Taming the Molecular Dance: Harnessing Statistical Mechanics to Quantitatively Characterize Allosteric Systems

Thesis by
Tal Einav

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

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2019
Defended June 6, 2019

ACKNOWLEDGEMENTS

I loved my time at Caltech. Every day was an extraordinary gift, and while it is impossible to distill six years' worth of transformative experiences and personal growth into a few pages, several pivotal moments cannot go unmentioned. In my first year of graduate school, I had a stroke of unbelievable luck when I met my future advisor, Rob Phillips, who in two hours did the impossible: he convinced a hard-line theoretical physicist to join an experimental biology lab. Physics, he claimed, has the best tools, but biology has the best problems. Like many graduate students, I picked my advisor on a whim; yet, in the following years, I found out that I had hit the jackpot. To this day, I continue to marvel at how well Rob leads our group. He promotes independence, encourages weirdness, and has a pure love of science. These traits have forged a close-knit and incredibly loyal family, and I am proud to be a part of this special group of people.

At the end of my first year, I took a sabbatical to pursue my dream of working at Wolfram Research (the makers of Mathematica). That year was rife with new experiences. Most importantly, I met Mike Perry, Rhonda Olds, and Erica Riddick, who made my time (through sunshine and polar vortices alike) unforgettable. I acquired orders of magnitude more proficiency at Mathematica juxtaposed with an awestruck understanding that, even now, I am only scratching the surface of what this program can do. My learning from that one year is embedded in all of the pages that follow – in the plots I present, the calculations I describe, and in the myriad unsuccessful computations you will not see, but that nevertheless formed an essential part of my scientific journey. Programming became one of my superpowers, enabling me to peer past the boundaries of our knowledge; in the words of Eddie Murphy, "I've got a dragon, and I'm not afraid to use it."

I returned to Caltech in the fall of 2014 and joined the Rob Phillips group, beginning my PhD trek in earnest. To my surprise, the skills required in graduate school were nearly orthogonal to those I had honed as an undergraduate. I learned to make figures, collaborate with others, and to continually question the depth of my knowledge on any subject. I came to understand that doing good science is, above all, about putting in consistent and relentless effort. The ability to sprint for a week quickly loses out to a pure love for a subject that makes work a reflex, something as natural as breathing.

In addition to my research, I had the privilege of holding two additional academic roles throughout graduate school. The first was a teaching role in the physics department. At Caltech, all freshman and sophomore classes are structured so that students receive two hours of lecture per week from a professor, complemented by an additional two hours each week led by a graduate student, and I leaped at this opportunity to teach some of my favorite subjects. Although physics is by no means a nascent field, I was shocked to discover that there were *zero guidelines* on what material to cover nor how to present it. In such cases, I follow the maxim that it is better to leap high and risk choking on greatness rather than to nibble on mediocrity. I loved every aspect of teaching, from the 15+ hours of preparation per lecture to the showmanship required to present the material to my postmortem examination of how each lecture could be improved the following year. Teaching was my gift to the next generation of scientists. I saw the freshman to whom I taught classical mechanics, special relativity, and electrodynamics mature into upperclassman, apply for jobs and graduate schools, and move onto the next great steps of their lives. It is a small world, and I look forward to meeting a professor or professional who remembers me as their TA at Caltech.

My other role came each summer when I joined Rob on his yearly sojourn to the Marine Biological Laboratory (MBL) at Woods Hole, where he served as one of the codirectors of the Physiology course. This program brings together thirty graduate students and postdocs from around the world to perform mini-research projects outside of their field of expertise. In three distinct two-week rotations, students choose a research project and pursue it relentlessly from 9am to 2am, Monday-*Saturday*! While this schedule appear deranged and borderline suicidal, it leaves out the most important element that consistently makes Woods Hole a highlight in any academic life: this experience rekindles your love of science. When arriving at Woods Hole, you leave behind your deadlines, your frustrations with collaborators, your scientific baggage, and you suffuse yourself with the exhilaration of the pursuit of knowledge. On top of that, what starts off as thirty complete strangers quickly coalesces into a tight-knit family permeated with trust, respect, and life-long bonds.

For four years, I have joined Rob as the teaching assistant for one of these two-week periods which we titled "Physical Biology X," and for which we invited theorists and experimentalists alike to join us in modeling a system of their choice. To aid in their modeling, Rob led a series of lectures discussing various theoretical tools (e.g. random walks, master equations, and statistical mechanics). Watching Rob in action has been awe-inspiring, as it reveals his seemingly endless knowledge on every conceivable topic in biology. After my first year teaching at Woods Hole, I remember leaving with a deep sense of wonder at the experience, mixed with the uncomfortable realization that I could never in a million years run the theory rotation like Rob does. I simply did not know enough.

After finishing my fourth MBL experience, I had an even greater reverence for the magic of this program. The other faculty had become my family, and the thrill of interacting with the new batch of students – working with someone until a breakthrough at 1am lights up their face – had not diminished. Yet intellectually, I was astonished by my own progress. My ability to help drive the student projects had skyrocketed over the past three years. I could keep up with the daily lecture series as well as in more casual science talks over meals. As I left Woods Hole, I realized that although I still did not know enough to teach the theory rotation like Rob does, for the first time I believed that I could do so in a few years.

My teaching at Woods Hole ended up having a much larger impact on my academics than I had originally anticipated. After each summer, Rob invited students to carry out post-course research at Caltech, and each of these collaborations blossomed into a published paper. I am continually awed by the effort and devotion that Rob and the other MBL students and faculty put into this program, and I am very thankful to have been part of these experiences.

Returning my attention to Caltech, I must thank the people who made my time there so special. Of many friends, I am especially grateful to Zofii Kaczmarek, Jenish Mehta, Betsy Medvedovsky, Peng (Brian) He, and Bassam Helou for countless hours of entertainment and shared experiences. I thank my thesis committee – Rob Phillips, Justin Bois, Jim Eisenstein, David Hsieh, and Olexei Motrunich – for their time, patience, and input on my academic progress. I owe special thanks to Nigel Orme whose beautiful scientific illustrations riddle these pages. The staff at Caltech was uniformly excellent. I give my deepest thanks to Jo Ann Hasbach, one of the nicest and most sincere individuals I have ever met, together with Michelle Vine, Sofie Leon, and the other incredible people who run the Caltech physics department, as well as Jonathan Gross and Celene Barrera who managed the Phillips group. Anywhere on campus, I was surrounded by family.

One of the joys of living on a college campus is the ability to interact with other students in many different capacities, either as their teacher in physics, as their fellow dancer during the annual Caltech dance show, as their peer in art class, or as a friend during dinner. I owe deep thanks to every member of Avery House from 2012-2019, who were part of my extended family at Caltech. But of that myriad of friends, I owe special thanks to three especially bright stars – Giacomo (G) Koszegi, Jiseon Min, and Gladia Hotan – I cannot wait to see what amazing feats they accomplish!

Each of my labmates has my undying respect; we matured into scientists together, and I love each and every single one of them. But I must pay special thanks to Manuel Razo-Mejia, whose love of science is so infectious that in my third year of graduate school, he convinced me to dedicate myself wholeheartedly to academia. In addition, Vahe Galstyan has been an amazing friend and labmate.

Caltech was a scientific mecca for brilliant minds, and there was no shortage of resources to tap into when a problem became challenging. Above all, I thank Marius Lemm, one of the smartest mathematical physicists I have ever met, who could solve any problem I had been tackling in my research within half an hour. I am also thankful for conversations with Anthony Bartolotta, William Chan, Peng (Brian) He, Noah Olsman, and so many others who have helped me develop my scientific skills. Forever near and dear to my heart are the Quantitative Biology students from the Kavli Institute of Theoretical Physics in 2015, a special group of people with whom I undertook a program comparable to the Physiology course at Woods Hole. I especially thank Stephanie Barnes, Tommaso Biancalani, Chico Camargo, Savita Chib, Amy Gilson, Yu-Ying (Phoebe) Hsieh, Michael Manhart, Ramya Purkanti, and Jorge Zanudo.

I finish with my two most heartfelt thanks to the individuals without whom none of this would have been possible. First, I owe the deepest of thanks to my parents, whom I appreciate ever more as the years flow by. Their steadfast love and support provides the foundation for everything I do. As Lois McMaster Bujold writes, "You don't pay back your parents. You can't. The debt you owe them gets collected by your children, who hand it down in turn." My fortune, favor, and positive attitude are all attributable to them, and I hope they find pride and joy in my accomplishments.

Lastly, I thank my advisor Rob Phillips. Joining his lab was one of the best decisions I have made in my life. I am told that being an advisor is one of the most challenging and rewarding aspects of academia, and I have found no deeper connection during my time at Caltech. He deepened my love of science and continually pushed me to exceed my boundaries. By the end of my PhD, Rob became a friend, a fellow scientist, my biggest supporter, and my best collaborator. I aspire to one day lead my own group with the same zeal and vigor.

ABSTRACT

The pace of biological research continues to grow at a staggering pace as high-throughput experimental techniques rapidly increase our ability to sequence DNA, quantify cell behavior, and image molecules of all types within the cellular milieu. Given this surge in experimental prowess, the time is ripe to examine how well our conceptual cartoons of biological phenomena can not only recapitulate the data but also successfully predict the outcomes of future experiments.

One of the fundamental challenges in biology is that the space of possible molecules is overwhelmingly large. The number of variants of a moderately-sized protein (20^{300}) is larger than the number of atoms in the universe, as is the space of possible bacterial genomes, protein interaction networks, and effector functions; progress in any of these fronts requires a theory-experiment dialogue that can extrapolate our small drop of data to explain large swaths of parameter space.

My thesis strives towards this goal by analyzing a number of central molecular players in biology including enzymes (biological catalysts that accelerate chemical reactions), transcription factors (proteins that bind to DNA and regulate its expression), and ion channels (signaling proteins that regulate ion transport). I develop a quantitative description in each context by harnessing the statistical mechanical Monod-Wyman-Changeux model of allostery which coarse-grains the behavior of a multi-state system into two effective states, demonstrating that these seemingly diverse molecules are all governed by the same fundamental equation.

Writ large, there are two overarching goals encompassed by these projects. The first is to translate our biological knowledge into concrete physical models, enabling us to quantitatively describe how the key molecular components in each system interact to carry out their function. The second goal is to analyze how mutations can be mapped into the fundamental biophysical parameters governing each system. In my opinion, predicting the effects of mutations remains one of the great unsolved problems in biology, and it has been incredibly exciting to make progress on this front.

Looking back at my amazing graduate school experience, one of the most surprising aspects of my PhD was how closely each of my projects revolved around experiments. I entered graduate school as a theoretical physicist expecting to work on esoteric mathematical models, yet the direct connection with data provided a window into

the exhilarating world of biology. While I have never physically manipulated these biological systems in the lab, my models allow me to push and prod and examine their behavior from the most mundane to the utterly extreme limits. Through modeling, I test our assumptions of how these systems work and tease out insights into their underlying biophysical mechanism. Most importantly, these models enable me to harness the incredible wealth of hard-won data to weave a few more threads of understanding into our tapestry of how these incredible living systems operate.

PUBLISHED CONTENT AND CONTRIBUTIONS

1. Einav T, Mazutis L, Phillips R. Statistical Mechanics of Allosteric Enzymes. The Journal of Physical Chemistry B. 2016;120(26):6021-6037. doi:10.1021/acs.jpcb.6b01911

2. Einav T, Phillips R. Monod-Wyman-Changeux Analysis of Ligand-Gated Ion Channel Mutants. The Journal of Physical Chemistry B. 2017;121(15):3813-3824. doi:10.1021/acs.jpcb.6b12672

3. Einav T, Duque J, Phillips R. Theoretical Analysis of Inducer and Operator Binding for Cyclic-AMP Receptor Protein Mutants. PLOS ONE (Public Library of Science). 2018;13(9):e0204275. doi:10.1371/journal.pone.0204275

4. Razo-Mejia M, Barnes S, Belliveau N, Chure G, Einav T, Mitchell L, Phillips R. Tuning Transcriptional Regulation through Signaling: A Predictive Theory of Allosteric Induction. Cell Systems. 2018;6(4):456-469. doi:10.1016/j.cels.2018.02.004

5. Galstyan V, Funk L, Einav T, Phillips R. Combinatorial Control through Allostery. The Journal of Physical Chemistry B. 2019;123(13):2792-2800. doi:10.1021/acs.jpcb.8b12517

6. Einav T, Phillips R. How the Avidity of Polymerase Binding to the -35/-10 Promoter Sites Affects Gene Expression. Proceedings of the National Academy of Sciences. *In press*

7. Einav T, Yazdi S, Coey A, Bjorkman P, Phillips R. Harnessing Avidity: Quantifying the Entropic and Energetic Effects of Linker Length and Rigidity Required for Multivalent Binding of Antibodies to HIV-1 Spikes. *In submission*

8. Chure G, Razo-Mejia M, Belliveau N, Einav T, Barnes S, Mitchell L, Phillips R. The Energetics of Molecular Adaptation in Transcriptional Regulation. *In submission*

In cases of first authorship, I performed the research and wrote the paper. In all other cases, I assisted with both the research and writing as needed.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Number *Page*

LIST OF TABLES

Chapter 0

INTRODUCTION

Physics and biology strive to explain how the world works, yet their underlying philosophies are markedly different. Physics promotes simple, unifying ideas. Biology embraces the unique and breathtaking complexity of life. This work empathizes with both views, but at its core we apply the physicist's mindset toward the broad goal of determining how well our current understanding of biological mechanisms, when translated into statistical mechanical models, can characterize biological phenomena.

The level of description we aim for is not atomistic, but rather a coarse-grained model of a system's behavior. For example, Fig. 0.1 shows a ubiquitous input-output response (black curve) that arise in many biological contexts where a protein is regulated by an effector, a small molecule that binds to the protein and modulates its activity. Later chapters contain many instances of such curves, and we develop theoretical models to characterize these responses and quantify their salient features, namely, their minimum (the *leakiness* of the response in the absence of effector), maximum (the *saturating* response), midpoint (called the $[EC_{50}]$), and slope (known as the *effective Hill coefficient*). The resulting formulas tie these properties to experimentally tunable parameters (e.g. concentrations of molecules or binding energies), making sharp predictions that can guide experimental efforts seeking a particular behavior (e.g. a sharp response with large saturation and low leakiness).

The majority of my thesis centers around the concept of allostery, in which macromolecules switch between multiple conformations, and its characterization through the MWC model, eponymously named for Jacques Monod, Jeffries Wyman, and Jean-Pierre Changeux who first developed it in the 1960s. This model provides an elegant way to coarse-grain molecular fluctuations; at the nano scale, the innards of a cell are a tumultuous maelstrom of activity, with individual molecules constantly bouncing around, unfolding, refolding, and assuming different conformational states. The MWC model quantifies the average behavior of this ensemble of states by grouping conformations together and assessing how frequently each occurs, enabling us to map intricate molecular responses onto a tractable two-state system. And much as the two-state Ising model has been applied to diverse areas

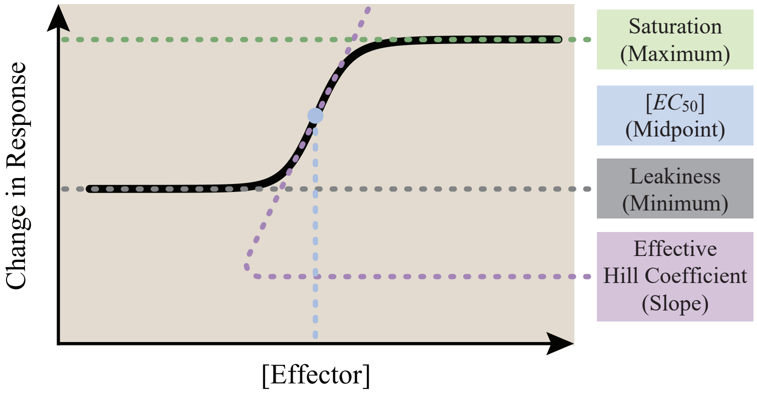


Figure 0.1: General input-output function of a biological system. Key biological properties of the response (black curve) are shown on the right.

of physics ranging from magnetism to evaporation to spin glasses, the breadth and scope of the following chapters – which includes molecular machinery involved in bacterial transcription as well as ion channels that transmit signals in our brain – are a testament to how a simple idea can describe complex and highly disparate systems.

The remainder of this chapter provides a roadmap that highlights the key objectives and major results of this work. In **Chapter 1**, I describe my first case study of allostery in the context of enzyme kinetics. Enzymes are molecular factories that catalyze chemical reactions, the rate of which depend on the concentrations of their substrates. Each cell in our body contains over 1000 different enzymes that are collectively responsible for a myriad of tasks ranging from digesting the food we eat to replicating our DNA to neutralizing invading pathogens. This process is often regulated by a suite of activators and inhibitors that can speed it up or slow it down.

I first describe a number of theoretical results that apply to many different types of enzymes. A key insight is that the activity of an enzyme in the presence of an allosteric regulator or a competitive inhibitor can be exactly recast into the activity of an enzyme with no regulator or inhibitor present, provided that parameters are appropriately rescaled. This demonstrates that the space of possible input-output responses cannot be increased by adding either of these effector molecules and offers a way to rigorously integrate out these degrees of freedom.

We then consider two seemingly counter-intuitive phenomena where enzymes exhibit non-monotonic activity as shown in Fig. 0.2. These peaked responses are impossible in classical Michaelis-Menten kinetics (a simple and commonly-used description of an enzyme), yet we demonstrate that an allosteric enzyme can exhibit both behaviors provided that its binding constants obey a specific relationship.

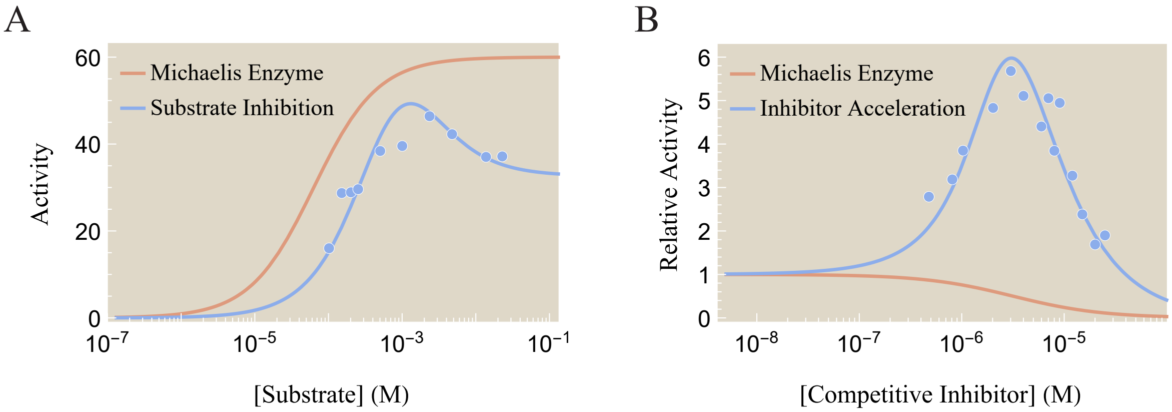


Figure 0.2: Non-monotonic enzyme activity. A peak in activity (blue curve) contradicts the canonical Michaelis-Menten description (tan curve). (A) Adding more substrate (the fuel for an enzyme's reaction) ordinarily makes an enzyme work faster rather than slower. (B) Adding a competitive inhibitor (which competes with substrate) decreases an enzyme's speed in the Michaelis-Menten model, in clear contradiction with the data.

A subsequent literature search found that the enzyme acetylcholinesterase obeys the mathematical criterion we predicted would assure non-monotonicity, and we confirmed that this enzyme does indeed exhibit a non-monotonic activity curve!

Chapter 2 represents an incredible collaboration within the Phillips group, in which I teamed up with four experimentalists (Stephanie Barnes, Nathan Belliveau, Griffin Chure, and Manuel Razo-Mejia) to test how accurately we can predict the behavior of a transcription factor. Specifically, we examined the Lac repressor which is responsible for regulating metabolism in the bacterium *Escherichia coli*. This organism prefers to use glucose as its source of energy, but when glucose is scarce and lactose is present, it switches tactics and begins to generate the lactose-digesting machinery. This switch is mediated by the Lac repressor, which ordinarily binds to and blocks expression of the genes necessary to import and break down lactose (Fig. 0.3A). But when lactose is present, the repressor assumes a different conformation in which it can no longer bind to DNA.

We developed a model of this process (Fig. 0.3B) that quantifies the fold-change, the ratio of gene expression in the presence and absence of a repressor (a quantity that must lie between 0 and 1 since the repressor inhibits expression). More precisely, our model relates fold-change to experimentally-manipulatable, physical parameters of the system including the number of repressors, the amount of the effector lactose, and the binding energy of the repressor.

We first inferred the values of all parameters in our model by measuring the system's response under one set of conditions (white points in Fig. 0.3D). This permitted us to predict how the system should behave when the repressor copy number, effector concentration, and repressor binding energy were varied (colored curves in Fig. 0.3C-E). As a theoretical physicist, I was immensely proud of this result because it exemplified our belief in the predictive power of our models, which stands as a strong counterpoint to the more common role of theory in biology, where models are primarily used to retrospectively analyze data.

The benefit of collaborating with four fantastic experimentalists is that they were able to create and measure the responses of these 17 new strains of bacteria. This project took an extraordinary amount of work, and the methods they developed could have merited an entirely separate paper, yet the final results shown in Fig. 0.4A-C

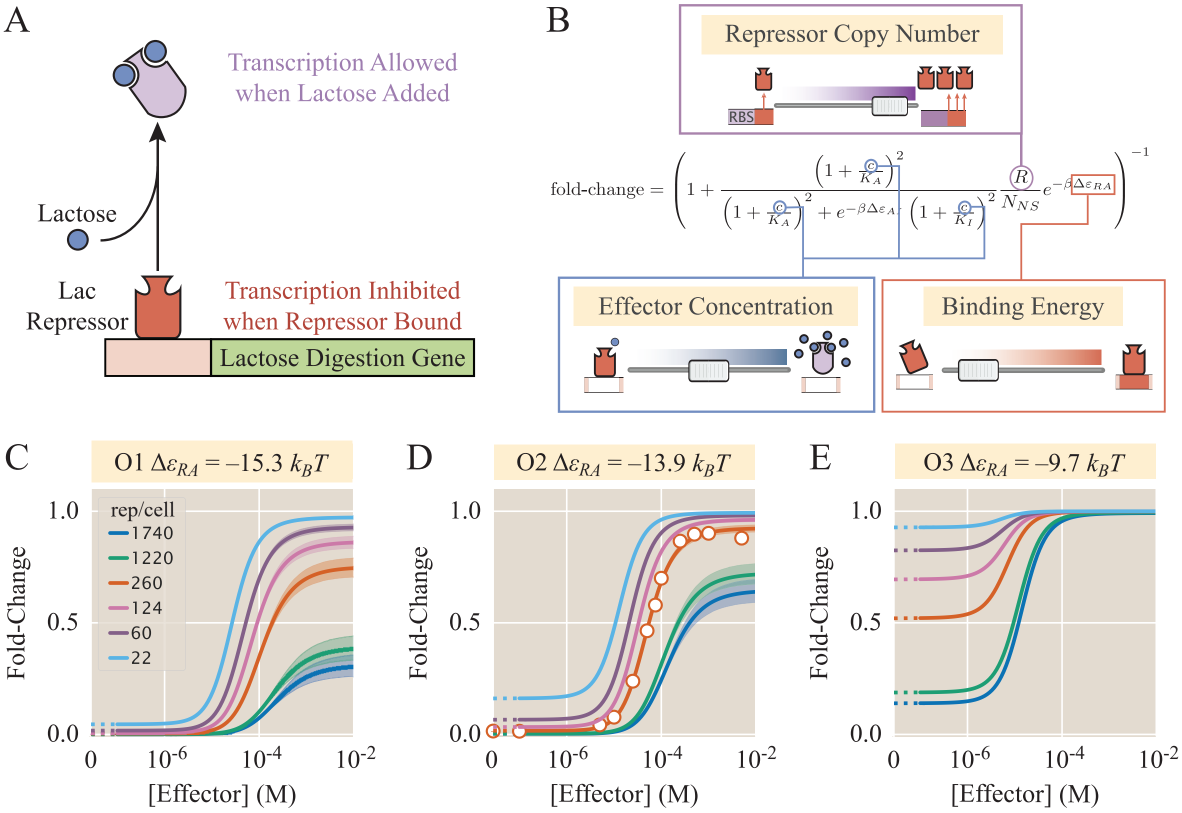


Figure 0.3: Predicting the behavior of a transcription factor. (A) The Lac repressor (red) normally inhibits the expression of lactose digesting machinery, but in the presence of lactose it assumes a different conformation (purple) and permits the gene to be expressed. (B) The model of this system showing the experimental variables that were tuned. (C-E) Using a single set of measurements (white points), we predicted the behavior of 17 other genetic circuits (colored curves).

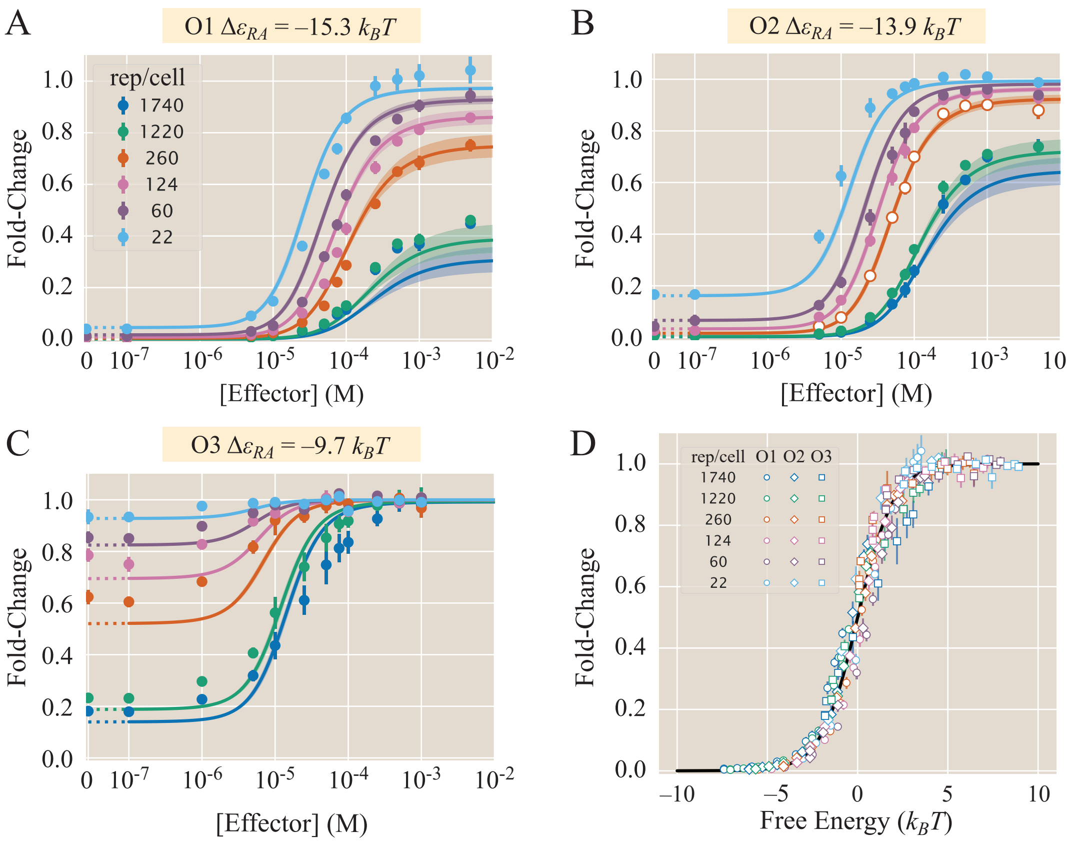


Figure 0.4: Demonstrating the predictive power of our model. (A-C) 17 strains of bacteria were subsequently designed and their response was measured to test the predictive power of our model. (D) The full suite of data from all 18 strains can be collapsed down into a one-dimensional parameterization using the free energy of repressor binding.

were well worth it. The data points fall beautifully on the line, representing the most careful and rigorous confirmation to-date of how predictive biology can be in the context of this induction process. The *coup de grace* is shown in Fig. 0.4D where we demonstrated that the responses of all 18 strains at every effector concentration, repressor copy number, and DNA binding energy collapse down into a one-dimensional curve governed by the free energy of the repressor's binding. This result embodies the physicist's credo that characterizing the underlying mechanism governing a process – from tossing a ball into the air to regulating the transcription of genetic elements – enables you to collapse seemingly disparate and complex behavior and understand it from a unifying perspective.

While the success of this project was immensely gratifying, it was also grounded upon decades of experiments that built up significant intuition about this system.

In contrast, **Chapter 3** abandons the comfortable bedrock of intuition and jumps to the wild frontier by analyzing how mutations of this same Lac repressor change its ability to inhibit gene expression. A mutation is an alteration in a DNA sequence that can change the composition of a protein. In a sense, much of biology is the study of mutations – they are the instrument by which change happens both on a local scale (as when our immune system fights a new virus) and on evolutionary time (the process by which organisms evolve). Yet studying mutations is daunting because: (1) the space of possible mutants is enormous and (2) because it is rarely possible to predict the effects of a mutation *a priori*.

To quantify the first point, an average protein has 300 amino acids, and with 20 variants for each slot there are 20^{300} permutations to explore (Fig. 0.5A). Since this breathtakingly enormous quantity is larger than the number of atoms in the universe, it is clear that even over evolutionary time, only a tiny fraction of all possible proteins have been created in nature, even when considering every living organism that has ever existed on our planet. This suggests that the search for an optimal protein-targeting drug or the design of a synthetic circuit may be never-ending.

Yet the second point is of paramount importance, since the array of possible mutants is only daunting when there is no model to extrapolate how the unknown multitudes will behave. Unfortunately, we have very few models capable of characterizing mutations. One confounding factor is that even a single amino acid mutation may cause a protein to stop functioning, as seen in the case of sickle cell anemia, where a change in one amino acid causes hemoglobin to misfold. Because even single mutations may have such large and unforeseen consequences, it is tempting to treat each mutant as a new intellectual adventure with no visible quantitative link to its unmutated form.

In this sense, modeling mutations is akin to working on understanding memory or the emergence of consciousness – it is an area of biology where some believe that no simple theory is possible. Amid this backdrop of uncertainty, Rob and I embarked on our journey to analyze a series of mutations in the effector- and DNA-binding domains of the Lac repressor (Fig. 0.5B). We attempted to calculate the first order "Taylor approximation" of each mutation by inferring the minimal set of thermodynamic parameters that differ between each mutant and the wild type protein.

Rob hypothesized that mutations in the effector binding domain should *only* affect the repressor-effector binding energy and mutations in the DNA binding domain

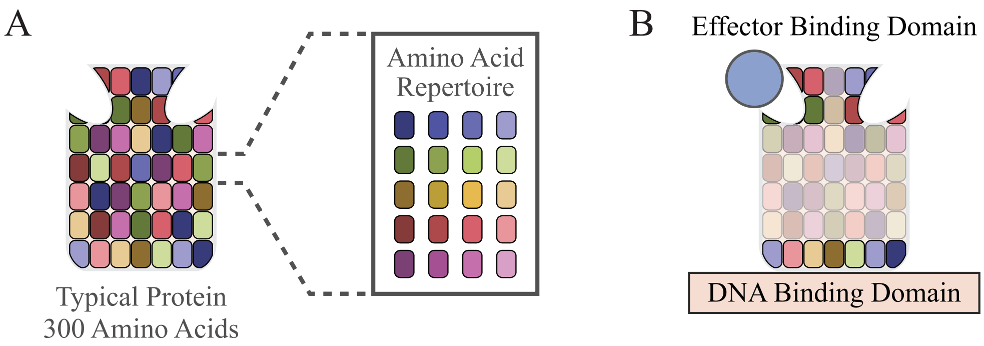


Figure 0.5: The arsenal of mutations within a typical protein. (A) The Lac repressor from Fig. 0.3A has approximately 20^{300} variants. (B) We analyzed a suite of mutations in amino acids surrounding the ligand biding domain or DNA binding domain.

should *only* alter the repressor-DNA binding energy, with no cross-talk between them. This lack of interference may sound innocuous, yet biology is replete with examples where a single mutation alters many aspects of a protein's function (for example, it could affect all four properties of the response shown in Fig. 0.1). We were curious whether seemingly diverse effects on an input-output function could decouple when viewed through the underlying parameters of the system. More precisely, does mutating the effector binding domain only affect the K_A and K_I dissociation constants while mutating the DNA binding domain only alter $\Delta\epsilon_{RA}$ (as depicted in Fig. 0.3B) relative to the wild type repressor?

Much to our surprise, we found that the 14 mutants we considered were all well characterized by our model. A key result was that the DNA and effector binding mutations have different signatures. Each mutant in the former class exhibits the same change in free energy when transitioning from zero to saturating concentrations of effector (the same length of horizontal lines above the right plot in Fig. 0.6A), while the latter category is characterized by the same leakiness (all curves in the left plot of Fig. 0.6B have small fold-change in the absence of effector). This remarkable result suggests that the enormous space of arbitrary mutations to either binding domain could be collapsed down into a one- or two-dimensional family of curves, providing an unprecedented insight into how we can view the spectrum of possible phenotypes.

Towards the end of this project, I made the casual observation that since the effects of DNA and effector mutations are decoupled within our model, we can predict the behavior of any double mutant with one DNA mutation and one effector mutation.

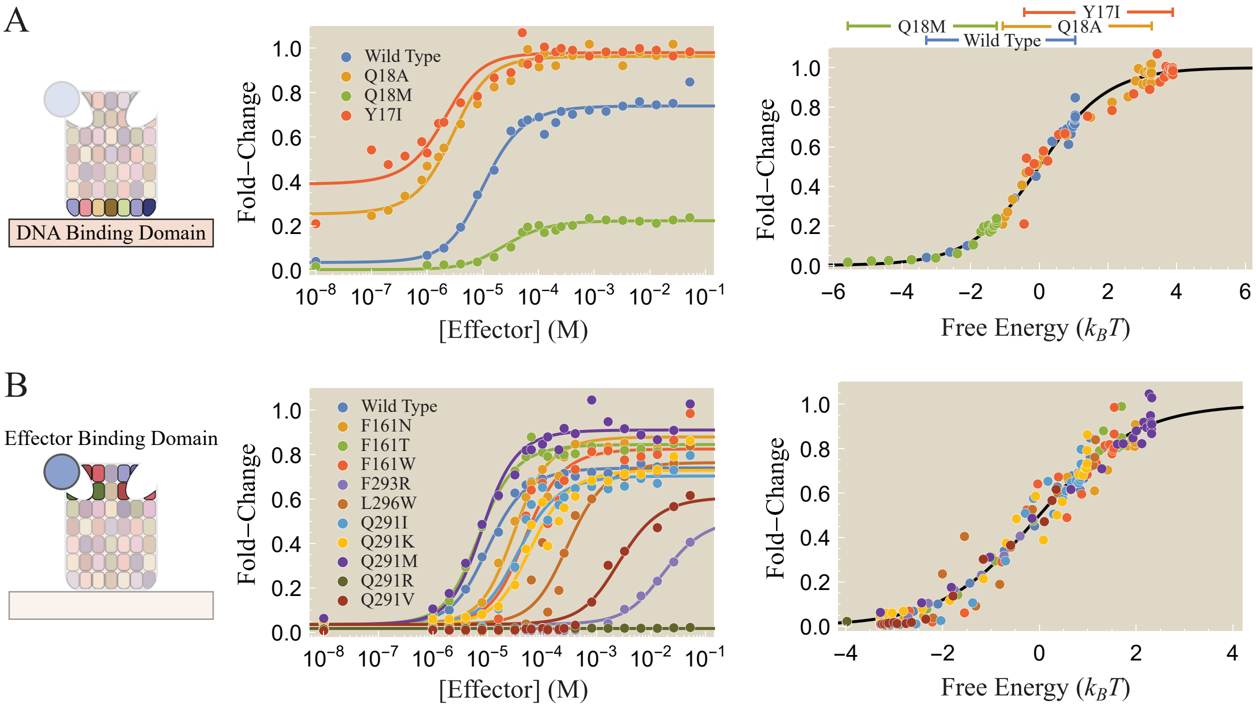


Figure 0.6: Characterizing Lac repressor mutants. (A) Mutations in the DNA binding domain were characterized by only altering their repressor-DNA interaction energy relative to the wild type. (B) Mutations in the effector binding domain only had their repressor-effector binding parameter changed from the wild type protein.

