collect all moments before the cell division up the $\langle m^x p^y \rangle_{t_d}$ moment, i.e.

$$\langle \mathbf{m^x p^y} \rangle_{t_d} = \left( \left\langle m^0 p^0 \right\rangle_{t_d}, \left\langle m^1 \right\rangle_{t_d}, \ldots, \langle m^x p^y \rangle_{t_d} \right). \tag{5.103}$$

Then any moment after cell division $\left\langle m^{x'} p^{y'} \right\rangle_{t_o}$ for $x' \leq x$ and $y' \leq y$ can be computed as

$$\left\langle m^{x'} p^{y'} \right\rangle_{t_o} = \mathbf{z}_{x'y'} \cdot \langle \mathbf{m^x p^y} \rangle_{t_d}, \tag{5.104}$$

where we define the vector $\mathbf{z}_{x'y'}$ as the vector containing all the coefficients that we obtain with the product of the two binomial distributions. For example, for the case of the third protein moment $\left\langle p^3 \right\rangle_{t_o}$ the vector $\mathbf{z}_{x'y'}$ would have zeros for all entries except for the corresponding entry for $\left\langle p^2 \right\rangle_{t_d}$ and for $\left\langle p^3 \right\rangle_{t_d}$, where it would have 3/8 and 1/8 accordingly.

If we want then to compute all the moments after the cell division up to $\langle m^x p^y \rangle_{t_o}$, let us define an equivalent vector

$$\langle \mathbf{m^x p^y} \rangle_{t_o} = \left( \left\langle m^0 p^0 \right\rangle_{t_o}, \left\langle m^1 \right\rangle_{t_o}, \ldots, \langle m^x p^y \rangle_{t_o} \right). \tag{5.105}$$

Then we need to build a square matrix $\mathbf{Z}$ such that each row of the matrix contains the corresponding vector $\mathbf{z}_{x'y'}$ for each of the moments. Having this matrix, we would simply compute the moments after the cell division as

$$\langle \mathbf{m^x p^x} \rangle_{t_o} = \mathbf{Z} \cdot \langle \mathbf{m^x p^x} \rangle_{t_d}. \tag{5.106}$$

In other words, the matrix $\mathbf{Z}$ will contain all the coefficients that we need to multiply by the moments before the cell division to obtain the moments after cell division. The matrix $\mathbf{Z}$ was then generated automatically using Python's analytical math library sympy [149].

Fig. 5.8 (adapted from Fig. 3.3(B)) shows how the first moment of both mRNA and protein changes over several cell cycles. The mRNA quickly relaxes to the steady-state corresponding to the parameters for both a single and two promoters copies. This is expected since the parameters for the mRNA production was determined in the first place under this assumption. We note that there is no apparent delay before reaching the steady-state of the mean mRNA count after the cell divides. This is because the mean mRNA count for the two promoter copies state is precisely twice the expected mRNA count for the single promoter state (see Sec. 5.1). Therefore, once the mean mRNA count is halved after the cell division, it is already at the steady-state value for the single promoter case. On the other hand, given that the degradation rate determines the relaxation time to steady-state, the mean protein count does not reach its corresponding steady-state value for either promoter copy number state. Interestingly, once a couple of cell cycles have passed, the first moment has a repetitive trajectory over cell cycles. We have observed this experimentally by tracking cells as they grow under the microscope. Comparing cells at the beginning of the cell cycle with the daughter cells that appear after cell division showed that, on average, all cells have the same amount of protein at the start of the cell cycle (see Fig. 18 of [16]), suggesting that this dynamical steady state takes place *in vivo*.

When measuring gene expression levels experimentally from an asynchronous culture, cells are sampled from any time point across their cell cycles. This means that the moments determined experimentally correspond to an average over the cell cycle. In the following section, we discuss how to account for the fact that cells are not uniformly distributed across the cell cycle to compute these averages.

**Exponentially Distributed Ages**

As mentioned in Sec. 5.2, cells in exponential growth have exponentially distributed ages across the cell cycle, having more young cells than old ones. Specifically, the probability of a cell being at any time point in the cell cycle is given by
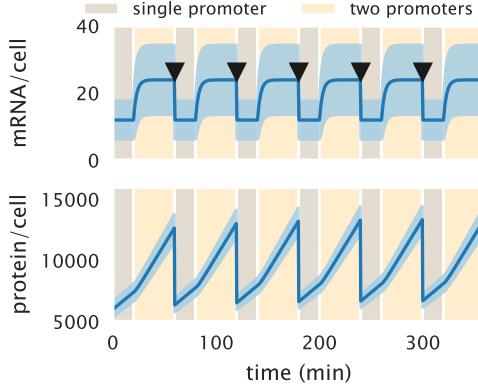
Figure 5.8: **First and second moment dynamics over the cell cycle.** Mean $\pm$ standard deviation mRNA (upper panel) and mean $\pm$ standard deviation protein copy number (lower panel) as the cell cycle progresses. The dark shaded region delimits the fraction of the cell cycle that cells spend with a single copy of the promoter. The light-shaded region delimits the fraction of the cell cycle that cells spend with two copies of the promoter. For a 100 min doubling time at the *galK* locus, cells spend 60% of the time with one copy of the promoter and the rest with two copies. The Python code (`ch5_fig08.py`) used to generate this figure can be found on the original paper's GitHub repository.

[120]

$$P(a) = (\ln 2) \cdot 2^{1-a}, \tag{5.107}$$

where $a \in [0, 1]$ is the stage of the cell cycle, with $a = 0$ being the start of the cycle and $a = 1$ being the cell division. In Sec. 5.10, we reproduce this derivation. It is a surprising result, but it can be intuitively thought as follows: if the culture is growing exponentially, that means that all the time, there is an increasing number of cells. That means, for example, that if in a time interval $\Delta t$ $N$ "old" cells divided, these produced $2N$ "young" cells. So at any point, there are always more younger than older cells.

Our numerical integration of the moment equations gave us a time evolution of the moments as cells progress through the cell cycle. Since experimentally we sample asynchronous cells that follow Eq. 5.107, each time point along the moment dynamic must be weighted by the probability of having sampled a cell at such a specific time point of the cell cycle. Without loss of generality, let us focus on the first mRNA moment $\langle m(t) \rangle$ (the same can be applied to all other moments). As mentioned before, to calculate the first moment across the entire cell cycle, we

must weigh each time point by the corresponding probability that a cell is found at such a point of its cell cycle. This translates to computing the integral

$$\langle m \rangle_c = \int_{\text{beginning cell cycle}}^{\text{end cell cycle}} \langle m(t) \rangle \, P(t) dt, \tag{5.108}$$

where $\langle m \rangle_c$ is the mean mRNA copy number averaged over the entire cell cycle trajectory, and $P(t)$ is the probability of a cell being at a time $t$ of its cell cycle.

If we set the time in units of the cell cycle length, we can use Eq. 5.107 and compute instead

$$\langle m \rangle = \int_0^1 \langle m(a) \rangle \, P(a) da, \tag{5.109}$$

where $P(a)$ is given by Eq. 5.107.

What Eq. 5.109 implies is that to compute the first moment (or any moment of the distribution), we must weigh each point in the moment dynamics by the corresponding probability of a cell being at that point along its cell cycle. That is why when computing a moment, we take the time trajectory of a single cell cycle as the ones shown in Fig. 5.8 and compute the average using Eq. 5.107 to weigh each time point. We perform this integral numerically for all moments using Simpson's rule.

**Reproducing the Equilibrium Picture**

Given the large variability of the first moments depicted in Fig. 5.8, it is worth considering why a simplistic equilibrium picture has shown to be very successful in predicting the mean expression level under diverse conditions [20,39,112,113]. This section compares the simple repression thermodynamic model with this dynamical picture of the cell cycle. But before diving into this comparison, it is worth recapping the assumptions that go into the equilibrium model.

**Steady-State Under the Thermodynamic Model**

For the thermodynamic model we can only describe the first moment's dynamics using this theoretical framework [51]. This is because these models are based on the probability distribution of the promoter microstates rather than the distri-

bution over the mRNA and protein counts. Let us only focus on the mRNA first moment $\langle m \rangle$. The same principles apply if we consider the protein first moment. We can write a dynamical system of the form

$$\frac{d\langle m \rangle}{dt} = r_m \cdot p_{\text{bound}} - \gamma_m \langle m \rangle, \tag{5.110}$$

where $r_m$ and $\gamma_m$ are the mRNA production and degradation rates, respectively, and $p_{\text{bound}}$ is the probability of finding the RNAP bound to the promoter [50]. This dynamical system is predicted to have a single stable fixed point that we can find by computing the steady-state. When we solve for the mean mRNA copy number at steady-state $\langle m \rangle_{ss}$, we find

$$\langle m \rangle_{ss} = \frac{r_m}{\gamma_m} p_{\text{bound}}. \tag{5.111}$$

Since we assume that the only effect that the repressor has over the promoter's regulation is the exclusion of the RNAP from binding to the promoter, we assume that only $p_{\text{bound}}$ depends on the repressor copy number $R$. Therefore when computing the fold-change in gene expression, we are left with

$$\text{fold-change} = \frac{\langle m(R \neq 0) \rangle_{ss}}{\langle m(R = 0) \rangle_{ss}} = \frac{p_{\text{bound}}(R \neq 0)}{p_{\text{bound}}(R = 0)}. \tag{5.112}$$

As derived in [20], this can be written in the language of equilibrium statistical mechanics as

$$\text{fold-change} = \left(1 + \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_r}\right)^{-1}, \tag{5.113}$$

where $\beta \equiv (k_B T)^{-1}$, $\Delta \varepsilon_r$ is the repressor-DNA binding energy, and $N_{NS}$ is the number of non-specific binding sites where the repressor can bind.

To arrive at Eq. 5.113, we ignore the physiological changes that occur during the cell cycle; one of the most important being the variability in gene copy number that we are exploring in this section. It is, therefore, worth thinking about whether or not the dynamical picture exemplified in Fig. 5.8 can be reconciled with the predictions made by Eq. 5.113 both at the mRNA and protein level.
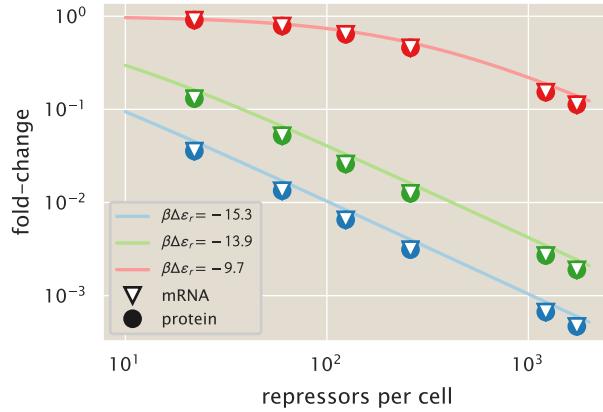
Figure 5.9: **Comparison of the equilibrium and kinetic repressor titration predictions.** The equilibrium model (solid lines) and the kinetic model with variation over the cell cycle (solid circles and white triangles) predictions are compared for varying repressor copy numbers and operator binding energy. The equilibrium model is directly computed from Eq. 5.113, while the kinetic model is computed by numerically integrating the moment equations over several cell cycles, and then averaging over the extent of the cell cycle as defined in Eq. 5.109 . The Python code (`ch5_fig09.py`) used to generate this figure can be found on the original paper's GitHub repository.

Fig. 5.9 compares the predictions of both theoretical frameworks for varying repressor copy numbers and repressor-DNA affinities. The solid lines are directly computed from Eq. 5.113. The hollow triangles and the solid circles represent the fold-change in mRNA and protein, respectively, as computed from the moment dynamics. To compute the fold-change from the kinetic picture, we first numerically integrate the moment dynamics for both the two- and the three-state promoter (see Fig. 5.8 for the unregulated case), and then average the time series accounting for the probability of cells being sampled at each stage of the cell cycle as defined in Eq. 5.109. The small systematic deviations between both models come partly from the simplifying assumption that the repressor copy number, and therefore the repressor on rate $k_{\text{on}}^{(r)}$ remains constant during the cell cycle. In principle, the gene producing the repressor protein itself is also subjected to the same duplication during the cell cycle, changing, therefore, the mean repressor copy number for both stages.
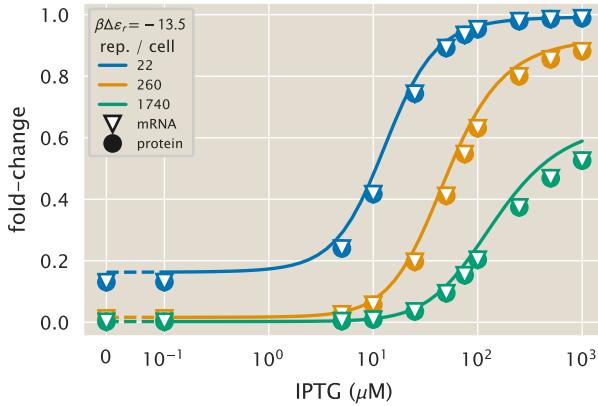
Figure 5.10: **Comparison of the equilibrium and kinetic inducer titration predictions.** The equilibrium model (solid lines) and the kinetic model with variation over the cell cycle (solid circles and white triangles) predictions are compared for varying repressor copy numbers and inducer concentrations. The equilibrium model is directly computed as Eq. 5 of reference [113] with repressor-DNA binding energy $\Delta\varepsilon_r = -13.5\,k_B T$, while the kinetic model is computed by numerically integrating the moment dynamics over several cell cycles, and then averaging over the extent of a single cell cycle as defined in Eq. 5.109 The Python code (`ch5_fig10.py`) used to generate this figure can be found on the original paper's GitHub repository.

For completeness, Fig. 5.10 compares the kinetic and equilibrium models for the extended model of [113] in which the inducer concentration enters into the equation. The solid line is directly computed from Eq. 2.5. The hollow triangles and solid points follow the same procedure as for Fig. 5.9, where the only effect that the inducer is assumed to have in the kinetics is an effective change in the number of active repressors, affecting, therefore, $k_{\text{on}}^{(r)}$.

**Comparison Between Single- and Multi-Promoter Kinetic Model**

After these calculations, it is worth questioning whether this change in gene dosage is drastically different from the more straightforward picture of a kinetic model that ignores the gene copy number variability during the cell cycle. To this end, we systematically computed the average moments for varying repressor copy numbers and repressor-DNA affinities. We then compare these results with the moments obtained from a single-promoter model and their corresponding parameters. The derivation of the steady-state moments of the distribution for the single-promoter model is detailed in Sec. 4.3.

Fig. 5.9 and Fig. 5.10 both suggest that, since the dynamic multi-promoter model can reproduce the results of the equilibrium model at the first-moment level, it must then also be able to reproduce the results of the single-promoter model at this level (see Sec. 5.2). The interesting comparison comes with higher moments. A useful metric to consider for gene expression variability is the noise in gene expression [142]. This quantity, defined as the standard deviation divided by the mean, is a dimensionless metric of how much variability there is with respect to the mean of a distribution. As we will show below, this quantity differs from the commonly used metric known as the Fano factor (variance/mean). For experimentally determined expression levels in arbitrary fluorescent units, the noise is a dimensionless quantity while the Fano factor is not.

Fig. 5.11 shows the comparison of the predicted protein noise between the single- (dashed lines) and the multi-promoter model (solid lines) for different operators and repressor copy numbers. A striking difference between both is that the single-promoter model predicts that, as the inducer concentration increases, the standard deviation grows much slower than the mean, giving a very small noise. In comparison, the multi-promoter model has a much higher floor for the lowest value of the noise, reflecting the expected result that the variability in gene copy number across the cell cycle should increase the cell-to-cell variability in gene expression [84,118]

**Comparison with Experimental Data**

Having shown that the kinetic model presented in this section can not only reproduce the results from the equilibrium picture at the mean level (see Fig. 5.9 and Fig. 5.10), but make predictions for the cell-to-cell variability as quantified by the noise (see Fig. 5.11), we can assess whether or not this model can predict experimental measurements of the noise. For this, we take the single-cell intensity measurements (see Methods) to compute the noise at the protein level.
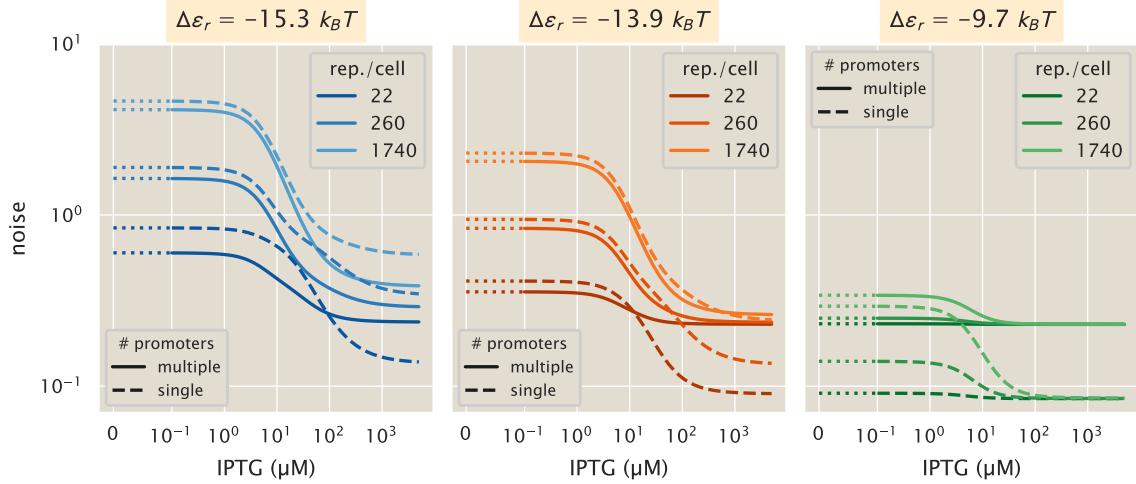
Figure 5.11: **Comparison of the predicted protein noise between a single- and a multi-promoter kinetic model.** Comparison of the noise (standard deviation/mean) between a kinetic model that considers a single promoter at all times (dashed line) and the multi-promoter model developed in this section (solid line) for different repressor operators. (A) Operator O1, $\Delta\varepsilon_r = -15.3\ k_B T$, (B) O2, $\Delta\varepsilon_r = -13.9\ k_B T$, (C) O3, $\Delta\varepsilon_r = -9.7\ k_B T$. The Python code (ch5_fig11.py) used to generate this figure can be found on the original paper's GitHub repository.

This metric differs from the Fano factor since the noise is a dimensionless quantity for arbitrary fluorescent units. To see why, consider that the noise is defined as

$$\text{noise} \equiv \frac{\sqrt{\langle p^2 \rangle - \langle p \rangle^2}}{\langle p \rangle}. \tag{5.114}$$

We assume that the intensity level of a cell $I$ is linearly proportional to the absolute protein count, i.e.,

$$I = \alpha p, \tag{5.115}$$

where $\alpha$ is the proportionality constant between arbitrary units and absolute protein number $p$. Substituting this definition on Eq. 5.114 gives

$$\text{noise} = \frac{\sqrt{\langle (\alpha I)^2 \rangle - \langle \alpha I \rangle^2}}{\langle \alpha I \rangle}. \tag{5.116}$$

Since $\alpha$ is a constant, it can be taken out of the average operator $\langle \cdot \rangle$, obtaining

$$\text{noise} = \frac{\sqrt{\alpha^2 \left( \langle I^2 \rangle - \langle I \rangle^2 \right)}}{\alpha \langle I \rangle} = \frac{\sqrt{\left( \langle I^2 \rangle - \langle I \rangle^2 \right)}}{\langle I \rangle}. \tag{5.117}$$

Notice that in Eq. 5.115, the linear proportionality between intensity and protein count has no intercept. This ignores the autofluorescence that cells without reporter would generate. To account for this, in practice, we compute

$$\text{noise} = \frac{\sqrt{\left(\langle(I - \langle I_{\text{auto}}\rangle)^2\rangle - \langle I - \langle I_{\text{auto}}\rangle\rangle^2\right)}}{\langle I - \langle I_{\text{auto}}\rangle\rangle}. \tag{5.118}$$

where $I$ is the intensity of the strain of interest and $\langle I_{\text{auto}}\rangle$ is the mean autofluorescence intensity, obtained from a strain that does not carry the fluorescent reporter gene.

Fig. 5.12 shows the comparison between theoretical predictions and experimental measurements for the unregulated promoter. The reason we split the data by operator despite the fact that, since these are unregulated promoters, they should, in principle, have identical expression profiles, is to make sure that this is the case precisely. We have found in the past that sequences downstream of the RNAP binding site can affect the expression level of constitutively expressed genes. We can see that both models, the single-promoter (gray dotted line) and the multi-promoter (black dashed line), underestimate the experimental noise to different degrees. The single-promoter model does a worse job predicting the experimental data since it does not account for the differences in gene dosage during the cell cycle. But still, we can see that accounting for this variability takes us to within a factor of two of the experimentally determined noise for these unregulated strains.

To further test the model's predictive power, we compare the predictions for the three-state regulated promoter. Fig. 5.13 shows the theoretical predictions for the single- and multi-promoter model for varying repressor copy numbers and repressor-DNA binding affinities as a function of the inducer concentration. Again, we can see that our zero-parameter fits systematically underestimate the noise for all strains and all inducer concentrations. We highlight that the $y$-axis is shown in a log-scale to emphasize this deviation more, but, as we will show in the next section, our predictions still fall within a factor of two from the experimental data.
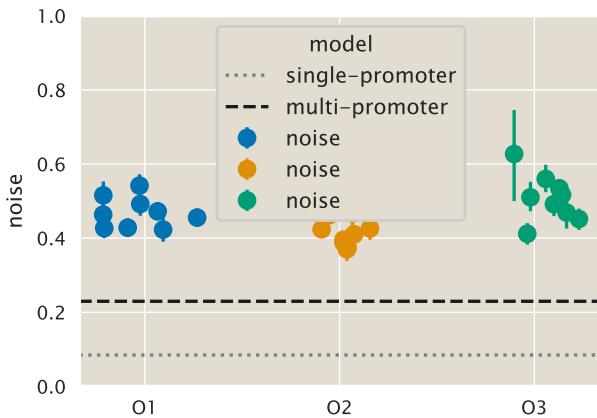
Figure 5.12: **Protein noise of the unregulated promoter.** Comparison of the experimental noise for different operators with the theoretical predictions for the single-promoter (gray dotted line) and the multi-promoter model (black dashed line). Each datum represents a single date measurement of the corresponding $\Delta lacI$ strain with $\geq 300$ cells. The points correspond to the median, and the error bars correspond to the 95% confidence interval as determined by 10,000 bootstrap samples. The Python code (`ch5_fig12.py`) used to generate this figure can be found on the original paper's GitHub repository.
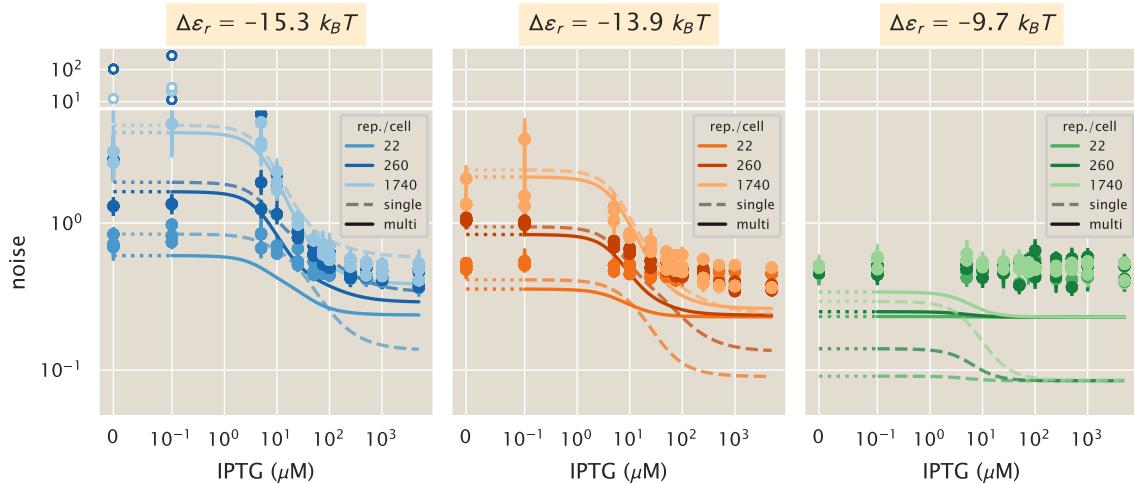


Figure 5.13: **Protein noise of the regulated promoter.** Comparison of the experimental noise for different operators ((A) O1, $\Delta\varepsilon_r = -15.3\ k_BT$, (B) O2, $\Delta\varepsilon_r = -13.9\ k_BT$, (C) O3, $\Delta\varepsilon_r = -9.7\ k_BT$) with the theoretical predictions for the single-promoter (dashed lines) and the multi-promoter model (solid lines). Points represent the experimental noise as computed from single-cell fluorescence measurements of different *E. coli* strains under 12 different inducer concentrations. The dotted line indicates the plot in linear rather than logarithmic scale. Each datum represents a single date measurement of the corresponding strain and IPTG concentration with $\geq 300$ cells. The points correspond to the median, and the error bars correspond to the 95% confidence interval as determined by 10,000 bootstrap samples. White-filled dots are plot at a different scale for better visualization. The Python code (`ch5_fig13.py`) used to generate this figure can be found on the original paper's GitHub repository.
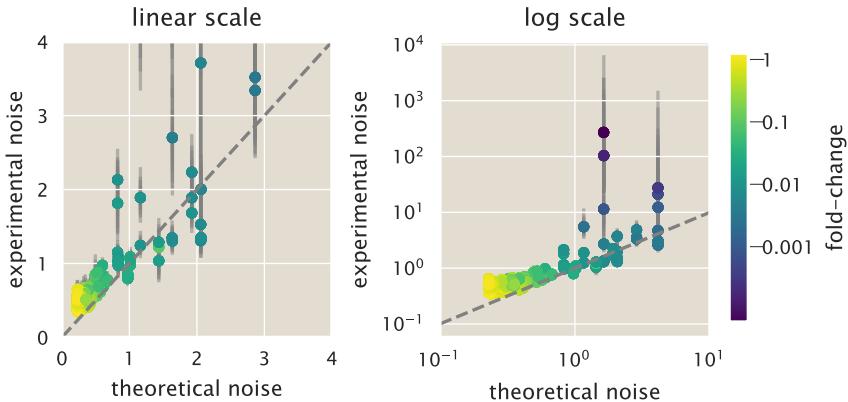
Figure 5.14: **Systematic comparison of theoretical vs. experimental noise in gene expression.** Theoretical vs. experimental noise both in linear (left) and log (right) scale. The dashed line shows the identity line of slope one and intercept zero. All data are colored by the corresponding experimental fold-changes value in gene expression, as indicated by the color bar. Each datum represents a single date measurement of the corresponding strain and IPTG concentration with $\geq 300$ cells. The points correspond to the median, and the error bars correspond to the 95% confidence interval as determined by 10,000 bootstrap samples. The Python code (`ch5_fig14.py`) used to generate this figure can be found on the original paper's GitHub repository.