I created Fig. 0.7 and stuck it into the Supplementary Information of our paper, but Rob quickly pointed out that this noteworthy figure deserved to be featured in the main text, as it represents a significant stride towards harnessing the predictive power of our model. Although we had narrowed the possible phenotypes of the DNA and effector mutants to a one- or two-dimensional family of curves, our model could not precisely predict how a new mutation at either site would behave *a priori*. But these double mutant predictions were parameter-free and hence completely determined by our model! This implies that after n DNA mutants and n effector mutants are characterized, you can predict the behavior of n^2 double mutants without doing a single additional experiment! The notion that one experiment buys you more than a single data point was extremely satisfying to me, and it generated significant interest in our lab.

In fact, the same four experimentalists from our collaboration in Chapter 2 found these results so intriguing that they set out to create these double mutants and validate my model, turning it from a retrospective theoretical analysis of existing data to yet another journey into predicting the unknown. It has been an absolute privilege

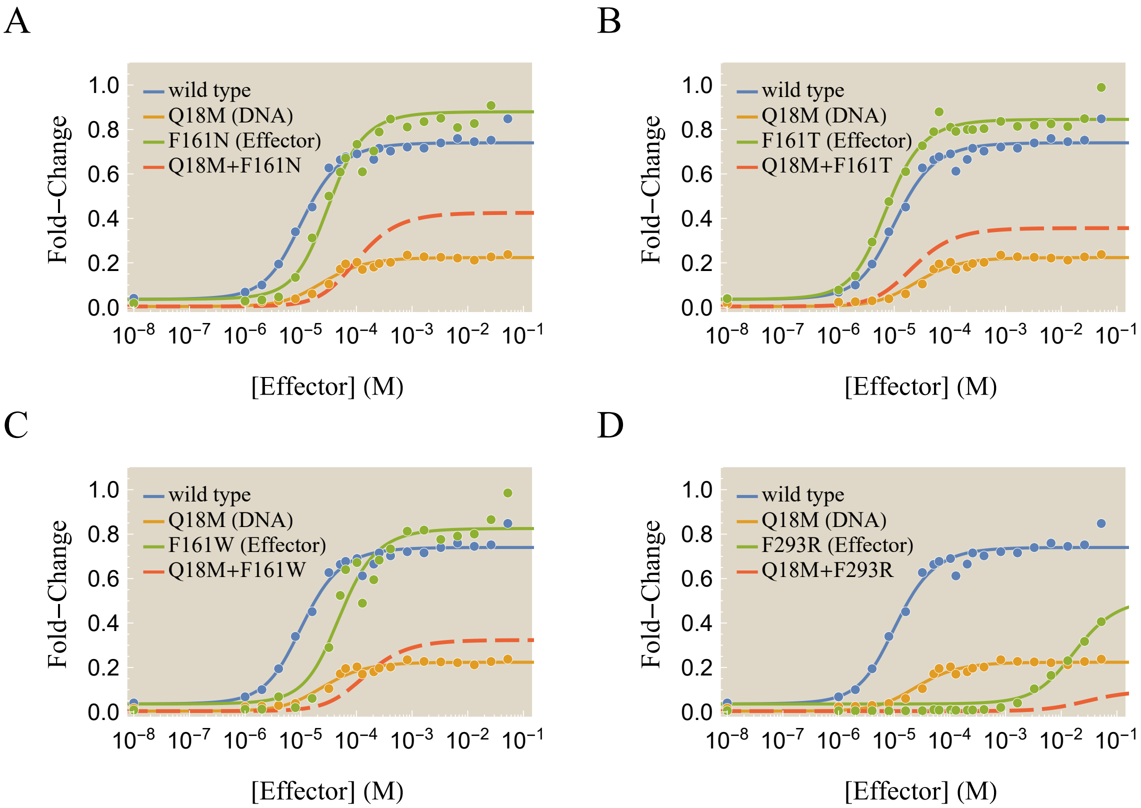


Figure 0.7: Predicting the effect of double mutants. Combining the disjoint effects of a DNA binding mutation and an effector binding mutation leads to parameter-free predictions of double mutants (dashed red curves). Shown here are four of the $3 \times 10 = 30$ possible double mutants utilizing the data in Fig. 0.6.

to have labmates willing to invest months (or even years) of effort based on this work. Three years later, this experiment is nearing completion, and our data have confirmed the key prediction regarding how these classes of mutations combine.

In the interim, I generalized our model of mutations to other biological systems. In **Chapter 4**, I turn to the transcriptional activator CRP shown in Fig. 0.8A. In many respects, this protein serves as a counterpoint to the Lac repressor discussed above. Both act on the lactose metabolism genes and both undergo a conformational change upon binding an effector (lactose for the Lac repressor and cAMP for CRP) that alters their ability to bind to DNA. But whereas the Lac repressor inhibits gene expression, CRP enhances it.

The primary goal of this project was to characterize how wild type CRP interacts with its effector to increase gene expression. But an exciting application of our model arose when Rodrigo Maillard's group in Georgetown University published

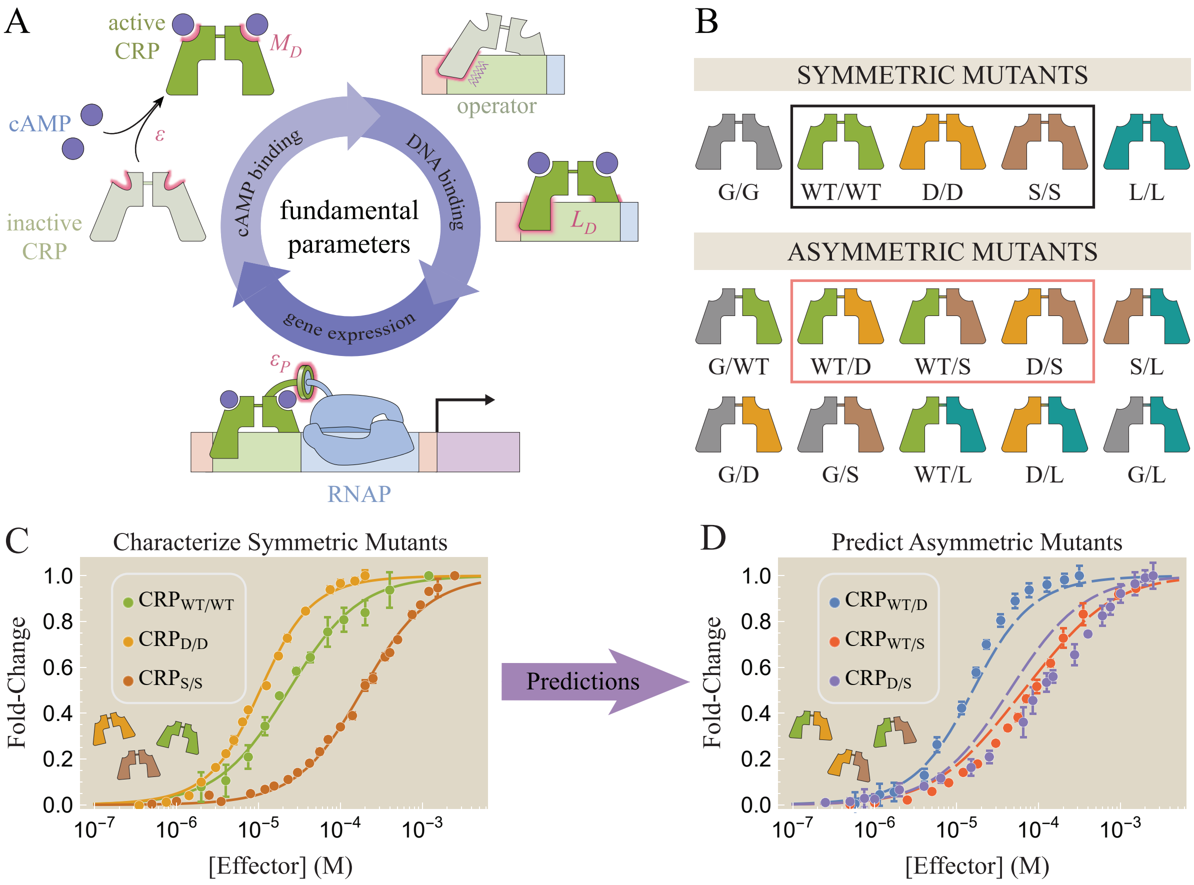


Figure 0.8: Predicting the behavior of the CRP mutant combinations. (A) CRP becomes allosterically active when bound by the effector cAMP, enabling the protein to bind to DNA and recruit RNA polymerase (RNAP). (B) We analyzed a series of mutations (denoted by D, S, G, and L) that could be introduced in either subunit. The six boxed mutants were created by Rodrigo Maillard's group and analyzed in this work. (C) We first characterized the symmetric mutants (WT/WT, D/D, S/S; solid curves). (D) We then predicted the responses of the asymmetric mutants (dashed lines) and compared our results to the experimental data.

an experimental paper in which they independently mutated the two CRP subunits, generating both symmetric and asymmetric mutants (Fig. 0.8B). This was an impressive accomplishment, since CRP is a homodimer composed of two identical wild type subunits (WT/WT). Altering the wild type gene to have the D mutation creates the symmetric double mutant (D/D) but precludes the possibility of making the asymmetric mutant with only one altered subunit (WT/D). If, instead of modifying the existing wild type gene, an additional D gene is introduced, a cell creates multiple CRP types (WT/WT, WT/D, and D/D), making it difficult to disentangle the effects of the asymmetric mutant. Instead, Rodrigo's group modified the gene to express both subunits tethered to one another, enabling them to create a homogeneous population of mutants where each subunit could be independently altered (shown by the six boxed molecules in Fig. 0.8B).

A key result from our analysis was that data from the symmetric CRP mutants (WT/WT, D/D, S/S) were sufficient to predict how the asymmetric mutants (WT/D, WT/S, D/S) would behave, as shown in Fig. 0.8C and D. As with the Lac repressor, this provides a way to harness the combinatorics of possible mutations, since characterizing n subunits enables us to predict how $\frac{n(n-1)}{2}$ combination mutants would behave (shown for $n = 5$ in Fig. 0.8B). These results confirmed yet again that a single experiment can convey more than a single data point of information, providing a way to build upon previous work and start navigating the vast space of possible mutations.

Chapter 5 represents an even broader application of our mutation model, this time in the context of ion channels. Upon ligand binding, these proteins permit ions to flow across cell membranes, mediating a diverse array of signaling pathways that enable our muscles to contract and our neurons to quickly transmit electric pulses. In this project, we considered one of the best-studied channels in biology, namely, the nicotinic acetylcholine receptor (nAChR) shown in Fig. 0.9A.

This project began when Rob sent me a paper with the data shown in Fig. 0.9B depicting the current flowing through nAChR channels in eukaryotic oocytes at different ligand concentrations. The data suggested that this channel becomes 10x more sensitive to ligand when any one of its five subunits acquires a specific mutation, as exhibited by the leftward shift of the curve representing $n = 1$ mutated subunits relative to the wild type $n = 0$ channel. Moreover, nAChR becomes 100x, 1000x, or 10 000x more sensitive when any two, three, or four of its subunits are mutated in this manner. It is peculiar to see such clean trends in mutational data, and we wondered whether our framework could explain this behavior.

It is worth mentioning that although research is more akin to a marathon than a sprint, requiring long and consistent effort, the brunt of this analysis was completed within *three days*! During that time, we successfully characterized the diverse set of ion channel mutants as a one-parameter family of curves, where only the free energy difference between the open and closed conformations was varied with each mutation (shown by the curves in Fig. 0.9B). More generally, we demonstrated that the full suite of data could be collapsed as a function of the free energy of ligand binding (Fig. 0.9C), as we had seen in the case of transcription factors.

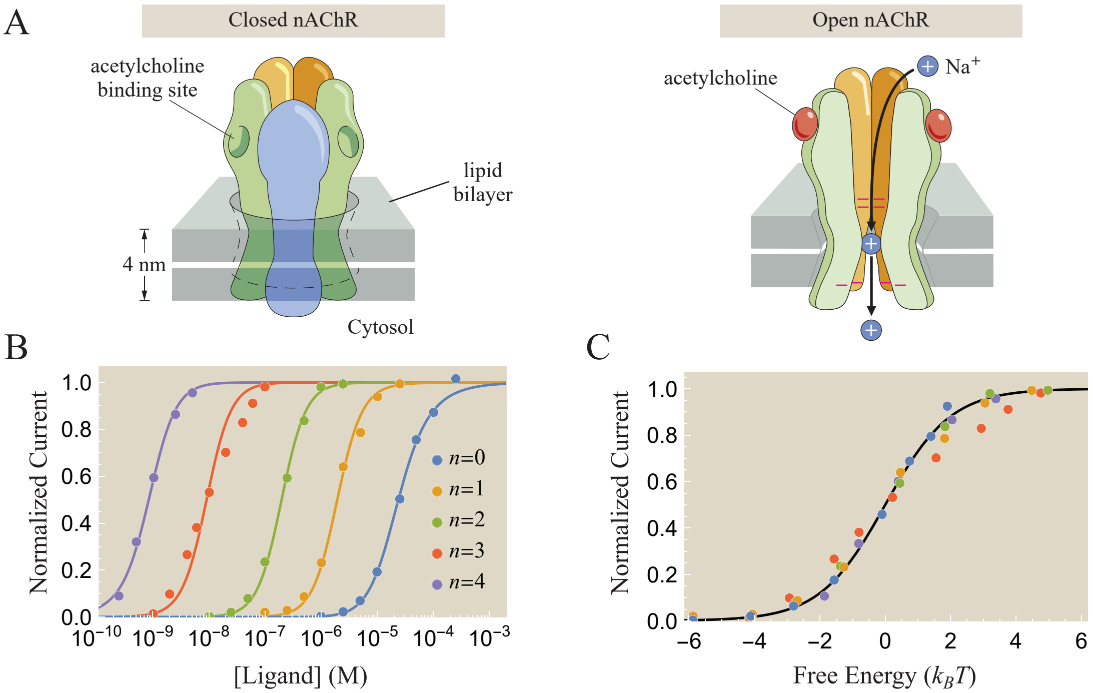


Figure 0.9: Schematic of the nAChR ion channel. (A) The nicotinic acetylcholine receptor (nAChR) is composed of five different subunits with two ligand binding sites for acetylcholine. The channel has a higher probability of being closed in the absence of ligand and open when bound to ligand. (B) The current through the ion channel with n mutated subunits (L251S). (C) The full suite of data can be collapsed in terms of the free energy between the open and closed states of the channel.

We also discovered that the normalization of the y-axis in Fig. 0.9B (which is stretched to run from 0 to 1) erased important information about the leakiness of each mutant. Indeed, when we back-calculated the unnormalized current, we discovered that the leakiness (the fraction of time that each channel is open in the absence of ligand) increases exponentially with the number of mutated subunits. Hence, while an ion channel with 0-3 mutations exhibits essentially no leakiness, a channel with 4 mutations was 5% leaky, and a channel with all 5 subunits mutated exhibits a leakiness greater than 50%. When Rob asked the paper's author, Caltech Professor Henry Lester, about this trend, Henry confirmed that channels with 4 mutations had some leakiness, and that cells with all five subunits mutated exhibited so much leakiness that they quickly died (which is why there is no $n = 5$ data). This example demonstrates how quantitative models can make penetrating insights into the underlying mechanisms and expected effects of mutations.

Finally, **Chapter 6** details my first brush with big data, where I examined whether statistical mechanical models can predict the behavior of tens of thousands of genetic constructs. More specifically, this project analyzed how different DNA sequences alter the binding affinity of RNA polymerase (RNAP), a molecular machine that performs the first step of transforming a gene into a protein. Nearly every cellular process relies on recruiting RNAP to a particular location on the genome, and understanding how RNAP interacts with its myriad target sequences lies at the heart of this critical task.

Fig. 0.10A shows the experimental setup, where over 10 000 promoters were constructed, representing all combinations of several sets of promoter elements (called Background (BG), UP, -35, Spacer, and -10). The gene expression of each construct was measured *in vivo*, an incredible feat that would have been nearly impossible as little as ten years ago. To analyze this wealth of data, the original authors from the Sri Kosuri lab utilized an energy matrix model which assumes that each promoter element contributes additively and independently to the RNAP binding energy. In their framework, gene expression is proportional to the Boltzmann factor

$$\text{Gene Expression} = r_{\text{max}} e^{-\beta(E_{\text{BG}}+E_{\text{UP}}+E_{\text{-35}}+E_{\text{Spacer}}+E_{\text{-10}})}. \tag{1}$$

The full suite of data was used to infer the free energy for each of the eight Backgrounds, three UPs, eight -35s, eight Spacers, and eight -10s, enabling them to predict the expression of each construct using 35 parameters (+1 parameter for the constant of proportionality r_{max}).

Fig. 0.10B shows that the resulting predictions match the experimental data poorly. Indeed, in their original paper, the authors showed this figure and suggested two ways to fix the model: (1) by introducing sixty-four interaction energies, one for each pair of -35 and -10 elements or (2) by training a neural network. Both models were able to characterize the data far better, yet we wondered whether a physically-motivated change that required fewer parameters could describe the data equally well.

I hypothesized that one of the key features missing from the existing model was the increase in binding affinity (also called avidity) arising from RNAP binding at multiple sites. More precisely, RNAP is known to bind most tightly to the -35 and -10 sites, and this two-site binding should enhance gene expression, as can be understood from both kinetic and thermodynamic standpoints. Kinetically speaking, if RNAP dissociates from one site, the likelihood of it quickly reassociating is increased because it is tethered in place by the other site. The thermodynamic advantage emerges because binding the first RNAP site results in losses of translational and

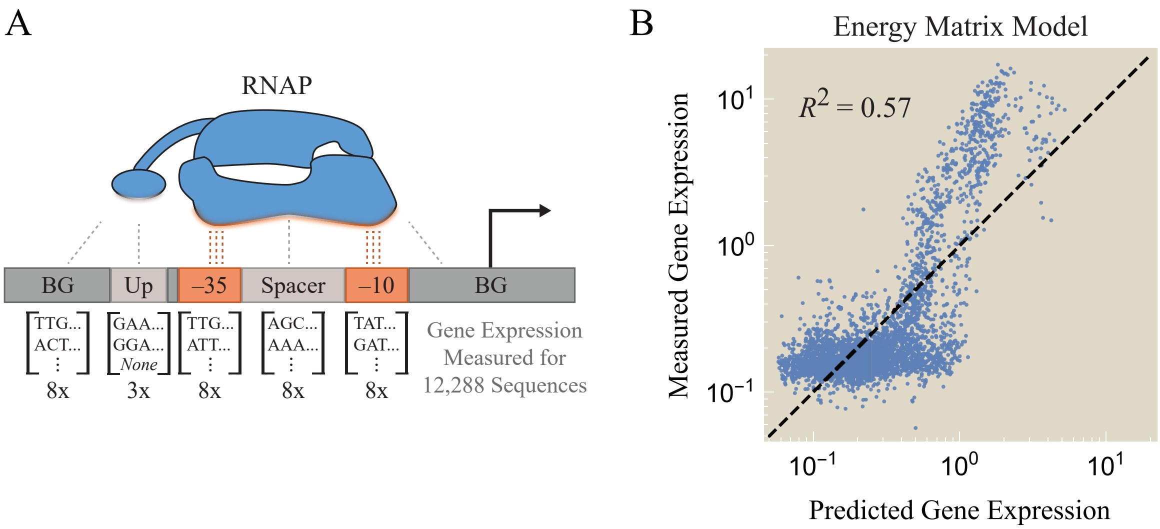


Figure 0.10: Gene expression from combinations of promoter elements. (A) A promoter is composed of Background (BG), UP, -35, Spacer, and -10 elements. Every combination of a set of these elements was created, and its gene expression was measured *in vivo*. (B) The energy matrix analysis in the original paper poorly characterized the data; a good characterization would lie along the diagonal line $y = x$ and have a large coefficient of determination R^2.

rotational degrees of freedom, but the subsequent binding of the second site incurs a smaller entropic cost, thereby boosting the likelihood of the bivalent state.

The simplest statistical mechanical model embodying this idea is shown in Fig. 0.11A, and it differs from the energy matrix model in three key ways: (1) there is a single interaction energy E_{int} (with the same value for all -35 and -10 elements) representing the boost in avidity in the bivalently bound state; (2) RNAP is assumed to express low levels of the gene when unbound or only bound at one site (a small r_0 corresponding to background noise or spurious transcription) but exhibits much higher expression when bivalently bound ($r_{\text{max}} \gg r_0$); and (3) the Boltzmann weight of each state must be normalized by the sum of all weights, namely,

$$\text{Gene Expression} = \frac{r_0 + e^{-\beta(E_{\text{BG}}+E_{\text{UP}}+E_{\text{Spacer}})}\left(r_0 e^{-\beta E_{\text{-35}}} + r_0 e^{-\beta E_{\text{-10}}} + r_{\text{max}} e^{-\beta(E_{\text{-35}}+E_{\text{-10}}+E_{\text{int}})}\right)}{1 + e^{-\beta(E_{\text{BG}}+E_{\text{UP}}+E_{\text{Spacer}})}\left(e^{-\beta E_{\text{-35}}} + e^{-\beta E_{\text{-10}}} + e^{-\beta(E_{\text{-35}}+E_{\text{-10}}+E_{\text{int}})}\right)}. \tag{2}$$

This model only requires two more parameters (r_0 and E_{int}) than the poor energy matrix model discussed above, yet it shows a marked improvement in its ability to characterize the 10 000 promoters (Fig. 0.11B). To test its predictive power, we inferred the parameter values on 25% of the data and used them to predict the gene expression of the remaining 75%, resulting in a significantly higher coefficient of

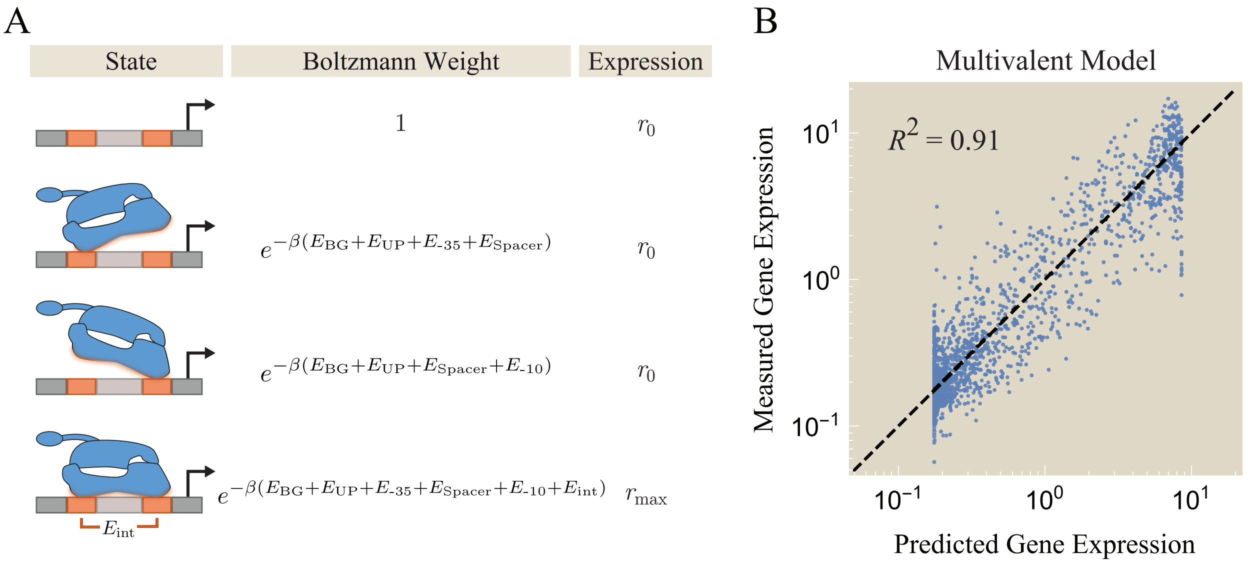


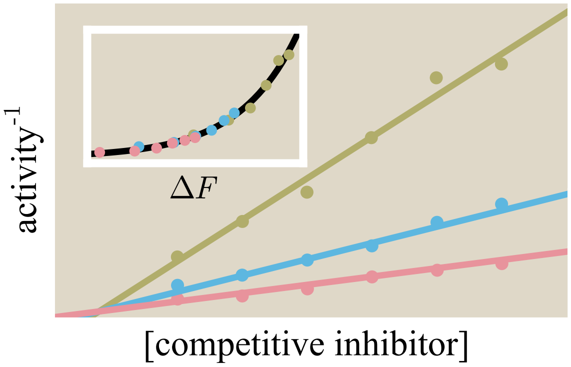
Figure 0.11: Modeling the avidity of the -35 and -10 sites. (A) Because RNAP binds most tightly to the -35 and -10 sites, we hypothesized that the avidity from this two-site binding is a critical component of gene expression models. The Boltzmann weight and level of gene expression for the four key states of the system are shown. (B) The resulting model is better able to characterize the data than the energy matrix model at the cost of only two additional, physically-motivated parameters.

determination $R^2 = 0.91$ compared to the energy matrix ($R^2 = 0.57$). To date, this project remains the broadest application of my modeling efforts in terms of the amount of data that was predicted and confirmed. It demonstrates that if we understand each part of a biological system, we can harness their combinatorial complexity to gain an explosion in our understanding.

To close, I invite you to contemplate the following question that my advisor Rob Phillips asked during each project: “When should we be satisfied with our work and declare victory?” My answer has been that theory must push our understanding of a system until we uncover a jaw-dropping conclusion. While this criterion obviously depends upon the spring constant of your particular jaw, it emphasizes my view that it is not enough to simply draw a curve that fits the data. A successful model may give intuition into a previously unexplained phenomenon, provide a mental map with which to contemplate new experiments, enable you to search the spectrum of possible behaviors and completely characterize the system, or significantly accelerate the pace of experimental work. The following chapters each aim to fulfill at least one of these criteria. Collectively, I hope these works demonstrate how the melding of theory and experiment creates a whole that is far greater than the sum of its parts.

Chapter 1

STATISTICAL MECHANICS OF ALLOSTERIC ENZYMES



Einav T, Mazutis L, Phillips R. Statistical Mechanics of Allosteric Enzymes. The Journal of Physical Chemistry B. 2016;120(26):6021-6037. doi:10.1021/acs.jpcb.6b01911

Every journey begins with a single step, and this paper marks my first milestone in science not merely because it was my first paper (when I learned how to make figures, write scientific prose, and interact with Rob), but because towards the end of the project I got stuck. We had developed a theory which predicted that an allosteric enzyme with certain properties can exhibit a peculiar phenomenon called substrate inhibition. Naturally, Rob asked me to search the literature for such an instance, and I spent a week crawling through biochemical databases and manuscripts. Having painstakingly searched through 30 papers, understanding very little and without finding our desired example, I returned to Rob and informed him that this search may be impossible for me. He suggested that we email his enzymology friends and capitalize on their decades of experience. We sent out emails, and the next day a response came from Jean-Pierre Changeux telling us about one of his paper where he specified that acetylcholinesterase had the exact properties we needed and exhibited substrate inhibition. I was ecstatic! I had not only learned the key piece of information needed to finish the paper, but I also tangibly felt that science was not (and should not) be an independent adventure, but rather that we work as part of a community where members support each other. It was a valuable lesson to learn early on.

1.1 Abstract

The concept of allostery in which macromolecules switch between two different conformations is a central theme in biological processes ranging from gene regulation to cell signaling to enzymology. Allosteric enzymes pervade metabolic processes, yet a simple and unified treatment of the effects of allostery in enzymes has been lacking. In this work, we take the first step towards this goal by modeling allosteric enzymes and their interaction with two key molecular players – allosteric regulators and competitive inhibitors. We then apply this model to characterize existing data on enzyme activity, comment on how enzyme parameters (such as substrate binding affinity) can be experimentally tuned, and make novel predictions on how to control phenomena such as substrate inhibition.

1.2 Introduction

All but the simplest of cellular reactions are catalyzed by enzymes, macromolecules that can increase the rates of reactions by many orders of magnitude. In some cases, such as phosphoryl transfer reactions, rate enhancements can be as large as 10^{20}-fold or more [1]. A deeper understanding of how enzymes work can provide insights into biological phenomena as diverse as metabolic regulation or the treatment of disease [2–4]. The basic principles of enzyme mechanics were first proposed by Michaelis and Menten [5] and later extended by others [6–8]. While the earliest models considered enzymes as single-state catalysts, experiments soon revealed that some enzymes exhibit richer dynamics [9, 10]. The concept of allosteric enzymes was introduced by Monod-Wyman-Changeux (MWC) and independently by Pardee and Gerhart [7, 11–13], providing a much broader framework for explaining the full diversity of enzyme behavior. Since then, the MWC concept in which macromolecules are thought of as having both an inactive and active state has spread into many fields, proving to be a powerful conceptual tool capable of explaining many biological phenomena [14–16].

Enzymology is a well studied field, and much has been learned both theoretically and experimentally about how enzymes operate [17–20]. With the vast number of distinct molecular players involved in enzymatic reactions (for example: mixed, competitive, uncompetitive, and non-competitive inhibitors as well as cofactors, allosteric effectors, and substrate molecules), it is not surprising that new discoveries continue to emerge about the subtleties of enzyme action [9, 21, 22]. In this paper, we use the MWC model to form a unifying framework capable of describing the broad array of behaviors available to allosteric enzymes.

Statistical mechanics is a field of physics that describes the collective behavior of large numbers of molecules. Historically developed to understand the motion of gases, statistical physics has now seen applications in many areas of biology and has provided unexpected connections between distinct problems such as how transcription factors are induced by signals from the environment, the function of the molecular machinery responsible for detecting small gradients in chemoattractants, the gating properties of ligand-gated ion channels, and even the accessibility of genomic DNA in eukaryotes which is packed into nucleosomes [23–29]. One of us (RP) owes his introduction to the many beautiful uses of statistical mechanics in biology to Bill Gelbart to whom this special issue is dedicated. During his inspiring career, Gelbart has been a passionate and creative developer of insights into a wide number of problems using the tools of statistical mechanics and we hope that our examples on the statistical mechanics of allosteric enzymes will please him.

The remainder of the paper is organized as follows. In section 1.3.1, we show how the theoretical treatment of the traditional Michaelis-Menten enzyme, an inherently non-equilibrium system, can be stated in a language remarkably similar to equilibrium statistical mechanics. This sets the stage for the remainder of the paper by introducing key notation and the states and weights formalism that serves as the basis for analyzing more sophisticated molecular scenarios. In section 1.3.2, we discuss how the states and weights formalism can be used to work out the rates for the simplest MWC enzyme, an allosteric enzyme with a single substrate binding site. This is followed by a discussion of how allosteric enzymes are modified by the binding of ligands, first an allosteric regulator in section 1.3.3 and then a competitive inhibitor in section 1.3.4. We next generalize to the much richer case of enzymes with multiple substrate binding sites in section 1.3.5. Lastly, we discuss how to combine the individual building blocks of allostery, allosteric effectors, competitive inhibitors, and multiple binding sites to analyze general enzymes in section 1.3.6. Having built up this framework, we then apply our model to understand observed enzyme behavior. In section 1.4.1, we show how disparate enzyme activity curves can be unified within our model and collapsed onto a single curve. We close by examining the exotic phenomenon of substrate inhibition in section 1.4.2 and show how the allosteric nature of some enzymes may be the key to understanding and controlling this phenomenon.

1.3 Models

1.3.1 Michaelis-Menten Enzyme

We begin by briefly introducing the textbook Michaelis-Menten treatment of enzymes [18]. This will serve both to introduce basic notation and to explain the states and weights methodology which we will use throughout the paper.

Many enzyme-catalyzed biochemical reactions are characterized by Michaelis-Menten kinetics. Such enzymes comprise a simple but important class where we can study the relationship between the traditional chemical kinetics based on reaction rates with a physical view dictated by statistical mechanics. According to the Michaelis-Menten model, enzymes are single-state catalysts that bind a substrate and promote its conversion into a product. Although this scheme precludes allosteric interactions, a significant fraction of non-regulatory enzymes (e.g. triosephosphate isomerase, bisphosphoglycerate mutase, adenylate cyclase) are well-described by Michaelis-Menten kinetics [18].

The key player in this reaction framework is a monomeric enzyme E that binds a substrate S at the substrate binding site (also called the active site or catalytic site), forming an enzyme-substrate complex ES. The enzyme then converts the substrate into product P which is subsequently removed from the system and cannot return to its original state as substrate. In terms of concentrations, this reaction can be written as

$$[E]+[S] \underset{koff}{\overset{kon}{\rightleftarrows}} [ES] \xrightarrow{kcat} [E]+[P] \tag{1.1}$$

where the rate of product formation equals

$$\frac{d[P]}{dt} = [ES]k_{cat}. \tag{1.2}$$

Briggs and Haldane assumed a time scale separation where the substrate and product concentrations ($[S]$ and $[P]$) slowly change over time while the free and bound enzyme states ($[E]$ and $[ES]$) changed much more rapidly [6]. This allows us to approximate this system over short time scales by assuming that the slow components (in this case $[S]$) remain constant and can therefore be absorbed into the k_{on} rate [30],

$$[E] \underset{koff + kcat}{\overset{kon[S]}{\rightleftarrows}} [ES]. \tag{1.3}$$

Assuming that the system 1.3 reaches steady-state (over the short time scale of this approximation) quickly enough that the substrate concentration does not appreciably

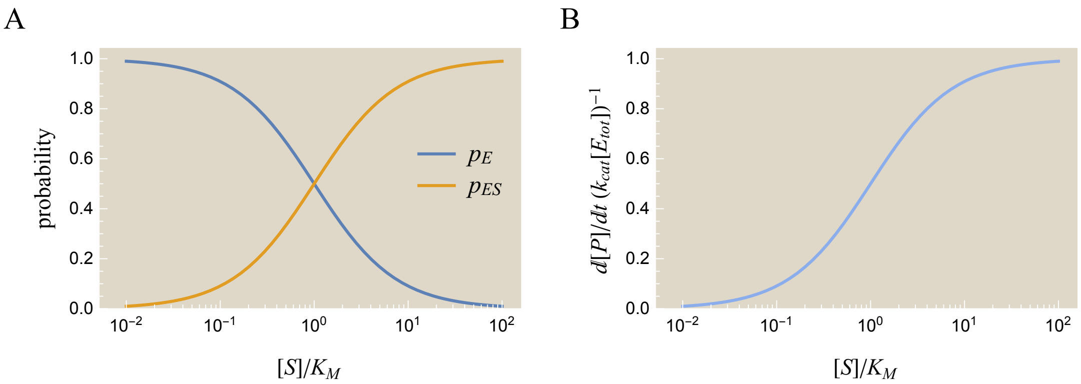


Figure 1.1: Dynamics of the Michaelis-Menten enzyme. (A) Probabilities of the free enzyme p_E and bound enzyme p_{ES} states as a function of substrate concentration. As the amount of substrate $[S]$ increases, more enzyme is found in the bound state rather than the free state. (B) The rate of product formation for a non-allosteric enzyme. The rate of product formation has the same functional form as the probability p_{ES} of the enzyme-substrate complex, as illustrated by Eqs. 1.2 and 1.7.

diminish, this implies

$$[E][S]k_{on} = [ES]\left(k_{off} + k_{cat}\right), \tag{1.4}$$

which we can rewrite as

$$\frac{[ES]}{[E]} = \frac{[S]k_{on}}{k_{off} + k_{cat}} \equiv \frac{[S]}{K_M} \tag{1.5}$$

where $K_M = \frac{k_{off}+k_{cat}}{k_{on}}$ is called the *Michaelis constant*. K_M incorporates the binding and unbinding of ligand as well as the conversion of substrate into product; in the limit $k_{cat} = 0$, K_M reduces to the familiar dissociation constant $K_D = \frac{k_{off}}{k_{on}}$. Using Eq. 1.5 and the fact that the total enzyme concentration is conserved, $[E] + [ES] = [E_{tot}]$, we can solve for $[E]$ and $[ES]$ separately as

$$[E] = [E_{tot}]\frac{1}{1 + \frac{[S]}{K_M}} \equiv [E_{tot}]p_E \tag{1.6}$$

$$[ES] = [E_{tot}]\frac{\frac{[S]}{K_M}}{1 + \frac{[S]}{K_M}} \equiv [E_{tot}]p_{ES}, \tag{1.7}$$

where $p_E = \frac{[E]}{[E_{tot}]}$ and $p_{ES} = \frac{[ES]}{[E_{tot}]}$ are the probabilities of finding an enzyme in the unbound and bound form, respectively. Substituting the concentration of bound enzymes $[ES]$ from Eq. 1.7 into the rate of product formation Eq. 1.2,

$$\frac{d[P]}{dt} = k_{cat}[E_{tot}]\frac{\frac{[S]}{K_M}}{1 + \frac{[S]}{K_M}}. \tag{1.8}$$

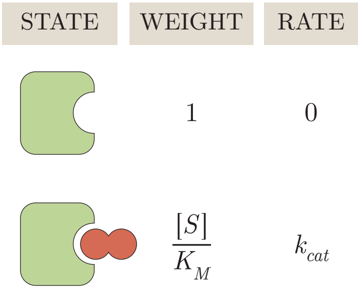


Figure 1.2: States and weights for the Michaelis-Menten enzyme. Each enzyme conformation is shown together with its weight and its catalytic rate. The probability of finding an enzyme (green) in either the free or bound state equals the weight of that state divided by the sum of all weights ($1 + \frac{[S]}{K_M}$) where $[S]$ is the concentration of substrate (dark red) and $K_M = \frac{k_{off}+k_{cat}}{k_{on}}$ is the Michaelis constant. At $[S] = K_M$, half of the enzyme population exists in the free form and half exists in the bound form. For $[S] > K_M$, more than half of all enzymes will be bound to substrate.

Fig. 1.1 shows the probability of free and bound enzyme as well as the rate of product formation. The two parameters k_{cat} and $[E_{tot}]$ scale $\frac{d[P]}{dt}$ vertically (if k_{cat} is increased by a factor of 10, the y-axis values in Fig. 1.1B will be multiplied by that same factor of 10), while K_M effectively rescales the substrate concentration $[S]$. Increasing K_M by a factor of 10 implies that 10 times as much substrate is needed to obtain the same rate of product formation; on the semi-log plots in Fig. 1.1, this corresponds to shifting all curves to the right by one power of 10.

We can visualize the microscopic states of the enzyme using a modified states and weights diagram shown in Fig. 1.2 [31]. The *weight* of each enzyme state is proportional to the probability of its corresponding state ($w_E \propto p_E$, $w_{ES} \propto p_{ES}$) – the constant of proportionality is arbitrary but must be the same for all weights. For example, from Eqs. 1.6 and 1.7 we can multiply the probability that the enzyme will be unbound (p_E) or bound to substrate (p_{ES}) by $1 + \frac{[S]}{K_M}$ which yields the weights

$$w_E = 1 \tag{1.9}$$

$$w_{ES} = \frac{[S]}{K_M}. \tag{1.10}$$

Given the weights of an enzyme state, we can proceed in the reverse direction and obtain the probability for each enzyme state using

$$p_E = \frac{w_E}{Z_{tot}} = \frac{1}{1 + \frac{[S]}{K_M}} \tag{1.11}$$

$$p_{ES} = \frac{w_{ES}}{Z_{tot}} = \frac{\frac{[S]}{K_M}}{1 + \frac{[S]}{K_M}} \tag{1.12}$$

where

$$Z_{tot} = w_E + w_{ES} \tag{1.13}$$

is the sum of all weights. Dividing by Z_{tot} ensures the total probability of all enzyme states equals unity, $p_E + p_{ES} = 1$. The rate of product formation Eq. 1.8 is given by the product of the enzyme concentration $[E_{tot}]$ times the average catalytic rate over all states, weighed by each state's (normalized) weights. In the following sections, we will find this trick of writing states and weights very useful for modeling other molecular players.

The weights in Fig. 1.2 allow us to easily understand Fig. 1.1A: when $[S] < K_M$, $w_E > w_{ES}$ so that an enzyme is more likely to be in the substrate-free state; when $[S] > K_M$, $w_E < w_{ES}$ and an enzyme is more likely to be found as an enzyme-substrate complex. Increasing K_M shifts the tipping point of how much substrate is needed before the bound ES enzyme state begins to dominate over the free E state.

It should be noted that the formal notion of states and weights employed in physics applies only to equilibrium systems. For example, a ligand binding to a receptor in equilibrium will yield states and weights similar to Fig. 1.2 but with the Michaelis constant K_M replaced by the dissociation constant K_D [32]. Yet the ligand-receptor states and weights can also be derived from the Boltzmann distribution (where the weight of any state j with energy E_j is proportional to $e^{-\beta E_j}$) while the enzyme states and weights cannot be derived from the Boltzmann distribution (because the enzyme system is not in equilibrium). Instead, the non-equilibrium kinetics of the system are described by the modified states and weights in Fig. 1.2, where the K_D for substrate must be replaced with K_M. These modified states and weights serve as a mathematical trick that compactly and correctly represents the behavior of the enzyme, enabling us to apply the well established tools and intuition of equilibrium statistical mechanics when analyzing the inherently non-equilibrium problem of enzyme kinetics. In the next several sections, we will show how to generalize this method of states and weights to MWC enzymes with competitive inhibitors, allosteric regulators, and multiple substrate binding sites.

1.3.2 MWC Enzyme

Many enzymes are not static entities, but dynamic macromolecules that constantly fluctuate between different conformational states. This notion was initially conceived by Monod-Wyman-Changeux (MWC) to characterize complex multi-subunit proteins such as hemoglobin and aspartate transcarbamoylase (ATCase) [7, 11, 12]. The authors suggested that the ATC enzyme exists in two supramolecular states: a relaxed "R" state, which has high-affinity for substrate and a tight "T" state, which has low-affinity for substrate. Although in the case of ATCase, the transition between the T and R states is induced by an external ligand, recent experimental advances have shown that many proteins intrinsically fluctuate between these different states even in the absence of ligand [33–35]. These observations imply that the MWC model can be applied to a wide range of enzymes beyond those with multi-subunit complexes.

We will designate an enzyme with two possible states (an $\underline{\text{A}}$ctive state E_A and an $\underline{\text{I}}$nactive state E_I) as an MWC enzyme. The kinetics of a general MWC enzyme are given by

$$\begin{tikzcd}
{[E_A]+[S]} \arrow[r, shift left, "k_{on}^{A}"] \arrow[d, shift right, "k_{trans}^{A}"'] & {[E_AS]} \arrow[l, shift left, "k_{off}^{A}"] \arrow[r, "k_{cat}^{A}"] \arrow[d, shift right, "k_{trans}^{AS}"'] & {[E_A]+[P]} \\
{[E_I]+[S]} \arrow[r, shift left, "k_{on}^{I}"] \arrow[u, shift right, "k_{trans}^{I}"'] & {[E_IS]} \arrow[l, shift left, "k_{off}^{I}"] \arrow[r, "k_{cat}^{I}"] \arrow[u, shift right, "k_{trans}^{IS}"'] & {[E_I]+[P],}
\end{tikzcd} \tag{1.14}$$

which relates the active and inactive enzyme concentrations ($[E_A]$, $[E_I]$) to the active and inactive enzyme-substrate complexes ($[E_AS]$, $[E_IS]$). In this two-state MWC model, similar to that explored by Howlett et al.[36], the rate of product formation is given by

$$\frac{d[P]}{dt} = k_{cat}^{A}[E_AS] + k_{cat}^{I}[E_IS]. \tag{1.15}$$

The active state will have a faster catalytic rate (often much faster) than the inactive state, $k_{cat}^{A} > k_{cat}^{I}$.

As in the case of a Michaelis-Menten enzyme, we will assume that all four forms of the enzyme (E_A, E_I, E_AS, and E_IS) quickly reach steady state on time scales so short that the substrate concentration $[S]$ remains nearly constant. Therefore, we can incorporate the slowly-changing quantities $[S]$ and $[P]$ into the rates, a step dubbed the *quasi-steady-state approximation* [30]. This allows us to rewrite the scheme

1.14 in the following form,

$$\begin{tikzcd}
{[E_A]} \arrow[r, shift left, "k_{on}^A{[S]}"] \arrow[d, shift right, "k_{trans}^A"'] & {[E_AS]} \arrow[l, shift left, "k_{off}^A+k_{cat}^A"] \arrow[d, shift right, "k_{trans}^{AS}"'] \\
{[E_I]} \arrow[r, shift left, "k_{on}^I{[S]}"] \arrow[u, shift right, "k_{trans}^I"'] & {[E_IS].} \arrow[l, shift left, "k_{off}^I+k_{cat}^I"] \arrow[u, shift right, "k_{trans}^{IS}"']
\end{tikzcd} \tag{1.16}$$

Assuming the quasi-steady-state approximation holds, the four enzyme states will rapidly attain steady-state values

$$\frac{d[E_AS]}{dt} = \frac{d[E_A]}{dt} = \frac{d[E_IS]}{dt} = \frac{d[E_I]}{dt} = 0. \tag{1.17}$$

In addition, a separate constraint on the system that is necessary and sufficient to apply the method of states and weights is given by the *cycle condition*: the product of rates going clockwise around any cycle must equal the product of rates going counterclockwise [30]. It should be noted that to violate the cycle condition, a system must continuously pay energy since at least one step in any cycle must be energetically unfavorable. We shall proceed with the assumption that there are no such cycles in our system. For the MWC enzyme 1.16, this implies

$$\left(k_{on}^A[S]\right) k_{trans}^{AS} \left(k_{off}^I + k_{cat}^I\right) k_{trans}^I = \left(k_{off}^A + k_{cat}^A\right) k_{trans}^A \left(k_{on}^I[S]\right) k_{trans}^{IS} \tag{1.18}$$

or equivalently

$$\underbrace{\frac{k_{on}^A[S]}{k_{off}^A + k_{cat}^A}}_{\frac{[E_AS]}{[E_A]}} \underbrace{\frac{k_{trans}^I}{k_{trans}^A}}_{\frac{[E_A]}{[E_I]}} = \underbrace{\frac{k_{on}^I[S]}{k_{off}^I + k_{cat}^I}}_{\frac{[E_IS]}{[E_I]}} \underbrace{\frac{k_{trans}^{IS}}{k_{trans}^{AS}}}_{\frac{[E_AS]}{[E_IS]}}. \tag{1.19}$$

The validity of both the quasi-steady-state approximation 1.17 and the cycle condition 1.19 will be analyzed in Appendix S1.1. Assuming both statements hold, we can invoke *detailed balance* – the ratio of concentrations between two enzyme states equals the inverse of the ratio of rates connecting these two states. For example, between the active states $[E_AS]$ and $[E_A]$ in 1.16,

$$\frac{[E_AS]}{[E_A]} = \frac{k_{on}^A[S]}{k_{off}^A + k_{cat}^A} \equiv \frac{[S]}{K_M^A} \tag{1.20}$$

where we have defined the Michaelis constant for the active state, K_M^A. Similarly, we can write the equation for detailed balance between the inactive states $[E_IS]$ and $[E_I]$ as

$$\frac{[E_IS]}{[E_I]} = \frac{k_{on}^I[S]}{k_{off}^I + k_{cat}^I} \equiv \frac{[S]}{K_M^I}. \tag{1.21}$$

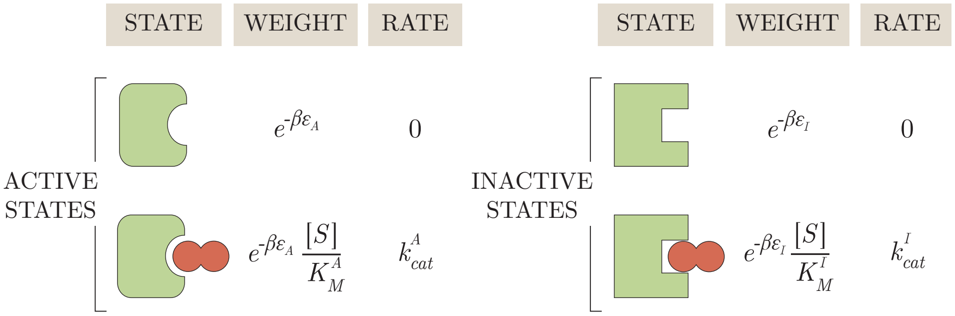

STATE
WEIGHT
RATE
STATE
WEIGHT
RATE
ACTIVE STATES
$e^{-\beta\varepsilon_A}$
0
$e^{-\beta\varepsilon_A}\frac{[S]}{K_M^A}$
k_{cat}^A
INACTIVE STATES
$e^{-\beta\varepsilon_I}$
0
$e^{-\beta\varepsilon_I}\frac{[S]}{K_M^I}$
k_{cat}^I

Figure 1.3: States and weights for an MWC enzyme. The energies ϵ_A and ϵ_I provide the free energy scale for the substrate-free conformations, dictating their relative probabilities. Decreasing the energy ϵ_A of the active state would increase the probability of all the active enzyme conformations relative to the inactive conformations. K_M^A denotes the substrate concentration at which half of the active enzymes are bound and half the active enzymes are unbound, as indicated by the crossing of the (p_{E_A}, blue) and (p_{E_AS}, gold) curves at $[S] = K_M^A$ in Fig. 1.4. K_M^I serves an analogous role for the inactive states.

An enzyme may have a different affinity for substrate or a different catalytic rate in the active and inactive forms. Typical measured values of K_M fall into the range $10^{-7} - 10^{-1}$ M [37]. Whether K_M^A or K_M^I is larger depends on the specific enzyme.

As a final link between the language of chemical rates and physical energies, we can recast detailed balance between $[E_A]$ and $[E_I]$ as

$$\frac{[E_A]}{[E_I]} = \frac{k_{trans}^I}{k_{trans}^A} \equiv e^{-\beta(\epsilon_A - \epsilon_I)}, \tag{1.22}$$

where ϵ_A and ϵ_I are the free energies of the enzyme in the active and inactive state, respectively, and $\beta = \frac{1}{k_B T}$ where k_B is Boltzmann's constant and T is the temperature of the system. Whether the active state energy is greater than or less than the inactive state energy depends on the enzyme. For example, $\epsilon_I < \epsilon_A$ in ATCase whereas the opposite holds true, $\epsilon_A < \epsilon_I$, in chemoreceptors [9, 32].

Using Eqs. 1.20-1.22, we can recast the cycle condition 1.19 (as shown in the underbraces) into a simple relationship between the steady-state enzyme concentrations. Additionally, we can use these equations to define the weights of each enzyme state in Fig. 1.3. Following section 1.3.1, the probability of each state equals its weight divided by the sum of all weights,

$$p_{E_A} = e^{-\beta \epsilon_A} \frac{1}{Z_{tot}} \tag{1.23}$$

$$p_{E_A S} = e^{-\beta \epsilon_A} \frac{\frac{[S]}{K_M^A}}{Z_{tot}} \tag{1.24}$$

$$p_{E_I} = e^{-\beta \epsilon_I} \frac{1}{Z_{tot}} \tag{1.25}$$

$$p_{E_I S} = e^{-\beta \epsilon_I} \frac{\frac{[S]}{K_M^I}}{Z_{tot}}, \tag{1.26}$$

where

$$Z_{tot} = e^{-\beta \epsilon_A} \left(1 + \frac{[S]}{K_M^A} \right) + e^{-\beta \epsilon_I} \left(1 + \frac{[S]}{K_M^I} \right). \tag{1.27}$$

Note that multiplying all of the weights by a constant c will also multiply Z_{tot} by c, so that the probability of any state will remain unchanged. That is why in Fig. 1.2 we could neglect the $e^{-\beta \epsilon}$ factor that was implicitly present in each weight.

The total amount of enzyme is conserved among the four enzyme states, $[E_{tot}] = [E_A] + [E_A S] + [E_I] + [E_I S]$. Using this fact together with Eqs. 1.20-1.22 enables us to solve for the concentrations of both types of bound enzymes, namely,

$$[E_A S] = [E_{tot}] \frac{e^{-\beta \epsilon_A} \frac{[S]}{K_M^A}}{e^{-\beta \epsilon_A} \left(1 + \frac{[S]}{K_M^A} \right) + e^{-\beta \epsilon_I} \left(1 + \frac{[S]}{K_M^I} \right)} = [E_{tot}] p_{E_A S} \tag{1.28}$$

$$[E_I S] = [E_{tot}] \frac{e^{-\beta \epsilon_I} \frac{[S]}{K_M^I}}{e^{-\beta \epsilon_A} \left(1 + \frac{[S]}{K_M^A} \right) + e^{-\beta \epsilon_I} \left(1 + \frac{[S]}{K_M^I} \right)} = [E_{tot}] p_{E_I S}. \tag{1.29}$$

Substituting these relations into 1.15 yields the rate of product formation,

$$\frac{d[P]}{dt} = [E_{tot}] \frac{k_{cat}^A e^{-\beta \epsilon_A} \frac{[S]}{K_M^A} + k_{cat}^I e^{-\beta \epsilon_I} \frac{[S]}{K_M^I}}{e^{-\beta \epsilon_A} \left(1 + \frac{[S]}{K_M^A} \right) + e^{-\beta \epsilon_I} \left(1 + \frac{[S]}{K_M^I} \right)}. \tag{1.30}$$

The probabilities 1.23-1.26 of the different states and the rate of product formation 1.30 are shown in Fig. 1.4. Although we use the same parameters from Fig. 1.1 for the active state, the p_{E_A} and $p_{E_A S}$ curves in Fig. 1.4A look markedly different from the p_E and p_{ES} Michaelis-Menten curves in Fig. 1.1A. This indicates that the activity of an MWC enzyme *does not* equal the activity of two independent Michaelis-Menten enzymes, one with the MWC enzyme's active state parameters and the other with the MWC enzyme's inactive state parameters. The interplay of the active and inactive states makes an MWC enzyme inherently more complex than a Michaelis-Menten enzyme.

When $[S] = 0$ the enzyme only exists in the unbound states E_A and E_I whose relative probabilities are given by $\frac{p_{E_A}}{p_{E_I}} = e^{-\beta(\epsilon_A - \epsilon_I)}$. When $[S] \to \infty$, the enzyme spends all of its time in the bound states $E_A S$ and $E_I S$ which have relative probabilities $\frac{p_{E_A S}}{p_{E_I S}} = e^{-\beta(\epsilon_A - \epsilon_I)} \frac{K_M^I}{K_M^A}$. The curves for the active states (for free enzyme p_{E_A} and bound enzyme $p_{E_A S}$) intersect at $[S] = K_M^A$ while the curves of the two inactive states intersect at $[S] = K_M^I$. For the particular parameters shown, even though the unbound *inactive* state (green) dominates at low substrate concentrations, the *active* state (gold) has the largest statistical weights as the concentration of substrate increases. Thus, adding substrate causes the enzyme to increasingly favor the active state.

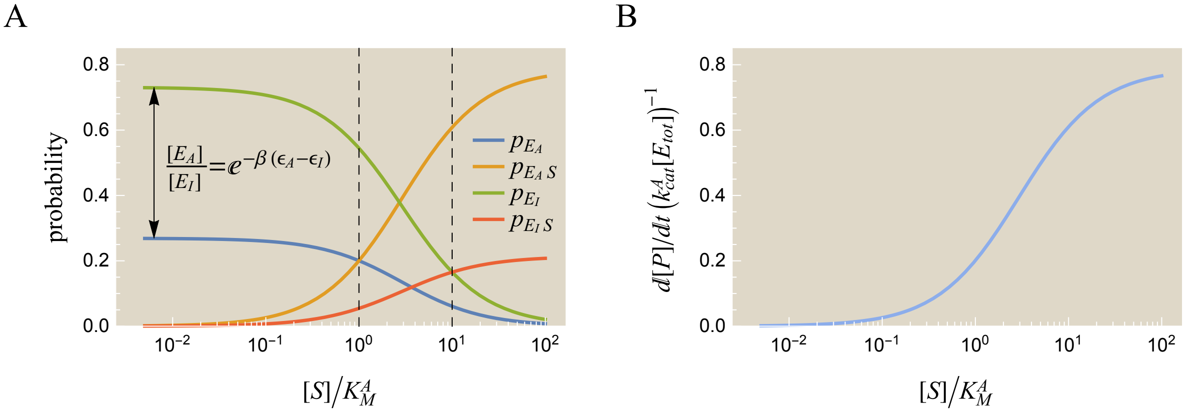


Figure 1.4: Quantitative description of an MWC enzyme. (A) Probabilities of each enzyme state. While the active state has the same catalytic rate k_{cat}^A and Michaelis constant K_M^A as the Michaelis-Menten enzyme in Fig. 1.1A, the inactive state significantly alters the forms of p_{E_A} and $p_{E_A S}$. The dashed vertical lines indicate where the substrate concentration equals K_M^A and K_M^I, respectively. (B) The rate of product formation, $\frac{d[P]}{dt}$. Assuming $\frac{k_{cat}^A}{k_{cat}^I} \gg 1$, $\frac{d[P]}{dt}$ (blue curve in (B)) is dominated by the active enzyme-substrate complex, $p_{E_A S}$ (gold curve in (A)). Parameters were chosen to reflect "typical" enzyme kinetics values: $\frac{k_{cat}^A}{k_{cat}^I} = 10^2$, $\frac{K_M^A}{K_M^I} = 10^{-1}$, and $e^{-\beta(\epsilon_A - \epsilon_I)} = e^{-1}$. Substrate concentrations are shown normalized relative to the active state parameter $\frac{[S]}{K_M^A}$, although the inactive state parameter $\frac{[S]}{K_M^I}$ could also have been used.

Using this framework, we can compute properties of the enzyme kinetics curve shown in Fig. 1.4(B). One important property is the dynamic range of an enzyme, the difference between the maximum and minimum rate of product formation. In the absence of substrate ($[S] \to 0$) and a saturating concentration of substrate ($[S] \to \infty$), the rate of product formation Eq. 1.30 becomes

$$\lim_{[S]\to 0} \frac{d[P]}{dt} = 0 \tag{1.31}$$

$$\lim_{[S]\to\infty} \frac{d[P]}{dt} = [E_{tot}] \frac{k_{cat}^A \frac{e^{-\beta\epsilon_A}}{K_M^A} + k_{cat}^I \frac{e^{-\beta\epsilon_I}}{K_M^I}}{\frac{e^{-\beta\epsilon_A}}{K_M^A} + \frac{e^{-\beta\epsilon_I}}{K_M^I}}. \tag{1.32}$$

From these two expressions, we can write the dynamic range as

$$\begin{aligned} \text{dynamic range} &= \left(\lim_{[S]\to\infty} \frac{d[P]}{dt} \right) - \left(\lim_{[S]\to 0} \frac{d[P]}{dt} \right) \\ &= [E_{tot}] k_{cat}^A \left(1 - \frac{1 - \frac{k_{cat}^I}{k_{cat}^A}}{1 + e^{-\beta(\epsilon_A - \epsilon_I)} \frac{K_M^I}{K_M^A}} \right) \end{aligned} \tag{1.33}$$

where every term in the fraction has been written as a ratio of the active and inactive state parameters. We find that the dynamic range increases as $\frac{k_{cat}^I}{k_{cat}^A}$, $e^{-\beta(\epsilon_A - \epsilon_I)}$, and $\frac{K_M^I}{K_M^A}$ increase (assuming $k_{cat}^A > k_{cat}^I$).

Another important property is the concentration of substrate at which the rate of product formation lies halfway between its minimum and maximum value, which we will denote as $[S_{50}]$. It is straightforward to show that the definition

$$\lim_{[S]\to[S_{50}]} \frac{d[P]}{dt} = \frac{1}{2} \left(\lim_{[S]\to\infty} \frac{d[P]}{dt} + \lim_{[S]\to 0} \frac{d[P]}{dt} \right) \tag{1.34}$$

is satisfied when

$$[S_{50}] = K_M^A \frac{e^{-\beta(\epsilon_A - \epsilon_I)} + 1}{e^{-\beta(\epsilon_A - \epsilon_I)} + \frac{K_M^A}{K_M^I}}. \tag{1.35}$$

With increasing $e^{-\beta(\epsilon_A - \epsilon_I)}$, the value of $[S_{50}]$ increases if $K_M^A > K_M^I$ and decreases otherwise. $[S_{50}]$ always decreases as $\frac{K_M^A}{K_M^I}$ increases. Lastly, we note that in the limit of a Michaelis-Menten enzyme, $\epsilon_I \to \infty$, we recoup the familiar results

$$\text{dynamic range} = [E_{tot}] k_{cat}^A \qquad (\epsilon_I \to \infty) \tag{1.36}$$

$$[S_{50}] = K_M^A \qquad (\epsilon_I \to \infty). \tag{1.37}$$

1.3.3 Allosteric Regulator

The catalytic activity of many enzymes is controlled by molecules that bind to regulatory sites which are often different from the active sites themselves. As a result of ligand-induced conformational changes, these molecules alter the substrate binding site which modifies the rate of product formation, $\frac{d[P]}{dt}$. Allosterically controlled enzymes represent important regulatory nodes in metabolic pathways and are often responsible for keeping cells in homeostasis. Some well-studied examples of allosteric control include glycogen phosphorylase, phosphofructokinase, glutamine synthetase, and aspartate transcarbamoylase (ATCase). In many cases the data from these systems are characterized phenomenologically using Hill functions, but the Hill coefficients thus obtained can be difficult to interpret [39]. In addition, Hill coefficients do not provide much information about the organization or regulation of an enzyme, nor do they reflect the relative probabilities of the possible enzyme conformations, although recent results have begun to address these issues [40]. In this section, we add one more layer of complexity to our statistical mechanics framework by introducing an allosteric regulator.

Consider an MWC enzyme with one site for an allosteric regulator R and a different site for a substrate molecule S that will be converted into product. We can define the effects of the allosteric regulator directly through the states and weights. As shown in Fig. 1.5, the regulator R contributes a factor $\frac{[R]}{R_D^A}$ when it binds to an active state and a factor $\frac{[R]}{R_D^I}$ when it binds to an inactive state where R_D^A and R_D^I are the dissociation constants between the regulator and the active and inactive states of the enzyme, respectively. Unlike the *Michaelis* constants K_M^A and K_M^I for the substrate, the *dissociation* constants R_D^A and R_D^I enter the states and weights because the regulator can only bind and unbind to the enzyme (and cannot be transformed into product). In other words, if we were to draw a rates diagram for this enzyme system, detailed balance between the two states where the regulator is bound and unbound would yield a dissociation constant ($\frac{k_{off}}{k_{on}}$) rather than a Michaelis constant ($\frac{k_{off}+k_{cat}}{k_{on}}$).

Using the states and weights in Fig. 1.5, we can compute the probability of each enzyme state. For example, the probabilities of the four states that form product are

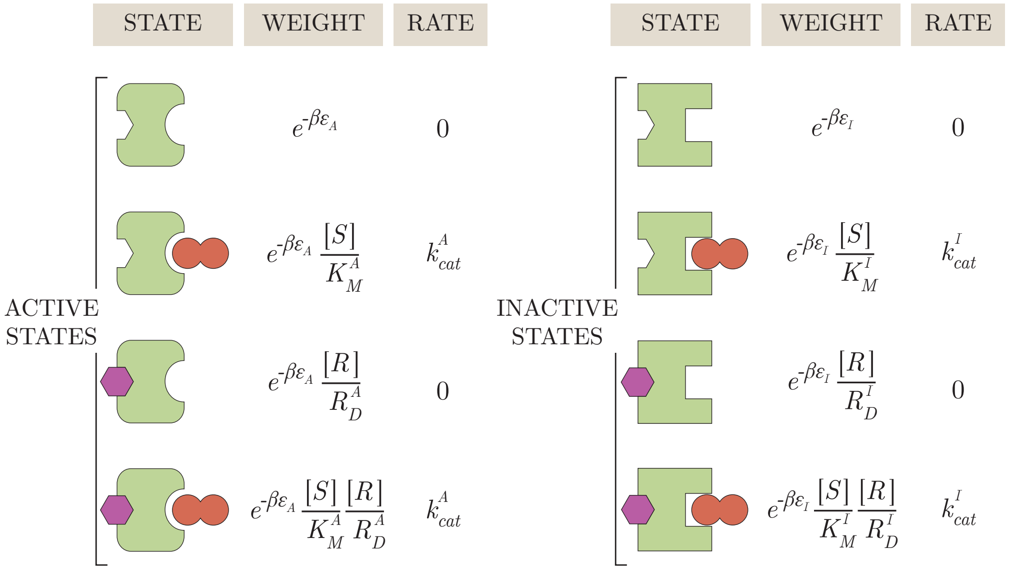

STATE
WEIGHT
RATE
STATE
WEIGHT
RATE
$e^{-\beta\varepsilon_A}$
0
$e^{-\beta\varepsilon_I}$
0
$e^{-\beta\varepsilon_A}\frac{[S]}{K_M^A}$
k_{cat}^A
$e^{-\beta\varepsilon_I}\frac{[S]}{K_M^I}$
k_{cat}^I
ACTIVE STATES
INACTIVE STATES
$e^{-\beta\varepsilon_A}\frac{[R]}{R_D^A}$
0
$e^{-\beta\varepsilon_I}\frac{[R]}{R_D^I}$
0
$e^{-\beta\varepsilon_A}\frac{[S]}{K_M^A}\frac{[R]}{R_D^A}$
k_{cat}^A
$e^{-\beta\varepsilon_I}\frac{[S]}{K_M^I}\frac{[R]}{R_D^I}$
k_{cat}^I

Figure 1.5: States and weights for an MWC enzyme with an allosteric regulator. The allosteric regulator (purple) does not directly interact with the substrate (dark red) but instead introduces a factor $\frac{[R]}{R_D}$ into the weights where R_D is a *dissociation* constant. Note that the regulator can only associate to and dissociate from the enzyme, whereas substrate can be turned into product as shown by the *Michaelis* constant K_M. An allosteric activator binds more tightly to the active state enzyme, $R_D^A < R_D^I$, which leads to an increased rate of product formation because the active state catalyzes substrate at a faster rate than the inactive state, $k_{cat}^A > k_{cat}^I$. An allosteric inhibitor would satisfy $R_D^A > R_D^I$.

given by

$$p_{E_A S} = e^{-\beta\epsilon_A}\frac{\frac{[S]}{K_M^A}}{Z_{tot}} \tag{1.38}$$

$$p_{E_A SR} = e^{-\beta\epsilon_A}\frac{\frac{[S]}{K_M^A}\frac{[R]}{R_D^A}}{Z_{tot}} \tag{1.39}$$

$$p_{E_I S} = e^{-\beta\epsilon_I}\frac{\frac{[S]}{K_M^I}}{Z_{tot}} \tag{1.40}$$

$$p_{E_I SR} = e^{-\beta\epsilon_I}\frac{\frac{[S]}{K_M^I}\frac{[R]}{R_D^I}}{Z_{tot}} \tag{1.41}$$

where

$$Z_{tot} = e^{-\beta\epsilon_A}\left(1+\frac{[S]}{K_M^A}\right)\left(1+\frac{[R]}{R_D^A}\right) + e^{-\beta\epsilon_I}\left(1+\frac{[S]}{K_M^I}\right)\left(1+\frac{[R]}{R_D^I}\right) \tag{1.42}$$

is the sum of all weights in Fig. 1.5. An allosteric activator has a smaller dissociation constant $R_D^A < R_D^I$ for binding to the active state enzyme, so that for larger $[R]$ the probability that the enzyme will be in the active state increases. Because the active state catalyzes substrate at a faster rate than the inactive state, $k_{cat}^A > k_{cat}^I$, adding an activator increases the rate of product formation $\frac{d[P]}{dt}$. An allosteric inhibitor has the flipped relation $R_D^A > R_D^I$ and hence causes the opposite effects.

Proceeding analogously to section 1.3.2, the total enzyme concentration $[E_{tot}]$ is a conserved quantity which equals the sum of all enzyme states ($[E_A]$, $[E_A S]$, $[E_A R]$, $[E_A SR]$, and their inactive state counterparts). Using the probabilities in Eqs. 1.38-1.41, we can write these concentrations as $[E_A S] = [E_{tot}] p_{E_A S}$, $[E_A SR] = [E_{tot}] p_{E_A SR}$, ... so that the rate of product formation is given by

$$\begin{aligned}\frac{d[P]}{dt} &= k_{cat}^A \left([E_A S] + [E_A SR]\right) + k_{cat}^I \left([E_I S] + [E_I SR]\right) \\ &= [E_{tot}] \frac{k_{cat}^A e^{-\beta\epsilon_A} \frac{[S]}{K_M^A}\left(1 + \frac{[R]}{R_D^A}\right) + k_{cat}^I e^{-\beta\epsilon_I} \frac{[S]}{K_M^I}\left(1 + \frac{[R]}{R_D^I}\right)}{e^{-\beta\epsilon_A}\left(1 + \frac{[S]}{K_M^A}\right)\left(1 + \frac{[R]}{R_D^A}\right) + e^{-\beta\epsilon_I}\left(1 + \frac{[S]}{K_M^I}\right)\left(1 + \frac{[R]}{R_D^I}\right)}. \end{aligned} \tag{1.43}$$

The rate of product formation 1.43 for different $[R]$ values is shown in Fig. 1.6. It is important to realize that by choosing the weights in Fig. 1.5, we have selected a particular model for the allosteric regulator, namely one in which the regulator binds equally well to an enzyme with or without substrate. There are many other possible models. For example, we could add an interaction energy between an allosteric regulator and a bound substrate. However, the simple model in Fig. 1.5 already possesses the important feature that adding more allosteric activator yields a larger rate of product formation $\frac{d[P]}{dt}$, as shown in Fig. 1.6.

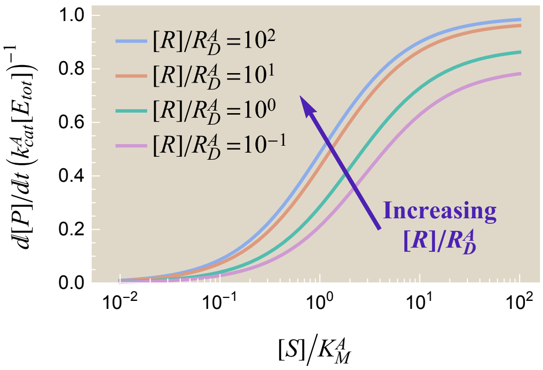


Figure 1.6: Effects of an allosteric regulator *R* on the rate of product formation $\frac{d[P]}{dt}$. The regulator's greater affinity for the active enzyme state increases the fraction of the active conformations and hence $\frac{d[P]}{dt}$. Parameters used were $\frac{R_D^A}{R_D^I} = 10^{-2}$ and the parameters from Fig. 1.4.

An allosteric regulator effectively tunes the energies of the active and inactive states. To better understand this, consider the probability of an active state enzyme-substrate complex (with or without a bound regulator). Adding Eqs. 1.38 and 1.39,

$$\begin{aligned} p_{E_AS} + p_{E_ASR} &= \frac{e^{-\beta\epsilon_A}\frac{[S]}{K_M^A}\left(1+\frac{[R]}{R_D^A}\right)}{e^{-\beta\epsilon_A}\left(1+\frac{[S]}{K_M^A}\right)\left(1+\frac{[R]}{R_D^A}\right)+e^{-\beta\epsilon_I}\left(1+\frac{[S]}{K_M^I}\right)\left(1+\frac{[R]}{R_D^I}\right)} \\ &\equiv \frac{e^{-\beta\tilde{\epsilon}_A}\frac{[S]}{K_M^A}}{e^{-\beta\tilde{\epsilon}_A}\left(1+\frac{[S]}{K_M^A}\right)+e^{-\beta\tilde{\epsilon}_I}\left(1+\frac{[S]}{K_M^I}\right)} \end{aligned} \tag{1.44}$$

where

$$\tilde{\epsilon}_A = \epsilon_A - \frac{1}{\beta}\log\left(1+\frac{[R]}{R_D^A}\right) \tag{1.45}$$

$$\tilde{\epsilon}_I = \epsilon_I - \frac{1}{\beta}\log\left(1+\frac{[R]}{R_D^I}\right). \tag{1.46}$$

We now compare the total probability that an active state enzyme will be bound to substrate in the presence of an allosteric regulator (Eq. 1.44) to this probability in the absence of an allosteric regulator (Eq. 1.24). These two equations show that an MWC enzyme in the presence of regulator concentration $[R]$ is equivalent to an MWC enzyme with no regulator provided that we use the new energies $\tilde{\epsilon}_A$ and $\tilde{\epsilon}_I$ for the active and inactive states. An analogous statement holds for all the conformations of the enzyme, so that the effects of a regulator can be completely absorbed into the energies of the active and inactive states! In other words, adding an allosteric regulator allows us to tune the parameters ϵ_A and ϵ_I of an allosteric enzyme, and thus change its rate of product formation, in a quantifiable manner. This simple result emerges from our assumptions that the allosteric regulator and substrate bind independently to the enzyme and that the allosteric regulator does not effect the rate of product formation.