## Systematic Deviation of the Noise in Gene Expression

Fig. 5.12 and Fig. 5.13 highlight that our model underestimates the cell-to-cell variability as measured by the noise. To further explore this systematic deviation, Fig. 5.14 shows the theoretical vs. experimental noise both in linear and log scale. As we can see, the data is systematically above the identity line. Their corresponding experimental fold-change values color the data. The data with the largest deviations from the identity line also corresponds to the data with the largest error bars and the smallest fold-change. This is because measurements with very small fold-changes correspond to intensities very close to the autofluorescence background. Therefore minimal changes when computing the noise are amplified given the ratio of std/mean. In Sec. 4.8, we will explore empirical ways to improve the agreement between our minimal and experimental data to guide future efforts to improve the minimal.

## 5.6 Maximum Entropy Approximation of Distributions

(Note: The Python code used for the calculations presented in this section can be found in the following link as an annotated Jupyter notebook.)

On the one hand, chemical master equations like the one here represent a hard mathematical challenge. Peccoud and Ycart derived a closed-form solution for the two-state promoter at the mRNA level [116]. In an impressive display of mathematical skills, Shahrezaei and Swain were able to derive an approximate solution for the one- (not considered in this work) and two-state promoter master equation at the protein level [142]. Nevertheless, both of these solutions do not give instantaneous insights about the distributions as they involve complicated terms such as confluent hypergeometric functions.

On the other hand, there has been a great deal of work to generate methods that can approximate the solution of these discrete state Markovian models [150–154]. In particular, for master equations like the one that concerns us here, whose moments can be easily computed, the moment expansion method provides a simple method to approximate the full joint distribution of mRNA and protein [154]. This section will explain the principles behind this method and show the implementation for our particular case study.

### The MaxEnt Principle

The principle of maximum entropy (MaxEnt), first proposed by E. T. Jaynes in 1957, tackles the question of, given limited information, what is the least biased inference one can make about a particular probability distribution [28]. In particular, Jaynes used this principle to show the correspondence between statistical mechanics and information theory, demonstrating, for example, that the Boltzmann distribution is the probability distribution that maximizes Shannon's entropy subject to a constraint that the average energy of the system is fixed.

To illustrate the principle, let us focus on a univariate distribution $P_X(x)$. The $n^{\text{th}}$ moment of the distribution for a discrete set of possible values of $x$ is given by

$$\langle x^n \rangle \equiv \sum_x x^n P_X(x). \tag{5.119}$$

Now assume that we have knowledge of the first $m$ moments $\langle \mathbf{x} \rangle_m = (\langle x \rangle, \langle x^2 \rangle, \ldots, \langle x^m \rangle)$. The question is then how we can use this information to build an estimator $P_H(x \mid \langle \mathbf{x} \rangle_m)$ of the distribution such that

$$\lim_{m \to \infty} P_H(x \mid \langle \mathbf{x} \rangle_m) \to P_X(x), \tag{5.120}$$

i.e., that the more moments we add to our approximation, the more the estimator distribution converges to the real distribution.

The MaxEnt principle tells us that our best guess for this estimator is to build it based on maximizing the Shannon entropy, constrained by the information we have about these $m$ moments. Shannon's entropy maximization guarantees that we are the least committed to information that we do not possess. The Shannon entropy for a univariate discrete distribution is given by [27]

$$H(x) \equiv -\sum_x P_X(x) \log P_X(x). \tag{5.121}$$

For an optimization problem subject to constraints, we make use of the method of the Lagrange multipliers. For this, we define the constraint equation $\mathcal{L}(x)$ as

$$\mathcal{L}(x) \equiv H(x) - \sum_{i=0}^{m} \left[ \lambda_i \left( \langle x^i \rangle - \sum_x x^i P_X(x) \right) \right], \tag{5.122}$$

where $\lambda_i$ is the Lagrange multiplier associated with the $i^{\text{th}}$ moment. The inclusion of the zeroth moment is an additional constraint to guarantee the normalization of the resulting distribution. Since $P_X(x)$ has a finite set of discrete values, when taking the derivative of the constraint equation with respect to $P_X(x)$, we chose a particular value of $X = x$. Therefore from the sum over all possible $x$ values, only a single term survives. With this in mind, we take the derivative of the constraint equation, obtaining

$$\frac{d\mathcal{L}}{dP_X(x)} = -\log P_X(x) - 1 - \sum_{i=0}^{m} \lambda_i x^i. \tag{5.123}$$

Equating this derivative to zero and solving for the distribution (that we now start calling $P_H(x)$, our MaxEnt estimator) gives

$$P_H(x) = \exp\left(-1 - \sum_{i=0}^{m} \lambda_i x^i\right) = \frac{1}{\mathcal{Z}} \exp\left(-\sum_{i=1}^{m} \lambda_i x^i\right), \quad (5.124)$$

where $\mathcal{Z}$ is the normalization constant that can be obtained by substituting this solution into the normalization constraint. This results in

$$\mathcal{Z} \equiv \exp\left(1 + \lambda_0\right) = \sum_x \exp\left(-\sum_{i=1}^{m} \lambda_i x^i\right). \quad (5.125)$$

Eq. 5.124 is the general form of the MaxEnt distribution for a univariate distribution. The computational challenge then consists of finding numerical values for the Lagrange multipliers $\{\lambda_i\}$ such that $P_H(x)$ satisfies our constraints. In other words, the Lagrange multipliers weigh the contribution of each term in the exponent such that when computing any of the moments, we recover the value of our constraint. Mathematically what this means is that $P_H(x)$ must satisfy

$$\sum_x x^n P_H(x) = \sum_x \frac{x^n}{\mathcal{Z}} \exp\left(-\sum_{i=1}^{m} \lambda_i x^i\right) = \langle x^n \rangle. \quad (5.126)$$

As an example of applying the MaxEnt principle, let us use a six-face die's classic problem. If we are only told that after a large number of die rolls, the mean value of the face is $\langle x \rangle = 4.5$ (note that a fair die has a mean of 3.5), what would the least biased guess for the distribution look like? The MaxEnt principle tells us that our best guess would be of the form

$$P_H(x) = \frac{1}{\mathcal{Z}} \exp\left(\lambda x\right). \quad (5.127)$$
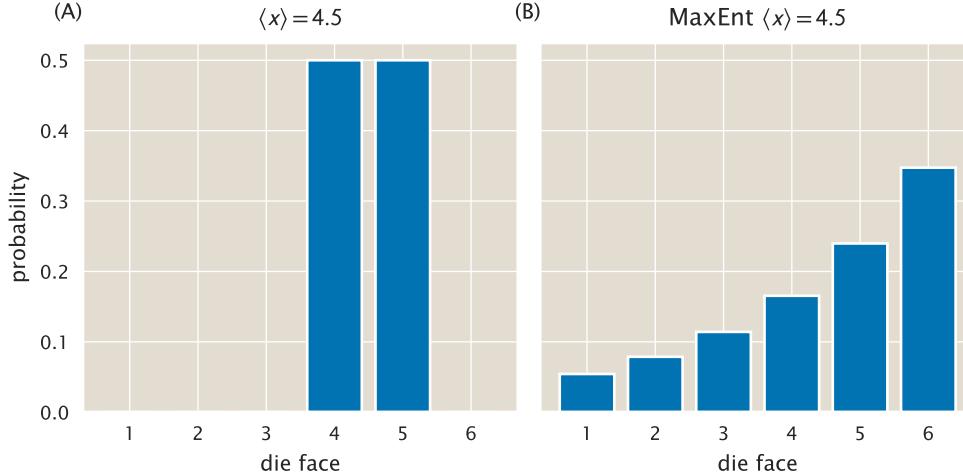
Figure 5.15: **Maximum entropy distribution of six-face die.** (A) Biased distribution consistent with the constraint $\langle x \rangle = 4.5$. (B) MaxEnt distribution also consistent with the constraint. The Python code (`ch5_fig15.py`) used to generate this figure can be found on the original paper's GitHub repository.

Using any numerical minimization package, we can easily find the value of the Lagrange multiplier $\lambda$ that satisfies our constraint. Fig. 5.15 shows two examples of distributions that satisfy the constraint. Panel (A) shows a distribution consistent with the 4.5 average where both 4 and 5 are equally likely. Nevertheless, in the information we got about the nature of the die, it was never stated that some of the faces were forbidden. In that sense, the distribution is committing to information about the process that we do not possess. Panel (B), by contrast, shows the MaxEnt distribution that satisfies this constraint. Since this distribution maximizes Shannon's entropy, it is guaranteed to be the least biased distribution given the available information.

**The mRNA and Protein Joint Distribution**

The MaxEnt principle can easily be extended to multivariate distributions. For our particular case, we are interested in the mRNA and protein joint distribution $P(m, p)$. The definition of a moment $\langle m^x p^y \rangle$ is a natural extension of Eq. 5.119 of the form

$$\langle m^x p^y \rangle = \sum_m \sum_p m^x p^y P(m, p). \tag{5.128}$$

As a consequence, the MaxEnt joint distribution $P_H(m, p)$ is of the form

$$P_H(m, p) = \frac{1}{\mathcal{Z}} \exp\left(-\sum_{(x,y)} \lambda_{(x,y)} m^x p^y\right),\tag{5.129}$$

where $\lambda_{(x,y)}$ is the Lagrange multiplier associated with the moment $\langle m^x p^y \rangle$, and again $\mathcal{Z}$ is the normalization constant, given by

$$\mathcal{Z} = \sum_m \sum_p \exp\left(-\sum_{(x,y)} \lambda_{(x,y)} m^x p^y\right).\tag{5.130}$$

Note that the sum in the exponent is taken over all available $(x, y)$ pairs that define the moment constraints for the distribution.

**The Bretthorst Rescaling Algorithm**

The Lagrange multipliers' determination suffers from a numerical underflow and overflow problem due to the difference in magnitude between the constraints. This becomes a problem when higher moments are taken into account. The resulting numerical values for the Lagrange multipliers end up being separated by several orders of magnitude. For routines such as Newton-Raphson or other minimization algorithms that can be used to find these Lagrange multipliers, these different scales become problematic.

To get around this problem, we implemented a variation to the algorithm due to G. Larry Bretthorst—Jaynes' last student. With a straightforward argument, we can show that linearly rescaling the constraints, the Lagrange multipliers, and the "rules" for computing each of the moments, i.e., each of the individual products that go into the moment calculation should converge to the same MaxEnt distribution. To see this, let us consider a univariate distribution $P_X(x)$ that we are trying to reconstruct given the first two moments $\langle x \rangle$, and $\langle x^2 \rangle$. The MaxEnt distribution can be written as

$$P_H(x) = \frac{1}{\mathcal{Z}} \exp\left(-\lambda_1 x - \lambda_2 x^2\right) = \frac{1}{\mathcal{Z}} \exp\left(-\lambda_1 x\right) \exp\left(-\lambda_2 x^2\right).\tag{5.131}$$

We can always rescale the terms in any way and obtain the same result. Assume that, for some reason, we want to rescale the quadratic terms by a factor $a$. We can define a new Lagrange multiplier $\lambda_2' \equiv \frac{\lambda_2}{a}$ that compensates for the rescaling of the terms, obtaining

$$P_H(x) = \frac{1}{\mathcal{Z}} \exp\left(-\lambda_1 x\right) \exp\left(-\lambda_2' a x^2\right). \tag{5.132}$$

Computationally it might be more efficient to find the numerical value of $\lambda_2'$ rather than $\lambda_2$ maybe because it is of the same order of magnitude as $\lambda_1$. Then we can always multiply $\lambda_2'$ by $a$ to obtain back the constraint for our quadratic term. This means that we can always rescale the MaxEnt problem to make it numerically more stable, then we can rescale it back to obtain the value of the Lagrange multipliers. The key to the Bretthorst algorithm lies in selecting what rescaling factor to choose to make the numerical inference more efficient.

Bretthorst's algorithm goes even further by further transforming the constraints and the variables to make the constraints orthogonal, making the computation much more effective. We now explain the algorithm's implementation for our joint distribution of interest $P(m, p)$.

**Algorithm Implementation**

Let the $M \times N$ matrix $\mathbf{A}$ contain all the factors used to compute the moments that serve as constraints, where each entry is of the form

$$A_{ij} = m_i^{x_j} \cdot p_i^{y_j}. \tag{5.133}$$

In other words, recall that to obtain any moment $\langle m^x p^y \rangle$, we compute

$$\langle m^x p^y \rangle = \sum_m \sum_p m^x p^y P(m, x). \tag{5.134}$$

If we have $M$ possible $(m, p)$ pairs in our truncated sample space (because we ca not include the sample space up to infinity) $\{(m, p)_1, (m, p)_2, \ldots (m, p)_N\}$, and we have $N$ exponent pairs $(x, y)$ corresponding to the $N$ moments used to constraint the maximum entropy distribution $\{(x, y)_1, (x, y)_2, \ldots, (x, y)_N\}$, then matrix $\mathbf{A}$ contains all the possible $M$ by $N$ terms of the form described in Eq. 5.133.

Let also **v** be a vector of length $N$ containing all the constraints with each entry of the form

$$v_j = \langle m^{x_j} p^{y_j} \rangle, \tag{5.135}$$

i.e., the information that we have about the distribution. That means that the constraint equation $\mathcal{L}$ to be used for this problem takes the form

$$\mathcal{L} = -\sum_i P_i \ln P_i + \lambda_0 \left( 1 - \sum_i P_i \right) + \sum_{j>0} \lambda_j \left( v_j - \sum_i A_{ij} P_i \right), \tag{5.136}$$

where $\lambda_0$ is the Lagrange multiplier associated with the normalization constraint and $\lambda_j$ is the Lagrange multiplier associated with the $j^{\text{th}}$ constraint. This constraint equation is equivalent to Eq. 5.122, but now all the details of how to compute the moments are specified in matrix **A**.

With this notation in hand, we now proceed to rescale the problem. The first step consists of rescaling the terms to compute the entries of the matrix **A**. As mentioned before, this is the crucial feature of the Bretthorst algorithm; the particular choice of rescaling factor used in the algorithm empirically promotes that the rescaled Lagrange multipliers are of the same order of magnitude. The rescaling takes the form

$$A'_{ij} = \frac{A_{ij}}{G_j}, \tag{5.137}$$

where $G_j$ serves to rescale the moments, providing numerical stability to the inference problem. Bretthorst proposes an empirical rescaling that satisfies

$$G_j^2 = \sum_i A_{ij}^2, \tag{5.138}$$

or, in terms of our particular problem,

$$G_j^2 = \sum_m \sum_p (m^{x_j} p^{y_j})^2. \tag{5.139}$$

What this indicates is that each pair $m_i^{x_j} p_i^{y_j}$ is normalized by the square root of the sum of all pairs of the same form squared.

Since we rescale the factors involved in computing the constraints, the constraints must also be rescaled simply as

$$v'_j = \langle m^{x_j} p^{y_j} \rangle' = \frac{\langle m^{x_j} p^{y_j} \rangle}{G_j}. \tag{5.140}$$

The Lagrange multipliers must compensate for this rescaling since the probability must add up to the same value at the end of the day. Therefore, we rescale the $\lambda_j$ terms as

$$\lambda'_j = \lambda_j G_j, \tag{5.141}$$

such that any $\lambda_j A_{ij} = \lambda'_j A'_{ij}$. If this empirical value for the rescaling factor makes the rescaled Lagrange multipliers $\lambda'_j$ be of the same order of magnitude, this by itself would already improve the algorithm convergence. Bretthorst proposes another linear transformation to make the optimization routine even more efficient. For this, we generate orthogonal constraints that make Newton-Raphson and similar algorithms converge faster. The transformation is as follows

$$A''_{ik} = \sum_j e_{jk} A'_{ij}, \tag{5.142}$$

for the entires of matrix $\mathbf{A}$, and

$$v''_k = \sum_j e_{jk} u'_j, \tag{5.143}$$

for entires of the constraint vector $\mathbf{v}$, finally

$$\lambda''_k = \sum_j e_{jk} \beta_j, \tag{5.144}$$

for the Lagrange multipliers. Here $e_{jk}$ is the $j^{\text{th}}$ component of the $k^{\text{th}}$ eigenvector of the matrix $\mathbf{E}$ with entries

$$E_{kj} = \sum_i A'_{ik} A'_{ij}. \tag{5.145}$$

This transformation guarantees that the matrix $\mathbf{A}''$ has the property

$$\sum_i A''_{ij} A''_{jk} = \beta_j \delta_{jk}, \tag{5.146}$$

where $\beta_j$ is the $j^{\text{th}}$ eigenvalue of the matrix $\mathbf{E}$ and $\delta_{jk}$ is the Kronecker delta function. This means that, as desired, the constraints are orthogonal to each other, improving the algorithm convergence speed.

**Predicting Distributions for Simple Repression Constructs**

Having explained the theoretical background and the practical difficulties, and a workaround strategy proposed by Bretthorst, we implemented the inference using the moments obtained from averaging over the variability along the cell cycle (see Sec. 5.4). Fig. 5.16 and Fig. 5.17 present these inferences for both mRNA and protein levels, respectively, for different values of the repressor-DNA binding energy and repressor copy numbers per cell. From these plots, we can easily appreciate that even though the mean of each distribution changes as the induction level changes, there is a lot of overlap between distributions. This, as a consequence, means that at the single-cell level, cells cannot perfectly resolve between different inputs.

**Comparison with Experimental Data**

Now that we have reconstructed an approximation of the probability distribution $P(m, p)$, we can compare this with our experimental measurements. But just as detailed in the single-cell microscopy, measurements are given in arbitrary units of fluorescence. Therefore, we cannot directly compare our predicted protein distributions with these values. To get around this issue, we use the fact that the fold-change in gene expression that we defined as the ratio of the gene expression level in the presence of the repressor and the expression level of a knockout strain is a non-dimensional quantity. Therefore, we normalize all of our single-cell measurements by the mean fluorescence value of the $\Delta lacI$ strain with the proper background fluorescence subtracted as explained in the noise measurements. In the case of the theoretical predictions of the protein distribution, we also normalize each protein value by the predicted mean protein level $\langle p \rangle$, having now non-dimensional scales that can be directly compared. Fig. 5.18 shows the experimental (color curves) and theoretical (dark dashed line) cumulative distribution functions for the three $\Delta lacI$ strains. As in Fig. 5.12, we do not expect differences between the operators, but we explicitly plot them separately to ensure that this is the case. We can see right away that as we would expect, given the model's limitations to
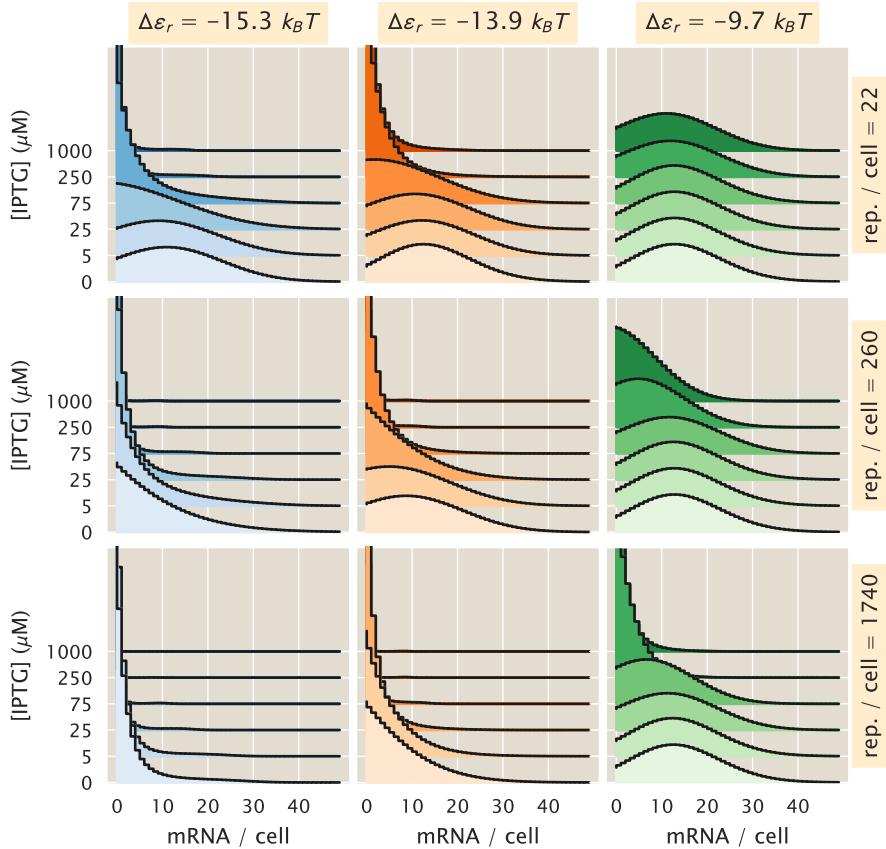
Figure 5.16: **Maximum entropy mRNA distributions for simple repression constructs.** mRNA distributions for different biophysical parameters. From left to right, the repressor-DNA affinity decreases as defined by the three lacI operators O1 ($-15.3\ k_BT$), O2 ($-13.9\ k_BT$), and O3 ($-9.7\ k_BT$). From top to bottom, the mean repressor copy number per cell increases. The curves on each plot represent different IPTG concentrations. Each distribution was fitted using the first three moments of the mRNA distribution. The Python code (`ch5_fig16.py`) used to generate this figure can be found on the original paper's GitHub repository..

predict the noise and skewness of the distribution accurately, the model does not accurately predict the data. Our model predicts a narrower distribution compared to what we measured with single-cell microscopy.

The same narrower prediction applies to the regulated promoters. Fig. 5.19, shows the theory-experiment comparison of the cumulative distribution functions for different repressor binding sites (different figures), repressor copy numbers (rows), and inducer concentrations (columns). In general, the predictions are systematically narrower compared to the actual experimental data.
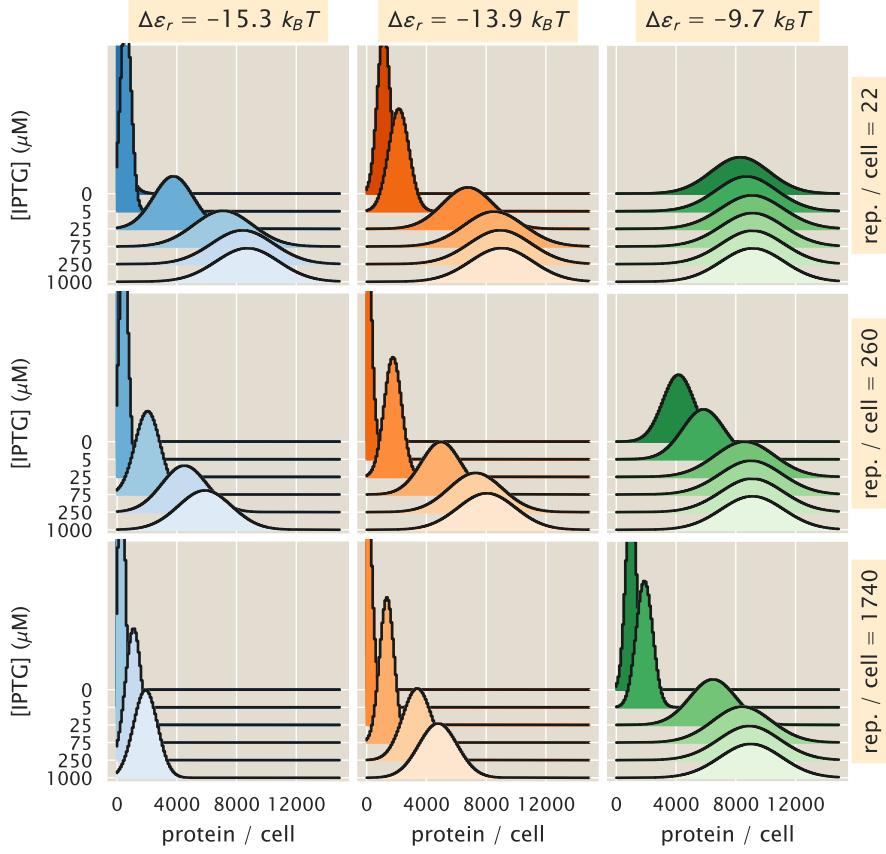
Figure 5.17: **Maximum entropy protein distributions for simple repression constructs.** Protein distributions for different biophysical parameters. From left to right, the repressor-DNA affinity decreases as defined by the three lacI operators O1 ($-15.3\ k_BT$), O2 ($-13.9\ k_BT$), and O3 ($-9.7\ k_BT$). From top to bottom, the mean repressor copy number per cell increases. The curves on each plot represent different IPTG concentrations. Each distribution was fitted using the first six moments of the protein distribution. The Python code (`ch5_fig17.py`) used to generate this figure can be found on the original paper's GitHub repository..

## 5.7  Gillespie Simulation of the Master Equation

(Note: The Python code used for the calculations presented in this section can be found in the following link as an annotated Jupyter notebook.)
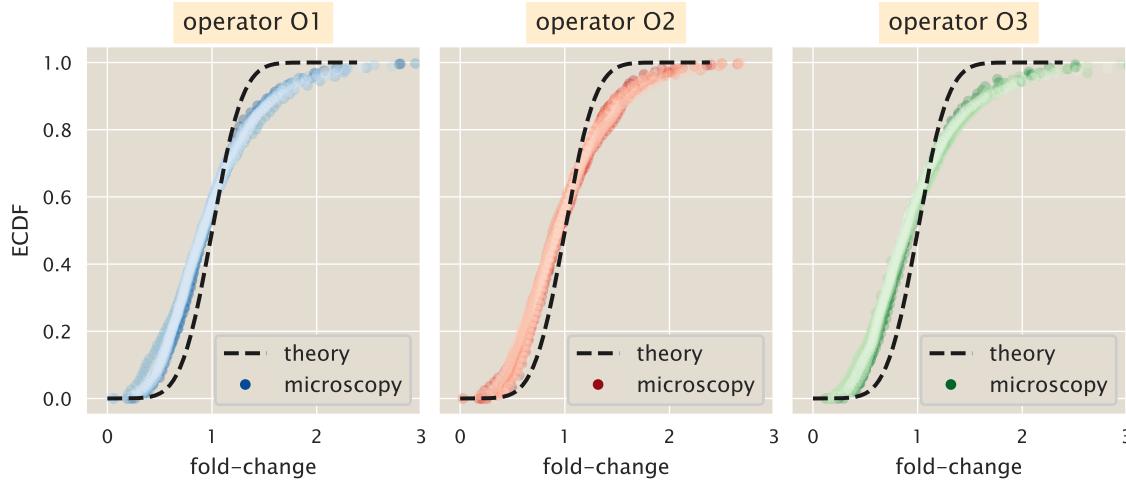
Figure 5.18: **Experiment vs. theory comparison for Δ*lacI* strain.** Example fold-change empirical cumulative distribution functions (ECDF) for strains with no repressors and different operators. The color curves represent single-cell microscopy measurements while the dashed black lines represent the theoretical distributions as reconstructed by the maximum entropy principle. The theoretical distributions were fitted using the first six moments of the protein distribution. The Python code (`ch5_fig18.py`) used to generate this figure can be found on the original paper's GitHub repository.

So far, we have generated a way to compute an approximated form of the joint distribution of protein and mRNA $P(m, p)$ as a function of the moments of the distribution $\langle m^x p^y \rangle$. This is a non-conventional form to work with the resulting distribution of the master equation. A more conventional approach to work with master equations whose closed-form solutions are not known or not computable is to use stochastic simulations, commonly known as Gillespie simulations. To benchmark our approach's performance based on distribution moments and maximum entropy, we implemented the Gillespie algorithm. Our implementation, as detailed in the corresponding Jupyter notebook, makes use of just-in-time compilation as implemented with the Python package numba.
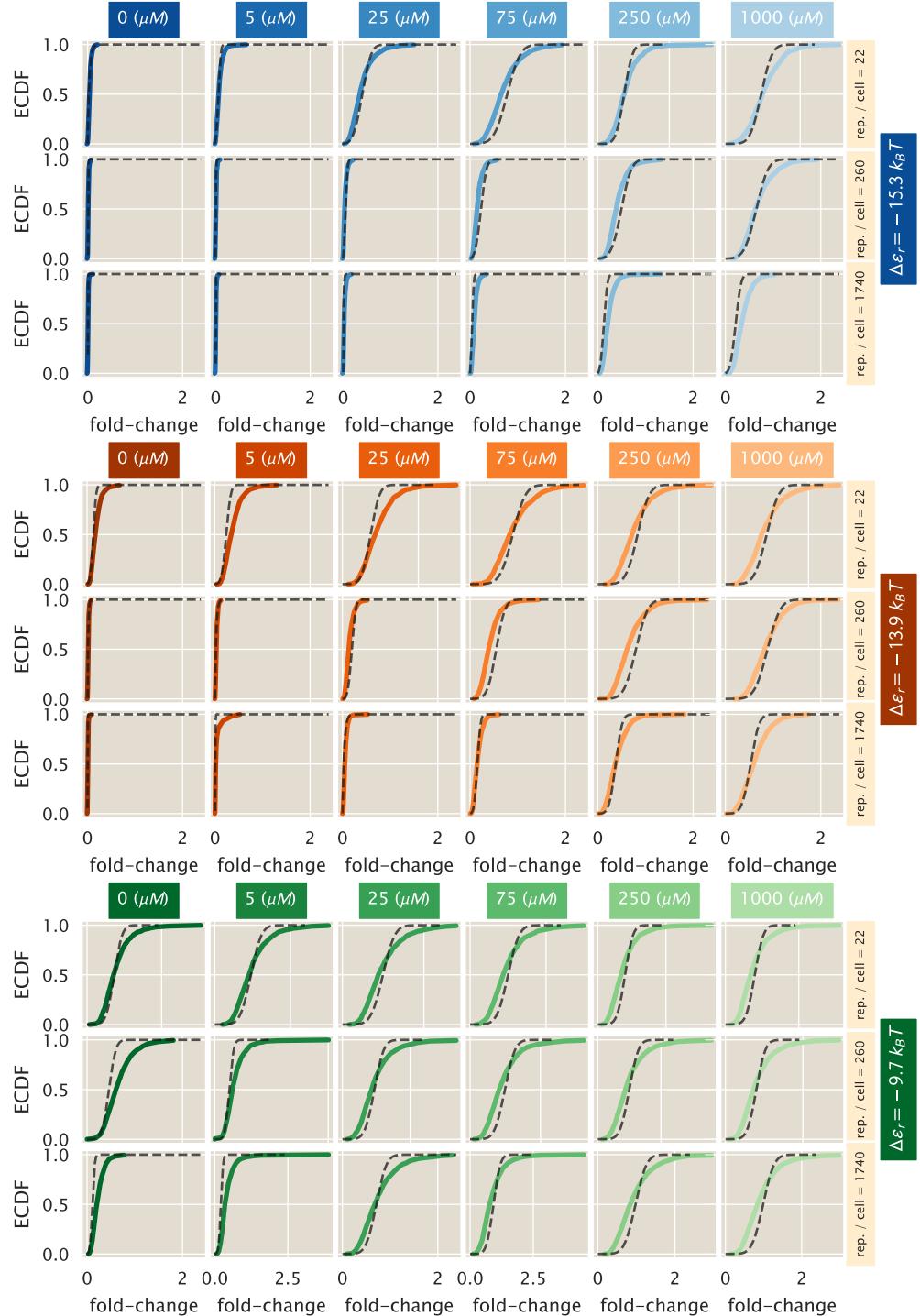
Figure 5.19: **Experiment vs. theory comparison for regulated promoters.** Example fold-change empirical cumulative distribution functions (ECDF) for regulated strains with the three operators (different colors) as a function of repressor copy numbers (rows) and inducer concentrations (columns). The color curves represent single-cell microscopy measurements, while the dashed black lines represent the theoretical distributions as reconstructed by the maximum entropy principle. The theoretical distributions were fitted using the first six moments of the protein distribution. The Python code (`ch5_fig19.py`) used to generate this figure can be found on the original paper's GitHub repository.
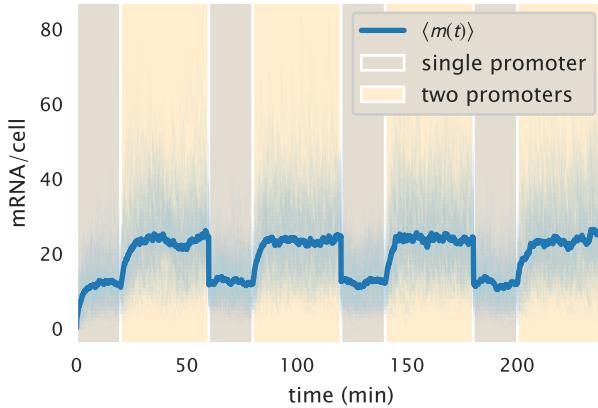
Figure 5.20: **Stochastic trajectories of mRNA counts.** 100 stochastic trajectories generated with the Gillespie algorithm for mRNA counts over time for a two-state unregulated promoter. Cells spend a fraction of the cell cycle with a single copy of the promoter (light brown) and the rest of the cell cycle with two copies (light yellow). When trajectories reach a new cell cycle, the mRNA counts undergo binomial partitioning to simulate the cell division. The Python code (`ch5_fig20.py`) used to generate this figure can be found on the original paper's GitHub repository.

### mRNA Distribution with Gillespie Simulations

To confirm that the Gillespie simulation's implementation was correct, we perform the simulation at the mRNA level, for which the closed-form solution of the steady-state distribution is known as detailed in Sec. 5.2. Fig. 5.20 shows example trajectories of mRNA counts. Each of these trajectories was computed over several cell cycles, where the cell division was implemented, generating a binomially distributed random variable that depended on the last mRNA count before the division event.

To check the implementation of our stochastic algorithm, we generated several of these stochastic trajectories to reconstruct the mRNA steady-state distribution. These reconstructed distributions for a single- and double-copy of the promoter can be compared with Eq. 5.10—the steady-state distribution for the two-state promoter. Fig. 5.21 shows the excellent agreement between the stochastic simulation and the analytical result, confirming that our implementation of the Gillespie simulation is correct.
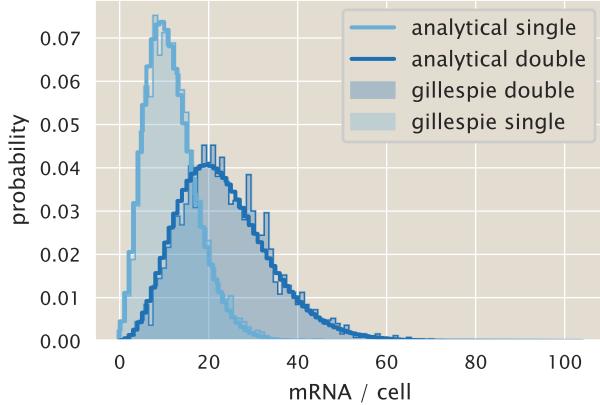
Figure 5.21: **Comparison of analytical and simulated mRNA distribution.** Solid lines show the steady-state mRNA distributions for one copy (light blue) and two copies of the promoter (dark blue) as defined by Eq. 5.10. Shaded regions represent the corresponding distribution obtained using 2500 stochastic mRNA trajectories and taking the last cell cycle to approximate the distribution. The Python code (`ch5_fig21.py`) used to generate this figure can be found on the original paper's GitHub repository.

## Protein Distribution with Gillespie Simulations

Having confirmed that our implementation of the Gillespie algorithm that includes the binomial partitioning of molecules reproduces analytical results, we extended the implementation to include protein counts. Fig. 5.22 shows representative trajectories for both mRNA and protein counts over several cell cycles. Especially for the protein, we can see that it takes several cell cycles for counts to converge to the dynamical steady-state observed with the deterministic moment equations. Once this steady-state is reached, the ensemble of trajectories between cell cycles looks very similar.

From these trajectories, we can compute the steady-state protein distribution, taking into account the cell-age distribution, as detailed in Sec. 5.5. Fig. 5.23 shows the comparison between this distribution and the one generated using the maximum entropy algorithm. Although the notorious differences between the distributions, the Gillespie simulation and the maximum entropy results are indistinguishable in terms of the mean, variance, and skewness of the distribution. We remind the reader that the maximum entropy approximates the distribution that gets better the more moments we add. We, therefore, claim that the approximation works
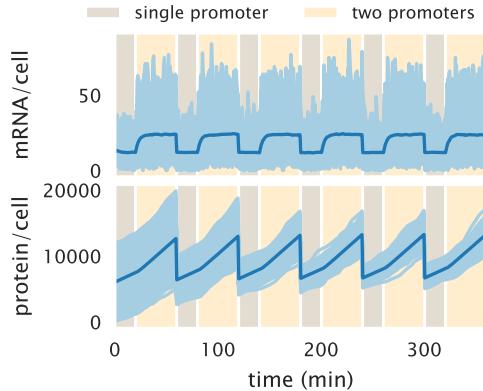
Figure 5.22: **Stochastic trajectories of mRNA and protein counts.** 2500 protein counts over time for a two-state unregulated promoter. Cells spend a fraction of the cell cycle with a single copy of the promoter (light brown) and the rest of the cell cycle with two copies (light yellow). When trajectories reach a new cell cycle, the molecule counts undergo binomial partitioning to simulate the cell division. The Python code (`ch5_fig22.py`) used to generate this figure can be found on the original paper's GitHub repository..

sufficiently well for our purpose. The enormous advantage of the maximum entropy approach comes from the computation time. For the number of distributions needed for our calculations, the Gillespie algorithm proved to be a very inefficient method given the ample sample space. Our maximum entropy approach reduces the computation time by several orders of magnitude, allowing us to explore different regulatory models' parameters extensively.

## 5.8 Computational Determination of the Channel Capacity

(Note: The Python code used for the calculations presented in this section can be found in the following link as an annotated Jupyter notebook.)
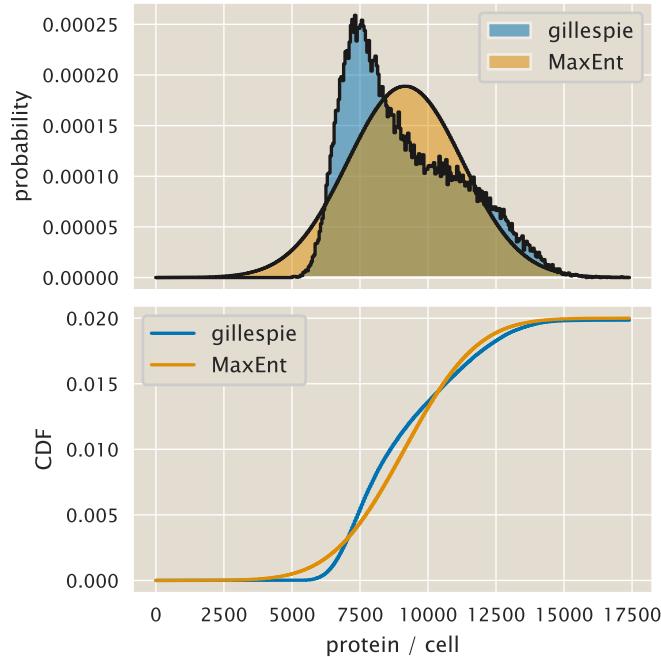
Figure 5.23: **Comparison of protein distributions.** Comparison of the protein distribution generated with Gillespie stochastic simulations (blue curve) and the maximum entropy approach (orange curve). The upper panel shows the probability mass function. The lower panel compares the cumulative distribution functions. The Python code (ch5_fig23.py) used to generate this figure can be found on the original paper's GitHub repository.

This section details the computation of the channel capacity of the simple genetic circuit shown in Fig. 3.5. The channel capacity is defined as the mutual information between input $c$ and output $p$ maximized over all possible input distributions $P(c)$ [27]. In principle, there is an infinite number of input distributions, so the task of finding $\hat{P}(c)$, the input distribution at channel capacity, requires an algorithmic approach that guarantees the convergence to this distribution. Tkačik, Callan, and Bialek developed an analytical approximation to find the $\hat{P}(c)$ distribution [108]. The validity of their so-called small noise approximation requires the standard deviation of the output distribution $P(p \mid c)$ to be much smaller than the distribution domain. For our particular case, such a condition is not satisfied given the spread of the inferred protein distributions shown in Fig. 3.4.

Fortunately, a numerical algorithm can approximate $\hat{P}(c)$ for discrete distributions. In 1972, Blahut and Arimoto independently came up with an algorithm mathematically shown to converge to $\hat{P}(c)$ [123]. To compute both the theoretical and the experimental channel capacity shown in Fig. 3.5, we implemented Blahut's algorithm. In the following section, we detail the definitions needed for the algorithm. Then we describe how to compute the experimental channel capacity when the distribution bins are not clear given the arbitrary intrinsic nature of microscopy fluorescence measurements.

**Blahut's algorithm**

Following [123], we implemented the algorithm to compute the channel capacity. We define $\mathbf{p_c}$ to be an array containing the probability of each of the input inducer concentrations (twelve concentrations, See Methods). Each entry $j$ of the array is then of the form

$$p_c^{(j)} = P(c = c_j), \tag{5.147}$$

with $j \in \{1, 2, \ldots, 12\}$. The objective of the algorithm is to find the entries $p_c^{(j)}$ that maximize the mutual information between inputs and outputs. We also define $\mathbf{Q}$ to be a $|\mathbf{p_c}|$ by $|\mathbf{p_{p|c}}|$ matrix, where $|\cdot|$ specifies the length of the array, and $\mathbf{p_{p|c}}$ is an array containing the probability distribution of an output given a specific value of the input. In other words, the matrix $\mathbf{Q}$ recollects all of the individual output distribution arrays $\mathbf{p_{p|c}}$ into a single object. Then each entry of the matrix $\mathbf{Q}$ is of the form

$$Q^{(i,j)} = P(p = p_i \mid c = c_j). \tag{5.148}$$

For the case of the theoretical predictions of the channel capacity (solid lines in Fig. 3.5), the entries of the matrix $\mathbf{Q}$ are given by the inferred maximum entropy distributions as shown in Fig. 3.4. In the next section, we will discuss how to define this matrix for the single-cell fluorescence measurements. Having defined these matrices, we proceed to implement the algorithm shown in Figure 1 of [123].

**Channel Capacity from Arbitrary Units of Fluorescence**

A difficulty when computing the channel capacity between inputs and outputs from experimental data is that ideally, we would like to compute

$$C(g;c) \equiv \sup_{P(c)} I(g;c), \tag{5.149}$$

where $g$ is the gene expression level, and $c$ is the inducer concentration. But in reality, we are computing

$$C(f(g);c) \equiv \sup_{P(c)} I(f(g);c), \tag{5.150}$$

where $f(g)$ is a function of gene expression that has to do with our mapping from the YFP copy number to some arbitrary fluorescent value as computed from the images taken with the microscope. The data processing inequality, as derived by Shannon himself, tells us that for a Markov chain of the form $c \to g \to f(g)$, it must be true that [27]

$$I(g;c) \geq I(f(g);c), \tag{5.151}$$

meaning that information can only be lost when mapping from the real relationship between gene expression and inducer concentration to a fluorescence value.

On top of that, given the limited number of samples that we have access to when computing the channel capacity, there is a bias in our estimate given this undersampling. The definition of accurate, unbiased descriptors of mutual information is still an area of active research. For our purposes, we will use the method described in [155]. The basic idea of the method is to write the mutual information as a series expansion in terms of inverse powers of the sample size, i.e.

$$I_{\text{biased}} = I_{\infty} + \frac{a_1}{N} + \frac{a_2}{N^2} + \cdots , \tag{5.152}$$

where $I_{\text{biased}}$ is the biased estimate of the mutual information as computed from experimental data, $I_\infty$ is the quantity we would like to estimate, being the unbiased mutual information when having access to an infinity number of experimental samples and the coefficients $a_i$ depend on the underlying distribution of the signal and the response. This is an empirical choice to be tested. Intuitively, this choice satisfies the limit that as the number of samples from the distribution grows, the empirical estimate of the mutual information $I_{\text{biased}}$ should get closer to the actual value $I_\infty$.

In principle, for a good number of data points, the terms of higher-order become negligible. So we can write the mutual information as

$$I_{\text{biased}} \approx I_\infty + \frac{a_1}{N} + \mathcal{O}(N^{-2}). \tag{5.153}$$

This means that if this particular arbitrary choice of functional form is a good approximation, when computing the mutual information for varying numbers of samples—by taking subsamples of the experimental data—we expect to find a linear relationship as a function of the inverse of these number of data points. From this linear relationship, the intercept is a bias-corrected estimate of the mutual information. Therefore, we can bootstrap the data by taking different sample sizes and then use the Blahut-Arimoto algorithm we implemented earlier to estimate the biased channel capacity. We can then fit a line and extrapolate when $1/N = 0$, which corresponds to our unbiased estimate of the channel capacity.