One application of this result is that we can easily compute the dynamic range of an enzyme as well as the concentration of substrate for half-maximal rate of product formation discussed in section 1.3.2. Both of these quantities follow from the analogous expressions for an MWC enzyme (Eqs. 1.33 and 1.35) using the effective energies $\tilde{\epsilon}_A$ and $\tilde{\epsilon}_I$, resulting in a dynamic range of the form

$$\text{dynamic range} = [E_{tot}]k_{cat}^A\left(1-\frac{1-\frac{k_{cat}^I}{k_{cat}^A}}{1+e^{-\beta(\epsilon_A-\epsilon_I)}\frac{1+[R]/R_D^A}{1+[R]/R_D^I}\frac{K_M^I}{K_M^A}}\right) \tag{1.47}$$

and an $[S_{50}]$ value of

$$[S_{50}] = K_M^A \frac{e^{-\beta(\epsilon_A - \epsilon_I)} \frac{1+[R]/R_D^A}{1+[R]/R_D^I} + 1}{e^{-\beta(\epsilon_A - \epsilon_I)} \frac{1+[R]/R_D^A}{1+[R]/R_D^I} + \frac{K_M^A}{K_M^I}}. \tag{1.48}$$

As expected, the dynamic range of an enzyme increases with regulator concentration $[R]$ for an allosteric activator ($R_D^A < R_D^I$). Adding more activator will shift $[S_{50}]$ to the left if $K_M^A < K_M^I$ (as shown in Fig. 1.6) or to the right if $K_M^A > K_M^I$. The opposite effects hold for an allosteric inhibitor ($R_D^I < R_D^A$).

1.3.4 Competitive Inhibitor

Another level of control found in many enzymes is inhibition. A competitive inhibitor C binds to the same active site as substrate S, yet unlike the substrate, the competitive inhibitor cannot be turned into product by the enzyme. An enzyme with a single active site can either exist in the unbound state E, as an enzyme-substrate complex ES, or as an enzyme-competitor complex EC. As more inhibitor is added to the system, it crowds out the substrate from the enzyme's active site which decreases product formation. Many cancer drugs (e.g. lapatinib, sorafenib, erlotinib) are competitive inhibitors for kinases involved in signaling pathways [41].

Starting from our model of an MWC enzyme in Fig. 1.3, we can introduce a competitive inhibitor by drawing two new states (an enzyme-competitor complex in the active and inactive forms) as shown in Fig. 1.7. Only the enzyme-substrate complex in the active (E_AS) and inactive (E_IS) states form product. The probabilities of each of these states is given by Eqs. 1.24 and 1.26 but using the new partition function (which includes the competitive inhibitor states),

$$Z_{tot} = e^{-\beta\epsilon_A}\left(1 + \frac{[S]}{K_M^A} + \frac{[C]}{C_D^A}\right) + e^{-\beta\epsilon_I}\left(1 + \frac{[S]}{K_M^I} + \frac{[C]}{C_D^I}\right). \tag{1.49}$$

Repeating the same analysis from section 1.3.2, we write the concentrations of bound enzymes as $[E_AS] = [E_{tot}]p_{E_AS}$ and $[E_IS] = [E_{tot}]p_{E_IS}$, where $[E_{tot}]$ is the total concentration of enzymes in the system and $p_{E_{A,I}S}$ is the weight of the bound (in)active state enzyme divided by the partition function, Eq. 1.49. Thus the rate of product formation equals

$$\begin{aligned}\frac{d[P]}{dt} &= k_{cat}^A[E_AS] + k_{cat}^I[E_IS] \\ &= [E_{tot}]\frac{k_{cat}^A e^{-\beta\epsilon_A}\frac{[S]}{K_M^A} + k_{cat}^I e^{-\beta\epsilon_I}\frac{[S]}{K_M^I}}{e^{-\beta\epsilon_A}\left(1 + \frac{[S]}{K_M^A} + \frac{[C]}{C_D^A}\right) + e^{-\beta\epsilon_I}\left(1 + \frac{[S]}{K_M^I} + \frac{[C]}{C_D^I}\right)}.\end{aligned} \tag{1.50}$$

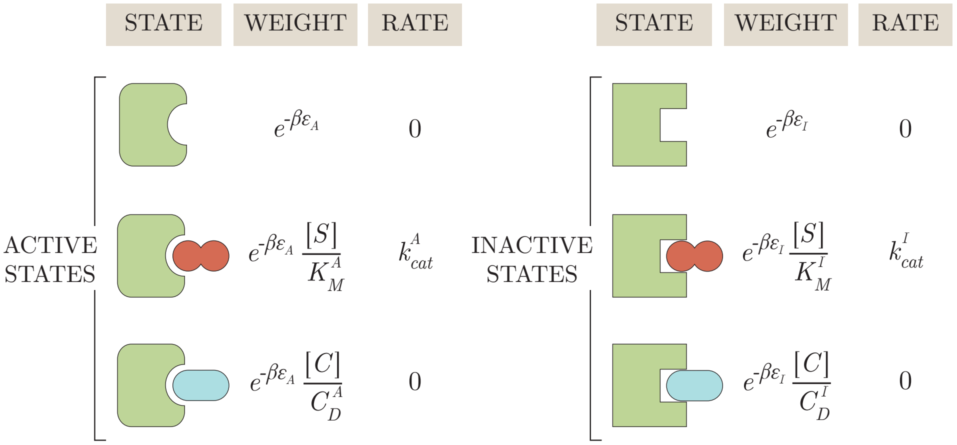

STATE
WEIGHT
RATE
STATE
WEIGHT
RATE
ACTIVE STATES
$e^{-\beta\varepsilon_A}$
0
$e^{-\beta\varepsilon_A}\frac{[S]}{K_M^A}$
k_{cat}^A
$e^{-\beta\varepsilon_A}\frac{[C]}{C_D^A}$
0
INACTIVE STATES
$e^{-\beta\varepsilon_I}$
0
$e^{-\beta\varepsilon_I}\frac{[S]}{K_M^I}$
k_{cat}^I
$e^{-\beta\varepsilon_I}\frac{[C]}{C_D^I}$
0

Figure 1.7: States and weights for an MWC enzyme with a competitive inhibitor. While the substrate S (dark red) can be transformed into product, the inhibitor C (light blue) can occupy the substrate binding site but cannot be catalyzed. As seen with the allosteric regulator in section 1.3.3, the competitive inhibitor contributes a factor $\frac{[C]}{C_D}$ to the statistical weight of a state where C_D is the inhibitor's dissociation constant.

Fig. 1.8 shows the rate of product formation for various inhibitor concentrations $[C]$. Adding more competitive inhibitor increases the probability of the inhibitor-bound states and thereby drains probability out of those states competent to form product, as expected. Similarly to our analysis of allosteric regulators, we can absorb the effects of the competitive inhibitor ($C_D^{A,I}$) in Eq. 1.50 into the enzyme parameters ($\epsilon_{A,I}$, $K_M^{A,I}$),

$$\begin{aligned}\frac{d[P]}{dt} &= [E_{tot}]\frac{k_{cat}^A e^{-\beta\epsilon_A}\left(1+\frac{[C]}{C_D^A}\right)\frac{[S]}{K_M^A\left(1+\frac{[C]}{C_D^A}\right)} + k_{cat}^I e^{-\beta\epsilon_I}\left(1+\frac{[C]}{C_D^I}\right)\frac{[S]}{K_M^I\left(1+\frac{[C]}{C_D^I}\right)}}{e^{-\beta\epsilon_A}\left(1+\frac{[C]}{C_D^A}\right)\left(1+\frac{[S]}{K_M^A\left(1+\frac{[C]}{C_D^A}\right)}\right)+e^{-\beta\epsilon_I}\left(1+\frac{[C]}{C_D^I}\right)\left(1+\frac{[S]}{K_M^I\left(1+\frac{[C]}{C_D^I}\right)}\right)}\\ &\equiv [E_{tot}]\frac{k_{cat}^A e^{-\beta\tilde{\epsilon}_A}\frac{[S]}{\tilde{K}_M^A}+k_{cat}^I e^{-\beta\tilde{\epsilon}_I}\frac{[S]}{\tilde{K}_M^I}}{e^{-\beta\tilde{\epsilon}_A}\left(1+\frac{[S]}{\tilde{K}_M^A}\right)+e^{-\beta\tilde{\epsilon}_I}\left(1+\frac{[S]}{\tilde{K}_M^I}\right)},\end{aligned} \tag{1.51}$$

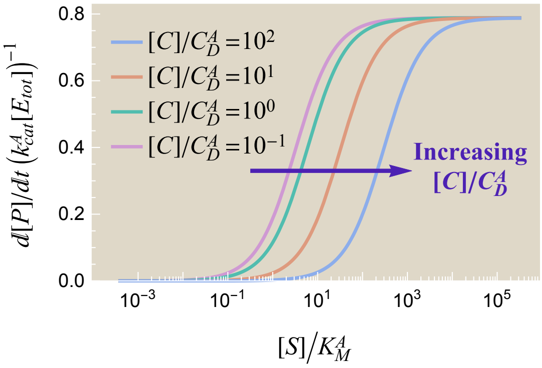


Figure 1.8: Effects of a competitive inhibitor C on the rate of product formation $\frac{d[P]}{dt}$. When $[C] \lesssim C^A_D, C^I_D$, the inhibitor cannot out-compete the substrate at high substrate concentrations while the free form of enzyme dominates at low substrate concentrations. Therefore increasing $[C]$ up to values of $\approx C^A_D$ or C^I_D has little effect on $\frac{d[P]}{dt}$. Once $[C] \gtrsim C^A_D, C^I_D$, the inhibitor can out-compete substrate at large concentrations, pushing the region where the enzyme-substrate complex dominates further to the right. Parameters used were $\frac{C^A_D}{C^I_D} = 1$ and the parameters from Fig. 1.4.

where we have defined the new energies and Michaelis constants,

$$\tilde{\epsilon}_A = \epsilon_A - \frac{1}{\beta} \log\left(1 + \frac{[C]}{C^A_D}\right) \tag{1.52}$$

$$\tilde{\epsilon}_I = \epsilon_I - \frac{1}{\beta} \log\left(1 + \frac{[C]}{C^I_D}\right) \tag{1.53}$$

$$\tilde{K}^A_M = K^A_M \left(1 + \frac{[C]}{C^A_D}\right) \tag{1.54}$$

$$\tilde{K}^I_M = K^I_M \left(1 + \frac{[C]}{C^I_D}\right). \tag{1.55}$$

Note that Eq. 1.51 has exactly the same form as the rate of product formation of an MWC enzyme without a competitive inhibitor, Eq. 1.30. In other words, a competitive inhibitor modulates both the effective energies and the Michaelis constants of the active and inactive states. Thus, an observed value of K_M may not represent a true Michaelis constant if an inhibitor is present. In the special case of a Michaelis-Menten enzyme ($e^{-\beta\epsilon_I} \to 0$), we recover the known result that a competitive inhibitor only changes the apparent Michaelis constant [17].

As shown for the allosteric regulator, the dynamic range and the concentration of substrate for half-maximal rate of product formation $[S_{50}]$ follow from the analogous expressions for an MWC enzyme (section 1.3.2, Eqs. 1.33 and 1.35) using the parameters $\tilde{\epsilon}_{A,I}$ and $\tilde{K}_M^{A,I}$. Hence an allosteric enzyme with one active site in the presence of a competitive inhibitor has a dynamic range given by

$$\text{dynamic range} = [E_{tot}]k_{cat}^A \left(1 - \frac{1 - \frac{k_{cat}^I}{k_{cat}^A}}{1 + e^{-\beta(\epsilon_A - \epsilon_I)} \frac{K_M^I}{K_M^A}} \right) \tag{1.56}$$

and an $[S_{50}]$ value of

$$[S_{50}] = K_M^A \frac{e^{-\beta(\epsilon_A - \epsilon_I)} \left(1 + \frac{[C]}{C_D^A}\right) + \left(1 + \frac{[C]}{C_D^I}\right)}{e^{-\beta(\epsilon_A - \epsilon_I)} + \frac{K_M^A}{K_M^I}}. \tag{1.57}$$

Notice that Eq. 1.56, the dynamic range of an MWC enzyme in the presence of a competitive inhibitor, is exactly the same as Eq. 1.33, the dynamic range in the absence of an inhibitor. This makes sense because in the absence of substrate ($[S] \to 0$) the rate of product formation must be zero and at saturating substrate concentrations ($[S] \to \infty$) the substrate completely crowds out any inhibitor concentration. Instead of altering the rate of product formation at these two limits, the competitive inhibitor shifts the $\frac{d[P]}{dt}$ curve, and therefore $[S_{50}]$, to the right as more inhibitor is added.

Said another way, adding a competitive inhibitor effectively rescales the concentration of substrate in a system. Consider an MWC enzyme in the absence of a competitive inhibitor at a measured substrate concentration $[S_{\text{no}\,[C]}]$. Now consider a separate system where an enzyme is in the presence of a competitive inhibitor at concentration $[C]$ and at a measured substrate concentration $[S_{\text{with}[C]}]$. It is straightforward to show that the rate of product formation $\frac{d[P]}{dt}$ is the same for both enzymes,

$$\begin{aligned} \frac{d[P]}{dt} &= [E_{tot}] \frac{k_{cat}^A e^{-\beta\epsilon_A} \frac{[S_{\text{no}\,[C]}]}{K_M^A} + k_{cat}^I e^{-\beta\epsilon_I} \frac{[S_{\text{no}\,[C]}]}{K_M^I}}{e^{-\beta\epsilon_A} \left(1 + \frac{[S_{\text{no}\,[C]}]}{K_M^A}\right) + e^{-\beta\epsilon_I} \left(1 + \frac{[S_{\text{no}\,[C]}]}{K_M^I}\right)} \\ &= [E_{tot}] \frac{k_{cat}^A e^{-\beta\epsilon_A} \frac{[S_{\text{with}[C]}]}{K_M^A} + k_{cat}^I e^{-\beta\epsilon_I} \frac{[S_{\text{with}[C]}]}{K_M^I}}{e^{-\beta\epsilon_A} \left(1 + \frac{[S_{\text{with}[C]}]}{K_M^A} + \frac{[C]}{C_D^A}\right) + e^{-\beta\epsilon_I} \left(1 + \frac{[S_{\text{with}[C]}]}{K_M^I} + \frac{[C]}{C_D^I}\right)}, \end{aligned} \tag{1.58}$$

provided that

$$[S_{\text{with}[C]}] = \frac{e^{-\beta(\epsilon_A - \epsilon_I)} \left(1 + \frac{[C]}{C_D^A}\right) + \left(1 + \frac{[C]}{C_D^I}\right)}{e^{-\beta(\epsilon_A - \epsilon_I)} + 1} [S_{\text{no}\,[C]}]. \tag{1.59}$$

For any fixed competitive inhibitor concentration $[C]$, this rescaling amounts to a constant multiplicative factor which results in a horizontal shift on a log scale of substrate concentration $[S]$, as is indeed shown in Fig. 1.8.

As we have seen, the effects of both an allosteric regulator and a competitive inhibitor can be absorbed into the parameters of an MWC enzyme. This suggests that experimental data from enzymes that titrate these ligands can be collapsed into a one-parameter family of curves where the single parameter is either the concentration of an allosteric effector or a competitive inhibitor. Indeed, in section 1.4.1 we shall find that this theory matches well with experimentally measured activity curves.

1.3.5 Multiple Substrate Binding Sites

In 1965, Gerhart and Schachman used ultracentrifugation to determine that ATCase can be separated into a large (100 kDa) catalytic subunit where substrate binds and a smaller (30 kDa) regulatory subunit which has binding sites for the allosteric regulators ATP and CTP [42]. Their measurements correctly predicted that ATCase had multiple active sites and multiple regulatory sites, although their actual numbers were off (they predicted 2 active sites and 4 regulatory sites, whereas ATCase has 6 active sites and 6 regulatory sites) [13]. Three years later, more refined sequencing by Weber and crystallographic measurements by Wiley and Lipscomb revealed the correct quaternary structure of ATCase [43–45].

Many enzymes are composed of multiple subunits that contain substrate binding sites (also called active sites or catalytic sites). Having multiple binding sites grants the substrate more locations to bind to an enzyme which increases the effective affinity between both molecules. A typical enzyme will have between 1 and 6 substrate binding sites, and bindings sites for allosteric regulators can appear with similar multiplicity. However, extreme cases exist such as hemocyanin which can have as many as 48 active sites. [46] Interestingly, across different species the same enzyme may possess different numbers of active or regulatory sites, as well as be affected by other allosteric regulators and competitive inhibitors [10, 47]. Furthermore, multiple binding sites may interact with each other in a complex and often uncharacterized manner [48].

We now extend the single-site model of an MWC enzyme introduced in Fig. 1.3 to an MWC enzyme with two substrate binding sites. Assuming that both binding sites are identical and independent, the states and weights of the system are shown in Fig. 1.9. When the enzyme is doubly occupied $E_A S^2$, we assume that it forms product twice as fast as a singly occupied enzyme $E_A S$.

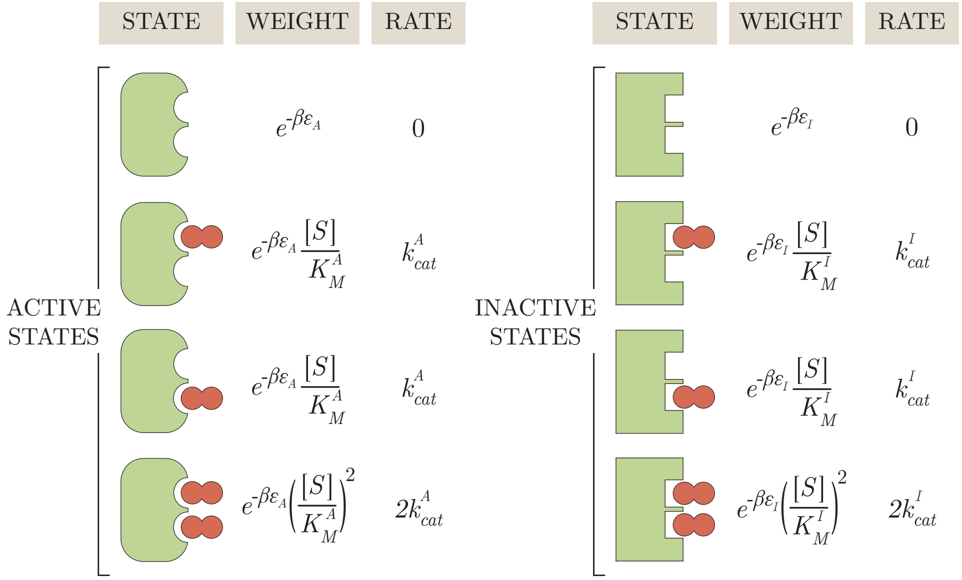

STATE
WEIGHT
RATE
STATE
WEIGHT
RATE
ACTIVE STATES
$e^{-\beta\varepsilon_A}$
0
$e^{-\beta\varepsilon_A}\frac{[S]}{K_M^A}$
k_{cat}^A
$e^{-\beta\varepsilon_A}\frac{[S]}{K_M^A}$
k_{cat}^A
$e^{-\beta\varepsilon_A}\left(\frac{[S]}{K_M^A}\right)^2$
$2k_{cat}^A$
INACTIVE STATES
$e^{-\beta\varepsilon_I}$
0
$e^{-\beta\varepsilon_I}\frac{[S]}{K_M^I}$
k_{cat}^I
$e^{-\beta\varepsilon_I}\frac{[S]}{K_M^I}$
k_{cat}^I
$e^{-\beta\varepsilon_I}\left(\frac{[S]}{K_M^I}\right)^2$
$2k_{cat}^I$

Figure 1.9: States and weights for an MWC enzyme with two substrate binding sites. Each binding site acts independently and the rate of product formation of a doubly bound state is twice the rate of the corresponding singly bound state.

It has been shown that in MWC models, explicit cooperative interaction energies are not required to accurately model biological systems; cooperativity is inherently built into the fact that all binding sites switch concurrently from an active state to an inactive state [16]. For example, suppose an inactive state enzyme with two empty catalytic sites binds with its inactive state affinity K_M^I to a single substrate, and that this binding switches the enzyme from the inactive to the active state. Then the second, still empty, catalytic site now has the active state affinity K_M^A, an effect which can be translated into cooperativity. Note that an explicit interaction energy, if desired, can be added to the model very simply.

As in the proceeding sections, we compute the probability and concentration of each enzyme conformation from the states and weights (see Eqs. 1.23-1.29). Because the active and inactive conformations each have two singly bound states and one doubly bound state with twice the rate, the enzyme's rate of product formation is given by

$$\begin{aligned}\frac{d[P]}{dt} &= k_{cat}^A\left(2p_{E_A S}\right) + 2k_{cat}^A\left(p_{E_A S^2}\right) + k_{cat}^I\left(2p_{E_I S}\right) + 2k_{cat}^I\left(p_{E_I S^2}\right)\\ &= 2[E_{tot}]\frac{k_{cat}^A e^{-\beta\epsilon_A}\frac{[S]}{K_M^A}\left(1+\frac{[S]}{K_M^A}\right) + k_{cat}^I e^{-\beta\epsilon_I}\frac{[S]}{K_M^I}\left(1+\frac{[S]}{K_M^I}\right)}{e^{-\beta\epsilon_A}\left(1+\frac{[S]}{K_M^A}\right)^2 + e^{-\beta\epsilon_I}\left(1+\frac{[S]}{K_M^I}\right)^2} \qquad (1.60)\end{aligned}$$

We will have much more to say about this model in section 1.4.2.2, where we will show that $\frac{d[P]}{dt}$ as a function of substrate concentration $[S]$ may form a peak. For now, we mention the well-known result that a Michaelis-Menten enzyme with two independent active sites will act identically to two Michaelis-Menten enzymes each with a single active site (as can be seen in the $\epsilon_I \to \infty$ limit of Eq. 1.60) [17]. It is intuitively clear that this result does not extend to MWC enzymes: $\frac{d[P]}{dt}$ for a two-site MWC enzyme, Eq. 1.60, does not equal twice the value of $\frac{d[P]}{dt}$ for a one-site MWC enzyme, Eq. 1.30.

1.3.6 Modeling Overview

The above sections allow us to model a complex enzyme with any number of substrate binding sites, competitive inhibitors, and allosteric regulators. Assuming that the enzyme is in steady state and that the cycle condition holds, we first enumerate its states and weights and then use those weights to calculate the rate of product formation. Our essential conclusions about the roles of the various participants in these reactions can be summarized as follows:

1. The (in)active state enzyme contributes a factor $(e^{-\beta\epsilon_I})$ $e^{-\beta\epsilon_A}$ to the weight. The mathematical simplicity of this model belies the complex interplay between the active and inactive states. Indeed, an MWC enzyme cannot be decoupled into two Michaelis-Menten enzymes (one for the active and the other for the inactive states).

2. Each bound substrate contributes a factor $(\frac{[S]}{K_M^I})$ $\frac{[S]}{K_M^A}$ in the (in)active state where $K_M = \frac{k_{off}+k_{cat}}{k_{on}}$ is a Michaelis constant between the substrate and enzyme. It is this Michaelis constant, and not the dissociation constant, which enters the states and weights diagram.

3. Each bound allosteric regulator or competitive inhibitor X contributes a factor $(\frac{[X]}{X_d^I})$ $\frac{[X]}{X_d^A}$ in the (in)active state where $X_D = \frac{k_{off}^X}{k_{on}^X}$ is the dissociation constant between X and the enzyme. An allosteric regulator R effectively tunes the energies of the active and inactive states as shown in Eqs. 1.45 and 1.46. A competitive inhibitor C effectively changes both the energies and Michaelis constants of the active and inactive states as described by Eqs. 1.52-1.55.

4. The simplest model for multiple binding sites assumes that each site is independent of the others. The MWC model inherently accounts for the cooperativity

between these sites, resulting in sigmoidal activity curves despite no direct interaction terms.

In Appendix S1.2, we simultaneously combine all of these mechanisms by analyzing the rate of product formation of ATCase (which has multiple binding sites) in the presence of substrate, a competitive inhibitor, and allosteric regulators.

Note that while introducing new components (such as a competitive inhibitor or an allosteric regulator) introduces new parameters into the system, increasing the number of sites does not. For example, an MWC enzyme with 1 (Fig. 1.3), 2 (Fig. 1.9), or more active sites would require the same five parameters: $e^{-\beta(\epsilon_A - \epsilon_I)}$, K_M^A, K_M^I, k_{cat}^A, and k_{cat}^I.

1.4 Applications

Having built a framework to model allosteric enzymes, we now turn to some applications of how this model can grant insights into observed enzyme behavior. Experimentally, the rate of product formation of an enzyme is often measured relative to the enzyme concentration, a quantity called *activity*,

$$A \equiv \frac{1}{[E_{tot}]} \frac{d[P]}{dt}. \tag{1.61}$$

Enzymes are often characterized by their activity curves as substrate, inhibitor, and regulator concentrations are titrated. Such data not only determines important kinetic constants but can also characterize the nature of molecular players such as whether an inhibitor is competitive, uncompetitive, mixed, or non-competitive [49–51]. After investigating several activity curves, we turn to a case study of the curious phenomenon of substrate inhibition, where saturating concentrations of substrate inhibit enzyme activity, and propose a new minimal mechanism for substrate inhibition caused solely by allostery.

1.4.1 Regulator and Inhibitor Activity Curves

We begin with an analysis of α-amylase, one of the simplest allosteric enzymes, which only has a single catalytic site. α-amylase catalyzes the hydrolysis of large polysaccharides (e.g. starch and glycogen) into smaller carbohydrates in human metabolism. It is competitively inhibited by isoacarbose [51] at the active site and is allosterically activated by Cl$^-$ ions at a distinct allosteric site [52].

Fig. 1.10 plots substrate concentration divided by activity, $[S]/A$, as a function of substrate $[S]$. Recall from section 1.3.3 that an enzyme with one active site and one

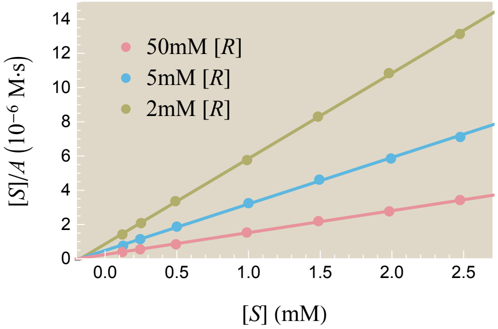


Figure 1.10: Theoretically and experimentally probing the effects of an allosteric regulator on activity. Data points show experimentally measured activity from Feller et al. for the enzyme α-amylase using substrate analog $[S]$ (EPS) and allosteric activator $[R]$ (NaCl). Best fit theoretical curves described by Eq. 1.63 are overlaid on the data. The best fit parameters are $e^{-\beta(\epsilon_A-\epsilon_I)} = 7.8 \times 10^{-4}$, $K_M^A = 0.6\,\text{mM}$, $K_M^I = 0.2\,\text{mM}$, $R_D^A = 0.03\,\text{mM}$, $R_D^I = 7.9\,\text{mM}$, $k_{cat}^A = 14\,\text{s}^{-1}$, and $k_{cat}^I = 0.01\,\text{s}^{-1}$.

allosteric site has activity given by Eq. 1.43,

$$A = \frac{k_{cat}^A e^{-\beta\epsilon_A} \frac{[S]}{K_M^A} \left(1 + \frac{[R]}{R_D^A}\right) + k_{cat}^I e^{-\beta\epsilon_I} \frac{[S]}{K_M^I} \left(1 + \frac{[R]}{R_D^I}\right)}{e^{-\beta\epsilon_A} \left(1 + \frac{[S]}{K_M^A}\right)\left(1 + \frac{[R]}{R_D^A}\right) + e^{-\beta\epsilon_I} \left(1 + \frac{[S]}{K_M^I}\right)\left(1 + \frac{[R]}{R_D^I}\right)}. \tag{1.62}$$

Thus we expect the $[S]/A$ curves in Fig. 1.10 to be linear in $[S]$,

$$\frac{[S]}{A} = \frac{e^{-\beta\epsilon_A} \left(1 + \frac{[S]}{K_M^A}\right)\left(1 + \frac{[R]}{R_D^A}\right) + e^{-\beta\epsilon_I} \left(1 + \frac{[S]}{K_M^I}\right)\left(1 + \frac{[R]}{R_D^I}\right)}{k_{cat}^A e^{-\beta\epsilon_A} \frac{1}{K_M^A} \left(1 + \frac{[R]}{R_D^A}\right) + k_{cat}^I e^{-\beta\epsilon_I} \frac{1}{K_M^I} \left(1 + \frac{[R]}{R_D^I}\right)}. \tag{1.63}$$

Fig. 1.10 shows that the experimental data is well characterized by the theory so that the rate of product formation at any other substrate and allosteric activator concentration can be predicted by this model. The fitting procedure is discussed in detail in Appendix S1.2.

In the special case of a Michaelis-Menten enzyme ($\epsilon_I \to \infty$), the above equation becomes

$$\frac{[S]}{A} = \frac{K_M^A + [S]}{k_{cat}^A} \qquad (\epsilon_I \to \infty). \tag{1.64}$$

The x-intercept of all lines in such a plot would intersect at the point $(-K_M^A, 0)$ which allows an easy determination of K_M^A. This is why plots of $[S]$ vs $[S]/A$, called Hanes plots, are often seen in enzyme kinetics data. Care must be taken, however, when extending this analysis to allosteric enzymes where the form of the x-intercept is more complicated.

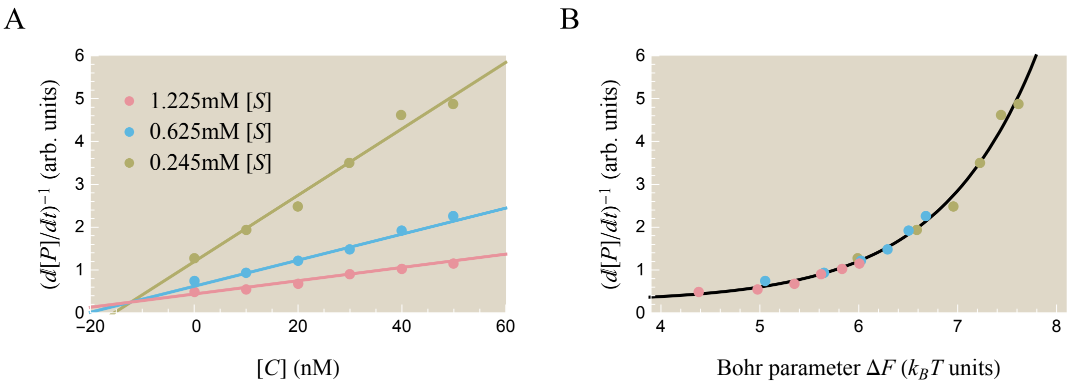


Figure 1.11: Theoretically and experimentally probing the effects of a competitive inhibitor on activity. (A) Data points show experimentally measured activity in arbitrary units from Li et al. for the enzyme α-amylase using substrate analog $[S]$ (α-maltotriosyl fluoride) and competitive inhibitor $[C]$ (isoacarbose). Best fit theoretical curves described by the inverse of Eq. 1.65 are overlaid on the data. The best fit parameters are $e^{-\beta(\epsilon_A-\epsilon_I)} = 36$, $K_M^A = 0.9\,\text{mM}$, $K_M^I = 2.6\,\text{mM}$, $C_D^A = 12\,\text{nM}$, $C_D^I = 260\,\text{nM}$, and $\frac{k_{cat}^A}{k_{cat}^I} = 1.4$. Note that the x-axis varies $[C]$ rather than $[S]$ as in most other plots. (B) A data collapse of the three curves using the Bohr parameter ΔF from Eq. 1.68 which encompasses the effects of both the substrate and inhibitor upon the system.

We now turn to competitive inhibition. Fig. 1.11(A) plots the inverse rate of product formation $\left(\frac{d[P]}{dt}\right)^{-1}$ of α-amylase as a function of the competitive inhibitor concentration $[C]$. The competitive inhibitor isoacarbose is titrated for three different concentrations of the substrate α-maltotriosyl fluoride (αG3F).

Recall from section 1.3.4, Eq. 1.50 that the rate of product formation for an allosteric enzyme with one active site in the presence of a competitive inhibitor is given by

$$\left(\frac{d[P]}{dt}\right)^{-1} = \frac{1}{[E_{tot}]} \frac{e^{-\beta\epsilon_A}\left(1+\frac{[S]}{K_M^A}+\frac{[C]}{C_D^A}\right)+e^{-\beta\epsilon_I}\left(1+\frac{[S]}{K_M^I}+\frac{[C]}{C_D^I}\right)}{k_{cat}^A e^{-\beta\epsilon_A}\frac{[S]}{K_M^A}+k_{cat}^I e^{-\beta\epsilon_I}\frac{[S]}{K_M^I}}, \tag{1.65}$$

so that the best fit $\left(\frac{d[P]}{dt}\right)^{-1}$ curves in Fig. 1.11(A) are linear functions of $[C]$. Rather than thinking of Eq. 1.65 as a function of the competitive inhibitor concentration $[C]$ and the substrate concentration $[S]$ separately, we can combine these two quantities into a single natural parameter for the system. This will enable us to collapse the different activity curves in Fig. 1.11(A) onto a single master curve as shown in

Fig. 1.11(B). Algebraically manipulating Eq. 1.65,

$$\begin{aligned}\frac{d[P]}{dt} &= [E_{tot}]\frac{\left(k_{cat}^{A} e^{-\beta(\epsilon_A-\epsilon_I)}\frac{K_M^I}{K_M^A} + k_{cat}^{I}\right)\frac{[S]}{K_M^I}}{\left(e^{-\beta(\epsilon_A-\epsilon_I)}\frac{K_M^I}{K_M^A}+1\right)\frac{[S]}{K_M^I} + e^{-\beta(\epsilon_A-\epsilon_I)}\left(1+\frac{[C]}{C_D^A}\right)+\left(1+\frac{[C]}{C_D^I}\right)} \\ &\equiv [E_{tot}]\frac{\left(k_{cat}^{A}K + k_{cat}^{I}\right)e^{-\beta\Delta F}}{(K+1)e^{-\beta\Delta F}+1} \end{aligned} \tag{1.66}$$

where

$$K = e^{-\beta(\epsilon_A-\epsilon_I)}\frac{K_M^I}{K_M^A} \tag{1.67}$$

$$\Delta F = -\frac{1}{\beta}\mathrm{Log}\left[\frac{\frac{[S]}{K_M^I}}{e^{-\beta(\epsilon_A-\epsilon_I)}\left(1+\frac{[C]}{C_D^A}\right)+\left(1+\frac{[C]}{C_D^I}\right)}\right]. \tag{1.68}$$

Therefore, $\left(\frac{d[P]}{dt}\right)^{-1}$ curves at any substrate and inhibitor concentrations can be compactly shown as data points lying on a single curve in terms of ΔF, which is called the *Bohr parameter*. Such a data collapse is also possible in the case of allosteric regulators or enzymes with multiples binding sites, although those data collapses may require more than one variable ΔF. In Appendix S1.3, we show that the Bohr parameter corresponds to a free energy difference between enzyme states and examine other cases of data collapse.

1.4.2 Substrate Inhibition

We now turn to a striking phenomenon observed in the enzyme literature: not all enzymes have a monotonically increasing rate of product formation. Instead peaks such as those shown schematically in Fig. 1.12 can arise in various enzymes, displaying behavior that is impossible within Michaelis-Menten kinetics. By exploring these two phenomena with the MWC model, we gain insight into their underlying mechanisms and can make quantifiable predictions as to how to create, amplify, or prevent such peaks.

In Fig. 1.12(A), the monotonically increasing Michaelis-Menten curve makes intuitive sense – a larger substrate concentration implies that at any moment the enzyme's active site is more likely to be occupied by substrate. Therefore, we expect that the activity, $A = \frac{1}{[E_{tot}]}\frac{d[P]}{dt}$, should increase with the substrate concentration $[S]$. Yet many enzymes exhibit a peak activity, a behavior called substrate inhibition [53].