Let us go through each of the steps to illustrate the method. Fig. 5.24 show a typical data set for a strain with an O2 binding site ($\Delta\varepsilon_r = -13.9\ k_B T$) and $R = 260$ repressors per cell. Each of the distributions in arbitrary units is binned into a specified number of bins to build matrix $\mathbf{Q}$.
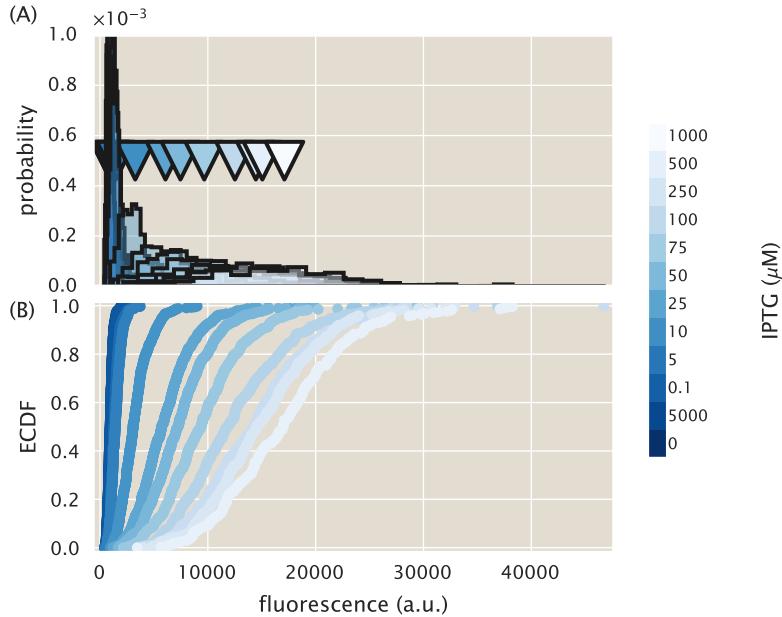
Figure 5.24: **Single-cell fluorescence distributions for different inducer concentrations.** Fluorescence distribution histogram (A) and cumulative distribution function (B) for a strain with 260 repressors per cell and a binding site with binding energy $\Delta\varepsilon_r = -13.9\ k_B T$. The different curves show the single-cell fluorescence distributions under the 12 different IPTG concentrations used throughout this work. The triangles in (A) show the mean of each of the distributions. The Python code (`ch5_fig24.py`) used to generate this figure can be found on the original paper's GitHub repository.

Given a specific number of bins used to construct **Q**, we subsample a fraction of the data and compute the channel capacity for such matrix using the Blahut-Arimoto algorithm. Fig. 5.25 shows an example where 50% of the data on each distribution from Fig. 5.24 was sampled and binned into 100 equal bins. The counts on each of these bins are then normalized and used to build matrix **Q** that is then fed to the Blahut-Arimoto algorithm. We can see that for these 200 bootstrap samples, the channel capacity varies by $\approx 0.1$ bits. Not a significant variability; nevertheless, we consider it essential to bootstrap the data multiple times to estimate the channel capacity better.
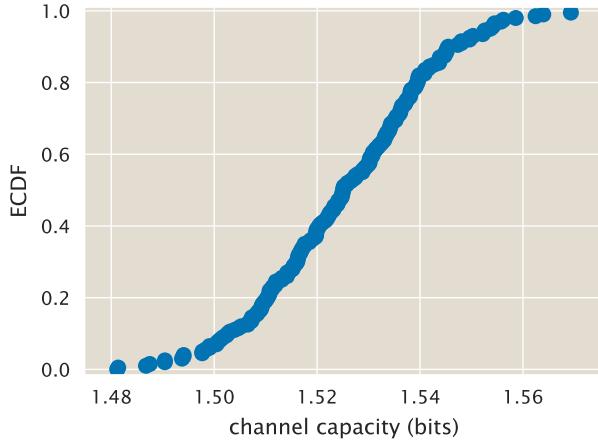
Figure 5.25: **Channel capacity bootstrap for experimental data.** The cumulative distribution function of the resulting channel capacity estimates obtained by subsampling 200 times 50% of each distribution shown in Fig. 5.24, binning it into 100 bins, and feeding the resulting **Q** matrix to the Blahut-Arimoto algorithm. The Python code (`ch5_fig25.py`) used to generate this figure can be found on the original paper's GitHub repository.

Eq. 5.153 tells us that if we subsample each of the distributions from Fig. 5.24 at different fractions and plot them as a function of the inverse sample size, we will find a linear relationship if the expansion of the mutual information is valid. To test this idea, we repeated the bootstrap estimate of Fig. 5.25 sampling 10%, 20%, and so on until taking 100% of the data. We repeated this for different numbers of bins since *a priori* for arbitrary units of fluorescence, we do not have a way to select the optimal number of bins. Fig. 5.26 shows the result of these estimates. We can see that the linear relationship proposed in Eq. 5.153 holds for all number of bins selected. We also note that the value of the linear regression intercept varies depending on the number of bins.
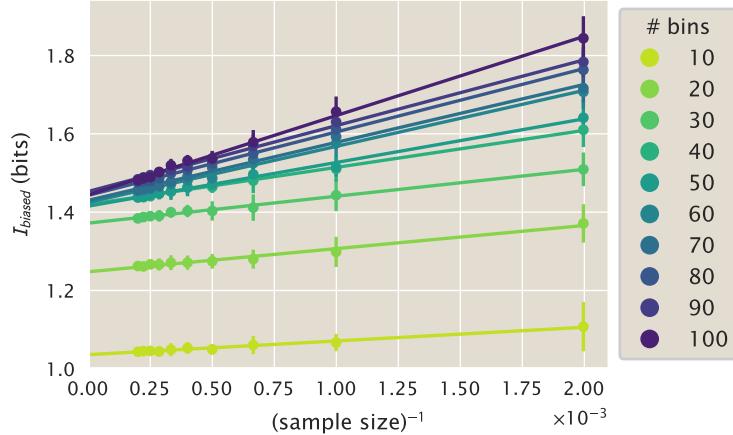
Figure 5.26: **Inverse sample size vs. channel capacity.** As indicated in Eq. 5.153, if the channel capacity obtained for different subsample sizes of the data are plotted against the inverse sample size, there must exist a linear relationship between these variables. Here, we perform 15 bootstrap samples of the data from Fig. 5.24, then we bin these samples using a different number of bins, and finally perform a linear regression (solid lines) between the bootstrap channel capacity estimates and the inverse sample size. The Python code (`ch5_fig26.py`) used to generate this figure can be found on the original paper's GitHub repository..

To address the variability in the estimates of the unbiased channel capacity $I_\infty$ we again follow the methodology suggested in [155]. We perform the data subsampling and computation of the channel capacity for a varying number of bins. As a control, we perform the same procedure with shuffled data, where the structure that connects the fluorescence distribution to the inducer concentration input is lost. The expectation is that this control should give a channel capacity of zero if the data is not "over-binned." Once the number of bins is too high, we expect some structure to emerge in the data that would cause the Blahut-Arimoto algorithm to return non-zero channel capacity estimates.

Fig. 5.27 shows the result of the unbiased channel capacity estimates obtained for the data shown in Fig. 5.24. For the blue curve, we can distinguish three phases: 1. A rapid increment from 0 bits to about 1.5 bits as the number of bins increases. 2. A flat region between $\approx$ 50 and 1000 bins. 3. A second rapid increment for a large number of bins.
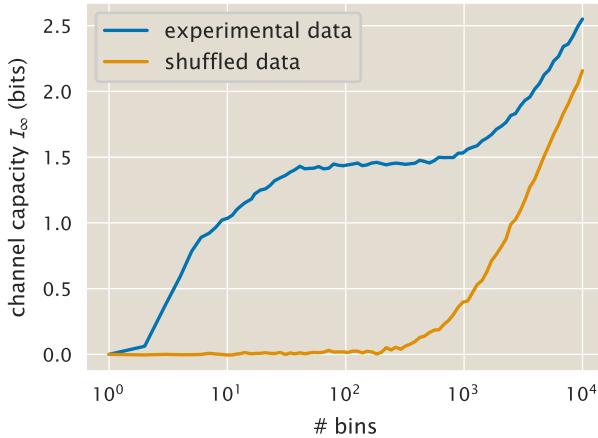
Figure 5.27: **Channel capacity as a function of the number of bins.** Unbiased channel capacity estimates we obtained from linear regressions as in Fig. 5.26. The blue curve shows the estimates obtained from the data shown in Fig. 5.24. The orange curve is generated from estimates where the same data is shuffled, losing the relationship between fluorescence distributions and inducer concentration. The Python code (`ch5_fig27.py`) used to generate this figure can be found on the original paper's GitHub repository.

We can see that the randomized data presents two phases only: 1. A flat region where there is, as expected, no information being processed since the structure of the data was lost when the data was shuffled. 2. A region with a fast growth of the channel capacity as the over-binning generates separated peaks on the distribution, making it look like there is a structure in the data.

We take the flat region of the experimental data ($\approx 100$ bins) to be our best unbiased estimate of the channel capacity from this experimental dataset.

**Assumptions Involved in the Computation of the Channel Capacity**

An interesting suggestion by Professor Gasper Tkacik was to dissect the different physical assumptions that went into the construction of the input-output function $P(p \mid c)$, and their relevance when comparing the theoretical channel capacities with the experimental inferences. In what follows, we describe the relevance of four important aspects that all affect the predictions of the information processing capacity.
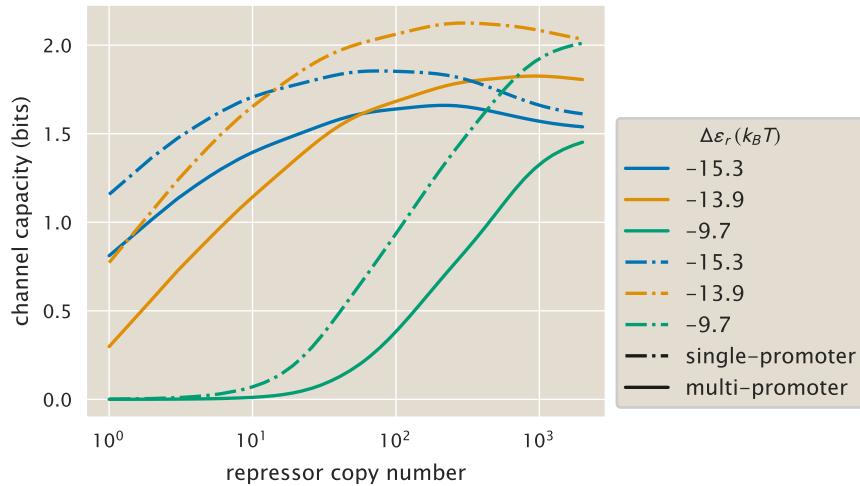
Figure 5.28: **Comparison of channel capacity predictions for single- and multi-promoter models.** Channel capacity for the multi-promoter model (solid lines) vs. the single-promoter steady-state model (dot-dashed lines) as a function of repressor copy numbers for different repressor-DNA binding energies. The single-promoter model assumes Poissonian protein degradation ($\gamma_p > 0$) and steady-state, while the multi-promoter model accounts for gene copy number variability during the cell cycle and has protein degradation as an effect due to dilution as cells grow and divide. The Python code (`ch5_fig28.py`) used to generate this figure can be found on the original paper's GitHub repository.

**(i) Cell Cycle Variability.**

We think that the inclusion of the gene copy number variability during the cell cycle and non-Poissonian protein degradation is crucial to our estimation of the input-output functions and channel capacity. This variability in gene copy number is an additional source of noise that systematically decreases the system's ability to resolve different inputs. The absence of the effects that the gene copy number variability and the protein partition have on the information processing capacity leads to an overestimate of the channel capacity, as shown in Fig. 5.28. When these noise sources are included in our inferences, we capture the experimental channel capacities with no additional fit parameters.

**(ii) Non-Gaussian Noise Distributions.**

For the construction of the probability distributions used in Chapter 3 (Fig. 3.4), we utilized the first six moments of the protein distribution. The maximum entropy

formalism tells us that the more constraints we include in the inference, the closer the maximum entropy distribution will be to the real distribution. But *a priori* there is no way of knowing how many moments should be included to capture the distribution's essence. In principle, two moments could suffice to describe the entire distribution as happens with the Gaussian distribution. To compare the effect of including more or fewer constraints on the maximum entropy inference, we constructed maximum entropy distributions using an increasing number of moments from 2 to 6. We then computed the Kullback-Leibler divergence $D_{KL}$ of the form

$$D_{KL}(P_6(p \mid c) || P_i(p \mid c)) = \sum_p P_6(p \mid c) \log_2 \frac{P_6(p \mid c)}{P_i(p \mid c)}, \qquad (5.154)$$

where $P_i(p \mid c)$ is the maximum entropy distribution constructed with the first $i$ moments, $i \in \{2, 3, 4, 5, 6\}$. Since the Kullback-Leibler divergence $D_{KL}(P || Q)$ can be interpreted as the amount of information lost by assuming the incorrect distribution $Q$ when the correct distribution is $P$, we used this metric as a way of how much information we would have lost by using fewer constraints compared to the six moments used in Chapter 3.

Fig. 5.29 shows this comparison for different operators and repressor copy numbers. We can see from here that using fewer moments as constraints gives the same result. This is because most of the values of the Kullback-Leibler divergence is significantly smaller than 0.1 bits. The entropy of these distributions is, in general, $> 10$ bits, so we would lose less than 1% of the information contained in these distributions by utilizing only two moments as constraints. Therefore the use of non-Gaussian noise is not an essential feature for our inferences.
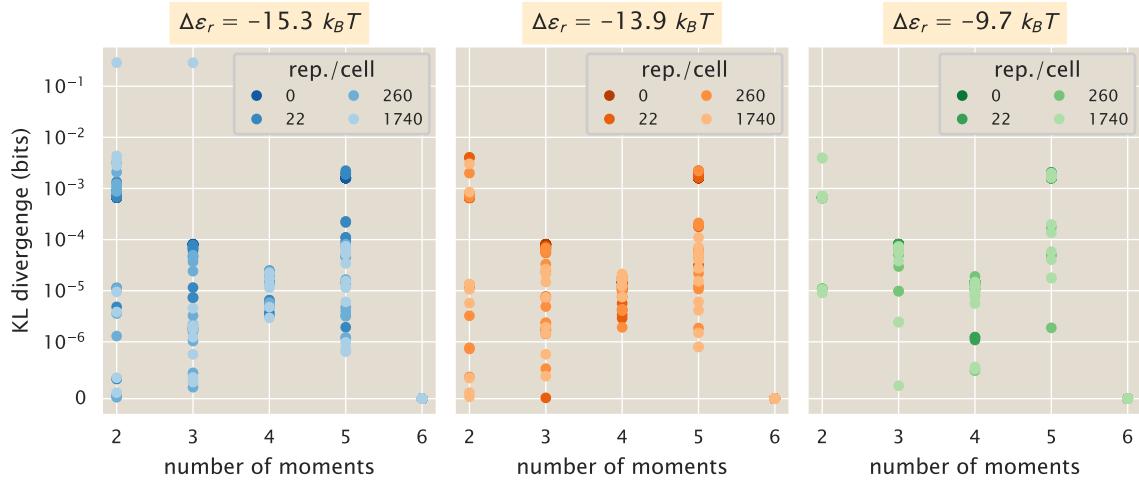
Figure 5.29: **Measuring the loss of information by using a different number of constraints.** The Kullback-Leibler divergence was computed between the maximum entropy distribution constructed using the first six moments of the distribution and a variable number of moments. The Python code (`ch5_fig29.py`) used to generate this figure can be found on the original paper's GitHub repository.

### (iii) Multi-State Promoter.

This particular point is something that we are still exploring from a theoretical perspective. We have shown that, to capture the single-molecule mRNA FISH data, a single-state promoter would not suffice. This model predicts a Poisson distribution as the steady-state, and the data shows super Poissonian noise. Given the bursty nature of gene expression, we opt to use a two-state promoter to reflect effective transcriptionally "active" and "inactive" states. We are currently exploring alternative formulations of this model to turn it into a single state with a geometrically distributed burst size.

**(iv) Optimal vs Log-Flat Distributions.**

The relevance of having to use the Blahut-Arimoto algorithm to predict the maximum mutual information between input and outputs was to understand the best-case scenario. We show the comparison between theoretical and experimental input-output functions $P(p \mid c)$ in Fig. 5.19. Given the good agreement between these distributions, we could compute the mutual information $I(c; p)$ for any arbitrary input distribution $P(c)$ and obtain a good agreement with the corresponding experimental mutual information.

The reason we opted to report the mutual information at the channel capacity was to put the results in context. By reporting the upper bound in performance of these genetic circuits, we can start to dissect how different molecular parameters such as repressor-DNA binding affinity or repressor copy number affect the ability of this genetic circuit to extract information from the environmental state.

## 5.9 Empirical Fits to Noise Predictions

(Note: The Python code used for the calculations presented in this section can be found in the following link as an annotated Jupyter notebook.)

In Fig. 3.3(C), we show that our minimal model has a systematic deviation on the gene expression noise predictions compared to the experimental data. This systematics will need to be addressed on an improved version of the minimal model presented in this work. To guide the insights into the origins of this systematic deviation in this appendix, we will explore the model's empirical modifications to improve the agreement between theory and experiment.

**Multiplicative Factor for the Noise**

The first option we will explore is to modify our noise predictions by a constant multiplicative factor. This means that we assume that the relationship between our minimal model predictions and the data for noise in gene expression are of the form

$$\text{noise}_{\text{exp}} = \alpha \cdot \text{noise}_{\text{theory}}, \tag{5.155}$$
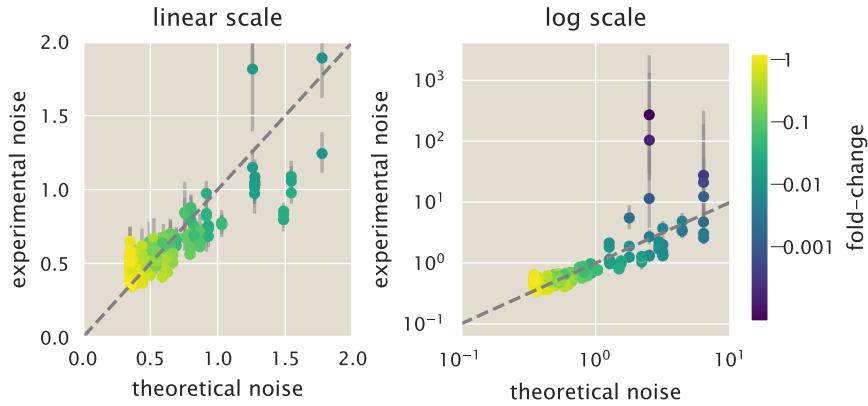
Figure 5.30: **Multiplicative factor in improving theoretical vs. experimental comparison of noise in gene expression.** Theoretical vs. experimental noise both in linear (left) and log (right) scale. The dashed line shows the identity line of slope 1 and intercept zero. All data are colored by the corresponding experimental fold-changes in gene expression as indicated by the color bar. The *x*-axis was multiplied by a factor of $\approx 1.5$ as determined by linear regression from the data in Fig. 5.11. Each datum represents a single date measurement of the corresponding strain and IPTG concentration with $\geq 300$ cells. The points correspond to the median, and the error bars correspond to the 95% confidence interval as determined by 10,000 bootstrap samples. The Python code (`ch5_fig30.py`) used to generate this figure can be found on the original paper's GitHub repository..

where $\alpha$ is a dimensionless constant to be fit from the data. The data, especially in Fig. 5.12, suggests that our predictions are within a factor of $\approx$ two from the experimental data. To further check that intuition, we performed a weighted linear regression between the experimental and theoretical noise measurements. The weight for each datum was proportional to the bootstrap errors in the noise estimate (this to have poorly determined noises weigh less during the linear regression). This regression with no intercept shows that a factor of two systematically improves the theoretical vs. experimental predictions. Fig. 5.30 shows the improved agreement when the noise's theoretical predictions are multiplied by $\approx 1.5$.

For completeness, Fig. 5.31 shows the noise in gene expression as a function of the inducer concentration, including this factor of $\approx 1.5$. Thus, overall a simple multiplicative factor improves the predictive power of the model.
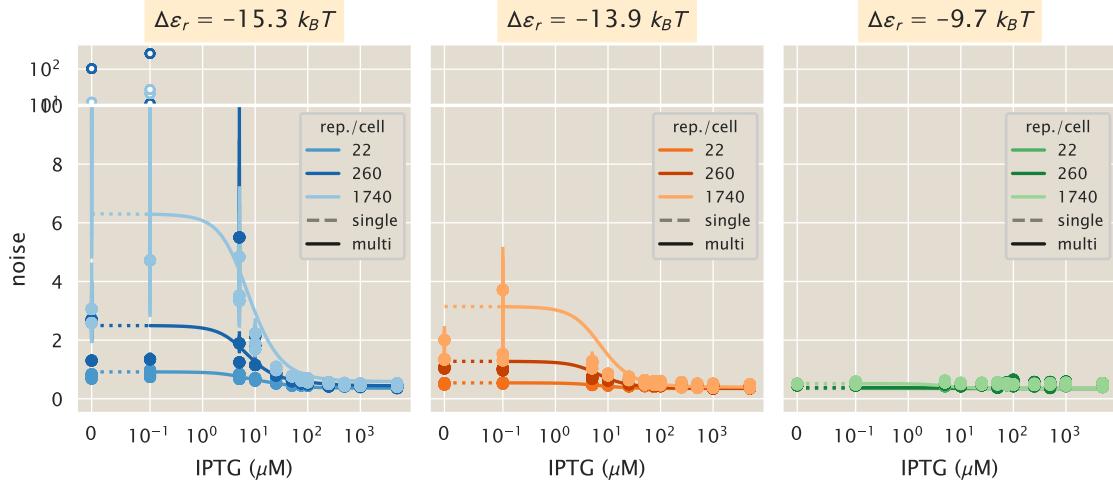
Figure 5.31: **Protein noise of the regulated promoter with multiplicative factor.** Comparison of the experimental noise for different operators ((A) O1, $\Delta\varepsilon_r = -15.3 \ k_B T$, (B) O2, $\Delta\varepsilon_r = -13.9 \ k_B T$, (C) O3, $\Delta\varepsilon_r = -9.7 \ k_B T$) with the theoretical predictions for the multi-promoter model. Linear regression revealed that multiplying the theoretical noise prediction by a factor of $\approx 1.5$ would improve agreement between theory and data. Points represent the experimental noise as computed from single-cell fluorescence measurements of different *E. coli* strains under 12 different inducer concentrations. The dotted line indicates the plot in linear rather than logarithmic scale. Each datum represents a single date measurement of the corresponding strain and IPTG concentration with $\geq 300$ cells. The points correspond to the median, and the error bars correspond to the 95% confidence interval as determined by 10,000 bootstrap samples. White-filled dots are plot at a different scale for better visualization. The Python code (`ch5_fig31.py`) used to generate this figure can be found on the original paper's GitHub repository.

## Additive Factor for the Noise

As an alternative way to empirically improve our model's predictions, we will now test the idea of an additive constant. What this means is that our minimal model underestimates the noise in gene expression as

$$\text{noise}_{\text{exp}} = \beta + \text{noise}_{\text{theory}}, \tag{5.156}$$

where $\beta$ is an additive constant to be determined from the data. As with the multiplicative constant, we performed a regression to determine this empirical additive constant, comparing experimental and theoretical gene expression noise values. We use the error in the 95% bootstrap confidence interval as a weight for the linear regression. Fig. 5.32 shows the resulting theoretical vs. experimental noise where $\beta \approx 0.2$. We can see a great improvement in the agreement between theory and experiment with this additive constant.
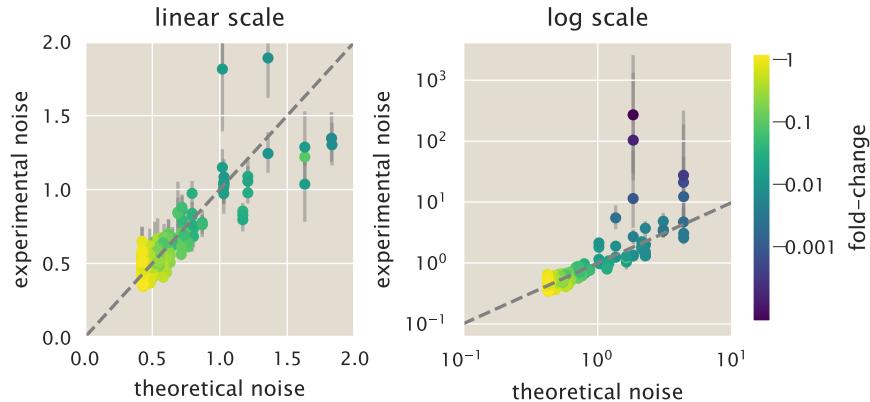
Figure 5.32: **Additive factor in improving theoretical vs. experimental comparison of noise in gene expression.** Theoretical vs. experimental noise both in linear (left) and log (right) scale. The dashed line shows the identity line of slope 1 and intercept zero. All data are colored by the corresponding experimental fold-change in gene expression as indicated by the color bar. A value of $\approx 0.2$ was added to all values in the $x$-axis as determined by linear regression from the data in Fig. 5.11. Each datum represents a single date measurement of the corresponding strain and IPTG concentration with $\geq 300$ cells. The points correspond to the median, and the error bars correspond to the 95% confidence interval as determined by 10,000 bootstrap samples. The Python code (ch5_fig32.py) used to generate this figure can be found on the original paper's GitHub repository..

For completeness, Fig. 5.33 shows the noise in gene expression as a function of the inducer concentration, including this additive factor of $\beta \approx 0.2$. If anything, the additive factor seems to improve the agreement between theory and data even more than the multiplicative factor.
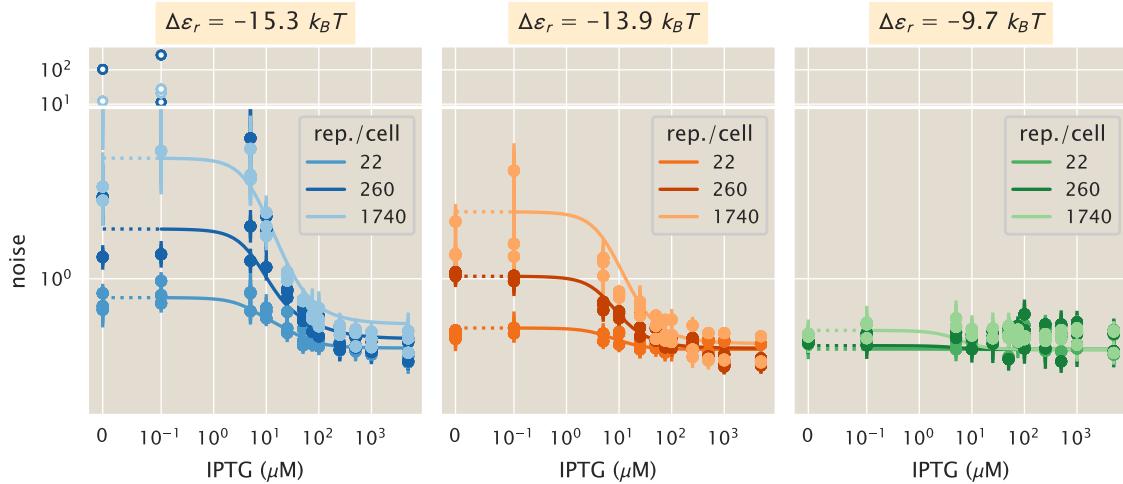
Figure 5.33: **Protein noise of the regulated promoter with an additive factor.** Comparison of the experimental noise for different operators ((A) O1, $\Delta\varepsilon_r = -15.3\ k_BT$, (B) O2, $\Delta\varepsilon_r = -13.9\ k_BT$, (C) O3, $\Delta\varepsilon_r = -9.7\ k_BT$) with the theoretical predictions for the multi-promoter model. Linear regression revealed that an additive factor of $\approx 0.2$ to the theoretical noise prediction would improve agreement between theory and data. Points represent the experimental noise as computed from single-cell fluorescence measurements of different *E. coli* strains under 12 different inducer concentrations. The dotted line indicates the plot in linear rather than logarithmic scale. Each datum represents a single date measurement of the corresponding strain and IPTG concentration with $\geq 300$ cells. The points correspond to the median, and the error bars correspond to the 95% confidence interval as determined by 10,000 bootstrap samples. White-filled dots are plot at a different scale for better visualization. The Python code (`ch5_fig33.py`) used to generate this figure can be found on the original paper's GitHub repository.

### Correction Factor for Channel Capacity with a Multiplicative Factor

A constant multiplicative factor can reduce the discrepancy between the model predictions and the data concerning the noise (standard deviation/mean) in protein copy numbers. Finding the equivalent correction for the channel capacity requires gaining insights from the so-called small noise approximation [108]. The small noise approximation assumes that the input-output function can be modeled as a Gaussian distribution in which the standard deviation is small. Using these assumptions, one can derive a closed-form for the channel capacity. Although our data and model predictions do not satisfy the small noise approximation requirements, we can gain some intuition for how the channel capacity would scale given a systematic deviation in the cell-to-cell variability predictions compared with the data.

Using the small noise approximation, one can derive the form of the input distribution at channel capacity $P^*(c)$. To do this, we use the fact that there is a deterministic relationship between the input inducer concentration $c$ and the mean output protein value $\langle p \rangle$, therefore we can work with $P(\langle p \rangle)$ rather than $P(c)$ since the deterministic relation allows us to write

$$P(c)dc = P(\langle p \rangle)d\langle p \rangle. \tag{5.157}$$

Optimizing over all possible distributions $P(\langle p \rangle)$ using calculus of variations results in a distribution of the form

$$P^*(\langle p \rangle) = \frac{1}{\mathcal{Z}}\frac{1}{\sigma_p(\langle p \rangle)}, \tag{5.158}$$

where $\sigma_p(\langle p \rangle)$ is the standard deviation of the protein distribution as a function of the mean protein expression, and $\mathcal{Z}$ is a normalization constant defined as

$$\mathcal{Z} \equiv \int_{\langle p(c=0) \rangle}^{\langle p(c \to \infty) \rangle} \frac{1}{\sigma_p(\langle p \rangle)}d\langle p \rangle. \tag{5.159}$$

Under these assumptions, the small noise approximation tells us that the channel capacity is of the form [108]

$$I = \log_2\left(\frac{\mathcal{Z}}{\sqrt{2\pi e}}\right). \tag{5.160}$$

From the theory-experiment comparison we know that the standard deviation predicted by our model is systematically off by a factor of two compared to the experimental data, i.e.,

$$\sigma_p^{\text{exp}} = 2\sigma_p^{\text{theory}}. \tag{5.161}$$

This then implies that the normalization constant $\mathcal{Z}$ between theory and experiment must follow a relationship of the form

$$\mathcal{Z}^{\text{exp}} = \frac{1}{2}\mathcal{Z}^{\text{theory}}. \tag{5.162}$$

With this relationship, the small noise approximation would predict that the difference between the experimental and theoretical channel capacity should be of the form

$$I^{\text{exp}} = \log_2\left(\frac{\mathcal{Z}^{\text{exp}}}{\sqrt{2\pi e}}\right) = \log_2\left(\frac{\mathcal{Z}^{\text{theory}}}{\sqrt{2\pi e}}\right) - \log_2(2). \tag{5.163}$$
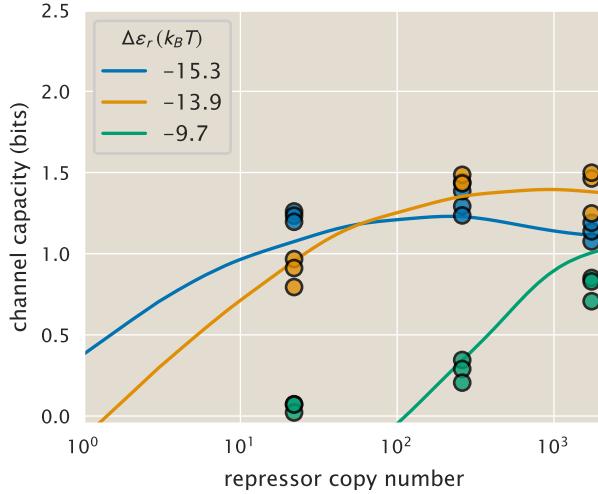
Figure 5.34: **Additive correction factor for channel capacity.** Solid lines represent the theoretical predictions of the channel capacity shown in (A). The dashed lines show the resulting predictions with a constant shift of -0.43 bits. Points represent single biological replicas of the inferred channel capacity. The Python code (`ch5_fig34.py`) used to generate this figure can be found on the original paper's GitHub repository.

Therefore under the small noise approximation, we would expect our predictions for the channel capacity to be off by a constant of 1 bit ($\log_2(2)$) of information. Again, the conditions for the small noise approximation do not apply to our data given the intrinsic level of cell-to-cell variability in the system; nevertheless, what this analysis tells us is that we expect that an additive constant should be able to explain the discrepancy between our model predictions and the experimental channel capacity. To test this hypothesis, we performed a "linear regression" between the model predictions and the experimental channel capacity with a fixed slope of 1. The intercept of this regression, -0.56 bits, indicates the systematic deviation we expect should explain the difference between our model and the data. Fig. 5.34 shows the comparison between the original predictions shown in Fig. 5.5(A) and the resulting predictions with this shift. Thus, other than the data with zero channel capacity, this shift can correct the systematic deviation for all data. We, therefore, conclude that our model ends up underestimating the experimentally determined channel capacity by a constant amount of 0.43 bits.
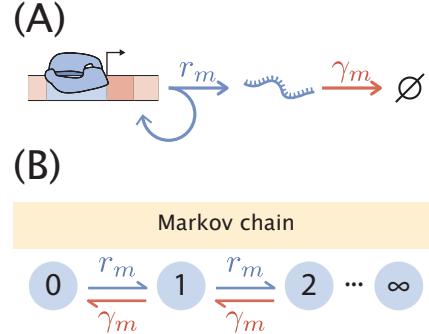
Figure 5.35: **One-state Poisson promoter.** (A) Schematic of the kinetics of the one-state promoter. mRNA is produced and degrade stochastically with a rate $r_m$ and $\gamma_m$, respectively. (B) Representation of the Markov chain for the state space that the promoter can be. The distribution $P(m, t)$ represents the probability of having a certain discrete number of mRNA $m$ at time $t$. The transition between states depends on the previously mentioned rates.

## 5.10 Derivation of the Steady-State mRNA Distribution

In this section, we will derive the two-state promoter mRNA distribution we quote in Sec. 5.2. For this method, we will make use of the so-called generating functions. Generating functions are mathematical objects on which we can encode a series of infinite numbers as coefficients of a power series. The power of generating functions comes from the fact that we can convert an infinite-dimensional system of coupled ordinary differential equations–in our case, the system of differential equations defining all probabilities $P(m, t)$ for $m \in \mathbb{Z}$–into a single partial differential equation that we can then solve to extract back the probability distributions.

To motivate the use of generating functions, we will begin with the simplest case: the one-state Poisson promoter.

### One-state Poisson promoter

We begin by defining the reaction scheme that defines the one-state promoter. Fig. 5.35 shows the schematic representation of the Poisson promoter as a simple cartoon (part (A)) and as the Markov chain that defines the state space of the system (part (B)).

The dynamics of the probability distribution $P(m,t)$ are governed by the chemical master equation

$$\frac{dP(m,t)}{dt} = \overbrace{r_m P(m-1,t)}^{m-1\to m} - \overbrace{r_m P(m,t)}^{m\to m+1} + \overbrace{\gamma_m(m+1)P(m+1,t)}^{m+1\to m} - \overbrace{\gamma_m m P(m,t)}^{m\to m-1}.$$

(5.164)

When solving for the distribution, our objective is to obtain the equation that defines $P(m,t)$ for all possible values of $m \in \mathbb{Z}$. The power of the generating functions is that these probability distribution values are used as a power series's coefficients. To make this clear, let us define the generating function $G(z,t)$ as

$$G(z,t) \equiv \sum_{m=0}^{\infty} z^m P(m,t),$$

(5.165)

where $z$ is a "dummy" variable that we do not care about. The reason this is useful is that if we find the closed-form solution for this generating function, and we are able to split the factor $z^m$ from its coefficient $P(m,t)$, then we will have to find the solution for the distribution. Furthermore, the generating function allows us to compute the moments of the distribution. For example, for the zeroth moment $\langle m^0 \rangle$, we know that

$$\langle m^0 \rangle = \sum_{m=0}^{\infty} m^0 P(m,t) = 1,$$

(5.166)

i.e., this is the normalization constraint of the distribution. From the definition of the generating function, we can then see that

$$G(1,t) = \sum_{m=0}^{\infty} 1^m P(m,t) = 1.$$

(5.167)

Furthermore, the first moment of the distribution is defined as

$$\langle m \rangle = \sum_{m=0}^{\infty} m P(m,t).$$

(5.168)

From the definition of the generating function, we can construct this quantity by computing

$$\left.\frac{\partial G(z,t)}{\partial z}\right|_{z=1} = \frac{\partial}{\partial z}\left[\sum_{m=0}^{\infty} z^m P(m,t)\right]_{z=1} = \sum_{m=0}^{\infty} m P(m,t).$$

(5.169)

Therefore we have that

$$\langle m \rangle = \left. \frac{\partial G(z,t)}{\partial z} \right|_{z=1}. \tag{5.170}$$

Similar constructions can be built for higher moments of the distribution.

Let us then apply the definition of the generating function to Eq. 5.164. For this, we multiply both sides by $z^m$ and sum over all values of $m$, obtaining

$$\sum_{m=0}^{\infty} z^m \frac{dP(m,t)}{dt} = \sum_{m=0}^{\infty} z^m \left[ r_m P(m-1,t) - r_m P(m,t) \right. \tag{5.171}$$
$$\left. + \gamma_m(m+1)P(m+1,t) - \gamma_m m P(m,t) \right].$$

Distributing the sum, we find

$$\frac{d}{dt} \sum_{m=0}^{\infty} z^m P(m,t) = \sum_{m=0}^{\infty} z^m r_m P(m-1,t) - \sum_{m=0}^{\infty} z^m r_m P(m,t) \tag{5.172}$$
$$+ \sum_{m=0}^{\infty} z^m \gamma_m(m+1)P(m+1,t) - \sum_{m=0}^{\infty} z^m \gamma_m m P(m,t).$$

We see that the terms involving $z^m P(m,t)$ can be directly substituted with Eq. 5.165. For the other terms, we have to be slightly more clever. The first trick will allow us to rewrite the term involving $z^m m P(m,t)$ as

$$\sum_m z^m \cdot m \cdot P(m,t) = \sum_m z \frac{\partial z^m}{\partial z} P(m,t),$$
$$= \sum_m z \frac{\partial}{\partial z} \left( z^m P(m,t) \right), \tag{5.173}$$
$$= z \frac{\partial}{\partial z} \left( \sum_m z^m P(m,t) \right),$$
$$= z \frac{\partial G(z,t)}{\partial z}.$$

Next, let us deal with the term involving $(m + 1)$. We first define $k = m + 1$. With this, we can write

$$\sum_{m=0}^{\infty} z^m \cdot (m+1) \cdot P(m+1, t) = \sum_{k=1}^{\infty} z^{k-l} \cdot k \cdot P(k, t),$$
$$= z^{-1} \sum_{k=1}^{\infty} z^k \cdot k \cdot P(k, t),$$
$$= z^{-1} \sum_{k=0}^{\infty} z^k \cdot k \cdot P(k, t), \qquad (5.174)$$
$$= z^{-1} \left( z \frac{\partial G(z)}{\partial z} \right),$$
$$= \frac{\partial G(z)}{\partial z},$$

where for the third step, we reindexed the sum to include $k = 0$ since it does not contribute to the total sum. Finally, for the term involving $P(m - 1, t)$, we define $k = m - 1$. This allows us to rewrite the term as

$$\sum_{m=0}^{\infty} z^m P(m-1, t) = \sum_{k=-1}^{\infty} z^{k+1} P(k, t),$$
$$= \sum_{k=0}^{\infty} z^{k+1} P(k, t), \qquad (5.175)$$
$$= z \sum_{k=0}^{\infty} z^k P(k, t),$$
$$= z G(z, t).$$

For the second step, we reindexed the sum from $-1$ to $0$ since $P(-1, t) = 0$.

All of these clever reindexing allows us to rewrite Eq. 5.172 as

$$\frac{\partial G(z, t)}{\partial t} = rz G(z, t) - r G(z, t) + \gamma \frac{\partial G(z, t)}{\partial z} - \gamma z \frac{\partial G(z, t)}{\partial z}. \qquad (5.176)$$

Factorizing terms, we have

$$\frac{\partial G(z, t)}{\partial t} = -r G(z, t)(1 - z) + \gamma \frac{\partial G(z, t)}{\partial z}(1 - z). \qquad (5.177)$$

Let us appreciate how beautiful this is: we took an infinite-dimensional system of ordinary differential equations—the master equation—and turned it into a single partial differential equation (PDE). All we have to do now is solve this PDE, and then transform the solution into a power series to extract the distribution.

Let us focus on the steady-state case. For this, we set the time derivative to zero. Doing this cancels the $(1 - z)$ term, leaving a straightforward ordinary differential equation for $G(z)$

$$\frac{dG(z)}{dz} = \frac{r}{\gamma} G(z). \tag{5.178}$$

Solving this equation by separation of variables results in

$$G(z) = C e^{\frac{r}{\gamma} z}. \tag{5.179}$$

To obtain the integration constant, we use the normalization condition of the probability distribution (Eq. 5.167), obtaining

$$1 = C e^{\frac{r}{\gamma}} \Rightarrow C = e^{-\frac{r}{\gamma}}. \tag{5.180}$$

This means that the generating function takes the form

$$G(z) = e^{-\frac{r}{\gamma}} e^{\frac{r}{\gamma} z}. \tag{5.181}$$

All we have left is trying to rewrite the generating function as a power series on $z$. If we succeed in doing so, we will have recovered the probability distribution $P(m, t)$. For this, we simply use the Taylor expansion of $e^x$, obtaining

$$G(z) = e^{-\frac{r}{\gamma}} \sum_{m=0}^{\infty} \frac{\left(\frac{r}{\gamma} z\right)^m}{m!}. \tag{5.182}$$

From this form, it becomes clear how to split the $z^m$ term from the coefficient that, by the definition of the generating function, is the probability distribution we are looking for. The separation takes the form

$$G(z) = \sum_{m=0}^{\infty} z^m \left[ \frac{e^{-r/\gamma} \left(\frac{r}{\gamma}\right)^m}{m!} \right], \tag{5.183}$$

where we can see that we recover the expected Poisson distribution for this one-state promoter

$$P(m) = e^{-r/\gamma} \frac{\left(\frac{r}{\gamma}\right)^m}{m!}. \tag{5.184}$$

**Two-state promoter**

Having shown the generating function's power, let us now turn our attention to the relevant equation we are after: the two-state mRNA distribution. This model assumes that the promoter can exist in two discrete states (see Fig. 5.36(A)): a transcriptionally active state $A$ from which transcription can take place at a constant rate $r_m$, and an inactive state $I$ where no transcription takes place. The mRNA is stochastically degraded with a rate $\gamma_m$ regardless of the state of the promoter. Fig. 5.36(B) shows the Markov chain that connects all of the possible states of the promoter. For this particular case, there are not only "horizontal" transitions where the mRNA copy number changes, but "vertical" transitions where only the promoter's state changes. Because of this, we need to define two coupled master equations that take the form

$$
\begin{aligned}
\frac{dP_A(m,t)}{dt} = {} & -k_{\text{off}}^{(p)} P_A(m,t) + k_{\text{on}}^{(p)} P_I(m,t) \\
& + \gamma_m(m+1)P_A(m+1,t) - \gamma_m m P_n(m,t) \\
& + r_m P_A(m-1,t) - r_m P_A(m,t)
\end{aligned}
\tag{5.185}
$$

for the active state, and

$$
\begin{aligned}
\frac{dP_I(m,t)}{dt} = {} & k_{\text{off}}^{(p)} P_A(m,t) - k_{\text{on}}^{(p)} P_I(m,t) \\
& + \gamma_m(m+1)P_I(m+1,t) - \gamma_m m P_I(m,t),
\end{aligned}
\tag{5.186}
$$

for the inactive state.

**Obtaining the partial differential equation for the generating function**

The first thing we must do is to transform this infinite-dimensional system of ordinary differential equations in $m$ to a single partial differential equation using the generating function. For this particular case, there are two generating functions of the form

$$
G_x(z,t) = \sum_{m=0}^{\infty} z^m P_x(m,t),
\tag{5.187}
$$
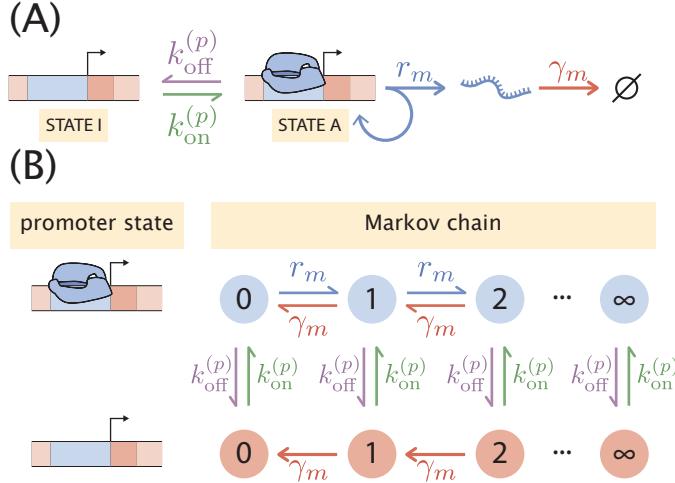
Figure 5.36: **Two-state Poisson promoter.** (A) Schematic of the kinetics of the two-state promoter. The promoter is imagined to exist in two states—a transcriptionally active state $A$ and an inactive state $I$. The transition between these states is governed by the rates $k_{on}^{(p)}$ and $k_{off}^{(p)}$ mRNA is produced and degrade stochastically with a rate $r_m$ and $\gamma_m$, respectively. (B) Representation of the Markov chain for the state space that the promoter can be in. The distribution $P(m, t)$ represents the probability of having a certain discrete number of mRNA $m$ at time $t$. The transition between states depends on the previously mentioned rates.

where $x \in \{A, I\}$. The probability of having $m$ mRNA at time $t$ regardless of the promoter state is given by

$$P(m, t) = P_A(m, t) + P_I(m, t). \tag{5.188}$$

Therefore, the corresponding generating function for the whole system is given by

$$G(z, t) = G_A(z, t) + G_I(z, t). \tag{5.189}$$

As with the one-state promoter case, let us transform our master equations by multiplying both sides by $z^m$ and sum over all $m$. For the active state $A$, we have

$$
\begin{aligned}
\sum_m z^m \frac{dP_A(m, t)}{dt} = \sum_m z^m & \left[ -k_{off}^{(p)} P_A(m, t) + k_{on}^{(p)} P_I(m, t) \right. \\
& + \gamma_m (m + 1) P_A(m + 1, t) - \gamma_m m P_m(m, t) \\
& \left. + r_m P_A(m - 1, t) - r_m P_A(m, t) \right].
\end{aligned}
\tag{5.190}
$$

After distributing the sum, we can use the tricks from the previous section, allowing us to write this as a partial differential equation of the form

$$\frac{\partial G_A(z,t)}{\partial t} = -k_{off}^{(p)} G_A(z,t) + k_{on}^{(p)} G_I(z,t)$$
$$- \gamma_m(z-1)\frac{\partial G_A(z,t)}{\partial z} + r_m(z-1)G_A(z,t). \tag{5.191}$$

An equivalent process can be done for the inactive state I, obtaining

$$\frac{\partial G_I(z,t)}{\partial t} = k_{off}^{(p)} G_A(z,t) - k_{on}^{(p)} G_I(z,t)$$
$$- \gamma_m(z-1)\frac{\partial GA(z,t)}{\partial z} + r_m(z-1)G_I(z,t). \tag{5.192}$$

We turned the infinite-dimensional system of ordinary differential equations into a system of two coupled partial differential equations. Let us transform the equations further. Since we have a common term $(z-1)$, it will be convenient to define $v \equiv (z-1)$. From the chain rule, it follows that

$$dv = d(z-1) = dz \Rightarrow \frac{\partial G}{\partial v} = \frac{\partial G}{\partial z}\frac{dz}{dv}. \tag{5.193}$$

Making this substitution in Eqs. 5.191 and 5.192 results in

$$\frac{\partial G_A(v,t)}{\partial t} = -k_{off}^{(p)} G_A(v,t) + k_{on}^{(p)} G_I(v,t)$$
$$- \gamma_m v \frac{\partial G_A(v,t)}{\partial v} + r_m v G_A(v,t) \tag{5.194}$$

for the inactive state, and

$$\frac{\partial GI(v,t)}{\partial t} = k_{off}^{(p)} G_A(v,t) - k_{on}^{(p)} G_I(v,t)$$
$$- r_m v \frac{\partial G_I(v,t)}{\partial v}, \tag{5.195}$$

for the active state.