Even more surprisingly, when a small amount of competitive inhibitor – a molecule whose very name implies that it competes with substrate and decreases activity –

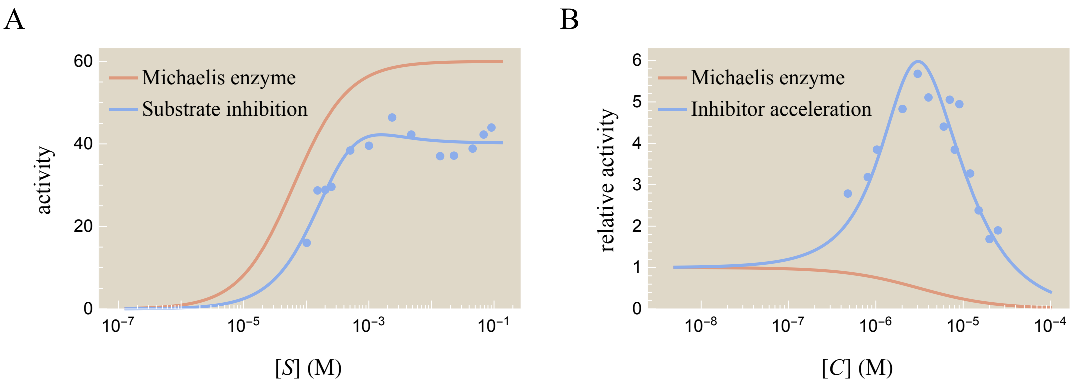


Figure 1.12: Enzyme activity curves do not need to be monotonic as predicted by Michaelis-Menten enzyme kinetics. (A) As many as 20% of enzymes exhibit substrate inhibition, where at high substrate concentrations activity decreases, in contrast to a Michaelis-Menten enzyme shown for reference. Activity for acetylcholinesterase is shown in units of (nanomoles product) $\cdot$ min^{-1} $\cdot$ $(\text{mL enzyme})^{-1}$. (B) Some enzymes exhibit inhibitor acceleration, where adding a small amount of a competitive inhibitor increases the rate of product formation. This generates a peak in activity, in stark contrast to a Michaelis-Menten enzyme which only decreases its activity as more competitive inhibitor is added. Relative activity is shown for ATCase, where relative activity equals activity at $[C]$ divided by the activity with no competitive inhibitor. The data and best fit parameters for the substrate inhibition and inhibitor acceleration curves are discussed in Appendix S1.3.

is mixed together with enzyme, it can *increase* the rate of product formation. This latter case, called inhibitor acceleration, is shown in Fig. 1.12(B) [10, 56]. In contrast, a Michaelis-Menten enzyme shows the expected behavior that adding more competitive inhibitor decreases activity. We will restrict our attention to the phenomenon of substrate inhibition and relegate a discussion of inhibitor acceleration to Appendix S1.4.

Using the MWC enzyme model, we can make predictions about which enzymes can exhibit substrate inhibition. We first formulate a relationship between the fundamental physical parameters of an enzyme that are required to generate such a peak and then consider what information about these underlying parameters can be gained by analyzing experimental data.

1.4.2.1 Single-Site Enzyme

As a preliminary exercise, we begin by showing that an enzyme with a *single* active site cannot exhibit substrate inhibition. Said another way, the activity, Eq. 1.61, of such an enzyme cannot have a peak as a function of substrate concentration

$[S]$. For the remainder of this paper, we will use the fact that all Michaelis and dissociation constants (K_M's, C_D's, and R_D's) are positive and assume that both catalytic constants (k_{cat}^A and k_{cat}^I) are strictly positive unless otherwise stated.

Consider the MWC enzyme with a single substrate binding site shown in Fig. 1.3. Using Eq. 1.30, it is straightforward to compute the derivative of activity with respect to substrate concentration $[S]$, namely,

$$\frac{dA}{d[S]} = \frac{(e^{-\beta\epsilon_A} + e^{-\beta\epsilon_I})\left(e^{-\beta\epsilon_A}\frac{k_{cat}^A}{K_M^A} + e^{-\beta\epsilon_I}\frac{k_{cat}^I}{K_M^I}\right)}{\left(e^{-\beta\epsilon_A}\left(1 + \frac{[S]}{K_M^A}\right) + e^{-\beta\epsilon_I}\left(1 + \frac{[S]}{K_M^I}\right)\right)^2}. \tag{1.69}$$

Since the numerator cannot equal zero, this enzyme cannot have a peak in its activity when $[S]$ is varied. Note that the numerator is positive, indicating that enzyme activity will always increase with substrate concentration.

The above results are valid for an arbitrary MWC enzyme with a single-site. In particular, in the limit $\epsilon_I \to \infty$, an MWC enzyme becomes a Michaelis-Menten enzyme. Therefore, a Michaelis-Menten enzyme with a single active site cannot exhibit a peak in activity. In Appendix S1.5, we discuss the generalization of this result: a Michaelis-Menten enzyme with an arbitrary number of catalytic sites cannot have a peak in activity. Yet as we shall now see, this generalization cannot be made for an MWC enzyme, which can indeed exhibit a peak in its activity when it has multiple binding sites.

1.4.2.2 Substrate Inhibition

As many as 20% of enzymes are believed to exhibit substrate inhibition, which can offer unique advantages to enzymes such as stabilizing their activity amid fluctuations, enhancing signal transduction, and increasing cellular efficiency [54]. This prevalent phenomenon has elicited various explanations, many of which rely on non-equilibrium enzyme dynamics, although some equilibrium mechanisms are known [53]. An example of this latter case is seen in the enzyme aspartate transcarbamoylase (ATCase) which catalyzes one of the first steps in the pyrimidine biosynthetic pathway. Before ATCase can bind to its substrate asparatate (Asp), an intermediate molecule carbamoyl phosphate (CP) must first bind to ATCase, inducing a change in the enzyme's shape and electrostatics which opens up the Asp binding slot [57, 58]. Because Asp can weakly bind to the CP binding pocket, at high concentrations Asp will outcompete CP and prevent the enzyme from working as efficiently, thereby causing substrate inhibition [59].

To the list of such mechanisms, we add the possibility that an enzyme may exhibit substrate inhibition without any additional effector molecules. In particular, an allosteric enzyme with two identical catalytic sites can exhibit a peak in activity when the substrate concentration $[S]$ is varied. We will first analyze the properties of this peak and then examine why it can occur. For simplicity, we will assume $k_{cat}^{I} = 0$ throughout this section and leave the general case for Appendix S1.5.

Using Eqs. 1.60 and 1.61, the activity of an MWC enzyme with two active sites is given by

$$A = \frac{1}{[E_{tot}]}\frac{d[P]}{dt} = \frac{2k_{cat}^{A} e^{-\beta\epsilon_A}\frac{[S]}{K_M^A}\left(1+\frac{[S]}{K_M^A}\right)}{e^{-\beta\epsilon_A}\left(1+\frac{[S]}{K_M^A}\right)^2 + e^{-\beta\epsilon_I}\left(1+\frac{[S]}{K_M^I}\right)^2}. \tag{1.70}$$

A peak will exist provided that $\frac{dA}{d[S]} = 0$ has a positive $[S]$ root. The details of differentiating and solving this equation are given in Appendix S1.5, the result of which is that a peak in activity A occurs as a function of $[S]$ provided that

$$1 + e^{-\beta(\epsilon_A - \epsilon_I)} < \left(\frac{K_M^A}{K_M^I} - 1\right)^2 \qquad (k_{cat}^{I} = 0). \tag{1.71}$$

The height of such a peak is given by

$$A_{peak} = k_{cat}^{A}\frac{K_M^I}{K_M^A - K_M^I}\left(\sqrt{1 + e^{-\beta(\epsilon_A - \epsilon_I)}} - 1\right). \tag{1.72}$$

Examples of peaks in activity are shown in Fig. 1.13 for various values of $e^{-\beta(\epsilon_A - \epsilon_I)}$. Substituting in the peak condition Eq. 1.71, the maximum peak height is at most

$$A_{peak} < k_{cat}^{A}\frac{\frac{K_M^A}{K_M^I} - 2}{\frac{K_M^A}{K_M^I} - 1}. \tag{1.73}$$

If we consider the maximum value of $e^{-\beta(\epsilon_A - \epsilon_I)}$ allowed by the peak condition Eq. 1.71, the peak height approaches k_{cat}^{A} for large $\frac{K_M^A}{K_M^I}$ (as seen by the green curve $e^{-\beta(\epsilon_A - \epsilon_I)} = 80$ in Fig. 1.13(B)). In this limit, the active bound state dominates over all the other enzyme states so that the activity reaches its largest possible value, k_{cat}^{A}. Although the "peak height" is maximum in this case, the activity curve is nearly sigmoidal, making the peak hard to distinguish. To that end, it is reasonable to compare the peak height to the activity at large substrate concentrations,

$$A_{[S]\to\infty} = 2k_{cat}^{A}\frac{e^{-\beta(\epsilon_A - \epsilon_I)}}{\left(\frac{K_M^A}{K_M^I}\right)^2 + e^{-\beta(\epsilon_A - \epsilon_I)}}. \tag{1.74}$$

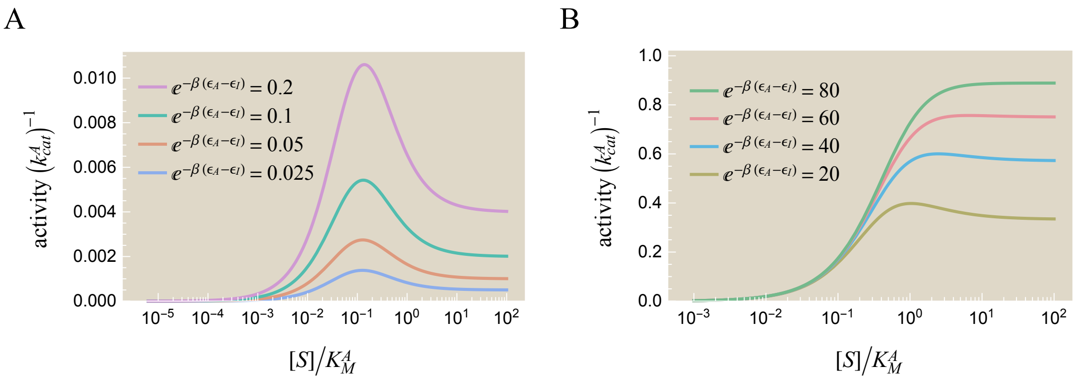


Figure 1.13: Peaks in enzyme activity $A = \frac{1}{E_{tot}}\frac{d[P]}{dt}$ as a function of substrate concentration $[S]$. Activity is shown in units of k_{cat}^A, which rescales the activity curves vertically. The peak for (A) small and (B) large ratios of the enzyme's energy in the active versus inactive state, $e^{-\beta(\epsilon_A-\epsilon_I)}$. The height of the peak increases with $e^{-\beta(\epsilon_A-\epsilon_I)}$. The activity is computed from Eq. 1.70 using the parameters $k_{cat}^I = 0$, $\frac{K_M^A}{K_M^I} = 10$, and the different values of $e^{-\beta(\epsilon_A-\epsilon_I)}$ shown. As predicted by Eq. 1.71, every value in the range $e^{-\beta(\epsilon_A-\epsilon_I)} < \left(\frac{K_M^A}{K_M^I} - 1\right)^2$ will yield a peak in activity. While the peak is more pronounced when the active state is energetically favorable ($e^{-\beta(\epsilon_A-\epsilon_I)} < 1$) in (A), the maximum peak height is much larger in (B) as seen by the different scale of the y-axis.

As the energy difference between the active and inactive state $e^{-\beta(\epsilon_A-\epsilon_I)}$ increases, the peak height A_{peak} monotonically increases but the relative peak height $\frac{A_{peak}}{A_{[S]\to\infty}}$ monotonically decreases. These relations might be used to design enzymes with particular activity curves; conversely, experimental data of substrate inhibition can be used to fix a relation between the parameters $e^{-\beta(\epsilon_A-\epsilon_I)}$ and $\frac{K_M^A}{K_M^I}$ of an enzyme.

We now turn to the explanation of how such a peak can occur. One remarkable fact is that a peak *cannot* happen without allostery. If we consider a Michaelis-Menten enzyme (by taking the limit $k_{cat}^I \to 0$ and $\epsilon_I \to \infty$), then the peak condition Eq. 1.71 cannot be satisfied.

To gain a qualitative understanding of how a peak can occur, consider an enzyme that inherently prefers the active state ($e^{-\beta(\epsilon_A-\epsilon_I)} > 1$) but with substrate that preferentially binds to the inactive state ($\frac{K_M^A}{K_M^I} > 1$). Such a system is realized in bacterial chemotaxis, where the chemotaxis receptors are active when unbound but inactive when bound to substrate [32]. This setup is shown schematically in Fig. 1.14. At low substrate concentrations, $[S] \ll K_M^A$, most enzymes will be unbound and therefore in the active state. At intermediate substrate concentrations, $[S] \approx K_M^A$,

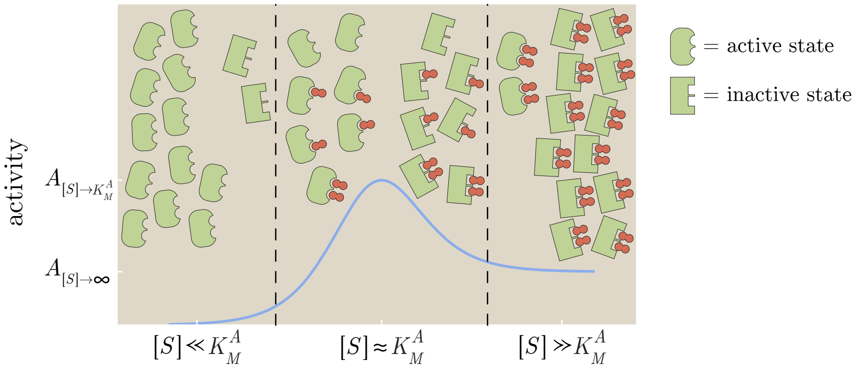


Figure 1.14: Mechanism underlying peak in activation by substrate *S*. At low substrate concentrations (left region), all enzymes are unbound and are mostly in the active form (rounded, green). As the amount of substrate is increased (middle region), the probability that an enzyme is singly bound and then doubly bound increase. Because the substrate prefers to bind to an inactive state (sharp, green) enzyme-substrate complex, binding more substrate pushes the enzymes into the inactive state. At medium substrate concentrations, more active state enzyme-substrate complexes exist than at high substrate concentrations (right region) which yields a peak. Each enzyme fluctuates between its different configurations, and the cartoons show the distributions of the most prevalent states.

many enzymes will be singly bound. Because $\frac{K_M^A}{K_M^I} > 1$, the substrate will pull these bound enzymes towards the inactive state. For large substrate concentrations, $[S] \gg K_M^A$, most of the enzymes will be doubly bound and hence will be predominantly in the inactive form. Because the inactive state does not catalyze substrate ($k_{cat}^I = 0$), only the number of substrate bound to active state enzymes increase the rate of product formation, and because more of these exist in the intermediate regime a peak forms.

To be more quantitative, the activity Eq. 1.70 at the medium substrate concentration ($[S] = K_M^A$) is given by

$$A_{[S]\to K_M^A} = k_{cat}^A \frac{4e^{-\beta(\epsilon_A-\epsilon_I)}}{\left(\frac{K_M^A}{K_M^I}+1\right)^2 + 4e^{-\beta(\epsilon_A-\epsilon_I)}}. \tag{1.75}$$

Comparing this to $A_{[S]\to\infty}$ in Eq. 1.74, we find that $A_{[S]\to K_M^A} > A_{[S]\to\infty}$ provided that

$$1 + e^{-\beta(\epsilon_A-\epsilon_I)} < \frac{1}{4}\left(\frac{K_M^A}{K_M^I} - 1\right)^2. \tag{1.76}$$

This is in close agreement with the peak condition Eq. 1.71, with the $\frac{1}{4}$ prefactor arising because the peak need not occur precisely at $[S] = K_M^A$.

Note that the peak condition Eq. 1.71 does not necessarily force the unbound enzyme to favor the active state ($e^{-\beta(\epsilon_A-\epsilon_I)} > 1$), since this condition can still be satisfied if $e^{-\beta(\epsilon_A-\epsilon_I)} < 1$. However, the peak condition does require that substrate preferentially binds to the inactive state enzyme (in fact, we must have $\frac{K_M^A}{K_M^I} > 2$ to satisfy the peak condition).

Recall that as many as 20% of enzymes exhibit substrate inhibition, and this particular mechanism will not apply in every instance. To be concrete, an allosteric enzyme that obeys the mode of substrate inhibition proposed above must: (1) have at least two catalytic sites and (2) must be driven towards the inactive state upon substrate binding. Therefore, an enzyme such as ATCase which exhibits substrate inhibition but where the substrate preferentially binds to the active state enzyme must have a different underlying mechanism [60]. Various alternative causes including the effects of pH due to substrate or product buildup [17, 61] or the sequestering effects of ions [62, 63] may also be responsible for substrate inhibition. Yet the mechanism of substrate inhibition described above exactly matches the conditions of acetylcholinesterase whose activity, shown in Fig. 1.12(A), is well categorized by the MWC model [55]. It would be interesting to test this theory by taking a well characterized enzyme, tuning the MWC parameters so as to satisfy the peak condition Eq. 1.71 (or an analogous relationship for an enzyme with more than two catalytic sites), and checking whether the system then exhibits substrate inhibition. Experimentally, tuning the parameters can be undertaken by introducing allosteric regulators or competitive inhibitors as described by Eqs. 1.45-1.46 and Eqs. 1.52-1.55, respectively. For example, in Appendix S1.5, we describe an enzyme system where introducing a competitive inhibitor induces a peak in activity.

1.5 Discussion

Allosteric molecules pervade all realms of biology from chemotaxis receptors to chromatin to enzymes [15, 64–66]. There are various ways to capture the allosteric nature of macromolecules, with the MWC model representing one among many [8, 67, 68]. In any such model, the simple insight that molecules exist in an active and inactive state opens a rich new realm of dynamics.

The plethora of molecular players that interact with enzymes serve as the building blocks to generate complex behavior. In this paper, we showed the effects of competitive inhibitors, allosteric regulators, and multiple binding sites, looking at each of these factors first individually and then combining separate aspects. This framework matched well with experimental data and enabled us to make quantifiable predictions on how the MWC enzyme parameters may be tuned upon the introduction of an allosteric regulator Eqs. 1.45-1.46 or a competitive inhibitor Eqs. 1.52-1.55.

As an interesting application, we used the MWC model to explore the unusual behavior of substrate inhibition, where past a certain point adding more substrate to a system decreases its rate of product formation. This mechanism implies that an enzyme activity curve may have a peak (see Fig. 1.12), a feat that is impossible for a Michaelis-Menten enzyme. We explored a novel minimal mechanism for substrate inhibition which rested upon the allosteric interactions of the active and inactive enzyme states, with suggestive evidence for such a mechanism in acetylecholinesterase.

The power of the MWC model stems from its simple description, far-reaching applicability, and its ability to unify the proliferation of data gained over the past 50 years of enzymology research. A series of activity curves at different concentrations of a competitive inhibitor all fall into a one-parameter family of curves, allowing us to predict the activity at any other inhibitor concentration. Such insights not only shed light on the startling beauty of biological systems but may also be harnessed to build synthetic circuits and design new drugs. We close by noting our gratitude and admiration to Prof. Bill Gelbart to whom this special is dedicated and who has inspired us with his clever use of ideas from statistical physics to understand biological systems.

1.6 Acknowledgements

The authors thank T. Biancalani, J.-P. Changeux, A. Gilson, J. Kondev, M. Manhart, R. Milo, M. Morrison, N. Olsman, and J. Theriot for helpful discussions and insights on this paper.

References

[1] Wendell Lim, Bruce Mayer TP. *Cell Signaling: Principles and Mechanisms*. Garland Science, 2014.

[2] Hidestrand M, Oscarson M, Salonen JS, Nyman L, Pelkonen O, Turpeinen M, et al. CYP2B6 and CYP2C19 as the Major Enzymes Responsible for the Metabolism of Selegiline, a Drug Used in the Treatment of Parkinson's Disease, as Revealed from Experiments with Recombinant Enzymes. Drug Metab. Dispos. 2001;29(11):1480–1484.

[3] Brattström L, Israelsson B, Norrving B, Bergqvist D, Thörne J, Hultberg B, et al. Impaired Homocysteine Metabolism in Early-Onset Cerebral and Peripheral Occlusive Arterial Disease Effects of Pyridoxine and Folic Acid Treatment. Atherosclerosis. 1990;81(1):51–60. doi:10.1016/0021-9150(90) 90058-Q.

[4] Zelezniak A, Pers TH, Soares S, Patti ME, Patil KR. Metabolic Network Topology Reveals Transcriptional Regulatory Signatures of Type 2 Diabetes. PLoS Comput. Biol. 2010;6(4):e1000729. doi:10.1371/journal.pcbi.1000729.

[5] Michaelis L, Menten ML. Die Kinetik Der Invertinwirkung. Biochem. Z. 1913;49:333–369.

[6] Briggs GE. A Further Note on the Kinetics of Enzyme Action. Biochem. J. 1925;19(6):1037–1038.

[7] Monod J, Wyman J, Changeux JP. On the Nature of Allosteric Transitions: A Plausible Model. J. Mol. Biol. 1965;12:88–118. doi:10.1016/S0022-2836(65)80285-6.

[8] Koshland DE, Némethy G, Filmer D. Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits. Biochem. 1966; 5(1):365–385. doi:10.1021/bi00865a047.

[9] Cockrell GM, Zheng Y, Guo W, Peterson AW, Truong JK, Kantrowitz ER. New Paradigm for Allosteric Regulation of *Escherichia coli* Aspartate Transcarbamoylase. Biochem. 2013;52(45):8036–8047. doi:10.1021/bi401205n.

[10] Wales ME, Madison LL, Glaser SS, Wild JR. Divergent Allosteric Patterns Verify the Regulatory Paradigm for Aspartate Transcarbamylase. J. Mol. Biol. 1999;294(5):1387–1400. doi:10.1006/jmbi.1999.3315.

[11] Monod J, Changeux JP, Jacob F. Allosteric Proteins and Cellular Control Systems. J. Mol. Biol. 1963;6:306–329.

[12] Gerhart JC, Pardee AB. The Enzymology of Control by Feedback Inhibition. J. Biol. Chem. 1962;237(3):891–896.

[13] Gerhart JC, Schachman HK. Distinct Subunits for the Regulation and Catalytic Activity of Aspartate Transcarbamylase. Biochem. 1965;4(6):1054–1062.

[14] Daber R, Sharp K, Lewis M. One Is Not Enough. J. Mol. Biol. 2009; 392(5):1133–1144.

[15] Changeux JP. Allostery and the Monod-Wyman-Changeux Model After 50 Years. Annu. Rev. Biophys. 2012;41:103–33. doi:10.1146/annurev-biophys-050511-102222.

[16] Marzen S, Garcia HG, Phillips R. Statistical Mechanics of Monod-Wyman-Changeux (MWC) Models. J. Mol. Biol. 2013;425(9):1433–1460. doi:10.1016/j.jmb.2013.03.013.

[17] Segel IH. *Enzyme Kinetics: Behavior and Analysis of Rapid Equilibrium and Steady-State Enzyme Systems*, volume 2. 1993.

[18] Cornish-Bowden A. *Fundamentals of Enzyme Kinetics*. Elsevier, 1979. doi:10.1016/B978-0-408-10617-7.50002-X.

[19] Fersht A. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. W. H. Freeman, 1999.

[20] Price NC, Stevens L. *Fundamentals of Enzymology: The Cell and Molecular Biology of Catalytic Proteins*. Oxford University Press, 1999.

[21] Reuveni S, Urbakh M, Klafter J. Role of Substrate Unbinding in Michaelis-Menten Enzymatic Reactions. Proc. Natl. Acad. Sci. USA. 2014; 111(12):4391–4396. doi:10.1073/Proc.Natl.Acad.Sci.USA.1318122111.

[22] Pinto MF, Estevinho BN, Crespo R, Rocha FA, Damas AM, Martins PM. Enzyme Kinetics: The Whole Picture Reveals Hidden Meanings. FEBS J. 2015;282(12):2309–16. doi:10.1111/febs.13275.

[23] Keymer JE, Endres RG, Skoge M, Meir Y, Wingreen NS. Chemosensing in *Escherichia coli*: Two regimes of two-state receptors. Proc. Natl. Acad. Sci. USA. 2006;103(6):1786–1791. doi:10.1073/Proc.Natl.Acad.Sci.USA.0507438103.

[24] Endres RG, Wingreen NS. Precise Adaptation in Bacterial Chemotaxis Through Assistance Neighborhoods. Proc. Natl. Acad. Sci. USA. 2006; 103(35):13040–4. doi:10.1073/Proc.Natl.Acad.Sci.USA.0603101103.

[25] Mello BA, Tu Y. An Allosteric Model for Heterogeneous Receptor Complexes: Understanding Bacterial Chemotaxis Responses to Multiple Stimuli. Proc. Natl. Acad. Sci. USA. 2005;102(48):17354–9. doi:10.1073/Proc.Natl.Acad.Sci.USA.0506961102.

[26] Hansen CH, Endres RG, Wingreen NS. Chemotaxis in *Escherichia coli*: a molecular model for robust precise adaptation. PLoS Comput. Biol. 2008; 4(1):e1. doi:10.1371/journal.pcbi.0040001.

[27] Phillips R. Napoleon Is in Equilibrium. Annu. Rev. Condens. Matter Phys. 2015;6(1):85–111. doi:10.1146/annurev-conmatphys-031214-014558.

[28] Mirny La. Nucleosome-Mediated Cooperativity Between Transcription Factors. Proc. Natl. Acad. Sci. USA. 2010;doi:10.1073/Proc.Natl.Acad.Sci.USA.0913805107.

[29] Narula J, Igoshin OA. Thermodynamic Models of Combinatorial Gene Regulation by Distant Enhancers. IET Syst. Biol. 2010;4(6):393–408. doi: 10.1049/iet-syb.2010.0010.

[30] Gunawardena J. A Linear Framework for Time-Scale Separation in Nonlinear Biochemical Systems. PLoS ONE. 2012;7(5):e36321. doi:10.1371/journal.pone.0036321.

[31] Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, et al. Transcriptional Regulation by the Numbers: Models. Curr Opin Genetics Dev. 2005; 15(2):116–124. doi:10.1016/j.gde.2005.02.007.

[32] Phillips R, Kondev J, Theriot J, Garcia H, Chasan B. *Physical Biology of the Cell*. Garland Science, 2nd edition, 2010.

[33] Gardino AK, Volkman BF, Cho HS, Lee SY, Wemmer DE, Kern D. The NMR Solution Structure of BeF3-Activated Spo0F Reveals the Conformational Switch in a Phosphorelay System. J. Mol. Biol. 2003;331(1):245–254. doi: 10.1016/S0022-2836(03)00733-2.

[34] Milligan G. Constitutive Activity and Inverse Agonists of G Protein-Coupled Receptors: A Current Perspective. Mol. Pharmacol. 2003;64(6):1271–6. doi:10.1124/mol.64.6.1271.

[35] Kern D, Zuiderweg ER. The Role of Dynamics in Allosteric Regulation. Curr. Opin. Struct. Biol. 2003;13(6):748–757. doi:10.1016/j.sbi.2003.10.008.

[36] Howlett GJ, Blackburn MN, Compton JG, Schachman HK. Allosteric Regulation of Aspartate Transcarbamoylase. Analysis of the Structural and Functional Behavior in Terms of a Two-State Model. Biochem. 1977;16(23):5091–5099. doi:10.1021/bi00642a023.

[37] Wolfenden R. Degrees of Difficulty of Water-Consuming Reactions in the Absence of Enzymes. Chem. Rev. 2006;106(8):3379–3396. doi:10.1021/cr050311y.

[38] Phillips R, Milo R. Rates and Duration. *Cell Biology by the Numbers*, chapter 4. 2015;.

[39] Frank S. Input-Output Relations in Biological Systems: Measurement, Information and the Hill Equation. Biol. Direct. 2013;8(1):31. doi:10.1186/1745-6150-8-31.

[40] Dyachenko A, Gruber R, Shimon L, Horovitz A, Sharon M. Allosteric Mechanisms Can Be Distinguished Using Structural Mass Spectrometry. Proc. Natl. Acad. Sci. USA. 2013;110(18):7235–9. doi:10.1073/Proc.Natl.Acad.Sci.USA.1302395110.

[41] Mellinghoff IK, Sawyers CL, editors. *Therapeutic Kinase Inhibitors*, volume 27. Springer Science & Business Media, 2012.

[42] Gerhart J. From Feedback Inhibition to Allostery: The Enduring Example of Aspartate Transcarbamoylase. FEBS J. 2014;281(2):612–620. doi:10.1111/febs.12483.

[43] Lipscomb WN, Kantrowitz ER. Structure and Mechanisms of *Escherichia coli* Aspartate Transcarbamoylase. Acc. Chem. Res. 2012;45(3):444–53. doi:10.1021/ar200166p.

[44] Weber K. New Structural Model of *E. coli* Aspartate Transcarbamylase and the Amino-Acid Sequence of the Regulatory Polypeptide Chain. Nature. 1968;218(5147):1116–1119. doi:10.1038/2181116a0.

[45] Wiley DC, Lipscomb WN. Crystallographic Determination of Symmetry of Aspartate Transcarbamylase. Nature. 1968;218(5147):1119–1121. doi:10.1038/2181119a0.

[46] Yokota E, Riggs AF. The Structure of the Hemocyanin from the Horseshoe Crab, Limulus Polyphemus. The Amino Acid Sequence of the Largest Cyanogen Bromide Fragment. J. Biol. Chem. 1984;259(8):4739–4749.

[47] Taylor AB, Hu G, Hart PJ, McAlister-Henn L. Allosteric Motions in Structures of Yeast NAD+-Specific Isocitrate Dehydrogenase. J. Biol. Chem. 2008;283(16):10872–10880. doi:10.1074/jbc.M708719200.

[48] Giroux E, Williams MK, Kantrowitz ER. Shared Active Sites of Fructose-1,6-bisphosphatase. Arginine 243 Mediates Substrate Binding and Fructose 2,6-Bisphosphate Inhibition. J. Biol. Chem. 1994;269(50):31404–9.

[49] Cornish-Bowden A. A Simple Graphical Method for Determining the Inhibition Constants of Mixed, Uncompetitive and Non-Competitive Inhibitors. Biochem. J. 1974;137(1):143–4.

[50] Berg JM, Tymoczko JL, Stryer L. *Enzymes Can Be Inhibited by Specific Molecules*. W H Freeman, 5th edition, 2002.

[51] Li C, Begum A, Numao S, Park KH, Withers SG, Brayer GD. Acarbose Rearrangement Mechanism Implied by the Kinetic and Structural Analysis of Human Pancreatic Alpha-Amylase in Complex with Analogues and Their Elongated Counterparts. Biochem. 2005;44(9):3347–57. doi:10.1021/bi048334e.

[52] Feller G, Bussy OL, Houssier C, Gerday C. Structural and Functional Aspects of Chloride Binding to Alteromonas Haloplanctis Alpha-Amylase. J. Biol. Chem. 1996;271(39):23836–23841. doi:10.1074/jbc.271.39.23836.

[53] Kaiser PM. Substrate Inhibition as a Problem of Non-Linear Steady State Kinetics with Monomeric Enzymes. J. Mol. Catal. 1980;8(4):431–442. doi:10.1016/0304-5102(80)80082-4.

[54] Reed MC, Lieb A, Nijhout HF. The Biological Significance of Substrate Inhibition: A Mechanism with Diverse Functions. BioEssays : news and reviews in molecular, cellular and developmental biology. 2010;32(5):422–9. doi:10.1002/bies.200900167.

[55] Changeux JP. Responses of Acetylcholinesterase from Torpedo Marmorata to Salts and Curarizing Drugs. Mol. Pharmacol. 1966;2(5):369–92.

[56] Miller OJ, Harrak AE, Mangeat T, Baret JC, Frenz L, Debs BE, et al. High-Resolution Dose-Response Screening Using Droplet-Based Microfluidics. Proc. Natl. Acad. Sci. USA. 2011;109(2):378–383. doi:10.1073/Proc. Natl.Acad.Sci.USA.1113324109.

[57] Wang J, Stieglitz KA, Cardia JP, Kantrowitz ER. Structural Basis for Ordered Substrate Binding and Cooperativity in Aspartate Transcarbamoylase. Proc. Natl. Acad. Sci. USA. 2005;102(25):8881–6. doi:10.1073/Proc.Natl.Acad. Sci.USA.0503742102.

[58] Hsuanyu Y, Wedler FC. Kinetic Mechanism of Native *Escherichia coli* Aspartate Transcarbamylase. Arch. Biochem. Biophys. 1987;259(2):316–330. doi:10.1016/0003-9861(87)90498-X.

[59] Harris KM, Cockrell GM, Puleo DE, Kantrowitz ER. Crystallographic Snapshots of the Complete Catalytic Cycle of the Unregulated Aspartate Transcarbamoylase from Bacillus Subtilis. J. Mol. Biol. 2011;411(1):190–200. doi:10.1016/j.jmb.2011.05.036.

[60] Wang J, Eldo J, Kantrowitz ER. Structural Model of the R State of *Escherichia coli* Aspartate Transcarbamoylase with Substrates Bound. J. Mol. Biol. 2007; 371(5):1261–73. doi:10.1016/j.jmb.2007.06.011.

[61] Masson P, Schopfer LM, Bartels CF, Froment MT, Ribes F, Nachon F, et al. Substrate Activation in Acetylcholinesterase Induced by Low PH or Mutation in the π-Cation Subsite. Biochim. Biophys. Acta (BBA) - Protein Structure and Molecular Enzymology. 2002;1594(2):313–324. doi:10.1016/S0167-4838(01)00323-5.

[62] Otero LH, Beassoni PR, Boetsch C, Lisa AT, Domenech CE. Different Effects of Mg and Zn on the Two Sites for Alkylammonium Compounds in *Pseudomonas Aeruginosa* Phosphorylcholine Phosphatase. Enzyme. Res. 2011; 2011:918283. doi:10.4061/2011/918283.

[63] Penner PE, Cohen LH. Effects of Adenosine Triphosphate and Magnesium Ions on the Fumarase Reaction. J. Biol. Chem. 1969;244(4):1070–1075.

[64] Kantrowitz ER, Lipscomb WN. *Escherichia coli* Aspartate Transcarbamylase: The Relation Between Structure and Function. Science (New York, N.Y.). 1988;241(4866):669–674.

[65] Sprang SR, Acharya KR, Goldsmith EJ, Stuart DI, Varvill K, Fletterick RJ, et al. Structural Changes in Glycogen Phosphorylase Induced by Phosphorylation. Nature. 1988;336(6196):215–221.

[66] Boettcher AJ, Wu J, Kim C, Yang J, Bruystens J, Cheung N, et al. Realizing the Allosteric Potential of the Tetrameric Protein Kinase A RIα Holoenzyme. Structure. 2011;19(2):265–276.

[67] Gill SJ, Robert CH, Coletta M, Di Cera E, Brunori M. Cooperative Free Energies for Nested Allosteric Models as Applied to Human Hemoglobin. Biophys. J. 1986;50(4):747–752.

[68] Yifrach O, Horovitz A. Nested Cooperativity in the ATPase Activity of the Oligomeric Chaperonin GroEL. Biochem. 1995;34(16):5303–5308.

Chapter S1

SUPPLEMENTARY INFORMATION FOR STATISTICAL MECHANICS OF ALLOSTERIC ENZYMES

S1.1 Validity of Approximations

In section 1.3.2, we showed the generalization of the Michaelis-Menten model by granting the enzyme access to an active and inactive conformation. We then analyzed this system using two assumptions: the quasi-steady-state approximation Eq. 1.17 and the cycle condition Eq. 1.19. In this section, we will formally determine when these approximations are valid for an MWC enzyme and discuss what happens when we relax these assumptions. It is straightforward to extend these results to the more complicated MWC enzyme models where we introduce allosteric regulators, add competitive inhibitors, and consider enzymes with multiple binding sites.

S1.1.1 Definitions

In section 1.3.2, we characterized an MWC enzyme using the reaction scheme

$$\begin{tikzcd}[E_A] \arrow[r, shift left, "k_{on}^A[S]"] \arrow[d, shift right, "k_{trans}^A"'] & {[E_A S]} \arrow[l, shift left, "k_{off}^A + k_{cat}^A"] \arrow[d, shift right, "k_{trans}^{AS}"'] \\ {[E_I]} \arrow[r, shift left, "k_{on}^I[S]"] \arrow[u, shift right, "k_{trans}^I"'] & {[E_I S]} \arrow[l, shift left, "k_{off}^I + k_{cat}^I"] \arrow[u, shift right, "k_{trans}^{IS}"']\end{tikzcd} \tag{S1.1}$$

which we will now discuss in detail. We will use the following definitions freely [1]:

- An *edge* of a reaction scheme denotes the value of an arrow from one enzyme state to another. The edges on the left of S1.1 are k_{trans}^A (linking $[E_A]$ to $[E_I]$) and k_{trans}^I (linking $[E_I]$ to $[E_A]$).
- A *path* along enzyme states is the product of edges along this path. For example, the path from $[E_I]$ to $[E_A]$ to $[E_A S]$ for the MWC scheme above is given by $k_{trans}^I k_{on}^A [S]$.
- A system is in *steady state* if the concentration of every enzyme conformation does not change over time. For the scheme above, this implies $\frac{d[E_A S]}{dt} = \frac{d[E_A]}{dt} = \frac{d[E_I S]}{dt} = \frac{d[E_I]}{dt} = 0$.