Since we care about the steady-state distribution, it is at this point that we set the time derivative of both equations to zero. Doing this results in

$$\gamma_m v \frac{dG_A(v)}{dv} = -k_{off}^{(p)} G_A(v) + k_{on}^{(p)} G_I(v). + r_m v G_A(v), \tag{5.196}$$

and

$$\gamma_m v \frac{dG_I(v)}{dv} = k_{off}^{(p)} G_A(v) - k_{on}^{(p)} G_I(v). \tag{5.197}$$

Adding Eqs. 5.196 and 5.197 gives a simple result

$$\gamma_m \frac{dG(v)}{dv} = r_m G_A(v). \tag{5.198}$$

Our objective is not to write Eqs. 5.196 and 5.197 as a function of only one of the generating functions, i.e., we want two independent differential equations. These equations are both functions of $G_A(v)$ and $G_I(v)$, but Eq. 5.198 tells us how to relate both generating functions via the first derivative. This suggests that taking another derivative of Eqs. 5.196 and 5.197 with respect to $z$ could be useful. Let us go ahead and compute these derivatives. For the active state, we find

$$\gamma_m \frac{dG_A(v)}{dv} + \gamma_m v \frac{d^2 G_A(v)}{dv^2} = -k_{\text{off}}^{(p)} \frac{dG_A(v)}{dv} + k_{\text{on}}^{(p)} \frac{d\sigma I(v)}{dv} + r_m G_A(v,t) + r_m v \frac{dG_A(v)}{dv}. \tag{5.199}$$

Upon simplification, we can write this equation as

$$\gamma_m v \frac{d^2 G_A}{dv^2} + \left( \gamma_m + k_{\text{off}}^{(p)} - r_m v \right) \frac{dG_A}{dv} - k_{\text{on}}^{(p)} \frac{dG_I}{dv} - r_m G_A(v) = 0. \tag{5.200}$$

From Eq. 5.198, we have that

$$\frac{G_I}{dv} = \frac{r_m}{\gamma_m} G_A(v) - \frac{dG_A}{dv}. \tag{5.201}$$

Substituting this into Eq. 5.200 results in

$$\gamma_m v \frac{d^2 G_A}{dv^2} + \left( \gamma_m + k_{\text{off}}^{(p)} + k_{\text{on}}^{(p)} - r_m v \right) \frac{dG_A}{dv} - r_m \left( 1 + \frac{k_{\text{on}}^{(p)}}{\gamma_m} \right) G_A(v) = 0. \tag{5.202}$$

For the inactive state, upon taking a derivative with respect to $v$, we find

$$\gamma_m v \frac{d^2 G_I}{dv^2} + \left( \gamma_m + k_{\text{on}}^{(p)} \right) \frac{dG_I}{dv} - k_{\text{off}}^{(p)} \frac{dG_A}{dv} = 0. \tag{5.203}$$

Again from Eq. 5.198, we have that

$$\frac{dG_A}{dv} = \frac{r_m}{\gamma_m} G_A - \frac{dG_I}{dv}. \tag{5.204}$$

Substituting this result into Eq. 5.203 gives

$$\gamma_m v \frac{d^2 G_I}{dv^2} + \left( \gamma_m + k_{\text{on}}^{(p)} + k_{\text{off}}^{(p)} \right) \frac{dG_I}{dv} - \frac{k_{\text{off}}^{(p)} r_m}{\gamma_m} G_A(v) = 0. \tag{5.205}$$

So far, we have not removed the dependence on $G_A(v)$. But we notice that from Eq. 5.197, we have that

$$G_A(v) = \frac{\gamma_m v}{k_{\text{off}}^{(p)}} \frac{dG_I}{dv} + \frac{k_{\text{on}}^{(p)}}{k_{\text{off}}^{(p)}} G_I. \tag{5.206}$$

Using this identity allows us to write Eq. 5.205 as

$$\gamma_m v \frac{d^2 G_I}{dv^2} + \left(\gamma_m + k_{\text{on}}^{(p)} + k_{\text{off}}^{(p)} - r_m v\right) \frac{dG_I}{dv} - \frac{k_{\text{on}}^{(p)} r_m}{\gamma_m} G_I = 0. \tag{5.207}$$

To obtain a single partial differential equation, we add Eqs. 5.202 and 5.207, obtaining

$$\gamma_m v \frac{d^2 G}{dv^2} + \left(\gamma_m + k_{\text{off}}^{(p)} + k_{\text{on}}^{(p)} - r_m v\right) \frac{dG}{dv} - \frac{r_m k_{\text{on}}^{(p)}}{\gamma_m} G(v) - r_m G_A(v) = 0, \tag{5.208}$$

where we substituted $G_A(v) + G_I(v) = G(v)$. To remove the last $G_A(v)$, we utilize again Eq. 5.198, obtaining

$$\gamma_m v \frac{d^2 G}{dv^2} + \left(k_{\text{off}}^{(p)} + k_{\text{on}}^{(p)} - r_m v\right) \frac{dG}{dv} - \frac{r_m k_{\text{on}}^{(p)}}{\gamma_m} G(v) = 0. \tag{5.209}$$

**Solving the partial differential equation**

Eq. 5.209 looks almost like the so-called Kummer's equation also known as the confluent hypergeometric differential equation—a second order differential equation of the form

$$z \frac{d^2 w}{dz^2} + (b - z) \frac{dw}{dz} - aw = 0. \tag{5.210}$$

The solution for the Kummer equation can be expressed as the sum of two functions: 1. The confluent hypergeometric function of the first kind, 2. The Tricomi function. This is written as

$$w(z) = A\,_1F_1(a, b, z) + B z^{1-b}\,_1F_1(a + 1 - b, 2 - b, z), \tag{5.211}$$

where $A$ and $B$ are constants, and $_1F_1$ is the confluent hypergeometric function of the first kind defined as

$$_1F_1(a, b, z) = \sum_{m=0}^{\infty} \frac{a^{(m)} z^n}{b^{(m)} m!}, \tag{5.212}$$

where $a^{(n)}$ and $b^{(n)}$ are the rising factorials, i.e.,

$$a^{(0)} = 1, \tag{5.213}$$

and

$$a^{(n)} = a(a+1)(a+2)\cdots(a+n-1). \tag{5.214}$$

To write Eq. 5.209 in the form of Eq. 5.210, we can define $s \equiv r_m v / \gamma_m$. The chain rule tells us that

$$ds = \frac{r_m}{\gamma_m} dv \Rightarrow \frac{dG}{ds} = \frac{dG}{dv}\frac{dv}{ds} = \frac{\gamma_m}{r_m}\frac{dG}{dv}. \tag{5.215}$$

From the chain rule, we also conclude that

$$\frac{d^2G}{ds^2} = \frac{d}{dv}\left(\frac{dG}{dv}\frac{dv}{ds}\right)\frac{dv}{ds} = \frac{\gamma_m^2}{r_m^2}\frac{d^2G}{dv^2}. \tag{5.216}$$

So the three relationships of $v$ with $s$ that we have derived take the form

$$v = \frac{\gamma_m}{r_m}s, \quad \frac{dG}{dv} = \frac{r_m}{\gamma_m}\frac{dG}{ds}, \text{ and } \frac{d^2G}{dv^2} = \frac{r_m^2}{\gamma_m^2}\frac{d^2G}{dv^2}. \tag{5.217}$$

Substituting these definitions results in

$$\gamma_m\left(\frac{\gamma_m}{r_m}s\right)\frac{r_m^2}{\gamma_m^2}\frac{d^2G}{ds^2} + \left[k_{\text{off}}^{(p)} + k_{\text{on}}^{(p)} - r_m\left(\frac{\gamma_m}{r_m}s\right)\right]\frac{r_m}{\gamma_m}\frac{dG}{ds} - \frac{r_m k_{\text{on}}^{(p)}}{\gamma_m}G(s) = 0. \tag{5.218}$$

Upon simplifying terms, we find an equation that is now in the form of Eq. 5.210

$$s\frac{d^2G}{ds^2} + \left(\frac{k_{\text{off}}^{(p)} + k_{\text{off}}^{(p)}}{\gamma_m} - s\right)\frac{dG}{ds} - \frac{k_{\text{on}}^{(p)}}{\gamma_m}G(s) = 0. \tag{5.219}$$

Having put this in the form of the Kummer Eq., we can use Eq. 5.211 to write $G(s)$ as

$$G(s) = A\,{}_1F_1\left(\frac{k_{\text{on}}^{(p)}}{\gamma_m}, \frac{k_{\text{on}}^{(p)} + k_{\text{off}}^{(p)}}{\gamma_m}, s\right)$$
$$+ Bs^{1-\frac{k_{\text{on}}^{(p)}+k_{\text{off}}^{(p)}}{\gamma_m}}\,{}_1F_1\left(\frac{k_{\text{on}}^{(p)}}{\gamma_m}+1-\frac{k_{\text{on}}^{(p)}+k_{\text{off}}^{(p)}}{\gamma_m}, 2-\frac{k_{\text{on}}^{(p)}+k_{\text{off}}^{(p)}}{\gamma_m}, s\right). \tag{5.220}$$

We can write down this solution in terms of the original variable of the generating function. We have that $s = r_m/\gamma_m v$, and $v = z - 1$. With this, we then write

$$G(z) = A \, _1F_1 \left( \frac{k_{\text{on}}^{(p)}}{\gamma_m}, \frac{k_{\text{on}}^{(p)} + k_{\text{off}}^{(p)}}{\gamma_m}, \frac{r_m}{\gamma_m}(z-1) \right) + B \left[ \frac{r_m}{\gamma_m}(z-1) \right]^{1 - \frac{k_{\text{on}}^{(p)} + k_{\text{off}}^{(p)}}{\gamma_m}} \times$$
$$_1F_1 \left( \frac{k_{\text{on}}^{(p)}}{\gamma_m} + 1 - \frac{k_{\text{on}}^{(p)} + k_{\text{off}}^{(p)}}{\gamma_m}, 2 - \frac{k_{\text{on}}^{(p)} + k_{\text{off}}^{(p)}}{\gamma_m}, \frac{r_m}{\gamma_m}(z-1) \right).$$
(5.221)

**Finding the Coefficients for the Solution**

We can now use the normalization condition for the generating function; this is,

$$G(1) = \sum_{m=0}^{\infty} 1^m P(m) = 1.$$
(5.222)

Evaluating $z = 1$ in Eq. 5.221 results in

$$G(1) = A \, _1F_1 \left( \frac{k_{\text{on}}^{(p)}}{\gamma_m}, \frac{k_{\text{on}}^{(p)} + k_{\text{off}}^{(p)}}{\gamma_m}, 0 \right).$$
(5.223)

Let us look at the hypergeometric function evaluated of the form $_1F_1(a,b,0)$. This takes the form

$$_1F_1(a,b,0) = \sum_{m=0}^{\infty} \frac{a^{(m)} 0^n}{b^{(m)} m!}.$$
(5.224)

All of the terms but one ($n = 0$) are zero. The first term involving $0^0$ is undefined. Taking the limit as $z \to 0$ from the positive side, we find

$$_1F_1(a,b,0) = \lim_{z \to 0^+} {_1F_1(a,b,z)} = \lim_{z \to 0^+} z^0 = 1.$$
(5.225)

Using this property in Eq. 5.223 tells us that $A = 1$.

We do not have another constraint for $B$. Nevertheless, recall that Eq. 5.170 tells us how to compute the first moment of the distribution from the generating function. For this, we need to compute the derivative of the confluent hypergeometric function. Let us derive this identity. Rather than computing the derivative directly, we will compute

$$z \frac{d}{dz} {_1F_1} = z \frac{d}{dz} \left[ \sum_{m=0}^{\infty} \frac{a^{(m)} z^m}{b^{(m)} m!} \right].$$
(5.226)

Taking the derivative inside the sum gives

$$z \frac{d}{dz} {}_1F_1 = z \left[ \sum_{m=0}^{\infty} \frac{a^{(m)}}{b^{(m)} m!} \frac{d}{dz} z^m \right] = \left[ \sum_{m=0}^{\infty} \frac{a^{(m)}}{b^{(m)}} \frac{mz^m}{m!} \right]. \tag{5.227}$$

Simplifying the term $m/m!$ gives

$$z \frac{d}{dz} {}_1F_1 = \left[ \sum_{m=0}^{\infty} \frac{a^{(m)}}{b^{(m)}} \frac{z^m}{(m-1)!} \right]. \tag{5.228}$$

Note that the rising factorials can be rewritten as

$$a^{(m)} = a(a+1)(a+2)\cdots(a+m-1)$$

$$= a \cdot (a+1)[(a+1)+1][(a+1)+2]\cdots[(a+1)+m-2] \tag{5.229}$$

$$= a \cdot (a+1)^{(m-1)}.$$

Therefore, we can rewrite Eq. 5.228 as

$$\sum_{m=0}^{\infty} \frac{a^{(m)}}{b^{(m)}} \frac{z^m}{(m-1)!} = \sum_{m=0}^{\infty} \frac{a \cdot (a+1)^{(m-1)}}{b \cdot (b+1)^{(m-1)}} \frac{z \cdot z^{(m-1)}}{(m-1)!},$$

$$= \frac{az}{b} \sum_{m=0}^{\infty} \frac{(a+1)^{(m-1)}}{(b+1)^{(m-1)}} \frac{z^{m-1}}{(m-1)!}. \tag{5.230}$$

If we define $m' = m - 1$, we have

$$\frac{az}{b} \sum_{m=0}^{\infty} \frac{(a+1)^{(m-1)}}{(b+1)^{(m-1)}} \frac{z^{m-1}}{(m-1)!} = \frac{az}{b} \sum_{m'=-1}^{\infty} \frac{(a+1)^{m'}}{(b+1)^{m'}} \frac{z^{m'}}{m'!}. \tag{5.231}$$

The term on the left is almost of the form of the confluent hypergeometric function again. The only difference is that the sum starts at $m' = -1$. This first term of the sum would then involve a term of the form $1/(-1)!$ But what does this even mean? To find this out, we can generalize the factorial function using the Gamma function such that

$$(x-1)! = \Gamma(x). \tag{5.232}$$

The Gamma function diverges as $x \to 0$, therefore $1/\Gamma(x) \to 0$ as $x \to 0$. This means that the first term of the sum is zero, so we can begin the sum at $m' = 0$, recovering a confluent hypergeometric function. With this, we find that

$$z \frac{d}{dz} {}_1F_1 = \frac{az}{b} \sum_{m=0}^{\infty} \frac{(a+1)^m}{(b+1)^m} \frac{z^m}{m!} = \frac{a}{b} z {}_1F_1(a+1, b+1, z), \tag{5.233}$$

therefore

$$\frac{d}{dz}{}_1F_1 = \frac{a}{b}{}_1F_1(a+1, b+1, z). \tag{5.234}$$

After this small but necessary detour, we can come back to computing the first moment of our distribution from the generating function. To evaluate Eq. 5.170 on Eq. 5.221, we first compute the derivative of the generating function. This can be easily evaluated using the relationship we derived for derivatives of ${}_1F_1$. The only thing to be aware of is that of the chain rule. In particular for the third entry of the function, we have $r_m/\gamma_m(z-1)$ rather than simply $z$ as we had in Eq. 5.234. This means that by the chain rule, we have that if we define $u = r_m/\gamma_m(z-1)$, we have

$$du = \frac{r_m}{\gamma_m}dz \Rightarrow \frac{dG}{dz} = \frac{dG}{du}\frac{du}{dz} = \frac{dG}{du}\frac{r_m}{\gamma_m}. \tag{5.235}$$

So there is an extra factor of $r_m/\gamma_m$ that will come along when we compute the derivative of our generating functions. Computing the derivative of Eq. 5.221 results in

$$
\begin{aligned}
\frac{dG}{dz} =\ & \frac{k_{on}^{(p)}}{k_{on}^{(p)} + k_{off}^{(p)}}\frac{r_m}{\gamma_m}{}_1F_1\left(\frac{k_{on}^{(p)}}{\gamma_m}+1, \frac{k_{on}^{(p)}+k_{off}^{(p)}}{\gamma_m}+1, \frac{r_m}{\gamma_m}(z-1)\right) \\
& + B\left(1 - \frac{k_{on}^{(p)}+k_{off}^{(p)}}{\gamma_m}\right)\left[\frac{r_m}{\gamma_m}(z-1)\right]^{\frac{k_{on}^{(p)}+k_{off}^{(p)}}{\gamma_m}} \times \\
& {}_1F_1\left(\frac{k_{on}^{(p)}}{\gamma_m}+1-\frac{k_{on}^{(p)}+k_{off}^{(p)}}{\gamma_m}, 2-\frac{k_{on}^{(p)}+k_{off}^{(p)}}{\gamma_m}, \frac{r_m}{\gamma_m}(z-1)\right) \\
& + B\left[\frac{r_m}{\gamma_m}(z-1)\right]^{1-\frac{k_{on}^{(p)}+k_{off}^{(p)}}{\gamma_m}}\left(\frac{k_{on}^{(p)}+\gamma_m}{k_{on}^{(p)}+k_{off}^{(p)}+\gamma_m}\right)\frac{r_m}{\gamma_m}\times \\
& {}_1F_1\left(\frac{k_{on}^{(p)}}{\gamma_m}+2-\frac{k_{on}^{(p)}+k_{off}^{(p)}}{\gamma_m}, 1-\frac{k_{on}^{(p)}+k_{off}^{(p)}}{\gamma_m}, \frac{r_m}{\gamma_m}(z-1)\right).
\end{aligned}
\tag{5.236}
$$

This rather convoluted result is enormously simplified upon evaluating the derivative at $z=1$ (see Eq. 5.170). This results in

$$\left.\frac{dG}{dz}\right|_{z=1} = \frac{k_{on}^{(p)}}{k_{on}^{(p)}+k_{off}^{(p)}}\frac{r_m}{\gamma_m}{}_1F_1\left(\frac{k_{on}^{(p)}}{\gamma_m}+1, \frac{k_{on}^{(p)}+k_{off}^{(p)}}{\gamma_m}+1, 0\right) = \frac{r_m}{\gamma_m}\frac{k_{on}^{(p)}}{k_{on}^{(p)}+k_{off}^{(p)}}, \tag{5.237}$$

which is precisely the mean mRNA copy number we derived before. Since $B$ does not contribute to the mean, we can safely assume that $B = 0$. This means that the final result for the generating function takes the much more compact form

$$G(z) = {}_1F_1\left(\frac{k_{\text{on}}^{(p)}}{\gamma_m}, \frac{k_{\text{on}}^{(p)} + k_{\text{off}}^{(p)}}{\gamma_m}, \frac{r_m}{\gamma_m}(z-1)\right). \tag{5.238}$$

**Extracting the steady-state mRNA distribution**

Let us quickly recapitulate where we are. We started with a system of infinite many ordinary differential equations, one for each promoter state and mRNA copy number that defined the master equation for our two-state promoter. We then used the generating function to transform this system into a single partial differential equation. The resulting differential equation for the generating function took the form of the so-called Kummer differential equation, which has as a solution the confluent hypergeometric function and the Tricomi function. After imposing the normalization condition on the generating function, we found that the confluent hypergeometric function's coefficient was $A = 1$. We then used the fact that the mean mRNA copy number $\langle m \rangle$ exists to show that the Tricomi function's coefficient is $B = 0$. All that effort lead us to Eq. 5.238, the generating function for the two-state promoter mRNA steady-state distribution. All we have left is trying to beat Eq. 5.238 into the form of a standard generating function to extract the probability distribution from it.

Let us begin this task by writing down Eq. 5.238 with the full definition of the confluent hypergeometric function. This gives us

$$G(z) = \sum_{m=0}^{\infty} \frac{\left(\frac{k_{\text{on}}^{(p)}}{\gamma_m}\right)^{(m)}}{\left(\frac{k_{\text{on}}^{(p)} + k_{\text{off}}^{(p)}}{\gamma_m}\right)^{(m)}} \frac{\left[\frac{r_m}{\gamma_m}(z-1)\right]^m}{m!}. \tag{5.239}$$

Let us now split apart the term $(z - 1)$, obtaining

$$G(z) = \sum_{m=0}^{\infty} \frac{\left(\frac{k_{\text{on}}^{(p)}}{\gamma_m}\right)^{(m)}}{\left(\frac{k_{\text{on}}^{(p)}+k_{\text{off}}^{(p)}}{\gamma_m}\right)^{(m)}} \frac{\left(\frac{r_m}{\gamma_m}\right)^m}{m!} (z - 1)^m. \qquad (5.240)$$

We now rewrite this last term $(z - 2)^m$ using the binomial expansion. This results in

$$G(z) = \sum_{m=0}^{\infty} \frac{\left(\frac{k_{\text{on}}^{(p)}}{\gamma_m}\right)^{(m)}}{\left(\frac{k_{\text{on}}^{(p)}+k_{\text{off}}^{(p)}}{\gamma_m}\right)^{(m)}} \frac{\left(\frac{r_m}{\gamma_m}\right)^m}{m!} \left[\sum_{n=0}^{m} \binom{m}{n} z^n (-1)^{m-n}\right]. \qquad (5.241)$$

We can take out the sum over the index $n$ to the front, obtaining

$$G(z) = \sum_{m=0}^{\infty} \sum_{n=0}^{n} \frac{\left(\frac{k_{\text{on}}^{(p)}}{\gamma_m}\right)^{(m)}}{\left(\frac{k_{\text{on}}^{(p)}+k_{\text{off}}^{(p)}}{\gamma_m}\right)^{(m)}} \frac{\left(\frac{r_m}{\gamma_m}\right)^m}{m!} \left[\binom{m}{n} z^n (-1)^{m-n}\right]. \qquad (5.242)$$

To make further progress, we must reindex the sum. The trick is to reverse the default order of the sums as

$$\sum_{m=0}^{\infty} \sum_{n=0}^{m} = \sum_{n=0}^{\infty} \sum_{m=n}^{\infty}. \qquad (5.243)$$

To see the logic of the sum, we point the reader to Fig. 5.37. The key is to notice that the double sum $\sum_{m=0}^{\infty} \sum_{n=0}^{m}$ is adding all possible pairs $(m, n)$ in the lower triangle, so we can add the terms vertically as the original sum indexing suggests, i.e.

$$\sum_{m=0}^{\infty} \sum_{n=0}^{m} x_{(m,n)} = x_{(0,0)} + x_{(1,0)} + x_{(1,1)} + x_{(2,0)} + x_{(2,1)} + x_{(2,2)} + \ldots, \qquad (5.244)$$

where the variable $x$ is just a placeholder to indicate the order in which the sum is taking place. But we can also add the terms horizontally as

$$\sum_{n=0}^{\infty} \sum_{m=n}^{\infty} x_{(m,n)} = x_{(0,0)} + x_{(1,0)} + x_{(2,0)} + \ldots + x_{(1,1)} + x_{(2,1)} + \ldots, \qquad (5.245)$$

which still adds all of the lower triangle terms.
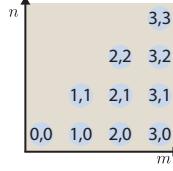
Figure 5.37: **Reindexing double sum.** Schematic for reindexing the sum $\sum_{m=0}^{\infty}\sum_{n=0}^{m}$. Blue circles depict the 2D grid of nonnegative integers restricted to the lower triangular part of the $m, n$ plane. The trick is that this double sum runs over all $(m, n)$ pairs with $n \leq m$. Summing $m$ first instead of $n$ requires determining the boundary: the upper boundary of the $n$-first double sum becomes the lower boundary of the $m$-first double sum.

Rewriting the sum in this way results in

$$G(z) = \sum_{n=0}^{\infty} \sum_{m=n}^{\infty} \frac{\left(\frac{k_{\text{on}}^{(p)}}{\gamma_m}\right)^{(m)}}{\left(\frac{k_{\text{on}}^{(p)}+k_{\text{off}}^{(p)}}{\gamma_m}\right)^{(m)}} \frac{\left(\frac{r_m}{\gamma_m}\right)^m}{m!} \left[\binom{m}{n} z^n (-1)^{m-n}\right]. \tag{5.246}$$

This allows us to separate the variable $z^n$ from the rest of the equation, leaving the standard format generating function to read the probability distribution $P(m)$. This looks as

$$G(z) = \sum_{n=0}^{\infty} z^n \left[\sum_{m=n}^{\infty} \frac{\left(\frac{k_{\text{on}}^{(p)}}{\gamma_m}\right)^{(m)}}{\left(\frac{k_{\text{on}}^{(p)}+k_{\text{off}}^{(p)}}{\gamma_m}\right)^{(m)}} \frac{\left(\frac{r_m}{\gamma_m}\right)^m}{m!} \binom{m}{n} (-1)^{m-n}\right]. \tag{5.247}$$

Given the "dummy" nature of $z$, it does not matter what the sum variable name is. We can simply rename $m = n$ and $n = m$ and conclude that our distribution takes the form

$$P(m) = \sum_{n=m}^{\infty} \frac{\left(\frac{k_{\text{on}}^{(p)}}{\gamma_m}\right)^{(n)}}{\left(\frac{k_{\text{on}}^{(p)}+k_{\text{off}}^{(p)}}{\gamma_m}\right)^{(n)}} \frac{\left(\frac{r_m}{\gamma_m}\right)^n}{n!} \frac{n!}{m!(n-m)!} (-1)^{n-m}. \tag{5.248}$$

We can simplify Eq. 5.248 further. First, we split the term $(-1)^{n-m} = (-1)^{-m}(-1)^n$. Furthermore, we absorb the $(-1)^n$ term on the $(r_m/\gamma_m)^n$ term. We also cancel the obvious $n!/n!$ term, obtaining

$$P(m) = \sum_{n=m}^{\infty} \frac{(-1)^{-m}}{m!} \frac{\left(\frac{k_{\text{on}}^{(p)}}{\gamma_m}\right)^{(n)}}{\left(\frac{k_{\text{on}}^{(p)}+k_{\text{off}}^{(p)}}{\gamma_m}\right)^{(n)}} \frac{\left(-\frac{r_m}{\gamma_m}\right)^n}{(n-m)!}. \tag{5.249}$$

We recognize in Eq. 5.249 that we have almost all the terms for a confluent hypergeometric function $_1F_1$. The problem is that the sum starts at $n = m$ rather than $n = 0$. Since the upper limit of the sum is $\infty$, we can simply define $u = n - m \Rightarrow n = m + u$. We can then use the following property of raising factorials

$$
\begin{aligned}
a^{(n)} &= a(a+1)(a+2)\cdots(a+n-1), \\
&= a(a+1)(a+2)\cdots(a+(u+m)-1), \\
&= a(a+1)\cdots(a+m-1)(a+m)(a+m+1)\cdots(a+m+u-1), \\
&= a^{(m)}(a+m)^{(u)}.
\end{aligned}
\tag{5.250}
$$

Making these substitutions results in

$$
P(m) = \sum_{u=0}^{\infty} \frac{(-1)^{-m}}{m!} \frac{\left(\frac{k_{on}^{(p)}}{\gamma_m}\right)^{(m)} \left(\frac{k_{on}^{(p)}}{\gamma_m}+m\right)^{(u)} \left(-\frac{r_m}{\gamma_m}\right)^u \left(-\frac{r_m}{\gamma_m}\right)^m}{\left(\frac{k_{on}^{(p)}+k_{off}^{(p)}}{\gamma_m}\right)^{(m)} \left(\frac{k_{on}^{(p)}+k_{off}^{(p)}}{\gamma_m}+m\right)^{(n)}} \frac{1}{u!}.
\tag{5.251}
$$

Taking out of the sum the terms that do not depend on $u$ gives

$$
P(m) = \frac{(-1)^{-m}}{m!} \frac{\left(\frac{k_{on}^{(p)}}{\gamma_m}\right)^{(m)}}{\left(\frac{k_{on}^{(p)}+k_{off}^{(p)}}{\gamma_m}\right)^{(m)}} \left(-\frac{r_m}{\gamma_m}\right)^m \left[ \sum_{u=0}^{\infty} \frac{\left(\frac{k_{on}^{(p)}}{\gamma_m}+m\right)^{(u)}}{\left(\frac{k_{on}^{(p)}+k_{off}^{(p)}}{\gamma_m}+m\right)^{(u)}} \frac{\left(-\frac{r_m}{\gamma_m}\right)^u}{u!} \right].
\tag{5.252}
$$

We recognize the term in the square brackets to be the necessary component for a confluent hypergeometric function. We can therefore write the mRNA steady-state distribution as

$$
P(m) = \frac{1}{m!} \frac{\left(\frac{k_{on}^{(p)}}{\gamma_m}\right)^{(m)}}{\left(\frac{k_{on}^{(p)}+k_{off}^{(p)}}{\gamma_m}\right)^{(m)}} \left(\frac{r_m}{\gamma_m}\right)^m {}_1F_1\left(\frac{k_{on}^{(p)}}{\gamma_m}+m, \frac{k_{on}^{(p)}+k_{off}^{(p)}}{\gamma_m}+m, -\frac{r_m}{\gamma_m}\right).
\tag{5.253}
$$

For the last ingredient, we remove the rising factorials using the identity

$$
\begin{aligned}
a^{(m)} &= (a)(a+1)(a+2)\cdots(a+m-1), \\
&= \frac{(a+m-1)\cdots(a)(a-1)\cdots(1)}{(a+1)\cdots(1)}, \\
&= \frac{(a+m-1)!}{(a-1)!}.
\end{aligned}
\tag{5.254}
$$

This allows us to write

$$
P(m) = \frac{1}{m!} \frac{\left( \frac{k_{\text{on}}^{(p)}}{\gamma_m} + m - 1 \right)!}{\left( \frac{k_{\text{on}}^{(p)}}{\gamma_m} - 1 \right)!} \frac{\left( \frac{k_{\text{on}}^{(p)}+k_{\text{off}}^{(p)}}{\gamma_m} - 1 \right)!}{\left( \frac{k_{\text{on}}^{(p)}+k_{\text{off}}^{(p)}}{\gamma_m} + m - 1 \right)!} \left( \frac{r_m}{\gamma_m} \right)^m
$$

$$
\times \,_1F_1 \left( \frac{k_{\text{on}}^{(p)}}{\gamma_m} + m, \frac{k_{\text{on}}^{(p)} + k_{\text{off}}^{(p)}}{\gamma_m} + m, -\frac{r_m}{\gamma_m} \right). \tag{5.255}
$$

Or in terms of Gamma functions, we obtain the final form of the steady-state mRNA distribution

$$
P(m) = \frac{1}{\Gamma(m+1)} \frac{\Gamma\left( \frac{k_{\text{on}}^{(p)}}{\gamma_m} + m \right)}{\Gamma\left( \frac{k_{\text{on}}^{(p)}}{\gamma_m} \right)} \frac{\Gamma\left( \frac{k_{\text{on}}^{(p)}+k_{\text{off}}^{(p)}}{\gamma_m} \right)}{\Gamma\left( \frac{k_{\text{on}}^{(p)}+k_{\text{off}}^{(p)}}{\gamma_m} + m \right)} \left( \frac{r_m}{\gamma_m} \right)^m \times {}_1F_1\left( \frac{k_{\text{on}}^{(p)}}{\gamma_m} + m, \frac{k_{\text{on}}^{(p)}+k_{\text{off}}^{(p)}}{\gamma_m} + m, -\frac{r_m}{\gamma_m} \right), \tag{5.256}
$$

The equation used to fit the kinetic parameters for the unregulated promoter.

## 5.11 Derivation of the Cell Age Distribution

E. O. Powell first derived in 1956 the cell age distribution for a cell population growing steadily in the exponential phase [120]. This distribution is of the form

$$
P(a) = \ln(2) \cdot 2^{1-a}, \tag{5.257}
$$

where $a \in [0,1]$ is the fraction of the cell cycle, 0 being the moment right after the mother cell divides, and 1 being the end of the cell cycle just before cell division. In this section, we will reproduce and expand the details on each of the steps of the derivation.

For an exponentially growing bacterial culture, the cells satisfy the growth law

$$
\frac{dn}{dt} = \mu n, \tag{5.258}
$$

where $n$ is the number of cells, and $\mu$ is the growth rate in units of time$^{-1}$. We begin by defining $P(a)$ to be the probability density function of a cell having age $a$. At time zero of a culture in exponential growth, i.e., when we start considering the growth, not the initial condition of the culture, there are $NP(a)da$ cells with an age range between $[a, a + da]$. In other words, for $N \gg 1$ and $da \ll a$

$$
NP(a \leq x \leq a + da) \approx NP(a)da. \tag{5.259}
$$

We now define

$$F(\tau) = \int_{\tau}^{\infty} f(\xi)d\xi, \tag{5.260}$$

as the fraction of cells whose division time is greater than $\tau$. This is because in principle, not all cells divide exactly after $\tau$ minutes, but there is a distribution function $f(\tau)$ for the division time after birth. Empirically it has been observed that a generalized Gamma distribution fits well to experimental data on cell division time, but we will worry about this specific point later on.

From the definition of $F(\tau)$, we can see that if a cell reaches an age $a$, the probability of surviving to an age $a + t$ without dividing is given by

$$P(\text{age} = (a+t) \mid \text{age} = a) = F(a+t \mid a) = \frac{F(a+t)}{F(a)}. \tag{5.261}$$

This result comes simply from the definition of conditional probability. Since $F(a)$ is the probability of surviving $a$ or more minutes without dividing, by the definition of conditional probability, we have that

$$F(a+t \mid a) = \frac{F(a, a+t)}{F(a)}, \tag{5.262}$$

where $F(a, a+t)$ is the joint probability of surviving $a$ minutes and $a+t$ minutes. But the probability of surviving $a + t$ minutes or more implies that the cell already survived $a$ minutes, therefore, the information is redundant, and we have

$$F(a, a+t) = F(a+t). \tag{5.263}$$

This explains Eq. 5.261. From this equation, we can find that out of the $NP(a)da$ cells with age $a$, only a fraction

$$[NP(a)da] \, F(a+t \mid a) = NP(a)\frac{F(a+t)}{F(a)}da \tag{5.264}$$

will survive without dividing until time $a + t$. During that time interval $t$, the culture has passed from $N$ cells to $Ne^{\mu t}$ cells, given the assumption that they are growing exponentially. The survivors $NP(a)F(a+t \mid a)da$ then represent a fraction of the total number of cells

$$\frac{\#\text{ survivors}}{\#\text{ total cells}} = \frac{[NP(a)da] \, F(a+t \mid a)}{Ne^{\mu t}} = P(a)\frac{F(a+t)}{F(a)}da\frac{1}{e^{\mu t}}, \tag{5.265}$$

and their ages lie in the range $[a + t, a + t + da]$. Since we assume that the culture is in a steady-state, then it follows that the fraction of cells that transitioned from age $a$ to age $a + t$ must be $P(a + t)da$. Therefore we have a difference equation—the discrete analogous of a differential equation—of the form

$$P(a + t)da = P(a)\frac{F(a + t)}{F(a)}e^{-\mu t}da. \tag{5.266}$$

What this equation shows is a relationship that connects the probability of having a lifetime of $a + t$ with a probability of having a shorter lifetime $a$ and the growth of the population. If we take $t$ to be very small, specifically if we assume $t \ll \mu^{-1}$, we can Taylor expand around $a$ the following terms:

$$F(a + t) \approx F(a) + \frac{dF}{da}t, \tag{5.267}$$

$$P(a + t) \approx P(a) + \frac{dP}{da}t, \tag{5.268}$$

and

$$e^{-\mu t} \approx 1 - \mu t. \tag{5.269}$$

Substituting these equations gives

$$P(a) + \frac{dP}{da}t = P(a)\left(\frac{F(a) + \frac{dF}{da}t}{F(a)}\right)(1 - \mu t). \tag{5.270}$$

This can be rewritten as

$$\frac{1}{P(a)}\frac{dP}{da} = \frac{1}{F(a)}\frac{dF}{da} - \mu - \frac{\mu t}{F(a)}\frac{dF}{da}. \tag{5.271}$$

Since we assumed $t \ll \mu^{-1}$, we approximate the last term to be close to zero. We can then simplify this result into

$$\frac{1}{P(a)}\frac{dP}{da} = \frac{1}{F(a)}\frac{dF}{da} - \mu. \tag{5.272}$$

Integrating both sides of the equation with respect to $a$ gives

$$\ln P(a) = \ln F(a) - \mu a + C, \tag{5.273}$$

where $C$ is the integration constant. Exponentiating both sides gives

$$P(a) = C'F(a)e^{-\mu a}, \tag{5.274}$$

where $C' \equiv e^C$. To obtain the unknown constant value, we recall that $F(0) = 1$ since the probability of having a life equal to or longer than zero must add up to one. Therefore we have that $P(0) = C'$. This gives then

$$P(a) = P(0)e^{-\mu a}F(a). \tag{5.275}$$

Substituting the definition of $F(a)$ gives

$$P(a) = P(0)e^{-\mu a}\int_a^\infty f(\xi)d\xi. \tag{5.276}$$

The last step of the derivation involves writing $P(0)$ and the growth rate $\mu$ in terms of the cell cycle length distribution $f(\tau)$.

The growth rate of the population cell number (not the growth of cell mass) is defined as the number of cell doublings per unit of time divided by the number of cells. This is more clear to see if we write as a finite difference

$$\frac{N(t + \Delta t) - N(t)}{\Delta t} = \mu N(t). \tag{5.277}$$

If the time $\Delta t$ is the time interval it takes to go from $N$ to $2N$ cells, we have

$$\frac{2N - N}{\Delta t} = \mu N. \tag{5.278}$$

Solving for $\mu$ gives

$$\mu = \overbrace{\frac{2N - N}{\Delta t}}^{\text{\# doubling events per unit time}} \overbrace{\frac{1}{N}}^{\frac{1}{\text{population size}}}. \tag{5.279}$$

We defined $F(a)$ to be the probability of a cell reaching an age $a$ or greater. For a cell to reach an age $a + da$, we can then write

$$F(a + da) = \int_{a+da}^\infty f(\xi)d\xi = \int_a^\infty f(\xi)d\xi - \int_a^{a+da} f(\xi)d\xi. \tag{5.280}$$

We can approximate the second term on the right-hand side to be

$$\int_a^{a+da} f(\xi)d\xi \approx f(a)da, \tag{5.281}$$

for $da \ll a$, obtaining

$$F(a + da) \approx F(a) - f(a)da. \tag{5.282}$$

What this means is that from the original fraction of cells $F(a)$ with age $a$ or greater, a fraction $f(a)da/F(a)$ will not reach age $(a + da)$ because they will divide. So, out of the $NP(a)$ cells that reached exactly age $a$, the number of doubling events on a time interval $da$ is given by

$$\text{\# doublings of cells of age } a \text{ on interval } da = \overbrace{NP(a)}^{\text{\# cells of age } a} \quad \overbrace{\frac{f(a)da}{F(a)}}^{\text{fraction of doublings per unit time}}. \tag{5.283}$$

The growth rate then is just the sum (integral) of each age contribution to the total number of doublings. This is

$$\mu = \frac{1}{N} \int_0^\infty NP(a)\frac{f(a)da}{F(a)}. \tag{5.284}$$

Substituting gives

$$\mu = \int_0^\infty [P(0)e^{-\mu a}F(a)]\frac{f(a)da}{F(a)} = \int_0^\infty P(0)e^{-\mu a}f(a)da. \tag{5.285}$$

We now have the growth rate $\mu$ written in terms of the cell cycle length probability distribution $f(a)$ and the probability $P(0)$. Since $P(a)$ is a probability distribution, it must be normalized, i.e.,

$$\int_0^\infty P(a)da = 1. \tag{5.286}$$

Substituting into this normalization constraint gives

$$\int_0^\infty P(0)e^{-\mu a}F(a)da = 1. \tag{5.287}$$

From here, we can integrate the left-hand side by parts. We note that given the definition of $F(a)$, the derivative with respect to $a$ is $-f(a)$ rather than $f(a)$. This is because if we write the derivative of $F(a)$, we have

$$\frac{dF(a)}{da} \equiv \lim_{da \to 0} \frac{F(a + da) - F(a)}{da}. \tag{5.288}$$

Substituting the definition of $F(a)$ gives

$$\frac{dF(a)}{da} = \lim_{da \to 0} \frac{1}{da} \left[ \int_{a+da}^\infty f(\xi)d\xi - \int_a^\infty f(\xi)d\xi \right]. \tag{5.289}$$

This difference in the integrals can be simplified to

$$\lim_{da \to 0} \frac{1}{da} \left[ \int_{a+da}^{\infty} f(\xi)d\xi - \int_{a}^{\infty} f(\xi)d\xi \right] \approx \frac{-f(a)da}{da} = -f(a). \tag{5.290}$$

Taking this into account, we now perform the integration by parts obtaining

$$P(0) \left[ \frac{e^{-\mu t}}{-\mu} F(a) \right]_{0}^{\infty} - P(0) \int_{0}^{\infty} \frac{e^{-\mu a}}{-\mu}(-f(a))da = 1. \tag{5.291}$$

On the first term on the left hand side, we have that, as $a \to \infty$, both terms $e^{-\mu a}$ and $F(a)$ go to zero. We also have that $e^{\mu 0} = 1$ and $F(0) = 1$. This results in

$$\frac{P(0)}{\mu} - P(0) \int_{0}^{\infty} \frac{e^{-\mu a}}{\mu} f(a)da = 1. \tag{5.292}$$

The second term on the left-hand side is equal to 1 since

$$\mu = \int_{0}^{\infty} P(0)e^{-\mu a}f(a)da \Rightarrow 1 = \int_{0}^{\infty} P(0)\frac{e^{-\mu a}}{\mu}f(a)da. \tag{5.293}$$

This implies that we have

$$\frac{P(0)}{\mu} - 1 = 1 \Rightarrow P(0) = 2\mu. \tag{5.294}$$

With this result in hand, we can rewrite it as

$$P(a) = 2\mu e^{-\mu a} \int_{a}^{\infty} f(\xi)d\xi. \tag{5.295}$$

Also, we can rewrite the result for the growth rate $\mu$ as

$$\mu = 2\mu \int_{0}^{\infty} e^{-\mu a}f(a)da \Rightarrow 2 \int_{0}^{\infty} e^{-\mu a}f(a)da = 1. \tag{5.296}$$

As mentioned before, the distribution $f(a)$ has been empirically fit to a generalized Gamma distribution. But if we assume that our distribution has almost negligible dispersion around the mean average doubling time $a = \tau_d$, we can approximate $f(a)$ as

$$f(a) = \delta(a - \tau_d), \tag{5.297}$$

a Dirac delta function. Applying this to Eq. 5.293 results in

$$2 \int_{0}^{\infty} e^{-\mu a}\delta(a - \tau_a)da = 1 \Rightarrow 2e^{-\mu \tau_d} = 1. \tag{5.298}$$

Solving for $\mu$ gives

$$\mu = \frac{\ln 2}{\tau_d}. \tag{5.299}$$

This delta function approximation for $f(a)$ has as a consequence that

$$F(a) = \begin{cases} 1 \text{ for } a \in [0, \tau_d], \\ 0 \text{ for } a > \tau_d. \end{cases} \tag{5.300}$$

Finally, we can rewrite it as

$$P(a) = 2\left(\frac{\ln 2}{\tau_d}\right) e^{-\frac{\ln 2}{\tau_d}a} \int_a^\infty \delta(\xi - \tau_d)d\xi \Rightarrow= 2\ln 2 \cdot 2^{\frac{-a}{\tau_d}}. \tag{5.301}$$

Simplifying this, we obtain

$$P(a) = \begin{cases} \ln 2 \cdot 2^{1 - \frac{a}{\tau_d}} \text{ for } a \in [0, \tau_d], \\ 0 \text{ otherwise.} \end{cases} \tag{5.302}$$

This is the equation we aimed to derive. The distribution of cell ages over the cell cycle.

# REFERENCES

[1] E. Schrödinger, *What is life?: With mind and matter and autobiographical sketches* (Cambridge University Press, 1992).

[2] R. Phillips, *Schrödinger' "What is Life?" at 75*, ArXiv 1 (2021).

[3] L. Cronin and S. I. Walker, *Beyond prebiotic chemistry*, Science **352**, 1174 (2016).

[4] P. Davies, *The Demon in the machine: How hidden webs of information are solving the mystery of life* (University of Chicago Press, 2019).

[5] C. Adami, *What is information?*, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **374**, 20150230 (2016).

[6] S. F. Taylor, N. Tishby, and W. Bialek, *Information and fitness*, ArXiv (2007).

[7] W. Bialek, *Biophysics: searching for principles* (Princeton University Press, 2012).

[8] D. F. Browning and S. J. W. Busby, *The regulation of bacterial transcription initiation*, Nature Reviews Microbiology **2**, 57 (2004).

[9] W. T. Ireland, S. M. Beeler, E. Flores-Bautista, N. S. McCarty, T. Röschinger, N. M. Belliveau, M. J. Sweredoski, A. Moradian, J. B. Kinney, and R. Phillips, *Deciphering the regulatory genome of Escherichia coli, one hundred promoters at a time*, Elife **9**, e55308 (2020).

[10] G. K. Ackers, A. D. Johnson, and M. A. Shea, *Quantitative model for gene regulation by lambda phage repressor*, Proceedings of the National Academy of Sciences **79**, 1129 (1982).

[11] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips, *Transcriptional regulation by the numbers: Applications*, Current Opinion in Genetics & Development **15**, 125 (2005).

[12] T. Kuhlman, Z. Zhang, M. H. Saier, and T. Hwa, *Combinatorial transcriptional control of the lactose operon of Escherichia coli*, Proceedings of the National Academy of Sciences **104**, 6043 (2007).

[13] S. H. Strogatz, *Nonlinear dynamics and chaos with student solutions manual: With applications to physics, biology, chemistry, and engineering* (CRC Press, 2018).

[14] K. Dill and S. Bromberg, *Molecular driving forces: statistical thermodynamics in biology, chemistry, physics, and nanoscience* (Garland Science, 2010).

[15] R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman lectures on physics; vol. i*, American Journal of Physics **33**, 750 (1965).

[16] R. Phillips, N. M. Belliveau, G. Chure, H. G. Garcia, M. Razo-Mejia, and C. Scholes, *Figure 1 theory meets figure 2 experiments in the study of gene expression*, Annual Review of Biophysics **48**, 121 (2019).

[17] H. Bremer and P. P. Dennis, *Modulation of chemical composition and other parameters of the cell by growth rate*, Escherichia Coli and Salmonella: Cellular and Molecular Biology (n.d.).

[18] A. Schmidt, K. Kochanowski, S. Vedelaar, E. Ahrne, B. Volkmer, L. Callipo, K. Knoops, M. Bauer, R. Aebersold, and M. Heinemann, *The quantitative and condition-dependent Escherichia coli proteome*, Nature Biotechnology **34**, 104 (2016).

[19] I. L. Grigorova, N. J. Phleger, V. K. Mutalik, and C. A. Gross, *Insights into transcriptional regulation and sigma competition from an equilibrium model of RNA polymerase binding to DNA*, Proceedings of the National Academy of Sciences **103**, 5332 (2006).

[20] H. G. Garcia and R. Phillips, *Quantitative dissection of the simple repression input-output function*, Proceedings of the National Academy of Sciences **108**, 12173 (2011).

[21] Y.-J. Chen, P. Liu, A. A. K. Nielsen, J. A. N. Brophy, K. Clancy, T. Peterson, and C. A. Voigt, *Characterization of 582 natural and synthetic terminators and quantification of their design constraints*, Nature Methods **10**, 659 (2013).

[22] A. Eldar and M. B. Elowitz, *Functional roles for noise in genetic circuits*, Nature **467**, 167 (2010).

[23] M. Voliotis, R. M. Perrett, C. McWilliams, C. a. McArdle, and C. G. Bowsher, *Information transfer by leaky, heterogeneous, protein kinase signaling systems*, Proceedings of the National Academy of Sciences **111**, E326 (2014).

[24] N. Q. Balaban, *Bacterial persistence as a phenotypic switch*, Science **305**, 1622 (2004).

[25] U. Gerland and T. Hwa, *On the selection and evolution of regulatory DNA motifs*, Journal of Molecular Evolution **55**, 386 (2002).

[26] A. Sanchez, S. Choubey, and J. Kondev, *Stochastic models of transcription: From single molecules to single cells*, Methods **62**, 13 (2013).

[27] C. E. Shannon, *A mathematical theory of communication*, Bell System Technical Journal **27**, 379 (1948).

[28] E. T. Jaynes, *Information theory and statistical mechanics*, Physical Review **106**, 620 (1957).

[29] E. T. Jaynes, *Probability theory: The logic of science* (Cambridge University Press, 2003).

[30] J. K. Blitzstein and J. Hwang, *Introduction to probability* (Chapman; Hall/CRC, 2019).

[31] J. E. Lindsley and J. Rutter, *Whence cometh the allosterome?*, Proceedings of the National Academy of Sciences **103**, 10533 (2006).

[32] J. G. Harman, *Allosteric regulation of the cAMP receptor protein*, Biochimica Et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology **1547**, 1 (2001).

[33] M. F. Lanfranco, F. Gárate, A. J. Engdahl, and R. A. Maillard, *Asymmetric configurations in a reengineered homodimer reveal multiple subunit communication pathways in protein allostery*, The Journal of Biological Chemistry **292**, 6086 (2017).