- The *cycle condition* states that the product of edges going clockwise around any cycle must equal the product of edges going counterclockwise. For scheme S1.1, the product of clockwise edges $\left(k_{on}^{A}[S]\right)\left(k_{trans}^{AS}\right)\left(k_{off}^{I}+k_{cat}^{I}\right)\left(k_{trans}^{I}\right)$ equals the counter-clockwise product $\left(k_{off}^{A}+k_{cat}^{A}\right)\left(k_{trans}^{A}\right)\left(k_{on}^{I}[S]\right)\left(k_{trans}^{IS}\right)$.
- *Detailed balance* implies that the flow between two enzyme states is the same in the forward and backwards direction. For the scheme above, if the flow of enzymes from the $[E_A]$ state to the $[E_AS]$ state (given by $[E_A][S]k_{on}^{A}$) equals the flow from $[E_AS]$ to $[E_A]$ (given by $[E_AS]\left(k_{off}^{A}+k_{cat}^{A}\right)$) then the pair of edges between $[E_A]$ and $[E_AS]$ obeys detailed balance. A reaction scheme is in *equilibrium* if and only if every edge obeys detailed balance which occurs if and only if the system is in steady state and obeys the cycle condition.

S1.1.2 Cycle Condition

In this section, we consider why the cycle condition is necessary to ensure that a system in steady state is in equilibrium. Assume the MWC enzyme scheme S1.1 is in steady state,

$$\frac{d[E_AS]}{dt}=\frac{d[E_A]}{dt}=\frac{d[E_IS]}{dt}=\frac{d[E_I]}{dt}=0. \tag{S1.2}$$

The cycle condition ensures that equilibrium holds around the cycle in S1.1 regardless of which path is traversed. For example, suppose the system is in equilibrium and we want to use detailed balance to determine the relation between E_AS and E_I. Detailed balance provides a relation between adjacent vertices (i.e. any two enzyme states connected by arrows) such as E_AS and E_IS or E_IS and E_I. Hence we can find a relation between two non-adjacent edges such as E_AS and E_I by following two different paths,

$$\begin{tikzcd}
{[E_A]} \arrow[r, shift left, "k_{on}^{A}[S]"] \arrow[d, shift right, "k_{trans}^{A}"'] & {[E_AS]} \arrow[l, shift left, "k_{off}^{A}+k_{cat}^{A}"] \arrow[d, shift right, "k_{trans}^{AS}"'] \\
{[E_I]} \arrow[u, shift right, "k_{trans}^{I}"'] \arrow[r, shift left, "k_{on}^{I}[S]"] & {[E_IS].} \arrow[l, shift left, "k_{off}^{I}+k_{cat}^{I}"] \arrow[u, shift right, "k_{trans}^{IS}"']
\end{tikzcd} \tag{S1.3}$$

We could travel clockwise and follow the blue path around S1.3, first using detailed balance between E_AS and E_IS and then between E_IS and E_I,

$$\frac{[E_AS]}{[E_I]}=\frac{[E_AS]}{[E_IS]}\frac{[E_IS]}{[E_I]}=\frac{k_{trans}^{IS}}{k_{trans}^{AS}}\frac{k_{on}^{I}[S]}{k_{off}^{I}+k_{cat}^{I}}. \tag{S1.4}$$

On the other hand, we could have moved counter-clockwise around S1.3 along the orange path, first using the relationship between E_AS and E_A and then between E_A

and E_I,

$$\frac{[E_A S]}{[E_I]} = \frac{[E_A S]}{[E_A]}\frac{[E_A]}{[E_I]} = \frac{k_{on}^A [S]}{k_{off}^A + k_{cat}^A}\frac{k_{trans}^I}{k_{trans}^A}. \tag{S1.5}$$

Setting Eqs. S1.4 and S1.5 equal to each other yields the cycle condition!

S1.1.3 Quasi-Steady-State Approximation

We will now consider the dynamics of the MWC enzyme,

$$\begin{tikzcd}
{[E_A]+[S]} \arrow[r, shift left, "k_{on}^A"] \arrow[d, shift right, "k_{trans}^A"'] & {[E_A S]} \arrow[l, shift left, "k_{off}^A"] \arrow[r, "k_{cat}^A"] \arrow[d, shift right, "k_{trans}^{AS}"'] & {[E_A]+[P]} \\
{[E_I]+[S]} \arrow[r, shift left, "k_{on}^I"] \arrow[u, shift right, "k_{trans}^I"'] & {[E_I S]} \arrow[l, shift left, "k_{off}^I"] \arrow[r, "k_{cat}^I"] \arrow[u, shift right, "k_{trans}^{IS}"'] & {[E_I]+[P].}
\end{tikzcd} \tag{S1.6}$$

At time $t = 0$, the enzyme and substrate are mixed together and the rate of product formation is measured over time. The system starts off with all enzymes in the unbound forms E_A or E_I and there are no enzyme-substrate complexes $E_A S$ or $E_I S$.

To gain some intuition into this system, we first consider Fig. S1.1 which shows how this MWC enzyme can behave over time for reasonable parameter values. On the long time scales in Fig. S1.1B, the substrate concentration will appreciably diminish to $1/e$ of its original value after a long time τ_S. On the other hand, Fig. S1.1A shows that within a time $\tau_E \ll \tau_S$ the enzymes reach $1/e$ of what appears to be a "steady state." Of course, this is not a true steady-state, since after a time τ_S the substrate concentration will appreciably decrease and the enzyme conformations will correspondingly change. Instead, we call the situation after one second a quasi-steady-state, meaning that the enzyme conformations have all reached a steady-state value *assuming the current substrate concentration is fixed.*

When τ_E is significantly smaller than τ_S (typically τ_E only needs to be roughly 100 times smaller than τ_S), the dynamics of the enzymes and substrate can be separated. In other words, we can assume that the fast step (where the enzymes equilibrate to the current concentration of substrate) happens instantly when considering the slow dynamics of the substrate concentration diminishing over time. This is the quasi-steady-state approximation that we formally made in Eq. 1.17 of section 1.3.2. We will next show what relationship between the rate constants must hold so that the quasi-steady-state approximation is valid.

We first calculate the time scale τ_E for the enzyme conformations to equilibrate. We will assume that the substrate concentration equals the constant value $[S_{tot}]$

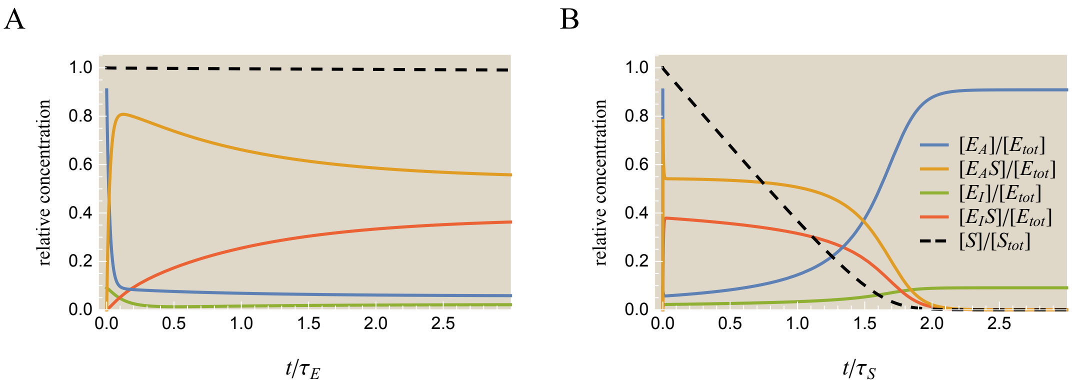


Figure S1.1: The quasi-steady-state approximation. (A) The fast dynamics of the system in Eq. S1.6 begins by mixing unbound enzymes (E_A and E_I) and substrate. The enzyme conformations quickly reach steady state on a time scale of $\tau_E \approx 0.04$ s. During this period, the substrate concentration remains very nearly constant. (B) The substrate changes appreciably over the much longer time scale $\tau_S \approx 11$ s. Over this longer time scale, we can assume the quasi-steady-state approximation: the enzymes conformations are always in quasi-steady-state with the slowly diminishing substrate concentration. Concentrations used were $[E_{tot}] = 1\,\mu$M, $[S_{tot}] = 1$ mM, $[E_AS] = [E_IS] = 0$, and $\frac{[E_A]}{[E_I]} = \frac{k^I_{trans}}{k^A_{trans}} \equiv e^{-\beta(\epsilon_A - \epsilon_I)}$. The rate constants used were $k^A_{on} = 1\,\text{s}^{-1}\text{M}^{-1}$, $k^I_{on} = 10^{-1}\,\text{s}^{-1}\text{M}^{-1}$, $k^A_{off} = 1\,\text{s}^{-1}$, $k^I_{off} = 10^{-3}\,\text{s}^{-1}$, $k^A_{cat} = 10^2\,\text{s}^{-1}$, $k^I_{cat} = 10\,\text{s}^{-1}$, $k^{AS}_{trans} = k^{IS}_{trans} = k^A_{trans} = 10\,\text{s}^{-1}$, and $k^I_{trans} = 10^2\,\text{s}^{-1}$.

throughout this short timescale (which, as shown in Fig. S1.1A, is reasonable) and then invoke a self-consistency condition to ensure that the actual change in substrate concentration during the period τ_E was negligible.

As a warm up, we first consider the Michaelis-Menten enzyme which we redraw here

$$[E]+[S] \underset{koff}{\overset{kon}{\rightleftarrows}} [ES] \xrightarrow{kcat} [E]+[P]. \tag{S1.7}$$

The Michaelis-Menten enzyme is governed by the multiple differential equations

$$\frac{d[E]}{dt} = [ES]\left(k_{off} + k_{cat}\right) - [E]\,[S_{tot}]\,k_{on} = -\frac{d[ES]}{dt} \tag{S1.8}$$

and the constraint $[E] + [ES] = [E_{tot}]$. As stated above, we fix the substrate concentration at $[S_{tot}]$ and assume that the system starts off with $[E] = [E_{tot}]$ and $[ES] = 0$. Solving the differential equation Eq. S1.8 yields

$$[E] = [E_{tot}]\,\frac{K_M + [S_{tot}]\,e^{-t/\tau}}{K_M + [S_{tot}]} \tag{S1.9}$$

$$[ES] = [E_{tot}]\,[S_{tot}]\frac{1 - e^{-t/\tau}}{K_M + [S_{tot}]} \tag{S1.10}$$

where $\tau = \frac{1}{k_{on}[S_{tot}]+k_{off}+k_{cat}}$ is the time scale for the system to equilibrate. Interestingly, $\frac{1}{\tau}$ equals the sum of all rates between the states $[E]$ and $[ES]$ (i.e. the sum of all time scales in this system). Furthermore, τ does not depend on the initial conditions of the system.

We now turn to the harder case of the MWC enzyme whose kinetics we describe using the scheme

$$\begin{tikzcd}
{[E_A]+[S]} \arrow[r, shift left, "k_{on}^{A}"] \arrow[d, shift right, "k_{trans}^{A}"'] & {[E_AS]} \arrow[l, shift left, "k_{off}^{A}"] \arrow[r, "k_{cat}^{A}"] \arrow[d, shift right, "k_{trans}^{AS}"'] & {[E_A]+[P]} \\
{[E_I]+[S]} \arrow[r, shift left, "k_{on}^{I}"] \arrow[u, shift right, "k_{trans}^{I}"'] & {[E_IS]} \arrow[l, shift left, "k_{off}^{I}"] \arrow[r, "k_{cat}^{I}"] \arrow[u, shift right, "k_{trans}^{IS}"'] & {[E_I]+[P].}
\end{tikzcd} \tag{S1.11}$$

As we just saw for the Michaelis-Menten enzyme, if we just considered any edge of the MWC enzyme separately, its corresponding time constant would be $\frac{1}{\text{sum of rates along this edge}}$: $\frac{1}{k_{on}^{A}[S_{tot}]+k_{off}^{A}+k_{cat}^{A}}$ between $[E_A]$ and $[E_AS]$ (blue); $\frac{1}{k_{trans}^{A}+k_{trans}^{I}}$ between $[E_A]$ and $[E_I]$ (red); $\frac{1}{k_{on}^{I}[S_{tot}]+k_{off}^{I}+k_{cat}^{I}}$ between $[E_I]$ and $[E_IS]$ (green); and $\frac{1}{k_{trans}^{AS}+k_{trans}^{IS}}$ between $[E_AS]$ and $[E_IS]$ (brown). We can approximate the time scale τ_E of this system as the maximum of these four time scales between adjacent edges,

$$\begin{aligned}
\tau_E &\approx \max\left(\frac{1}{k_{on}^{A}[S_{tot}]+k_{off}^{A}+k_{cat}^{A}}, \frac{1}{k_{trans}^{A}+k_{trans}^{I}}, \frac{1}{k_{on}^{I}[S_{tot}]+k_{off}^{I}+k_{cat}^{I}}, \frac{1}{k_{trans}^{AS}+k_{trans}^{IS}}\right) \\
&= \frac{1}{\min\left(k_{trans}^{A}+k_{trans}^{I}, k_{on}^{A}\left[S_{tot}\right]+k_{cat}^{A}+k_{off}^{A}, k_{trans}^{AS}+k_{trans}^{IS}, k_{on}^{I}\left[S_{tot}\right]+k_{cat}^{I}+k_{off}^{I}\right)}.
\end{aligned} \tag{S1.12}$$

This result is very similar (and in fact overestimates) the exact derivation of τ_E discussed in the next section, Appendix S1.1.4.

With this form of τ_E in hand, we could proceed in several ways to determine when the quasi-steady-state approximation holds. For example, we could compute the time scale τ_S for the substrate to diminish and then enforce $\tau_E \ll \tau_S$ as the quasi-steady-state approximation. However, Segel and Slemrod [2] determined a tighter constraint by demanding that the amount of substrate converted into product during the transient period $0 < t < \tau_E$ only amounts to a tiny fraction of the initial substrate concentration. The amount of substrate turned into product $\Delta[S]$ after time τ_E can be overestimated as

$$\Delta[S] \approx \left|\frac{d[S]}{dt}\right|_{\max} \tau_E \tag{S1.13}$$

so that the quasi-steady-state approximation can be written as

$$\frac{\Delta[S]}{[S_{tot}]} \approx \frac{1}{[S_{tot}]}\left|\frac{d[S]}{dt}\right|_{\max} \tau_E \ll 1. \tag{S1.14}$$

From S1.6, the rate of change of substrate concentration for the MWC enzyme is

$$\frac{d[S]}{dt} = -[E_A]\,[S]k_{on}^A - [E_I]\,[S]k_{on}^I + [E_AS]\,k_{off}^A + [E_IS]\,k_{off}^I. \tag{S1.15}$$

Recall that at $t = 0$, the system starts off with all enzymes unbound: $[E_AS] = [E_IS] = 0$ and $[E_A] + [E_I] = [E_{tot}]$. Then $\left|\frac{d[S]}{dt}\right|_{\max}$ occurs at $t = 0$ (when $[S] = [S_{tot}]$) and an upper bound is given by

$$\left|\frac{d[S]}{dt}\right|_{\max} = [S_{tot}]\left([E_A]\,k_{on}^A + [E_I]\,k_{on}^I\right) \leq [E_{tot}]\,[S_{tot}]\max\left(k_{on}^A, k_{on}^I\right). \tag{S1.16}$$

Substituting this result and the time scale Eq. S1.12 into Eq. S1.14, we find a sufficient condition for the quasi-steady state approximation to hold for an MWC enzyme:

$$[E_{tot}]\frac{\max\left(k_{on}^A, k_{on}^I\right)}{\min\left(k_{trans}^A + k_{trans}^I, k_{on}^A\,[S_{tot}] + k_{cat}^A + k_{off}^A, k_{trans}^{AS} + k_{trans}^{IS}, k_{on}^I\,[S_{tot}] + k_{cat}^I + k_{off}^I\right)} \ll 1. \tag{S1.17}$$

We could repeat this analysis for a Michaelis-Menten enzyme where only the E_A and E_AS states exist. This is equivalent to disregarding all terms except for k_{on}^A, k_{off}^A, and k_{cat}^A in the max and min of Eq. S1.17, so that the quasi-steady-state conditions reduces to $[E_{tot}]\frac{k_{on}^A}{k_{on}^A[S_{tot}]+k_{cat}^A+k_{off}^A} = \frac{[E_{tot}]}{[S_{tot}]+K_M^A} \ll 1$ which is identical to the condition found by Segel [2].

S1.1.4 Time Constants for the Quasi-Steady-State Approximation

In this section, we derive an exact expression for the time constant for which the MWC enzyme 1.16 will attain its steady state for each enzyme conformation assuming that the substrate concentration $[S] = [S_{tot}]$ remains fixed. The rate of change of each enzyme conformation can be written in matrix form (with bold denoting vectors and matrices) as

$$\frac{d\boldsymbol{E}}{dt} = \boldsymbol{K}\boldsymbol{E} \tag{S1.18}$$

where

$$\boldsymbol{K} = \begin{pmatrix} -k_{cat}^{A} - k_{off}^{A} - k_{trans}^{AS} & k_{on}^{A}\,[S_{tot}] & k_{trans}^{IS} & 0 \\ k_{cat}^{A} + k_{off}^{A} & -k_{on}^{A}\,[S_{tot}] - k_{trans}^{A} & 0 & k_{trans}^{I} \\ k_{trans}^{AS} & 0 & -k_{cat}^{I} - k_{off}^{I} - k_{trans}^{IS} & k_{on}^{I}\,[S_{tot}] \\ 0 & k_{trans}^{A} & k_{cat}^{I} + k_{off}^{I} & -k_{trans}^{I} - k_{on}^{I}\,[S_{tot}] \end{pmatrix}, \boldsymbol{E} = \begin{pmatrix} [E_A S] \\ [E_A] \\ [E_I S] \\ [E_I] \end{pmatrix}. \tag{S1.19}$$

This matrix can be decomposed as

$$\boldsymbol{K} = \boldsymbol{V}^{-1}\boldsymbol{\Lambda}\boldsymbol{V} \tag{S1.20}$$

where $\boldsymbol{V}$'s columns are the eigenvectors of $\boldsymbol{K}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix whose entries are the eigenvalues of $\boldsymbol{K}$. In general, it is known that the eigenvalues of such a matrix $\boldsymbol{K}$ representing the dynamics of any graph such as 1.16 from the text has one eigenvalue that is 0 while the remaining eigenvalues are non-zero and have negative real parts [3]. (Indeed, because all of the columns of $\boldsymbol{K}$ add up to zero, $\boldsymbol{K}$ is not full rank and hence one of its eigenvalues must be zero.) Defining the vector

$$\tilde{\boldsymbol{E}} \equiv \boldsymbol{V}\boldsymbol{E} = \begin{pmatrix} \tilde{E}_1 \\ \tilde{E}_2 \\ \tilde{E}_3 \\ \tilde{E}_4 \end{pmatrix}, \tag{S1.21}$$

Eq. S1.18 can be rewritten as

$$\frac{d\tilde{\boldsymbol{E}}}{dt} = \boldsymbol{\Lambda}\tilde{\boldsymbol{E}}. \tag{S1.22}$$

If the eigenvalues of Λ are λ_1, λ_2, λ_3, and 0, then $\tilde{E}_j = c_j e^{\lambda_j t}$ for $j = 1, 2, 3$ and $\tilde{E}_4 = c_4$ where the c_j's are constants determined by initial conditions. Since the $\tilde{E}_j$'s are linear combinations of $[E_A S], [E_A], [E_I S]$, and $[E_I]$, this implies that the $-\frac{1}{\lambda_1}$,$-\frac{1}{\lambda_2}$, and $-\frac{1}{\lambda_3}$ (or $-\frac{1}{\Re(\lambda_j)}$ if the eigenvalues are complex) are the time scales for the system to come to equilibrium. Therefore, we can compute the overall time scale for the system to come to equilibrium as

$$\tau_E^{(exact)} = \max\left(-\frac{1}{\lambda_1}, -\frac{1}{\lambda_2}, -\frac{1}{\lambda_3}\right). \tag{S1.23}$$

Although the eigenvalues of this matrix can be calculated in closed form, they are long and complicated expressions that contribute less intuition than the approximation

$$\tau_E = \max\left(\frac{1}{k_{on}^{A}[S] + k_{off}^{A} + k_{cat}^{A}}, \frac{1}{k_{trans}^{A} + k_{trans}^{I}}, \frac{1}{k_{on}^{I}[S] + k_{off}^{I} + k_{cat}^{I}}, \frac{1}{k_{trans}^{AS} + k_{trans}^{IS}}\right) \tag{S1.24}$$

used in Eq. S1.12 in the text. However, given the exact form, we can compare how well our approximation Eq. S1.24 matches the exact form Eq. S1.23.

When the four time scales in Eq. S1.24 are comparable to each other, the approximation is very close to the exact form. However, when at least one pair of edges in the MWC enzyme rates diagram,

$$\begin{tikzcd}[E_A] \arrow[r, shift left, "k_{on}^A[S]"] \arrow[d, shift right, "k_{trans}^A"'] & {[E_AS]} \arrow[l, shift left, "k_{off}^A+k_{cat}^A"] \arrow[d, shift right, "k_{trans}^{AS}"'] \\ {[E_I]} \arrow[r, shift left, "k_{on}^I[S]"] \arrow[u, shift right, "k_{trans}^I"'] & {[E_IS],} \arrow[l, shift left, "k_{off}^I+k_{cat}^I"] \arrow[u, shift right, "k_{trans}^{IS}"']\end{tikzcd} \tag{S1.25}$$

is very small the approximation tends to overshoot the exact value of τ_E. For example, if $k_{trans}^A \approx k_{trans}^I \approx 0$), Eq. S1.24 implies $\tau_E \to \infty$ whereas Eq. S1.23 can remain finite.

S1.1.5 Generalizing the Cycle Condition

We now consider what happens if an enzyme does not obey the cycle condition. Provided that the quasi-steady-state approximation holds, then on the long time scales the enzyme conformations quickly equilibrate to the current substrate concentration. From S1.6, the rate of change of each enzyme species obeys

$$\frac{d\,[E_AS]}{dt} = 0 = [E_A]\,[S]k_{on}^A - [E_AS]\left(k_{off}^A + k_{cat}^A + k_{trans}^{AS}\right) + [E_IS]\,k_{trans}^{IS} \tag{S1.26}$$

$$\frac{d\,[E_A]}{dt} = 0 = [E_AS]\left(k_{off}^A + k_{cat}^A\right) - [E_A]\,[S]k_{on}^A - [E_A]\,k_{trans}^A + [E_I]\,k_{trans}^I \tag{S1.27}$$

$$\frac{d\,[E_IS]}{dt} = 0 = [E_I]\,[S]k_{on}^I - [E_IS]\left(k_{off}^I + k_{cat}^I + k_{trans}^{IS}\right) + [E_AS]\,k_{trans}^{AS} \tag{S1.28}$$

$$\frac{d\,[E_I]}{dt} = 0 = [E_IS]\left(k_{off}^I + k_{cat}^I\right) - [E_I]\,[S]k_{on}^I - [E_I]\,k_{trans}^I + [E_A]\,k_{trans}^A. \tag{S1.29}$$

This system of equations, together with the conservation of total enzyme, $[E_{tot}] = [E_AS] + [E_IS] + [E_A] + [E_I]$, can be solved to obtain the quasi-steady-state values of each enzyme species. Using the Michaelis constants $K_M^A = \frac{k_{off}^A + k_{cat}^A}{k_{on}^A}$ and $K_M^I =$

$\frac{k_{off}^I + k_{cat}^I}{k_{on}^I}$, we can write the solutions as the three ratios

$$\frac{[E_A S]}{[E_A]} = \frac{[S]}{K_M^A} \frac{\left(K_M^I k_{on}^I + k_{trans}^I \gamma + [S] k_{on}^I \gamma\right) + k_{trans}^A \alpha\gamma}{\left(K_M^I k_{on}^I + k_{trans}^I \gamma + [S] k_{on}^I \gamma\right) + k_{trans}^A \alpha\beta\delta} \tag{S1.30}$$

$$\frac{[E_I S]}{[E_I]} = \frac{[S]}{K_M^I} \frac{\left(K_M^A k_{on}^A + k_{trans}^A \delta + [S] k_{on}^A \delta\right) + k_{trans}^I \frac{\delta}{\alpha}}{\left(K_M^A k_{on}^A + k_{trans}^A \delta + [S] k_{on}^A \delta\right) + k_{trans}^I \frac{\gamma}{\alpha\beta}} \tag{S1.31}$$

$$\frac{[E_A]}{[E_I]} = \frac{k_{trans}^I}{k_{trans}^A} \frac{\left(K_M^I k_{on}^I + k_{trans}^I \gamma + k_{trans}^A \alpha\beta\delta\right) + [S] k_{on}^I \gamma}{\left(K_M^I k_{on}^I + k_{trans}^I \gamma + k_{trans}^A \alpha\beta\delta\right) + [S] k_{on}^I \beta\delta} \tag{S1.32}$$

where we have defined $\alpha \equiv \frac{k_{on}^I}{k_{on}^A}$, $\beta \equiv \frac{K_M^I}{K_M^A}$, $\gamma \equiv \frac{k_{trans}^{IS}}{k_{trans}^I}$, $\delta \equiv \frac{k_{trans}^{AS}}{k_{trans}^A}$ to simplify the results. Notice that the terms in parenthesis in the numerator and denominator of these three ratios are the same. Indeed, the large fractions in all three equations equal 1 if we set $\gamma = \beta\delta$ so that

$$\frac{[E_A S]}{[E_A]} = \frac{[S]}{K_M^A} \tag{S1.33}$$

$$\frac{[E_I S]}{[E_I]} = \frac{[S]}{K_M^I} \tag{S1.34}$$

$$\frac{[E_A]}{[E_I]} = \frac{k_{trans}^I}{k_{trans}^A}. \tag{S1.35}$$

This fortuitous choice of γ is equivalent to the cycle condition Eq. 1.19, and so it is no surprise that these three ratios match Eqs. 1.20-1.22.

Invoking the cycle condition is a theoretical convenience which greatly simplifies our equations. If the cycle condition does not hold, we can follow our same procedure to turn Eqs. S1.30-S1.32 into a more general result for states and weights by only assuming the quasi-steady-state approximation. While this more general procedure is straightforward to implement numerically, it comes at the cost of introducing more parameters into the model (for example, values for k_{on}^I and k_{trans}^I must now be explicitly given whereas before we only needed to determine the ratios $\frac{k_{on}^I}{k_{off}^I + k_{cat}^I}$ and $\frac{k_{trans}^I}{k_{trans}^A}$) and the parameters will now depend upon the substrate concentration.

Finally, we note that the cycle condition need not be invoked if a model does not contain any cycles. In other words, if we instead defined an MWC enzyme using the rates diagram

$$\begin{tikzcd}
{[E_A]+[S]} \arrow[r, shift left, "k_{on}^{A}"] \arrow[d, shift right, "k_{trans}^{A}"'] & {[E_AS]} \arrow[l, shift left, "k_{off}^{A}"] \arrow[r, "k_{cat}^{A}"] & {[E_A]+[P]} \\
{[E_I]+[S]} \arrow[r, shift left, "k_{on}^{I}"] \arrow[u, shift right, "k_{trans}^{I}"'] & {[E_IS]} \arrow[l, shift left, "k_{off}^{I}"] \arrow[r, "k_{cat}^{I}"] & {[E_I]+[P],}
\end{tikzcd} \tag{S1.36}$$

our analysis would proceed identically without needing to invoke the cycle condition. Therefore, the cycle condition ensures that the system S1.6 has the right value of $\frac{k_{trans}^{AS}}{k_{trans}^{IS}}$ so that it can operate identically to S1.36.

S1.2 General Enzyme Models

In this section, we discuss the procedure used to fit the experimental enzyme kinetics data to the theoretical framework we have developed for allosteric enzymes. We then discuss the individual fits for each enzyme considered throughout the paper.

All fitting was done using nonlinear regression (NonlinearModelFit in *Mathematica*) using the realistic constraints $K_M, C_D, R_D \in [10^{-2}\mu\text{M}, 10^{6}\mu\text{M}]$, $k_{cat} \in [10^{-2}\,\text{s}^{-1}, 10^{5}\,\text{s}^{-1}]$, and $e^{-\beta(\epsilon_A-\epsilon_I)} \in [-10, 10]$ [4]. Initial conditions for the nonlinear regression were chosen randomly from this parameter space until a sufficiently good fit ($R^2 > 0.99$) was found. A notebook carrying out these calculations can be found in the supplement of the online publication.

It must be noted that, as with nearly all models, there are serious ambiguities in the best fit values since multiple sets of best fits values yield nearly identical curves. In point of fact, if the nonlinear regression would be performed without any constraints, it nearly always lands outside of the physically relevant parameter space (although the qualitative form of the best fit curves may be nearly indistinguishable from those that we show below). This attribute of models, dubbed as “sloppiness,” is well known [5]. One of its implications may be that a biological system can more easily evolve whichever activity profile it requires to maximize fitness, since the system is more likely the stumble across the best possible activity profile if it exists for numerous sets of parameters.

With this in mind, our results below demonstrate that our framework is *sufficient* to describe the complex interactions of allosteric enzymes, but that the individual parameter values (i.e. K_M, C_D, R_D values) are *not tightly determined* by these fits.

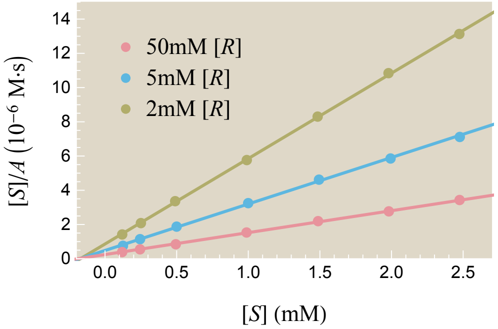


Figure S1.2: Theoretically and experimentally probing the effects of an allosteric regulator on activity. Data points show experimentally measured activity from Feller et al. for the enzyme α-amylase using substrate analog $[S]$ (EPS) and allosteric activator $[R]$ (NaCl), overlaid by theoretical curves of the form given in Eq. S1.37. Reproduced from Fig. 1.10 in the main text.

S1.2.1 Fitting α-Amylase and Allosteric Regulator Chlorine

Fig. S1.2 shows three activity curves for *A. haloplanctis* α-amylase titrating substrate at different concentrations of the allosteric activator NaCl. This enzyme has one substrate binding site and one allosteric site for binding chlorine ions. As discussed in section 1.4.1 of the main text, the $[S]/A$ curves are linear in $[S]$,

$$\frac{[S]}{A} = \frac{e^{-\beta\epsilon_A}\left(1+\frac{[S]}{K_M^A}\right)\left(1+\frac{[R]}{R_D^A}\right)+e^{-\beta\epsilon_I}\left(1+\frac{[S]}{K_M^I}\right)\left(1+\frac{[R]}{R_D^I}\right)}{k_{cat}^A e^{-\beta\epsilon_A}\frac{1}{K_M^A}\left(1+\frac{[R]}{R_D^A}\right)+k_{cat}^I e^{-\beta\epsilon_I}\frac{1}{K_M^I}\left(1+\frac{[R]}{R_D^I}\right)}. \tag{S1.37}$$

Note that we are fitting the 7 parameters from this equation into a linear form with 2 parameters (i.e. slope and intercept). Therefore, the individual parameters are not themselves reliable; instead, these fits are intended to show that the MWC model can account for the observed enzyme behavior. One possible set of parameters that matches the data is given by $e^{-\beta(\epsilon_A-\epsilon_I)} = 7.8\times10^{-4}$, $K_M^A = 0.6\,\text{mM}$, $K_M^I = 0.2\,\text{mM}$, $R_D^A = 0.03\,\text{mM}$, $R_D^I = 7.9\,\text{mM}$, $k_{cat}^A = 14\,\text{s}^{-1}$, and $k_{cat}^I = 0.01\,\text{s}^{-1}$. To find the value of an individual parameter, we would instead setup an experiment where only that single parameter varies and fit the resulting data.

S1.2.2 Fitting α-Amylase and Competitive Inhibitor Isoacarbose

Fig. S1.3 shows three activity curves of human pancreatic α-amylase titrating competitive inhibitor at different substrate concentrations. This enzyme has one active site which the substrate or competitive inhibitor can bind to. As discussed in section 1.4.1 of the main text, the activity curves all take the form

$$\left(\frac{d[P]}{dt}\right)^{-1} = \frac{1}{[E_{tot}]}\frac{e^{-\beta\epsilon_A}\left(1+\frac{[S]}{K_M^A}+\frac{[C]}{C_D^A}\right)+e^{-\beta\epsilon_I}\left(1+\frac{[S]}{K_M^I}+\frac{[C]}{C_D^I}\right)}{k_{cat}^A e^{-\beta\epsilon_A}\frac{[S]}{K_M^A}+k_{cat}^I e^{-\beta\epsilon_I}\frac{[S]}{K_M^I}} \tag{S1.38}$$

which is linear in $[C]$.

As noted above, in fitting 6 parameters to a linear form, the best fit parameter values are not reliable, but are only intended to show that the MWC model can account for the observed enzyme behavior. One possible set of parameters that matches the data is given by $e^{-\beta(\epsilon_A-\epsilon_I)} = 36$, $K_M^A = 0.9\,\text{mM}$, $K_M^I = 2.6\,\text{mM}$, $C_D^A = 12\,\text{nM}$, $C_D^I = 260\,\text{nM}$, and $\frac{k_{cat}^A}{k_{cat}^I} = 1.4$. Because units for activity were not included in original data, we instead fit the dimensionless quantity $[E_{tot}]k_{cat}^A\left(\frac{d[P]}{dt}\right)^{-1}$ which rescales the y-axis but does not change the form of the activity curves [7].

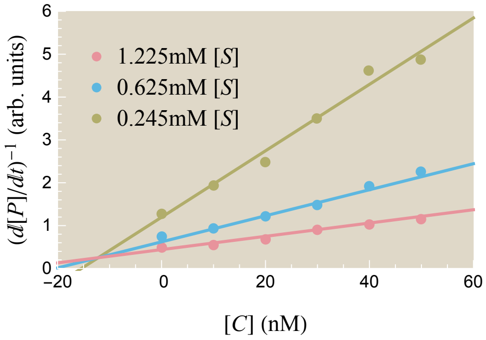


Figure S1.3: Theoretically and experimentally probing the effects of a competitive inhibitor on activity. Data points show experimentally measured activity in arbitrary units from Li et al. for the enzyme α-amylase using substrate analog $[S]$ (α-maltotriosyl fluoride) and competitive inhibitor $[C]$ (isoacarbose), overlaid by theoretical curves of the form given by Eq. S1.38. Best fit theoretical curves described by the inverse of Eq. 1.65 are overlaid on the data. Reproduced from Fig. 1.11(A) in the main text.

S1.2.3 Fitting Acetylcholinesterase Data

The acetylcholinesterase data in Fig. S1.4 was taken from *Torpedo marmorata* [8]. Using our framework from section 1.3.6, activity is given by

$$A = N\frac{e^{-\beta(\epsilon_A-\epsilon_I)}k_{cat}^A\frac{[S]}{K_M^A}\left(1+\frac{[S]}{K_M^A}\right)^{N-1}+k_{cat}^I\frac{[S]}{K_M^I}\left(1+\frac{[S]}{K_M^I}\right)^{N-1}}{e^{-\beta(\epsilon_A-\epsilon_I)}\left(1+\frac{[S]}{K_M^A}\right)^{N}+\left(1+\frac{[S]}{K_M^I}\right)^{N}} \tag{S1.39}$$

where $N = 2$ is the number of active sites.

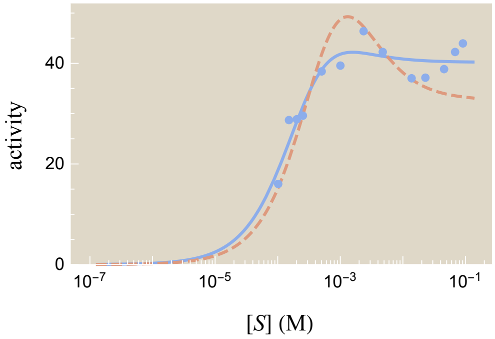


Figure S1.4: The activity of acetylcholinesterase exhibits a peak. The theoretical best-fit curve is shown (light blue) together with another theory curve which ignores the last three data points but better captures the height of the peak in the data (dashed, red).