[34] Y. Setty, A. E. Mayo, M. G. Surette, and U. Alon, *Detailed map of a cis-regulatory input function*, Proceedings of the National Academy of Sciences **100**, 7702 (2003).

[35] F. J. Poelwijk, M. G. J. deVos, and S. J. Tans, *Tradeoffs and optimality in the evolution of gene regulation*, Cell **146**, 462 (2011).

[36] J. M. G. Vilar and L. Saiz, *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 (2013).

[37] J. K. Rogers, C. D. Guzman, N. D. Taylor, S. Raman, K. Anderson, and G. M. Church, *Synthetic biosensors for precise gene control and real-time monitoring of metabolites*, Nucleic Acids Research **43**, 7648 (2015).

[38] J. Rohlhill, N. R. Sandoval, and E. T. Papoutsakis, *Sort-Seq approach to engineering a formaldehyde-inducible promoter for dynamically regulated Escherichia coli growth on methanol*, ACS Synthetic Biology (2017).

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

[40] F. M. Weinert, R. C. Brewster, M. Rydenfelt, R. Phillips, and W. K. Kegel, *Scaling of gene expression with transcription-factor fugacity*, Physical Review Letters **113**, 1 (2014).

[41] J. Monod, J. Wyman, and J.-P. Changeux, *On the nature of allosteric transitions: A plausible model*, Journal of Molecular Biology **12**, 88 (1965).

[42] H. G. Garcia, H. J. Lee, J. Q. Boedicker, and R. Phillips, *Comparison and calibration of different reporters for quantitative analysis of gene expression*, Biophysical Journal **101**, 535 (2011).

[43] R. C. Brewster, D. L. Jones, and R. Phillips, *Tuning promoter strength through RNA polymerase binding site design in Escherichia coli*, PLoS Computational Biology **8**, e1002811 (2012).

[44] J. Q. Boedicker, H. G. Garcia, and R. Phillips, *Theoretical and experimental dissection of DNA loop-mediated repression*, Physical Review Letters **110**, 1 (2013).

[45] S. J. James Boedicker Hernan Garcia, *DNA sequence-dependent mechanics and protein-assisted bending in repressor-mediated loop formation*, Physical Biology **10**, 066005 (2013).

[46] Z. Huang, L. Zhu, Y. Cao, G. Wu, X. Liu, Y. Chen, Q. Wang, T. Shi, Y. Zhao, Y. Wang, W. Li, Y. Li, H. Chen, G. Chen, and J. Zhang, *ASD: A comprehensive database of allosteric proteins and modulators*, Nucleic Acids Research **39**, D663 (2011).

[47] G.-W. Li, D. Burkhardt, C. Gross, and J. S. Weissman, *Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources*, Cell **157**, 624 (2014).

[48] N. E. Buchler, U. Gerland, and T. Hwa, *On schemes of combinatorial transcription logic*, Proceedings of the National Academy of Sciences **100**, 5136 (2003).

[49] J. M. G. Vilar and S. Leibler, *DNA looping and physical constraints on transcription regulation*, Journal of Molecular Biology **331**, 981 (2003).

[50] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips, *Transcriptional regulation by the numbers: Models*, Current Opinion in Genetics & Development **15**, 116 (2005).

[51] R. Phillips, *Napoleon is in equilibrium*, Annual Review of Condensed Matter Physics **6**, 85 (2015).

[52] R. Daber, M. A. Sochor, and M. Lewis, *Thermodynamic analysis of mutant Lac repressors*, Journal of Molecular Biology **409**, 76 (2011).

[53] S. Klumpp and T. Hwa, *Growth-rate-dependent partitioning of RNA polymerases in bacteria*, Proceedings of the National Academy of Sciences **105**, 20245 (2008).

[54] S. Marzen, H. G. Garcia, and R. Phillips, *Statistical mechanics of Monod-Wyman-Changeux (MWC) models*, Journal of Molecular Biology **425**, 1433 (2013).

[55] R. B. O'Gorman, J. M. Rosenberg, O. B. Kallai, R. E. Dickerson, K. Itakura, A. D. Riggs, and K. S. Matthews, *Equilibrium binding of inducer to lac repressor-operator DNA complex*, Journal of Biological Chemistry **255**, 10107 (1980).

[56] K. F. Murphy, G. Balázsi, and J. J. Collins, *Combinatorial promoter design for engineering noisy gene expression*, Proceedings of the National Academy of Sciences **104**, 12726 (2007).

[57] R. Daber, K. Sharp, and M. Lewis, *One is not enough*, Journal of Molecular Biology **392**, 1133 (2009).

[58] K. F. Murphy, R. M. Adams, X. Wang, G. Balázsi, and J. J. Collins, *Tuning and controlling gene expression noise in synthetic gene networks*, Nucleic Acids Research **38**, 2712 (2010).

[59] M. A. Sochor, *In vitro transcription accurately predicts lac repressor phenotype in vivo in Escherichia coli*, PeerJ **2**, e498 (2014).

[60] M. Rydenfelt, R. S. Cox, H. Garcia, and R. Phillips, *Statistical mechanical model of coupled transcription from multiple promoters due to transcription factor titration*, Physical Review E **89**, 012702 (2014).

[61] D. Sivia and J. Skilling, *Data analysis: a Bayesian tutorial* (OUP Oxford, 2006).

[62] S. Oehler, M. Amouyal, P. Kolkhof, B. von Wilcken-Bergmann, and B. Müller-Hill, *Quality and position of the three lac operators of E. coli define efficiency of repression*, The EMBO Journal **13**, 3348 (1994).

[63] M. Scott, C. W. Gunderson, E. M. Mateescu, Z. Zhang, and T. Hwa, *Interdependence of cell growth and gene expression: origins and consequences*, Science **330**, 1099 (2010).

[64] J. A. N. Brophy and C. A. Voigt, *Principles of genetic circuit design*, Nature Methods **11**, 508 (2014).

[65] D. L. Shis, F. Hussain, S. Meinhardt, L. Swint-Kruse, and M. R. Bennett, *Modular, multi-input transcriptional logic Gating with orthogonal LacI/GalR family chimeras*, ACS Synthetic Biology **3**, 645 (2014).

[66] B. M. C. Martins and P. S. Swain, *Trade-offs and constraints in allosteric sensing*, PLoS Computational Biology **7**, 1 (2011).

[67] V. Sourjik and H. C. Berg, *Receptor sensitivity in bacterial chemotaxis*, Proceedings of the National Academy of Sciences **99**, 123 (2002).

[68] J. E. Keymer, R. G. Endres, M. Skoge, Y. Meir, and N. S. Wingreen, *Chemosensing in Escherichia coli: Two regimes of two-state receptors*, Proceedings of the National Academy of Sciences **103**, 1786 (2006).

[69] L. R. Swem, D. L. Swem, N. S. Wingreen, and B. L. Bassler, *Deducing receptor signaling parameters from in vivo analysis: LuxN/AI-1 quorum sensing in Vibrio harveyi*, Cell **134**, 461 (2008).

[70] L. A. Mirny, *Nucleosome-mediated cooperativity between transcription factors*, Proceedings of the National Academy of Sciences **107**, 22534 (2010).

[71] T. Einav, L. Mazutis, and R. Phillips, *Statistical mechanics of allosteric enzymes*, The Journal of Physical Chemistry B **121**, (2016).

[72] H. G. Garcia, A. Sanchez, J. Q. Boedicker, M. Osborne, J. Gelles, J. Kondev, and R. Phillips, *Operator sequence alters gene expression independently of transcription factor occupancy in bacteria*, Cell Reports **2**, 150 (2012).

[73] J. Monod, J.-P. Changeux, and F. Jacob, *Allosteric proteins and cellular control systems*, Journal of Molecular Biology **6**, 306 (1963).

[74] A. Auerbach, *Thinking in cycles: MWC is a good model for acetylcholine receptor-channels*, The Journal of Physiology **590**, 93 (2012).

[75] A. Velyvis, Y. R. Yang, H. K. Schachman, and L. E. Kay, *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 (2007).

[76] M. Canals, J. R. Lane, A. Wen, P. J. Scammells, P. M. Sexton, and A. Christopoulos, *A Monod-Wyman-Changeux mechanism can explain G protein-coupled receptor (GPCR) allosteric modulation*, Journal of Biological Chemistry **287**, 650 (2012).

[77] R. Milo, J. H. Hou, M. Springer, M. P. Brenner, and M. W. Kirschner, *The relationship between evolutionary and physiological variation in hemoglobin*, Proceedings of the National Academy of Sciences **104**, 16998 (2007).

[78] M. Levantino, A. Spilotros, M. Cammarata, G. Schirò, C. Ardiccioni, B. Vallone, M. Brunori, and A. Cupane, *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 (2012).

[79] R. Lutz and H. Bujard, *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 (1997).

[80] T. S. Moon, C. Lou, A. Tamsir, B. C. Stanton, and C. A. Voigt, *Genetic programs constructed from layered logic gates in single cells*, Nature **491**, 249 (2012).

[81] L. Saiz and J. M. G. Vilar, *Ab initio thermodynamic modeling of distal multisite transcription regulation*, Nucleic Acids Research **36**, 726 (2008).

[82] S. Tungtur, H. Skinner, H. Zhan, L. Swint-Kruse, and D. Beckett, *In vivo tests of thermodynamic models of transcription repressor function*, Biophysical Chemistry **159**, 142 (2011).

[83] S. Forsén and S. Linse, *Cooperativity: over the Hill*, Trends in Biochemical Sciences **20**, 495 (1995).

[84] D. L. Jones, R. C. Brewster, and R. Phillips, *Promoter architecture dictates cell-to-cell variability in gene expression*, Science **346**, 1533 (2014).

[85] A. Eldar and M. Elowitz, *Functional roles for noise in genetic circuits*, Nature **467**, 167 (2010).

[86] J. Berg, S. Willmann, and M. Lässig, *Adaptive evolution of transcription factor binding sites*, BMC Evolutionary Biology **4**, 42 (2004).

[87] K. B. Zeldovich and E. I. Shakhnovich, *Understanding protein evolution: From protein physics to Darwinian selection*, Annual Review of Physical Chemistry **59**, 105 (2008).

[88] S. K. Sharan, L. C. Thomason, S. G. Kuznetsov, and D. L. Court, *Recombineering: a homologous recombination-based method of genetic engineering*, Nature Protocols **4**, 206 (2009).

[89] H. M. Salis, E. A. Mirsky, and C. A. Voigt, *Automated design of synthetic ribosome binding sites to control protein expression*, Nature Biotechnology **27**, 946 (2009).

[90]  L. C. Thomason, N. Costantino, and D. L. Court, *E. coli genome manipulation by P1 transduction*, Current Protocols in Molecular Biology **Chapter 1**, Unit 1.17 (2007).

[91]  A. Fernández-Castané, C. E. Vine, G. Caminal, and J. López-Santín, *Evidencing the role of lactose permease in IPTG uptake by Escherichia coli in fed-batch high cell density cultures*, Journal of Biotechnology **157**, 391 (2012).

[92]  M. Lewis, G. Chang, N. C. Horton, M. A. Kercher, H. C. Pace, M. A. Schumacher, R. G. Brennan, and P. Lu, *Crystal structure of the lactose operon repressor and its complexes with DNA and inducer*, Science **271**, 1247 (1996).

[93]  H. T. Maecker, A. Rinfret, P. D'Souza, J. Darden, E. Roig, C. Landry, P. Hayes, J. Birungi, O. Anzala, M. Garcia, A. Harari, I. Frank, R. Baydo, M. Baker, J. Holbrook, J. Ottinger, L. Lamoreaux, C. L. Epling, E. Sinclair, M. A. Suni, K. Punt, S. Calarota, S. El-Bahi, G. Alter, H. Maila, E. Kuta, J. Cox, C. Gray, M. Altfeld, N. Nougarede, J. Boyer, L. Tussey, T. Tobery, B. Bredt, M. Roederer, R. Koup, V. C. Maino, K. Weinhold, G. Pantaleo, J. Gilmour, H. Horton, and R. P. Sekaly, *Standardization of cytokine flow cytometry assays*, BMC Immunology **6**, 13 (2005).

[94]  K. Lo, R. R. Brinkman, and R. Gottardo, *Automated gating of flow cytometry data via robust model-based clustering*, Cytometry Part A **73A**, 321 (2008).

[95]  N. Aghaeepour, G. Finak, The2ptFlowCAP2ptConsortium, The2ptDREAM2ptConsortium, H. Hoos, T. R. Mosmann, R. Brinkman, R. Gottardo, and R. H. Scheuermann, *Critical assessment of automated flow cytometry data analysis techniques*, Nature Methods **10**, 228 (2013).

[96]  I. Nemenman, *Information theory and adaptation*, in *Quantitative Biology: From Molecular to Cellular Systems*, Vol. 30322 (Taylor; Francis, 2010), pp. 1–12.

[97]  D. J. MacKay, *Information theory, inference and learning algorithms* (Cambridge University Press, 2003).

[98]  J. B. Kinney, A. Murugan, C. G. Callan, and E. C. Cox, *Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence*, Proceedings of the National Academy of Sciences **107**, 9158 (2010).

[99]  W. Bialek and S. Setayeshgar, *Physical limits to biochemical signaling*, Proceedings of the National Academy of Sciences **102**, 10040 (2005).

[100]  T. Gregor, D. W. Tank, E. F. Wieschaus, and W. Bialek, *Probing the limits to positional information*, Cell **130**, 153 (2007).

[101]  G. Tkacik, C. G. Callan, and W. Bialek, *Information flow and optimization in transcriptional regulation*, Proceedings of the National Academy of Sciences **105**, 12265 (2008).

[102]  J. O. Dubuis, G. Tkacik, E. F. Wieschaus, T. Gregor, and W. Bialek, *Positional information, in bits*, Proceedings of the National Academy of Sciences **110**, 16301 (2013).

[103]  M. D. Petkova, G. Tkačik, W. Bialek, E. F. Wieschaus, and T. Gregor, *Optimal decoding of cellular identities in a genetic network*, Cell **176**, 844 (2019).

[104]  G. Rieckh and G. Tkačik, *Noise and information transmission in promoters with multiple internal states*, Biophysical Journal **106**, 1194 (2014).

[105]  E. Ziv, I. Nemenman, and C. H. Wiggins, *Optimal signal processing in small stochastic biochemical networks*, PLoS ONE **2**, e1077 (2007).

[106]  F. Tostevin and P. R. ten Wolde, *Mutual information between input and output trajectories of biochemical networks*, Physical Review Letters **102**, 218101 (2009).

[107]  G. Tkačik and A. M. Walczak, *Information transmission in genetic regulatory networks: a review*, Journal of Physics: Condensed Matter **23**, 153102 (2011).

[108]  G. Tkačik, C. G. Callan, and W. Bialek, *Information capacity of genetic regulatory elements*, Physical Review E **78**, 011910 (2008).

[109]  O. P. Tabbaa and C. Jayaprakash, *Mutual information and the fidelity of response of gene regulatory models*, Physical Biology **11**, 046004 (2014).

[110]  I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox, *Real-time kinetics of gene activity in individual bacteria*, Cell **123**, 1025 (2005).

[111]  H. Xu, L. A. Sepúlveda, L. Figard, A. M. Sokac, and I. Golding, *Combining protein and mRNA quantification to decipher transcriptional regulation*, Nature Methods **12**, 739 (2015).

[112]  S. L. Barnes, N. M. Belliveau, W. T. Ireland, J. B. Kinney, and R. Phillips, *Mapping DNA sequence to transcription factor binding energy in vivo*, PLoS Computational Biology **15**, e1006226 (2019).

[113]  M. Razo-Mejia, S. L. Barnes, N. M. Belliveau, G. Chure, T. Einav, M. Lewis, and R. Phillips, *Tuning transcriptional regulation through signaling: A predictive theory of allosteric induction*, Cell Systems **6**, 456 (2018).

[114] M. Rydenfelt, H. G. Garcia, R. S. Cox, and R. Phillips, *The influence of promoter architectures and regulatory motifs on gene expression in Escherichia coli*, PLoS ONE **9**, 1 (2014).

[115] M. R. Maurizi, *Proteases and protein degradation in Escherichia coli*, Experientia **48**, 178 (1992).

[116] J. Peccoud and B. Ycart, *Markovian modeling of gene-product synthesis*, Theoretical Population Biology **48**, 222 (1995).

[117] Q. Cui and M. Karplus, *Allostery and cooperativity revisited*, Protein Science **17**, 1295 (2008).

[118] J. R. Peterson, J. A. Cole, J. Fei, T. Ha, and Z. A. Luthey-Schulten, *Effects of DNA replication on mRNA noise*, Proceedings of the National Academy of Sciences **112**, 15886 (2015).

[119] H. Dong and C. G. Kurland, *Ribosome mutants with altered accuracy translate with reduced processivity*, Journal of Molecular Biology **248**, 551 (1995).

[120] E. O. Powell, *Growth rate and generation time of bacteria, with special reference to continuous culture*, Journal of General Microbiology **15**, 492 (1956).

[121] R. Grah, B. Zoller, and G. Tkačik, *Nonequilibrium models of optimal enhancer function*, Proceedings of the National Academy of Sciences **117**, 31614 (2020).

[122] C. G. Bowsher and P. S. Swain, *Environmental sensing, information transfer, and cellular decision-making*, Current Opinion in Biotechnology **28**, 149 (2014).

[123] R. Blahut, *Computation of channel capacity and rate-distortion functions*, IEEE Transactions on Information Theory **18**, 460 (1972).

[124] C. T. Bergstrom and M. Lachmann, *Shannon information and biological fitness*, in *Information Theory Workshop. IEEE, 2004* (IEEE, 2004), pp. 50–54.

[125] D. Polani, *Information: Currency of life?*, HFSP Journal **3**, 307 (2009).

[126] O. Rivoire and S. Leibler, *The value of information for populations in varying environments*, Journal of Statistical Physics **142**, 1124 (2011).

[127] E. Libby, T. J. Perkins, and P. S. Swain, *Noisy information processing through transcriptional regulation*, Proceedings of the National Academy of Sciences **104**, 7151 (2007).

[128] A. Rhee, R. Cheong, and A. Levchenko, *The application of information theory to biochemical signaling systems*, Physical Biology **9**, 045011 (2012).

[129] C. Scholes, A. H. DePace, and Á. Sánchez, *Combinatorial gene regulation through kinetic control of the transcription cycle*, Cell Systems **4**, 97 (2017).

[130] S. Choubey, J. Kondev, and A. Sanchez, *Distribution of initiation times reveals mechanisms of transcriptional regulation in single cells*, Biophysical Journal **114**, 2072 (2018).

[131] M. Lässig, V. Mustonen, and A. M. Walczak, *Predicting evolution*, Nature Ecology & Evolution **1**, 0077 (2017).

[132] A. K. Gardino, B. F. Volkman, H. S. Cho, S.-Y. Lee, D. E. Wemmer, and D. Kern, *The NMR Solution structure of BeF3-activated Spo0F reveals the conformational switch in a phosphorelay system*, Journal of Molecular Biology **331**, 245 (2003).

[133] S. Boulton and G. Melacini, *Advances in NMR methods to map allosteric sites: From models to translation*, Chemical Reviews **116**, 6267 (2016).

[134] S. Oehler, S. Alberti, and B. Müller-Hill, *Induction of the lac promoter in the absence of DNA loops and the stoichiometry of induction*, Nucleic Acids Research **34**, 606 (2006).

[135] A. Schmidt, K. Kochanowski, S. Vedelaar, E. Ahrné, B. Volkmer, L. Callipo, K. Knoops, M. Bauer, R. Aebersold, and M. Heinemann, *The quantitative and condition-dependent Escherichia coli proteome*, Nature Biotechnology **34**, 104 (2015).

[136] Chroma2ptTechnology2ptCorporation, *Chroma spectra viewer*, (2016).

[137] A. D. Edelstein, M. A. Tsuchida, N. Amodaj, H. Pinkard, R. D. Vale, and N. Stuurman, *Advanced methods of microscope control using μManager software*, Journal of Biological Methods **1**, 10 (2014).

[138] D. Marr and E. Hildreth, *Theory of edge detection*, Proceedings of the Royal Society B: Biological Sciences **207**, 187 (1980).

[139] S. Frank, *Input-output relations in biological systems: measurement, information and the Hill equation*, Biology Direct **8**, 31 (2013).

[140] J. N. Weiss, *The Hill equation revisited: uses and misuses*, The FASEB Journal **11**, 835 (1997).

[141] S. Gama-Castro, H. Salgado, A. Santos-Zavaleta, D. Ledezma-Tejeida, L. Muñiz-Rascado, J. S. García-Sotelo, K. Alquicira-Hernández, I. Martínez-Flores, L. Pannier, J. A. Castro-Mondragón, A. Medina-Rivera, H. Solano-Lira, C. Bonavides-Martínez, E. Pérez-Rueda, S. Alquicira-Hernández, L. Porrón-Sotelo, A. López-Fuentes, A. Hernández-Koutoucheva, V. D. Moral-Chávez, F. Rinaldi, and J. Collado-Vides, *RegulonDB version 9.0: high-level integration of gene regulation, co-expression, motif clustering and beyond*, Nucleic Acids Research **44**, D133 (2016).

[142] V. Shahrezaei and P. S. Swain, *Analytical distributions for stochastic gene expression*, Proceedings of the National Academy of Sciences **105**, 17256 (2008).

[143] M. K. Transtrum, B. B. Machta, K. S. Brown, B. C. Daniels, C. R. Myers, and J. P. Sethna, *Perspective: Sloppiness and emergent theories in physics, biology, and beyond*, The Journal of Chemical Physics **143**, 010901 (2015).

[144] J. Yu, *Probing gene expression in live cells, one protein molecule at a time*, Science **311**, 1600 (2006).

[145] J. M. G. Vilar and L. Saiz, *Control of gene expression by modulated self-assembly*, Nucleic Acids Research **39**, 6854 (2011).

[146] J. L. Radzikowski, S. Vedelaar, D. Siegel, Á. D. Ortega, A. Schmidt, and M. Heinemann, *Bacterial persistence is an active σ S stress response to metabolic flux limitation*, Molecular Systems Biology **12**, 882 (2016).

[147] P. Hammar, M. Walldén, D. Fange, F. Persson, Ö. Baltekin, G. Ullman, P. Leroy, and J. Elf, *Direct measurement of transcription factor dissociation excludes a simple operator occupancy model for gene regulation*, Nature Genetics **46**, 405 (2014).

[148] U. Moran, R. Phillips, and R. Milo, *SnapShot: key mumbers in biology*, Cell **141**, 1262 (2010).

[149] A. Meurer, C. P. Smith, M. Paprocki, O. Čertík, S. B. Kirpichev, M. Rocklin, A. Kumar, S. Ivanov, J. K. Moore, S. Singh, T. Rathnayake, S. Vig, B. E. Granger, R. P. Muller, F. Bonazzi, H. Gupta, S. Vats, F. Johansson, F. Pedregosa, M. J. Curry, A. R. Terrel, Š. Roučka, A. Saboo, I. Fernando, S. Kulal, R. Cimrman, and A. Scopatz, *SymPy: Symbolic Computing in Python*, PeerJ Computer Science **3**, e103 (2017).

[150] A. Ale, P. Kirk, and M. P. H. Stumpf, *A general moment expansion method for stochastic kinetic models*, The Journal of Chemical Physics **138**, 174101 (2013).

[151] A. Andreychenko, L. Bortolussi, R. Grima, P. Thomas, and V. Wolf, *Modeling Cellular Systems*, Vol. 11 (Cham, 2017).

[152] F. Fröhlich, P. Thomas, A. Kazeroonian, F. J. Theis, R. Grima, and J. Hasenauer, *Inference for stochastic chemical kinetics using moment equations and system size expansion*, PLoS Computational Biology **12**, e1005030 (2016).

[153] D. Schnoerr, G. Sanguinetti, and R. Grima, *Approximation and inference methods for stochastic biochemical kinetics—a tutorial review*, Journal of Physics A: Mathematical and Theoretical **50**, 093001 (2017).

[154] P. Smadbeck and Y. N. Kaznessis, *A closure scheme for chemical master equations*, Proceedings of the National Academy of Sciences **110**, 14261 (2013).

[155] R. Cheong, A. Rhee, C. J. Wang, I. Nemenman, and A. Levchenko, *Information transduction capacity of noisy biochemical signaling networks*, Science **334**, 354 (2011).