Activity is shown in units of (nanomoles product)$\cdot$min$^{-1}\cdot$(mL enzyme)$^{-1}$. Using the density 3.6 $\frac{\text{mg}}{\text{mL}}$ and molecular weight 2.3×10^5 $\frac{\text{g}}{\text{mol}}$ of the enzyme [8], 1 mL enzyme = 1.6×10^{-8} mol. Therefore, 1 unit on the y-axis of the figure corresponds to 10^{-3} sec^{-1}.

The best fit parameters (light blue curve in Fig. S1.4) were $e^{-\beta(\epsilon_A-\epsilon_I)} = 0.5$, $K_M^A = 6.1 \times 10^{-3}$ M, $K_M^I = 2.8 \times 10^{-4}$ M, $k_{cat}^A = 3.1\,\text{s}^{-1}$, and $k_{cat}^I = 3.7 \times 10^{-2}\,\text{s}^{-1}$. The fitting is made difficult by two factors. First, the data points are not evenly spaced, and the three data points clumped together near $[S] = 2 \times 10^{-4}$ M have more weight on the fit than other points. Second, we suspect that the final three data points in this figure have a significant amount of error and should not curve back up – indeed, none of the other acetylcholinesterase substrate inhibition curves from the same source exhibit this feature [8]. To that end, we also show another theoretical curve (dashed, red) in order to exemplify that the MWC model can capture the height of the peak in the data. This latter curve has the parameters $e^{-\beta(\epsilon_A-\epsilon_I)} = 0.7$, $K_M^A = 7.4 \times 10^{-3}$ M, $K_M^I = 5.9 \times 10^{-4}$ M, $k_{cat}^A = 2.9\,\text{s}^{-1}$, and $k_{cat}^I = 2.0 \times 10^{-2}\,\text{s}^{-1}$.

S1.2.4 Further Example Data

In this section, we present data on ATCase (not discussed in the main text) which provides an excellent opportunity to combine all of the molecular players and enzyme features we have analyzed – allosteric regulators, competitive inhibitors, multiple substrate binding sites – in one complete model.

The ATCase data in Fig. S1.5 was taken from *Escherichia coli* [9]. ATCase is an allosteric enzyme with 6 active sites and 6 allosteric regulator sites. A competitive

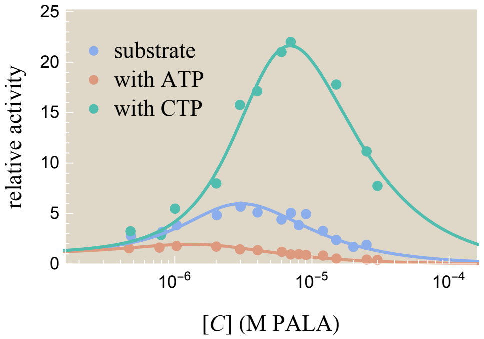


Figure S1.5: Inhibitor activation in aspartate carbamoyltransferase (ATCase). Activity curves from *E. coli* ATCase are shown in the absence (blue circles) and the presence of allosteric effectors, either the activator ATP (yellow squares) or the inhibitor CTP (green diamonds) as a function of the competitive inhibitor *N*-(phosphonacetyl)-L-aspartate (PALA). Data reproduced from Wales et al. and fit to an MWC model.

inhibitor PALA is titrated, and the experiment is then repeated in the presence of the allosteric activator ATP and the allosteric repressor CTP. Using our framework from section 1.3.6, the rate of product formation equals

$$\frac{d[P]}{dt} = N[E_{tot}]\frac{e^{-\beta(\epsilon_A-\epsilon_I)}k_{cat}^A\frac{[S]}{K_M^A}\left(1+\frac{[S]}{K_M^A}+\frac{[C]}{C_D^A}\right)^{N-1}+k_{cat}^I\frac{[S]}{K_M^I}\left(1+\frac{[S]}{K_M^I}+\frac{[C]}{C_D^I}\right)^{N-1}}{e^{-\beta(\epsilon_A-\epsilon_I)}\left(1+\frac{[S]}{K_M^A}+\frac{[C]}{C_D^A}\right)^{N}+\left(1+\frac{[S]}{K_M^I}+\frac{[C]}{C_D^I}\right)^{N}} \tag{S1.40}$$

where $N = 6$ is the number of active sites. The plot in Fig. S1.5 shows relative activity, which is defined as

$$\text{relative activity} = \frac{\frac{d[P]}{dt}}{\left(\frac{d[P]}{dt}\right)_{[C]\to 0}}. \tag{S1.41}$$

All three curves were carried out at a substrate concentration $[S] = 5\,\text{mM}$ of aspartate. In the absence of allosteric effectors (blue curve), the best fit parameters were $e^{-\beta(\epsilon_A-\epsilon_I)} = 0.005$, $K_M^A = 1.1\,\text{mM}$, $K_M^I = 1.8\,\text{mM}$, $k_{cat}^A = 400\,\text{s}^{-1}$, $k_{cat}^I = 0.02\,\text{s}^{-1}$, $C_D^A = 0.3\,\mu\text{M}$, and $C_D^I = 1.8\,\mu\text{M}$. As per the theoretical framework developed in section 1.3.3, an allosteric regulator such as ATP or CTP can be modeled by changing $e^{-\beta(\epsilon_A-\epsilon_I)} \to e^{-\beta(\epsilon_A-\epsilon_I)}\left(\frac{1+\frac{[R]}{R_D^A}}{1+\frac{[R]}{R_D^I}}\right)^N$ in Eq. S1.40. From [9], the concentrations of ATP (gold curve) and CTP (green curve) were $[R] = 2\,\text{mM}$. Using the same MWC parameters as in the blue curve, the best fit parameters for the allosteric activator ATP were $R_D^A = 0.07\,\text{mM}$ and $R_D^I = 0.10\,\text{mM}$; the best fit parameters for the allosteric inhibitor CTP were $R_D^A = 0.14\,\text{mM}$ and $R_D^I = 0.10\,\text{mM}$.

S1.3 Data Collapse

In this section, we analyze the concept of data collapse, which allows us to map the result of multiple activity curves onto a single curve using natural parameters of the system. In section S1.3.1, we start by reviewing the simplest case (presented in the main text) of an MWC enzyme with one active site in the presence of a competitive inhibitor. We show that such an enzyme admits a data collapse using a single parameter, so that all activity curves can be collapsed onto a single curve. In section S1.3.2, we next consider the simplest MWC enzyme in the presence of an allosteric regulator, with one active site and one allosteric site. This case requires two parameters for a data collapse, and we show the resulting collapse onto a sheet. We end with a general discussion of data collapse theory in section S1.3.3 which enables us to extend these results to more complex enzymes (e.g. enzymes with more catalytic sites in the presence of multiple species of allosteric regulators and competitive inhibitors).

S1.3.1 Special Case: Enzyme with 1 Active Site and a Competitive Inhibitor

We start with a recap of the data collapse (discussed in section 1.4.1) of an enzyme with a single active site in the presence of a competitive inhibitor whose states and weights diagram is redrawn in Fig. S1.6. The activity $A = \frac{1}{[E_{tot}]}\frac{d[P]}{dt}$ for such an enzyme is given by

$$A = \frac{k_{cat}^A e^{-\beta\Delta\epsilon}\frac{[S]}{K_M^A} + k_{cat}^I \frac{[S]}{K_M^I}}{e^{-\beta\Delta\epsilon}\left(1 + \frac{[S]}{K_M^A} + \frac{[C]}{C_D^A}\right) + \left(1 + \frac{[S]}{K_M^I} + \frac{[C]}{C_D^I}\right)} \tag{S1.42}$$

where $e^{-\beta\Delta\epsilon} = e^{-\beta(\epsilon_A - \epsilon_I)}$. Dividing the numerator and denominator by $e^{-\beta\Delta\epsilon}\left(1 + \frac{[C]}{C_D^A}\right) + \left(1 + \frac{[C]}{C_D^I}\right)$,

$$\begin{aligned} A &= \frac{k_{cat}^A\left(\frac{e^{-\beta\Delta\epsilon}\frac{[S]}{K_M^A}}{e^{-\beta\Delta\epsilon}\left(1+\frac{[C]}{C_D^A}\right)+\left(1+\frac{[C]}{C_D^I}\right)}\right) + k_{cat}^I\left(\frac{\frac{[S]}{K_M^I}}{e^{-\beta\Delta\epsilon}\left(1+\frac{[C]}{C_D^A}\right)+\left(1+\frac{[C]}{C_D^I}\right)}\right)}{\frac{e^{-\beta\Delta\epsilon}\frac{[S]}{K_M^A}}{e^{-\beta\Delta\epsilon}\left(1+\frac{[C]}{C_D^A}\right)+\left(1+\frac{[C]}{C_D^I}\right)} + \frac{\frac{[S]}{K_M^I}}{e^{-\beta\Delta\epsilon}\left(1+\frac{[C]}{C_D^A}\right)+\left(1+\frac{[C]}{C_D^I}\right)} + 1} \\ &= \frac{k_{cat}^A e^{-\beta\Delta F_{13}} + k_{cat}^I e^{-\beta\Delta F_{23}}}{e^{-\beta\Delta F_{13}} + e^{-\beta\Delta F_{23}} + 1} \end{aligned} \tag{S1.43}$$

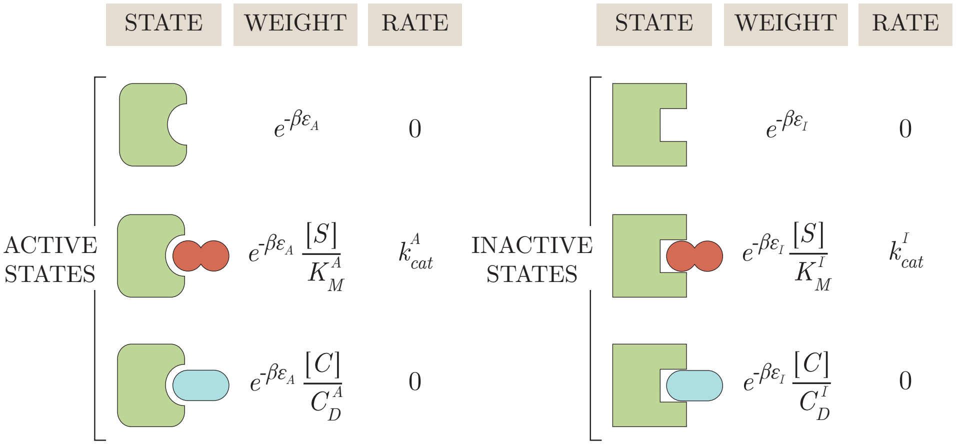

STATE
WEIGHT
RATE
STATE
WEIGHT
RATE
ACTIVE STATES
$e^{-\beta\varepsilon_A}$
0
$e^{-\beta\varepsilon_A}\frac{[S]}{K_M^A}$
k_{cat}^A
$e^{-\beta\varepsilon_A}\frac{[C]}{C_D^A}$
0
INACTIVE STATES
$e^{-\beta\varepsilon_I}$
0
$e^{-\beta\varepsilon_I}\frac{[S]}{K_M^I}$
k_{cat}^I
$e^{-\beta\varepsilon_I}\frac{[C]}{C_D^I}$
0

Figure S1.6: States and weights for an MWC enzyme with an allosteric regulator. Redrawn from Fig. 1.5 in the main text.

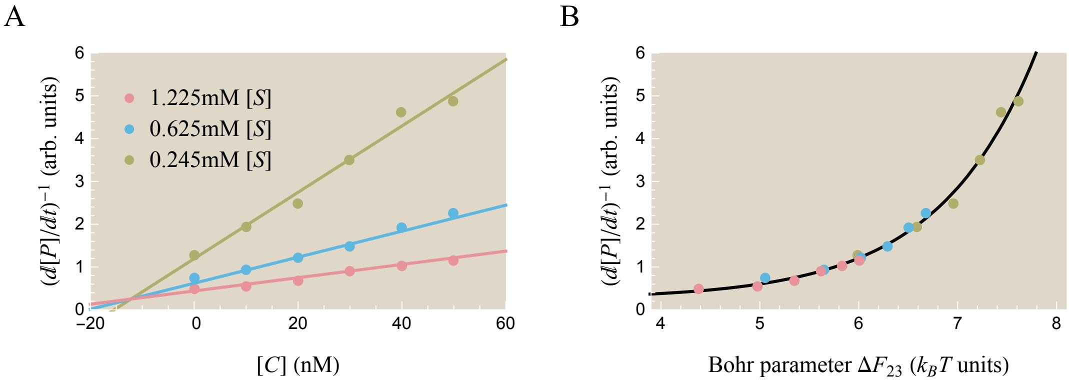


Figure S1.7: Data from Li et al. showing the effects of a competitive inhibitor C on the rate of product formation $\frac{d[P]}{dt}$. (A) Individual activity curves are shown at various concentrations of the substrate α-maltotriosyl fluoride (αG3F). (B) Curves are all data collapsed onto a single curve using the Bohr parameter ΔF_{23} from Eq. S1.49.

where we have defined the two *Bohr parameters*,

$$\Delta F_{13} = -\frac{1}{\beta}\text{Log}\left[\frac{e^{-\beta\Delta\epsilon}\frac{[S]}{K_M^A}}{e^{-\beta\Delta\epsilon}\left(1+\frac{[C]}{C_D^A}\right)+\left(1+\frac{[C]}{C_D^I}\right)}\right] \tag{S1.44}$$

$$\Delta F_{23} = -\frac{1}{\beta}\text{Log}\left[\frac{\frac{[S]}{K_M^I}}{e^{-\beta\Delta\epsilon}\left(1+\frac{[C]}{C_D^A}\right)+\left(1+\frac{[C]}{C_D^I}\right)}\right]. \tag{S1.45}$$

Because both ΔF_{13} and ΔF_{23} have the exact same dependence on $[S]$ and $[C]$, we can characterize the system by a single natural variable. For example, since

$$e^{-\beta \Delta F_{13}} = e^{-\beta \Delta \epsilon} \frac{K_M^I}{K_M^A} e^{-\beta \Delta F_{23}} \tag{S1.46}$$

we can rewrite Eq. 3.1 using only ΔF_{23},

$$A = \frac{k_{cat}^A e^{-\beta \Delta \epsilon} \frac{K_M^I}{K_M^A} e^{-\beta \Delta F_{23}} + k_{cat}^I e^{-\beta \Delta F_{23}}}{e^{-\beta \Delta \epsilon} \frac{K_M^I}{K_M^A} e^{-\beta \Delta F_{23}} + e^{-\beta \Delta F_{23}} + 1}. \tag{S1.47}$$

For cleanliness, we can group the constants using

$$K \equiv e^{-\beta \Delta \epsilon} \frac{K_M^I}{K_M^A}, \tag{S1.48}$$

so that the activity becomes

$$A = \frac{\left(k_{cat}^A K + k_{cat}^I\right) e^{-\beta \Delta F_{23}}}{(K+1) e^{-\beta \Delta F_{23}} + 1}, \tag{S1.49}$$

matching Eq. 1.66 from the text. As discussed in the text, this form allows us to map any number of activity curves onto a single curve of activity A versus the natural variable of the system ΔF_{23}. We redraw such a plot from the main text in Fig. S1.7.

S1.3.2 Special Case: Enzyme with 1 Active Site and an Allosteric Regulator

Consider an enzyme with one active site and one allosteric site in the presence of an allosteric regulator. The states and weights for such an enzyme are redrawn in Fig. S1.8. The activity of such an enzyme is given by

$$A = \frac{k_{cat}^A e^{-\beta \Delta \epsilon} \frac{[S]}{K_M^A} \left(1 + \frac{[R]}{R_D^A}\right) + k_{cat}^I \frac{[S]}{K_M^I} \left(1 + \frac{[R]}{R_D^I}\right)}{e^{-\beta \Delta \epsilon} \left(1 + \frac{[S]}{K_M^A}\right) \left(1 + \frac{[R]}{R_D^A}\right) + \left(1 + \frac{[S]}{K_M^I}\right) \left(1 + \frac{[R]}{R_D^I}\right)}. \tag{S1.50}$$

where $e^{-\beta \Delta \epsilon} = e^{-\beta (\epsilon_A - \epsilon_I)}$. We rewrite the numerator as

$$A = \frac{A_1 e^{-\beta \Delta \epsilon} \frac{[S]}{K_M^A} \left(1 + \frac{[R]}{R_D^A}\right) + A_2 \frac{[S]}{K_M^I} \left(1 + \frac{[R]}{R_D^I}\right)}{e^{-\beta \Delta \epsilon} \left(1 + \frac{[S]}{K_M^A}\right) \left(1 + \frac{[R]}{R_D^A}\right) + \left(1 + \frac{[S]}{K_M^I}\right) \left(1 + \frac{[R]}{R_D^I}\right)} \tag{S1.51}$$

where

$$A_1 = k_{cat}^A \tag{S1.52}$$

$$A_2 = k_{cat}^I. \tag{S1.53}$$

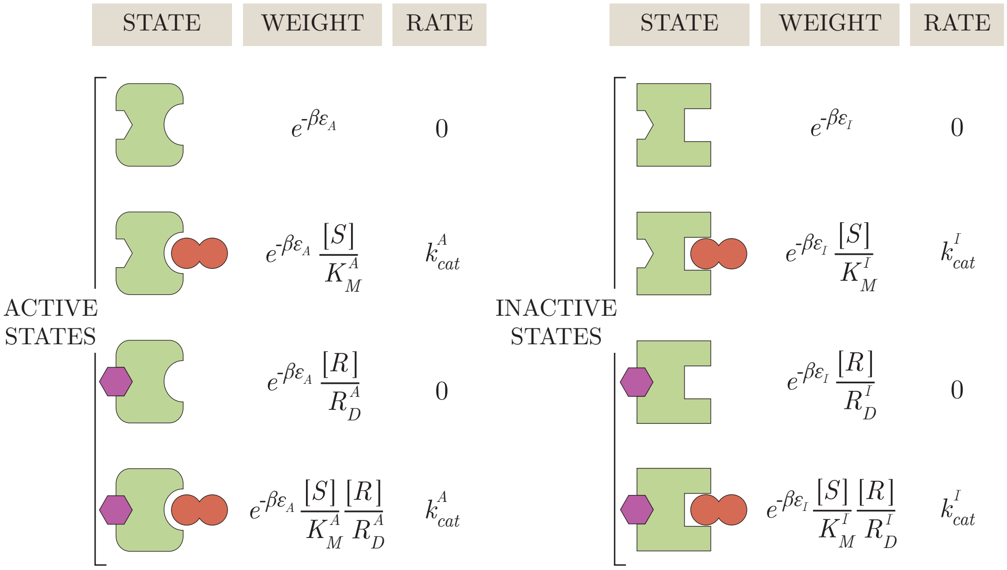

STATE
WEIGHT
RATE
STATE
WEIGHT
RATE
ACTIVE STATES
INACTIVE STATES
$e^{-\beta\varepsilon_A}$
0
$e^{-\beta\varepsilon_A}\frac{[S]}{K_M^A}$
k_{cat}^A
$e^{-\beta\varepsilon_A}\frac{[R]}{R_D^A}$
0
$e^{-\beta\varepsilon_A}\frac{[S]}{K_M^A}\frac{[R]}{R_D^A}$
k_{cat}^A
$e^{-\beta\varepsilon_I}$
0
$e^{-\beta\varepsilon_I}\frac{[S]}{K_M^I}$
k_{cat}^I
$e^{-\beta\varepsilon_I}\frac{[R]}{R_D^I}$
0
$e^{-\beta\varepsilon_I}\frac{[S]}{K_M^I}\frac{[R]}{R_D^I}$
k_{cat}^I

Figure S1.8: States and weights for an MWC enzyme with a competitive inhibitor. Redrawn from Fig. 1.7 in the main text.

Dividing the numerator and denominator by $\left(1+\frac{[R]}{R_D^A}\right)+\left(1+\frac{[R]}{R_D^I}\right)$, we can rewrite the activity using four natural variables,

$$A = \frac{A_1 e^{-\beta\Delta F_{13}} + A_2 e^{-\beta\Delta F_{23}}}{e^{-\beta\Delta F_{13}} + e^{-\beta\Delta F_{23}} + 1} \tag{S1.54}$$

where

$$\Delta F_{13} = -\frac{1}{\beta}\text{Log}\left[\frac{e^{-\beta\Delta\epsilon}\frac{[S]}{K_M^A}\left(1+\frac{[R]}{R_D^A}\right)}{e^{-\beta\Delta\epsilon}\left(1+\frac{[R]}{R_D^A}\right)+\left(1+\frac{[R]}{R_D^I}\right)}\right] \tag{S1.55}$$

$$\Delta F_{23} = -\frac{1}{\beta}\text{Log}\left[\frac{\frac{[S]}{K_M^I}\left(1+\frac{[R]}{R_D^I}\right)}{e^{-\beta\Delta\epsilon}\left(1+\frac{[R]}{R_D^A}\right)+\left(1+\frac{[R]}{R_D^I}\right)}\right]. \tag{S1.56}$$

In this case, the two natural variables have a fundamentally different dependence on $[R]$ and hence cannot be combined as in the case of a competitive inhibitor. With two parameters, any number of activity curves can be collapsed down upon a surface as shown in Fig. S1.9.

S1.3.3 General Theory

We now abstract the procedure used in the previous sections in order to understand how to obtain a data collapse for any enzyme system. Suppose we enumerate all of the states and weights of an enzyme, and that all of the states pooled together only

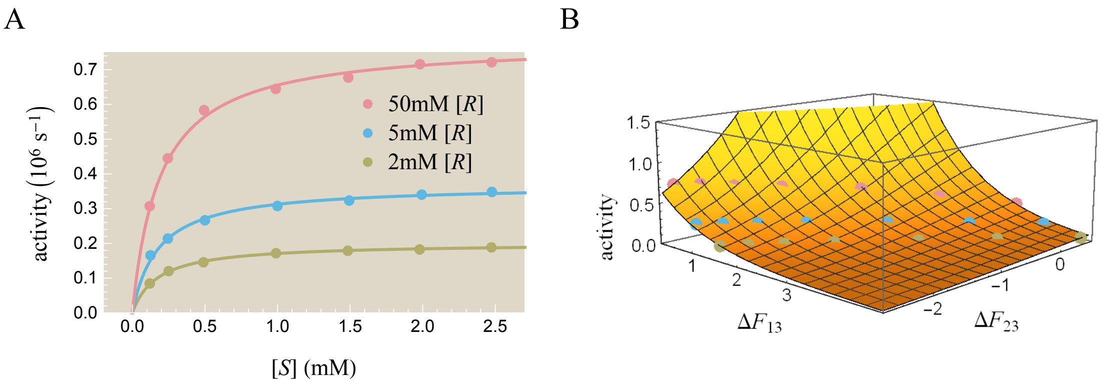


Figure S1.9: Data from Feller et al. demonstrating the rate of product formation $\frac{d[P]}{dt}$ in the presence of an allosteric activator NaCl. (A) Individual activity curves of α-amylase are shown at various concentrations of a substrate analog (EPS). Curves reproduced from Fig. 1.10 in main text but with the y-axis showing $\frac{d[P]}{dt}$ rather than $\frac{[S]}{d[P]/dt}$. (B) Curves are all data collapsed onto a surface using the Bohr parameters ΔF_{13} and ΔF_{23} from Eqs. S1.55 and S1.56.

have three distinct catalytic rates A_1, A_2, and A_3. (It is straightforward to generalize this argument to any number other than three.)

Define S_1, S_2, and S_3 to be the states that have catalytic rates A_1, A_2, and A_3. Then the activity of the enzyme is given by

$$A = \frac{A_1 \sum_{j \in S_1} e^{-\beta E_j} + A_2 \sum_{j \in S_2} e^{-\beta E_j} + A_3 \sum_{j \in S_3} e^{-\beta E_j}}{\sum_{j \in S_1} e^{-\beta E_j} + \sum_{j \in S_2} e^{-\beta E_j} + \sum_{j \in S_3} e^{-\beta E_j}}. \tag{S1.57}$$

Defining the free energies

$$e^{-\beta F_1} \equiv \sum_{j \in S_1} e^{-\beta E_j} \tag{S1.58}$$

$$e^{-\beta F_2} \equiv \sum_{j \in S_2} e^{-\beta E_j} \tag{S1.59}$$

$$e^{-\beta F_3} \equiv \sum_{j \in S_3} e^{-\beta E_j} \tag{S1.60}$$

allows us to rewrite the activity as

$$\begin{aligned} A &= \frac{A_1 e^{-\beta F_1} + A_2 e^{-\beta F_2} + A_3 e^{-\beta F_3}}{e^{-\beta F_1} + e^{-\beta F_2} + e^{-\beta F_3}} \\ &= \frac{A_1 e^{-\beta \Delta F_{13}} + A_2 e^{-\beta \Delta F_{23}} + A_3}{e^{-\beta \Delta F_{13}} + e^{-\beta \Delta F_{23}} + 1} \end{aligned} \tag{S1.61}$$

where $\Delta F_{13} \equiv F_1 - F_3$ and $\Delta F_{23} \equiv F_2 - F_3$ are the two minimal parameters defining the system. Here, we see explicitly that each Bohr parameter corresponds to a free energy difference between combinations of states with the same activity (hence the notation ΔF).

For example, in section S1.3.1 above, the activity of an enzyme with one active site in the presence of a competitive inhibitor is given by

$$A = \frac{k_{cat}^{A} e^{-\beta\Delta\epsilon} \frac{[S]}{K_M^A} + k_{cat}^{I} \frac{[S]}{K_M^I}}{e^{-\beta\Delta\epsilon}\left(1 + \frac{[S]}{K_M^A} + \frac{[C]}{C_D^A}\right) + \left(1 + \frac{[S]}{K_M^I} + \frac{[C]}{C_D^I}\right)}. \tag{S1.62}$$

To match the form of Eq. S1.61, we rewrite this equation as

$$A = \frac{k_{cat}^{A}\left(e^{-\beta\Delta\epsilon} \frac{[S]}{K_M^A}\right) + k_{cat}^{I}\left(\frac{[S]}{K_M^I}\right) + 0\left(e^{-\beta\Delta\epsilon}\left\{1 + \frac{[C]}{C_D^A}\right\} + \left\{1 + \frac{[C]}{C_D^I}\right\}\right)}{e^{-\beta\Delta\epsilon}\left(1 + \frac{[S]}{K_M^A} + \frac{[C]}{C_D^A}\right) + \left(1 + \frac{[S]}{K_M^I} + \frac{[C]}{C_D^I}\right)}, \tag{S1.63}$$

with $A_1 = k_{cat}^{A}$, $A_1 = k_{cat}^{I}$, and $A_3 = 0$. Dividing the numerator and denominator by $e^{-\beta\Delta\epsilon}\left(1 + \frac{[C]}{C_D^A}\right) + \left(1 + \frac{[C]}{C_D^I}\right)$ yields the data collapse equation

$$\begin{aligned} A &= \frac{k_{cat}^{A}\left(\frac{e^{-\beta\Delta\epsilon}\frac{[S]}{K_M^A}}{e^{-\beta\Delta\epsilon}\left(1+\frac{[C]}{C_D^A}\right)+\left(1+\frac{[C]}{C_D^I}\right)}\right) + k_{cat}^{I}\left(\frac{\frac{[S]}{K_M^I}}{e^{-\beta\Delta\epsilon}\left(1+\frac{[C]}{C_D^A}\right)+\left(1+\frac{[C]}{C_D^I}\right)}\right)}{\frac{e^{-\beta\Delta\epsilon}\frac{[S]}{K_M^A}}{e^{-\beta\Delta\epsilon}\left(1+\frac{[C]}{C_D^A}\right)+\left(1+\frac{[C]}{C_D^I}\right)} + \frac{\frac{[S]}{K_M^I}}{e^{-\beta\Delta\epsilon}\left(1+\frac{[C]}{C_D^A}\right)+\left(1+\frac{[C]}{C_D^I}\right)} + 1} \\ &= \frac{k_{cat}^{A} e^{-\beta\Delta F_{13}} + k_{cat}^{I} e^{-\beta\Delta F_{23}}}{e^{-\beta\Delta F_{13}} + e^{-\beta\Delta F_{23}} + 1} \end{aligned} \tag{S1.64}$$

with the two Bohr parameters

$$\Delta F_{13} = -\frac{1}{\beta}\text{Log}\left[\frac{e^{-\beta\Delta\epsilon}\frac{[S]}{K_M^A}}{e^{-\beta\Delta\epsilon}\left(1 + \frac{[C]}{C_D^A}\right) + \left(1 + \frac{[C]}{C_D^I}\right)}\right] \tag{S1.65}$$

$$\Delta F_{23} = -\frac{1}{\beta}\text{Log}\left[\frac{\frac{[S]}{K_M^I}}{e^{-\beta\Delta\epsilon}\left(1 + \frac{[C]}{C_D^A}\right) + \left(1 + \frac{[C]}{C_D^I}\right)}\right]. \tag{S1.66}$$

S1.4 Inhibitor Acceleration: ATCase

This section will examine the phenomenon of inhibitor acceleration. The analysis will closely follow section 1.4.2 in the text. We first demonstrate that inhibitor acceleration (having a peak in activity as a function of competitive inhibitor concentration) cannot occur for any enzyme with one active site and then show that it can occur for an MWC enzyme with two (or more) active sites.

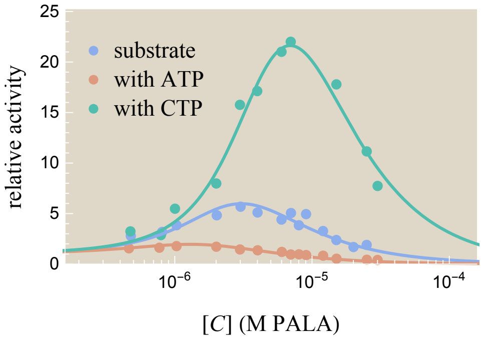


Figure S1.10: The activity of aspartate carbamoyltransferase (ATCase) exhibits a peak. Reproduced from Fig. S1.5.

S1.4.1 Inhibitor Acceleration Does Not Occur for an Enzyme with One Active Site

Consider an enzyme with a single active site in the presence of a competitive inhibitor, as in Fig. 1.7. We start by rewriting the activity for such an enzyme from Eq. 1.50,

$$A = \frac{1}{[E_{tot}]}\frac{d[P]}{dt} = \frac{k_{cat}^A e^{-\beta\epsilon_A}\frac{[S]}{K_M^A} + k_{cat}^I e^{-\beta\epsilon_I}\frac{[S]}{K_M^I}}{e^{-\beta\epsilon_A}\left(1+\frac{[S]}{K_M^A}+\frac{[C]}{C_D^A}\right)+e^{-\beta\epsilon_I}\left(1+\frac{[S]}{K_M^I}+\frac{[C]}{C_D^I}\right)}. \tag{S1.67}$$

The derivative of activity with respect to inhibitor concentration $[C]$ is given by

$$\frac{dA}{d[C]} = -\frac{\left(e^{-\beta\epsilon_A}\frac{1}{C_D^A}+e^{-\beta\epsilon_I}\frac{1}{C_D^I}\right)\left(e^{-\beta\epsilon_A}\frac{k_{cat}^A}{K_M^A}+e^{-\beta\epsilon_I}\frac{k_{cat}^I}{K_M^I}\right)[S]}{\left(e^{-\beta\epsilon_A}\left(1+\frac{[S]}{K_M^A}+\frac{[C]}{C_D^A}\right)+e^{-\beta\epsilon_I}\left(1+\frac{[S]}{K_M^I}+\frac{[C]}{C_D^I}\right)\right)^2}. \tag{S1.68}$$

Since the numerator cannot equal zero for any value of $[C]$, a peak cannot occur when the competitive inhibitor is added. Instead, $\frac{dA}{d[C]}$ is negative, indicating that adding more competitive inhibitor will decrease the activity, as is typically expected from an inhibitor.

S1.4.2 Inhibitor Acceleration for an Enzyme with Two Active Sites

Some allosteric enzymes exhibit an increase in activity when a small amount of competitive inhibitor C is introduced, as shown in Fig. S1.10. The simplest enzyme model which allows such a peak has two substrate binding sites and includes allostery. For simplicity, we work in the limit $k_{cat}^I = 0$. Combining the results from

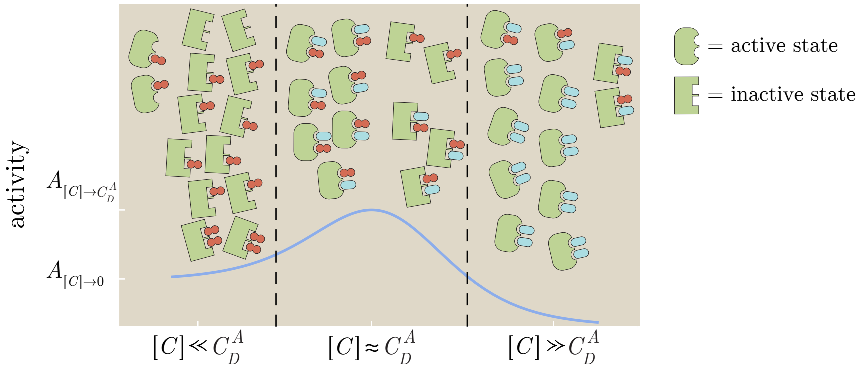


Figure S1.11: Mechanism underlying peak in activation by a competitive inhibitor $\boldsymbol{C}$. At low inhibitor concentrations, $[C] \ll C_D^A$, most enzymes are in the inactive form (sharp, green). As the amount of inhibitor increases, it will begin to compete with the substrate for active sites. At medium concentrations, $[C] \gg C_D^A$, some enzymes will have one site filled with a competitive inhibitor which prefers to bind in an active-state (rounded, green) enzyme complex. This increased probability of having active-state enzyme-substrate complexes (albeit with one enzyme site filled with an inhibitor) yields a larger activity compared to the low inhibitor concentrations. At large inhibitor concentrations, $[C] \gg C_D^A$, the inhibitor outcompetes the substrate for active sites and enzyme activity is suppressed.

sections 1.3.4 and 1.3.5, the activity for such an enzyme is given by

$$A = k_{cat}^A e^{-\beta\epsilon_A} \frac{2\frac{[S]}{K_M^A}\left(1 + \frac{[C]}{C_D^A} + \frac{[S]}{K_M^A}\right)}{e^{-\beta\epsilon_A}\left(1 + \frac{[S]}{K_M^A} + \frac{[C]}{C_D^A}\right)^2 + e^{-\beta\epsilon_I}\left(1 + \frac{[S]}{K_M^I} + \frac{[C]}{C_D^I}\right)^2}. \tag{S1.69}$$

A peak will occur provided that $\frac{dA}{d[C]} = 0$ for a positive value of $[C]$. For now, we skip the details of solving such a root (discussed in Appendix S1.5.2) and move straight to the results. Eq. S1.69 will have a positive root for $[C]$ provided the following relation holds,

$$e^{-\beta(\epsilon_A - \epsilon_I)} < \left(\frac{1 + \frac{[S]}{K_M^I}}{1 + \frac{[S]}{K_M^A}}\right)^2 - 2\frac{C_D^A}{C_D^I}\frac{1 + \frac{[S]}{K_M^I}}{1 + \frac{[S]}{K_M^A}} \qquad (k_{cat}^I = 0). \tag{S1.70}$$

Acceleration by an inhibitor has historically been explained by a competitive inhibitor binding to one active site of an enzyme, forcing it into the active state [10]. This is indeed part of the story. Consider an enzyme that natively favors the inactive state when no inhibitor is present, as shown in the $[C] \ll C_D^A$ region of Fig. S1.11. As $[C]$ increases, many enzymes will bind inhibitor in one active site, leaving the remaining active site free to bind substrate. If the inhibitor favors binding to the active-state

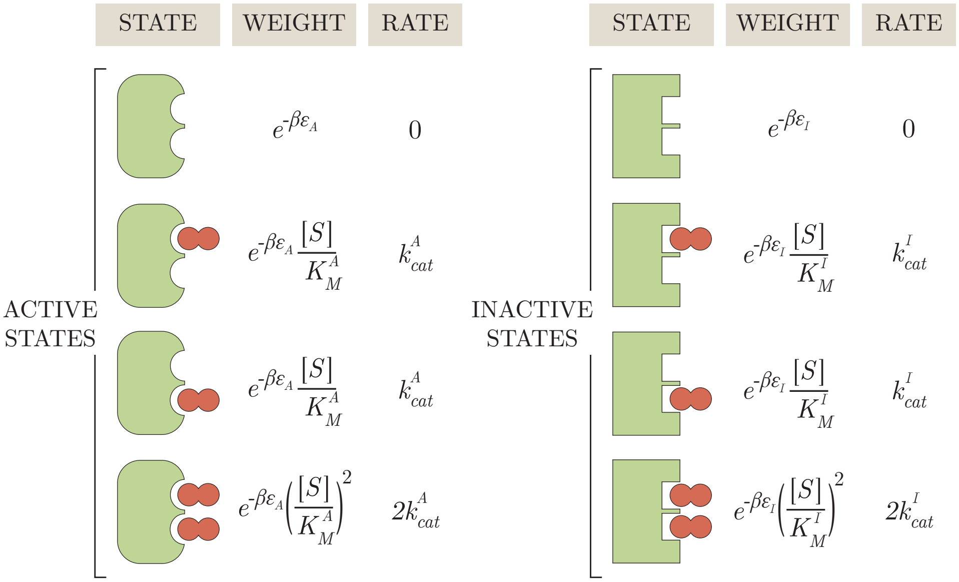

STATE
WEIGHT
RATE
STATE
WEIGHT
RATE
ACTIVE STATES
$e^{-\beta\varepsilon_A}$
0
$e^{-\beta\varepsilon_A}\frac{[S]}{K_M^A}$
k_{cat}^A
$e^{-\beta\varepsilon_A}\frac{[S]}{K_M^A}$
k_{cat}^A
$e^{-\beta\varepsilon_A}\left(\frac{[S]}{K_M^A}\right)^2$
$2k_{cat}^A$
INACTIVE STATES
$e^{-\beta\varepsilon_I}$
0
$e^{-\beta\varepsilon_I}\frac{[S]}{K_M^I}$
k_{cat}^I
$e^{-\beta\varepsilon_I}\frac{[S]}{K_M^I}$
k_{cat}^I
$e^{-\beta\varepsilon_I}\left(\frac{[S]}{K_M^I}\right)^2$
$2k_{cat}^I$

Figure S1.12: States and weights for an MWC enzyme with two substrate binding sites. Reproduced from Fig. 1.9.

enzyme, the ratio of active to inactive enzymes will increase which will generate a peak in activity. When $[C] \gg C_D^A$, the inhibitor will fill nearly all active sites and quash product formation. This story suggests that having a smaller $\frac{C_D^A}{C_D^I}$ value (i.e. having an inhibitor which strongly prefers binding to an active-state enzyme) will increase the likelihood of generating a peak. This is confirmed by the peak condition Eq. S1.70 where decreasing $\frac{C_D^A}{C_D^I}$ increases the right-hand side of the inequality.

However, the complete story behind activation by inhibitor is more nuanced. To gain some intuition, we first consider the limit $\frac{C_D^A}{C_D^I} \approx 0$ where the inhibitor binds exclusively to the active rather than the inactive state. This limit maximizes the right-hand side of Eq. S1.70 which we can rewrite as

$$e^{-\beta\epsilon_A}\left(1+\frac{[S]}{K_M^A}\right)^2 < e^{-\beta\epsilon_I}\left(1+\frac{[S]}{K_M^I}\right)^2 \qquad \left(k_{cat}^I = 0,\ \frac{C_D^A}{C_D^I} = 0\right). \qquad \text{(S1.71)}$$

This inequality tells us about the nature of the enzyme. Let us return momentarily to the states and weights of an allosteric enzyme with two substrate binding sites in the absence of competitive inhibitor which we reproduce here in Fig. S1.12. The total weights of the enzyme being in any active state is given by the sum of the weights in the left column,

$$w_A = e^{-\beta\epsilon_A} + e^{-\beta\epsilon_A}\frac{[S]}{K_M^A} + e^{-\beta\epsilon_A}\frac{[S]}{K_M^A} + e^{-\beta\epsilon_A}\left(\frac{[S]}{K_M^A}\right)^2 = e^{-\beta\epsilon_A}\left(1+\frac{[S]}{K_M^A}\right)^2. \quad \text{(S1.72)}$$

Similarly, the total weight of the enzyme being in any inactive state is given by

$$w_I = e^{-\beta\epsilon_I} + e^{-\beta\epsilon_I}\frac{[S]}{K_M^I} + e^{-\beta\epsilon_I}\frac{[S]}{K_M^I} + e^{-\beta\epsilon_I}\left(\frac{[S]}{K_M^I}\right)^2 = e^{-\beta\epsilon_I}\left(1+\frac{[S]}{K_M^I}\right)^2. \quad \text{(S1.73)}$$

Therefore, the relation Eq. S1.71 states that the total weight of the active states is smaller than the total weight of the inactive states, $w_A < w_I$, or equivalently that the enzyme (in the absence of a competitive inhibitor) is more likely to be in an inactive state.

We now return to the more general case when $\frac{C_D^A}{C_D^I} > 0$. Recall that as $\frac{C_D^A}{C_D^I}$ increases, so does the relative affinity of the competitive inhibitor to the inactive states over the active states. We can rewrite the peak condition when $\frac{C_D^A}{C_D^I} > 0$ from Eq. S1.70 as

$$e^{-\beta\epsilon_A}\left(1+\frac{[S]}{K_M^A}\right)^2 < e^{-\beta\epsilon_I}\left(1+\frac{[S]}{K_M^I}\right)^2 - \left\{2e^{-\beta\epsilon_I}\frac{C_D^A}{C_D^I}\left(1+\frac{[S]}{K_M^I}\right)\left(1+\frac{[S]}{K_M^A}\right)\right\} \qquad (k_{cat}^I = 0). \quad \text{(S1.74)}$$

The term in curly braces $\{\cdots\}$ on the right is positive and increases with $\frac{C_D^A}{C_D^I}$. Compared to the special case $\frac{C_D^A}{C_D^I} = 0$ in Eq. S1.71, an enzyme satisfying Eq. S1.74 must favor the inactive states over the active states to a greater extent. More formally, the maximal ratio $\frac{w_A}{w_I}$ of the active state weights to inactive state weights that permits a peak decreases as $\frac{C_D^A}{C_D^I}$ increases.

Second, consider the limit $C_D^A = C_D^I$ where the competitive inhibitor equally favors the active and inactive states. According to Eq. S1.70, a peak can still occur provided that

$$1 + e^{-\beta(\epsilon_A-\epsilon_I)} < \left(\frac{\frac{[S]}{K_M^A}}{1+\frac{[S]}{K_M^A}}\right)^2\left(\frac{K_M^A}{K_M^I} - 1\right)^2 \qquad (k_{cat}^I = 0,\ C_D^A = C_D^I). \quad \text{(S1.75)}$$

It may seem surprising that an inhibitor that binds equally well to the active and inactive enzyme states can increase the amount of active state enzymes as per Fig. S1.11. However, Eq. S1.71 shows that any enzyme that exhibits inhibitor acceleration must favor the inactive states more in the absence of inhibitor. Relative to this pool of enzyme which are mostly in the inactive states, the presence of an inhibitor with $C_D^A = C_D^I$ will increase the fraction of enzymes in the active states.

Finally, we consider the case where introducing a competitor keeps the same fraction of enzymes in the active and inactive states, and we expect that this case cannot generate a peak in activity. Drawing on the states and weights in Fig. 1.7 (but recalling that our enzyme has two active sites), the dissociation constants C_D^A and C_D^I of such a competitive inhibitor must satisfy

$$\frac{e^{-\beta\epsilon_A}\left(1+\frac{[S]}{K_M^A}+\frac{[C]}{C_D^A}\right)^2}{e^{-\beta\epsilon_I}\left(1+\frac{[S]}{K_M^I}+\frac{[C]}{C_D^I}\right)^2}=\frac{e^{-\beta\epsilon_A}\left(1+\frac{[S]}{K_M^A}\right)^2}{e^{-\beta\epsilon_I}\left(1+\frac{[S]}{K_M^I}\right)^2}. \tag{S1.76}$$

The only solution to this equation occurs when

$$\frac{C_D^A}{C_D^I}=\frac{1+\frac{[S]}{K_M^I}}{1+\frac{[S]}{K_M^A}}, \tag{S1.77}$$

which upon substitution into Eq. S1.70 yields the expected result that a peak cannot occur when the competitive inhibitor does not change the balance between the active and inactive states. One might expect that for all values of $\frac{C_D^A}{C_D^I}$ smaller than this (where the inhibitor does push more enzymes into the active state), a peak could occur. However, Eq. S1.70 indicates that a can only occur provided that a stronger constraint holds, namely

$$2\frac{C_D^A}{C_D^I}<\frac{1+\frac{[S]}{K_M^I}}{1+\frac{[S]}{K_M^A}}. \tag{S1.78}$$

Having analyzed these specific cases, we now turn to some general characteristics of this peak. Having calculated the concentration $[C]_0$ in Appendix S1.5.2 where the peak occurs, it is straightforward to compute the maximum height of the activity curve,

$$A_{peak}=k_{cat}^A\frac{[S]}{K_M^A}\frac{\left(\sqrt{\left(\frac{C_D^A}{C_D^I}\right)^2+e^{-\beta(\epsilon_A-\epsilon_I)}}-\frac{C_D^A}{C_D^I}\right)}{\left(1+\frac{[S]}{K_M^I}\right)-\frac{C_D^A}{C_D^I}\left(1+\frac{[S]}{K_M^A}\right)}. \tag{S1.79}$$

Substituting in the peak condition Eq. S1.70 we obtain

$$A_{peak}<k_{cat}^A\frac{\frac{[S]}{K_M^A}}{1+\frac{[S]}{K_M^A}}. \tag{S1.80}$$

The enzyme can approach the maximum possible activity k_{cat}^A in the limit $1 \ll \frac{[S]}{K_M^A}$ when the active state enzyme dominates, analogous to the result for substrate inhibition Eq. 1.73. We can also compare the peak height to the activity when no inhibitor is present,

$$A_{[C]\to 0}=2k_{cat}^A\frac{e^{-\beta(\epsilon_A-\epsilon_I)}\left(\frac{[S]}{K_M^A}+\frac{[S]}{K_M^A}^2\right)}{e^{-\beta(\epsilon_A-\epsilon_I)}\left(1+\frac{[S]}{K_M^A}\right)^2+\left(1+\frac{[S]}{K_M^I}\right)^2}. \tag{S1.81}$$

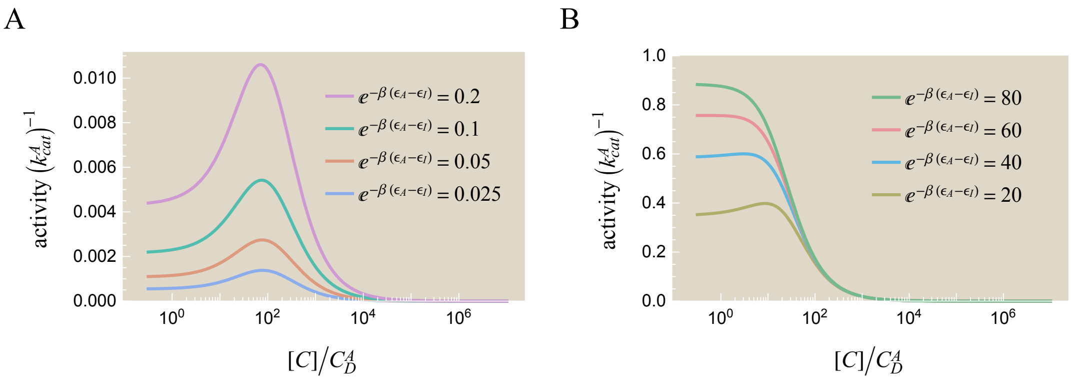


Figure S1.13: Peak in enzyme activity $A = \frac{1}{E_{tot}}\frac{d[P]}{dt}$ as a function of *competitive inhibitor* concentration $[C]$. As shown in Fig. 1.12B, with Michaelis-Menten kinetics adding a competitive inhibitor can only slow down activity, but an MWC enzyme can be activated by an inhibitor which results in a peak. Peak are shown for (A) small and (B) large ratios of the enzyme's energy in the active versus inactive state, $e^{-\beta(\epsilon_A-\epsilon_I)}$. As in the case of substrate inhibition, the height of the peak increases with $e^{-\beta(\epsilon_A-\epsilon_I)}$. The activity is computed from Eq. S1.69 using the parameters $\frac{[S]}{K_M^A} = 10$, $\frac{C_D^A}{C_D^I} = 10^{-2}$, the parameters from Fig. 1.13, and the different values of $e^{-\beta(\epsilon_A-\epsilon_I)}$ shown. As predicted by Eq. S1.70, for the parameters chosen every value in the range $e^{-\beta(\epsilon_A-\epsilon_I)} < 65$ will yield a peak in activity.

Examples of such peaks are shown in Fig. S1.13. As in the case of substrate inhibition, the peak height A_{peak} monotonically increases and the relative peak height $\frac{A_{peak}}{A_{[C]\to 0}}$ monotonically decreases with the energy difference between the active and inactive state, $e^{-\beta(\epsilon_A-\epsilon_I)}$.

The enzyme ATCase offers an example of inhibitor acceleration. ATCase is an allosteric enzyme with 6 active sites and 6 regulatory sites [11]. In the absence of ligand, ATCase exists in an equilibrium between the unbound active and unbound inactive states, the latter being more energetically favorable [12]. When the inhibitor PALA binds to ATCase, it strongly induces a transition from inactive to active state [13], in line with our theoretical prediction. It has been shown that by adding allosteric regulators, the peak in ATCase activity can be increased or prevented altogether [9]. It would be interesting to undertake the converse experiment and induce inhibitor activation in an enzyme that typically does not show a peak in activity.

S1.5 Derivations

S1.5.1 Substrate Inhibition

We now derive the general peak condition for substrate inhibition without the extra assumption $k_{cat}^{I} = 0$ used in the text. Recall that we define the active state of an enzyme as the state with the greater catalytic rate so that $k_{cat}^{A} > k_{cat}^{I}$. We start by rewriting the full form of the activity equation 1.70 from section 1.4.2.2,

$$A = \frac{2k_{cat}^{A} e^{-\beta\epsilon_A} \frac{[S]}{K_M^A}\left(1 + \frac{[S]}{K_M^A}\right) + 2k_{cat}^{I} e^{-\beta\epsilon_I} \frac{[S]}{K_M^I}\left(1 + \frac{[S]}{K_M^I}\right)}{e^{-\beta\epsilon_A}\left(1 + \frac{[S]}{K_M^A}\right)^2 + e^{-\beta\epsilon_I}\left(1 + \frac{[S]}{K_M^I}\right)^2}, \tag{S1.82}$$

we derive the peak condition Eq. 1.71. We define the numerator and denominator of the activity as

$$A \equiv \frac{Z_S}{Z_{tot}} \tag{S1.83}$$

where, from states and weights in Fig. 1.9,

$$Z_S = 2k_{cat}^{A} e^{-\beta\epsilon_A} \frac{[S]}{K_M^A}\left(1 + \frac{[S]}{K_M^A}\right) + 2k_{cat}^{I} e^{-\beta\epsilon_I} \frac{[S]}{K_M^I}\left(1 + \frac{[S]}{K_M^I}\right) \tag{S1.84}$$

is the sum of all weights multiplied by their rate of product formation and

$$Z_{tot} = e^{-\beta\epsilon_A}\left(1 + \frac{[S]}{K_M^A}\right)^2 + e^{-\beta\epsilon_I}\left(1 + \frac{[S]}{K_M^I}\right)^2 \tag{S1.85}$$

is the sum of all weights. By varying the substrate concentration $[S]$, we find a peak in the activity A provided that

$$\frac{dA}{d[S]} = \frac{\frac{dZ_S}{d[S]} Z_{tot} - Z_S \frac{dZ_{tot}}{d[S]}}{Z_{tot}^2} = 0. \tag{S1.86}$$

Thus, a peak occurs if the numerator $\frac{dZ_S}{d[S]} Z_{tot} - Z_S \frac{dZ_{tot}}{d[S]}$ equals zero. Because Z_S and Z_{tot} are quadratic in $[S]$, the terms $\frac{dZ_S}{d[S]} Z_{tot}$ and $Z_S \frac{dZ_{tot}}{d[S]}$ in the numerator are cubic in $[S]$. However, the cubic terms exactly cancel each other, so that Eq. S1.86 becomes a quadratic equation,

$$0 = \frac{dZ_S}{d[S]} Z_{tot} - Z_S \frac{dZ_{tot}}{d[S]} \equiv 2\left(K_M^A K_M^I\right)^4 \left(a[S]^2 + b[S] + c\right), \tag{S1.87}$$

where we have pulled out the prefactor $2\left(K_M^A K_M^I\right)^4$ for convenience and

$$a = \left(e^{-\beta\epsilon_A} + e^{-\beta\epsilon_I}\right)\left(\frac{e^{-\beta\epsilon_A}k_{cat}^A}{\left(K_M^A\right)^3} + \frac{e^{-\beta\epsilon_I}k_{cat}^I}{\left(K_M^I\right)^3}\right) \tag{S1.88}$$

$$-e^{-\beta\epsilon_A}e^{-\beta\epsilon_I}\left(\frac{1}{K_M^A} - \frac{1}{K_M^I}\right)^2\left(\frac{k_{cat}^A}{K_M^A} + \frac{k_{cat}^I}{K_M^I}\right) \tag{S1.89}$$

$$b = 2\left(e^{-\beta\epsilon_A} + e^{-\beta\epsilon_I}\right)\left(\frac{e^{-\beta\epsilon_A}k_{cat}^A}{\left(K_M^A\right)^2} + \frac{e^{-\beta\epsilon_I}k_{cat}^I}{\left(K_M^I\right)^2}\right) \tag{S1.90}$$

$$c = \left(e^{-\beta\epsilon_A} + e^{-\beta\epsilon_I}\right)\left(\frac{e^{-\beta\epsilon_A}k_{cat}^A}{K_M^A} + \frac{e^{-\beta\epsilon_I}k_{cat}^I}{K_M^I}\right). \tag{S1.91}$$

The roots of this equation are given by

$$[S]_0 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{S1.92}$$

Since $b, c > 0$, there will only be a positive real root $[S]_0 > 0$ if

$$a < 0. \tag{S1.93}$$

Writing this inequality out as

$$\left(e^{-\beta\epsilon_A} + e^{-\beta\epsilon_I}\right)\left(\frac{e^{-\beta\epsilon_A}k_{cat}^A}{\left(K_M^A\right)^3} + \frac{e^{-\beta\epsilon_I}k_{cat}^I}{\left(K_M^I\right)^3}\right) < e^{-\beta\epsilon_A}e^{-\beta\epsilon_I}\left(\frac{1}{K_M^I} - \frac{1}{K_M^A}\right)^2\left(\frac{k_{cat}^A}{K_M^A} + \frac{k_{cat}^I}{K_M^I}\right), \tag{S1.94}$$

we multiply by $\frac{\left(K_M^A\right)^3}{e^{-\beta\epsilon_A}e^{-\beta\epsilon_I}k_{cat}^I}$ to obtain

$$\left(1 + \frac{e^{-\beta\epsilon_A}}{e^{-\beta\epsilon_I}}\right)\left(\frac{k_{cat}^A}{k_{cat}^I} + \frac{e^{-\beta\epsilon_I}}{e^{-\beta\epsilon_A}}\left(\frac{K_M^A}{K_M^I}\right)^3\right) < \left(\frac{K_M^A}{K_M^I} - 1\right)^2\left(\frac{k_{cat}^A}{k_{cat}^I} + \frac{K_M^A}{K_M^I}\right) \tag{S1.95}$$

and move the $\frac{k_{cat}^A}{k_{cat}^I}$ terms to one side,

$$\left(\frac{K_M^A}{K_M^I}\right)^3\left(\left(1 + \frac{e^{-\beta\epsilon_I}}{e^{-\beta\epsilon_A}}\right) - \left(\frac{K_M^I}{K_M^A} - 1\right)^2\right) < \frac{k_{cat}^A}{k_{cat}^I}\left(\left(\frac{K_M^A}{K_M^I} - 1\right)^2 - \left(1 + \frac{e^{-\beta\epsilon_A}}{e^{-\beta\epsilon_I}}\right)\right). \tag{S1.96}$$

There are now two cases to consider. If the term on the right hand side is positive,

$$1 + \frac{e^{-\beta\epsilon_A}}{e^{-\beta\epsilon_I}} < \left(\frac{K_M^A}{K_M^I} - 1\right)^2, \tag{S1.97}$$

then we can divide by this term on both sides to obtain the peak condition

$$-\frac{\left(1+\frac{e^{-\beta\epsilon_I}}{e^{-\beta\epsilon_A}}\right)-\left(\frac{K_M^I}{K_M^A}-1\right)^2}{\left(1+\frac{e^{-\beta\epsilon_A}}{e^{-\beta\epsilon_I}}\right)-\left(\frac{K_M^A}{K_M^I}-1\right)^2}\left(\frac{K_M^A}{K_M^I}\right)^3<\frac{k_{cat}^A}{k_{cat}^I}. \tag{S1.98}$$

On the other hand, if the term on the right-hand side of Eq. S1.96 is negative, then the term on the left-hand side must also be negative,

$$1+\frac{e^{-\beta\epsilon_A}}{e^{-\beta\epsilon_I}}>\left(\frac{K_M^A}{K_M^I}-1\right)^2 \tag{S1.99}$$

$$1+\frac{e^{-\beta\epsilon_I}}{e^{-\beta\epsilon_A}}<\left(\frac{K_M^I}{K_M^A}-1\right)^2, \tag{S1.100}$$

and because $e^{-\beta\epsilon_A}, e^{-\beta\epsilon_I}, K_M^A, K_M^I > 0$ this implies

$$0<\frac{K_M^A}{K_M^I}<\frac{1}{2}. \tag{S1.101}$$

Solving Eq. S1.96 for $\frac{k_{cat}^A}{k_{cat}^I}$ (and flipping the sign of the inequality because of Eq. S1.99) yields the relation

$$-\frac{\left(1+\frac{e^{-\beta\epsilon_I}}{e^{-\beta\epsilon_A}}\right)-\left(\frac{K_M^I}{K_M^A}-1\right)^2}{\left(1+\frac{e^{-\beta\epsilon_A}}{e^{-\beta\epsilon_I}}\right)-\left(\frac{K_M^A}{K_M^I}-1\right)^2}\left(\frac{K_M^A}{K_M^I}\right)^3>\frac{k_{cat}^A}{k_{cat}^I}. \tag{S1.102}$$

Assuming Eq. S1.101, the term on the left-hand side can be at most $\frac{1}{2}$, so that for an enzyme that satisfies $k_{cat}^A > k_{cat}^I$ Eq. S1.102 can never be satisfied. Hence, for a two substrate binding site enzyme assuming $k_{cat}^I < k_{cat}^A$, a peak in activity as a function of substrate concentration $[S]$ will occur if and only if

$$\left(1+\frac{e^{-\beta\epsilon_A}}{e^{-\beta\epsilon_I}}\right)<\left(\frac{K_M^A}{K_M^I}-1\right)^2 \tag{S1.103}$$

$$-\frac{\left(1+\frac{e^{-\beta\epsilon_I}}{e^{-\beta\epsilon_A}}\right)-\left(\frac{K_M^I}{K_M^A}-1\right)^2}{\left(1+\frac{e^{-\beta\epsilon_A}}{e^{-\beta\epsilon_I}}\right)-\left(\frac{K_M^A}{K_M^I}-1\right)^2}\left(\frac{K_M^A}{K_M^I}\right)^3<\frac{k_{cat}^A}{k_{cat}^I}. \tag{S1.104}$$

In the text, we assumed $k_{cat}^I = 0$ so that the second condition Eq. S1.104 is automatically satisfied and Eq. S1.103 became the only necessary condition for a peak. In the general case when k_{cat}^I is not negligible, the second constraint Eq. S1.104

ensures that the contribution of product formation from the inactive state does not destroy the peak which would be formed by the active states alone.

Activity curves that exhibit a peak with a non-zero k_{cat}^{I} value are shown in Fig. S1.14. Although these curves look very similar to those shown in Fig. 1.13 for the case $k_{cat}^{I} = 0$, one important difference is that given $K_{M}^{A,I}$ and $k_{cat}^{A,I}$ values, there is now a *lower* bound for $e^{-\beta(\epsilon_A-\epsilon_I)}$ given by the second peak condition Eq. S1.104.

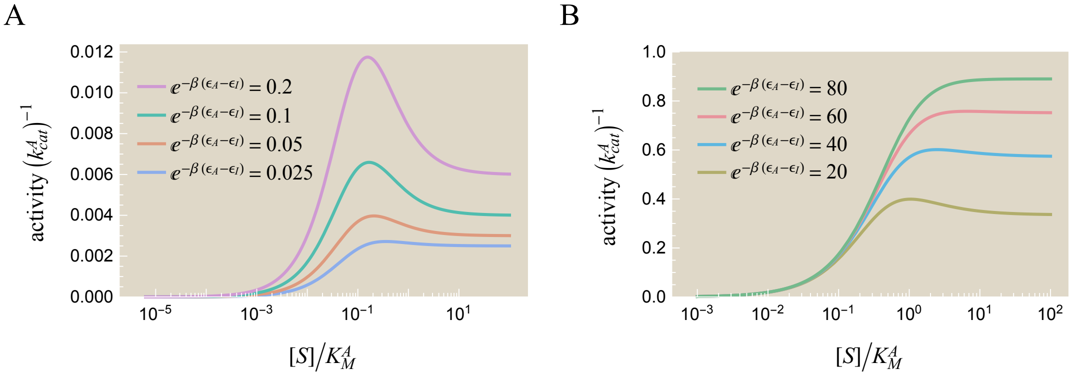


Figure S1.14: Peak in enzyme activity $A = \frac{1}{E_{tot}}\frac{d[P]}{dt}$ as a function of *substrate* concentration $[S]$. As shown in Fig. 1.12A, with Michaelis-Menten kinetics adding substrate can only increase enzyme activity, but an MWC enzyme can exhibit a peak due to the interactions between the active and inactive state. Peaks are shown for (A) small and (B) large ratios of the enzyme's energy in the active versus inactive state, $e^{-\beta(\epsilon_A-\epsilon_I)}$. The activity is computed from Eq. S1.82 using the same parameter values from Fig. 1.13 except that $\frac{k_{cat}^{A}}{k_{cat}^{I}} = 10^3$. The curves with small $e^{-\beta(\epsilon_A-\epsilon_I)}$ values shown in (A) vary appreciably from those in Fig. 1.13 (where $k_{cat}^{I} = 0$) because the inactive state catalyzes substrate. This changes both the shape and the height of the activity curves.

It is straightforward to substitute the positive root for substrate concentration Eq. S1.92 into the activity Eq. S1.82 to find the height of the peak, resulting in

$$A_{peak} = \frac{k_{cat}^{I}K_{M}^{A} - k_{cat}^{A}K_{M}^{I} + \sqrt{\left(\frac{1}{e^{-\beta\epsilon_I}} + \frac{1}{e^{-\beta\epsilon_A}}\right)\left(e^{-\beta\epsilon_I}\left(k_{cat}^{I}K_{M}^{A}\right)^2 + e^{-\beta\epsilon_A}\left(k_{cat}^{A}K_{M}^{I}\right)^2\right)}}{K_{M}^{A} - K_{M}^{I}}. \tag{S1.105}$$

In the limit $k_{cat}^{I} = 0$ discussed in the text, this simplifies to

$$A_{peak} = k_{cat}^{A}\frac{K_{M}^{I}}{K_{M}^{A} - K_{M}^{I}}\left(\sqrt{1 + \frac{e^{-\beta\epsilon_A}}{e^{-\beta\epsilon_I}}} - 1\right). \tag{S1.106}$$

Lastly, we note that adding a fixed amount of competitive inhibitor $[C]$ to a system may induce a peak in activity as a function of substrate concentration $[S]$, as shown in Fig. S1.15. In the language of the MWC model (Eqs. 1.52-1.55 in the text), adding the inhibitor tunes the MWC parameters so that the peak conditions Eqs. S1.103 and S1.104 apply.

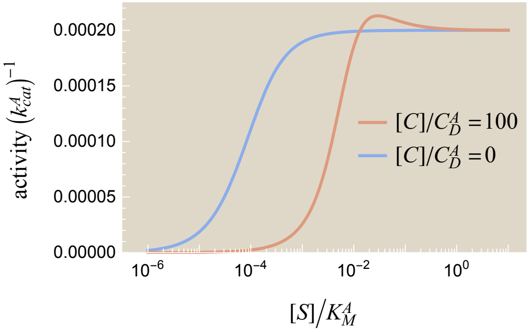


Figure S1.15: Peaks in activity can be induced by a competitive inhibitor. Adding a competitive inhibitor can induce a peak in activity $\frac{d[P]}{dt}$ versus substrate concentration $[S]$. Curves are shown for an enzyme with two active sites using the parameters $\frac{k_{cat}^A}{k_{cat}^I} = 10^4$, $\frac{K_M^A}{K_M^I} = 10^4$, $\frac{C_D^A}{C_D^I} = 10^{-1}$, and $e^{-\beta(\epsilon_A - \epsilon_I)} = \frac{1}{2}$.

S1.5.2 Inhibitor Acceleration

We now derive the peak condition for Inhibitor Acceleration discussed in Appendix S1.4.2. For an enzyme with two substrate binding sites and a competitive inhibitor C, enzyme activity is given by

$$\begin{aligned} A = {} & k_{cat}^A \left(2p_{E_A S}\right) + k_{cat}^A \left(2p_{E_A SC}\right) + 2k_{cat}^A \left(p_{E_A S^2}\right) \\ & + k_{cat}^I \left(2p_{E_I S}\right) + k_{cat}^I \left(2p_{E_I SC}\right) + 2k_{cat}^I \left(p_{E_I S^2}\right). \end{aligned} \tag{S1.107}$$

Assuming $k_{cat}^I = 0$ for simplicity, this equation takes the form

$$\begin{aligned} A &= 2k_{cat}^A e^{-\beta\epsilon_A} \frac{\frac{[S]}{K_M^A} + \frac{[S]}{K_M^A}\frac{[C]}{C_D^A} + \left(\frac{[S]}{K_M^A}\right)^2}{e^{-\beta\epsilon_A}\left(1 + \frac{[S]}{K_M^A} + \frac{[C]}{C_D^A}\right)^2 + e^{-\beta\epsilon_I}\left(1 + \frac{[S]}{K_M^I} + \frac{[C]}{C_D^I}\right)^2} \\ &\equiv 2k_{cat}^A e^{-\beta\epsilon_A} \frac{Z_C}{Z_{tot}} \end{aligned} \tag{S1.108}$$

where

$$Z_C = \frac{[S]}{K_M^A} + \frac{[S]}{K_M^A}\frac{[C]}{C_D^A} + \left(\frac{[S]}{K_M^A}\right)^2 \tag{S1.109}$$

$$Z_{tot} = e^{-\beta\epsilon_A}\left(1 + \frac{[S]}{K_M^A} + \frac{[C]}{C_D^A}\right)^2 + e^{-\beta\epsilon_I}\left(1 + \frac{[S]}{K_M^I} + \frac{[C]}{C_D^I}\right)^2. \tag{S1.110}$$

A peak in activity will occur provided that

$$\frac{dA}{d[C]} = 2k_{cat}^A e^{-\beta\epsilon_A} \frac{\frac{dZ_C}{d[C]} Z_{tot} - Z_C \frac{dZ_{tot}}{d[C]}}{Z_{tot}^2} = 0, \tag{S1.111}$$

or equivalently that the numerator $\frac{dZ_C}{d[C]} Z_{tot} - Z_C \frac{dZ_{tot}}{d[C]}$ equals zero. We can rewrite the numerator as

$$0 = \frac{dZ_C}{d[C]} Z_{tot} - Z_C \frac{dZ_{tot}}{d[C]} \equiv \frac{[S]}{K_M^A} \left(a[C]^2 + b[C] + c \right) \tag{S1.112}$$

where

$$a = -\frac{1}{C_D^A} \left(\frac{e^{-\beta\epsilon_A}}{\left(C_D^A\right)^2} + \frac{e^{-\beta\epsilon_I}}{\left(C_D^I\right)^2} \right) \tag{S1.113}$$

$$b = -2 \left(\frac{e^{-\beta\epsilon_A}}{\left(C_D^A\right)^2} + \frac{e^{-\beta\epsilon_I}}{\left(C_D^I\right)^2} \right) \left(1 + \frac{[S]}{K_M^A} \right) [C] \tag{S1.114}$$

$$c = \frac{e^{-\beta\epsilon_I}}{C_D^A} \left(1 + \frac{[S]}{K_M^I} \right)^2 - \frac{e^{-\beta\epsilon_A}}{C_D^A} \left(1 + \frac{[S]}{K_M^A} \right)^2 - 2 \frac{e^{-\beta\epsilon_I}}{C_D^I} \left(1 + \frac{[S]}{K_M^A} \right) \left(1 + \frac{[S]}{K_M^I} \right). \tag{S1.115}$$

The roots of this equation are given by

$$[C]_0 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{S1.116}$$

Since $a, b < 0$, there will only be a positive real root $[C]_0 > 0$ if

$$c > 0. \tag{S1.117}$$

Therefore, the peak condition can be written as

$$2 \frac{C_D^A}{C_D^I} \left(1 + \frac{[S]}{K_M^A} \right) \left(1 + \frac{[S]}{K_M^I} \right) < \left(1 + \frac{[S]}{K_M^I} \right)^2 - \frac{e^{-\beta\epsilon_A}}{e^{-\beta\epsilon_I}} \left(1 + \frac{[S]}{K_M^A} \right)^2 \tag{S1.118}$$

or equivalently,

$$\frac{e^{-\beta\epsilon_A}}{e^{-\beta\epsilon_I}} < \left(\frac{1 + \frac{[S]}{K_M^I}}{1 + \frac{[S]}{K_M^A}} \right)^2 - 2 \frac{C_D^A}{C_D^I} \frac{1 + \frac{[S]}{K_M^I}}{1 + \frac{[S]}{K_M^A}} \tag{S1.119}$$

which matches Eq. S1.70, as desired.

S1.5.3 Michaelis-Menten Enzymes Do Not Exhibit Peaks

In this section, we show that a Michaelis-Menten enzyme with an arbitrary number of substrate binding sites cannot exhibit substrate inhibition nor inhibitor acceleration. This implies that the interplay between the active and inactive MWC states were necessary to produce the peaks in activity discussed in section 1.4.2.2 and Appendix S1.4.2.

Consider a Michaelis-Menten enzyme with N binding sites where either a substrate S or a competitive inhibitor C can bind. Using the general formulation from section 1.3.6, we will assume that the enzyme only has an active state and drop the A superscripts. Each binding site can be either be empty, occupied by substrate, or occupied by competitor, which would contribute a factor of 1, $\frac{[S]}{K_M}$, or $\frac{[C]}{C_D}$, respectively, to its weight. A state with j bound substrates forms product at a rate of jk_{cat}. Therefore, the activity $A = \frac{1}{E_{tot}}\frac{d[P]}{dt}$ equals

$$\begin{aligned} A =& \frac{\sum_{j=0}^{N}\sum_{k=0}^{N-j}\left(jk_{cat}\right)\frac{N!}{j!k!(N-j-k)!}\left(\frac{[S]}{K_M}\right)^j\left(\frac{[C]}{C_D}\right)^k}{\left(1+\frac{[S]}{K_M}+\frac{[C]}{C_D}\right)^N} \\ =& Nk_{cat}\frac{\frac{[S]}{K_M}\left(1+\frac{[C]}{C_D}+\frac{[S]}{K_M}\right)^{N-1}}{\left(1+\frac{[S]}{K_M}+\frac{[C]}{C_D}\right)^N} \\ =& Nk_{cat}\frac{\frac{[S]}{K_M}}{1+\frac{[S]}{K_M}+\frac{[C]}{C_D}}. \end{aligned} \tag{S1.120}$$

Taking the derivative of the activity with respect to the substrate concentration $[S]$ and the inhibitor concentration $[C]$,

$$\frac{dA}{d[S]} = \frac{Nk_{cat}}{K_M}\frac{1+\frac{[C]}{C_D}}{\left(1+\frac{[S]}{K_M}+\frac{[C]}{C_D}\right)^2} \tag{S1.121}$$

and

$$\frac{dA}{d[C]} = -\frac{Nk_{cat}}{C_D}\frac{\frac{[S]}{K_M}}{\left(1+\frac{[S]}{K_M}+\frac{[C]}{C_D}\right)^2}, \tag{S1.122}$$

we find that neither derivative can be zero. Therefore, inhibitor acceleration cannot occur for a non-MWC enzyme.

References

[1] Gunawardena J. A Linear Framework for Time-Scale Separation in Nonlinear Biochemical Systems. PLoS ONE. 2012;7(5):e36321. doi:10.1371/journal.pone.0036321.

[2] Segel LA, Marshall S. The Quasi-Steady-State Assumption: A Case Study in Perturbation. Soc. Ind. Appl. Math. 2012;31(3):446–477.

[3] Mirzaev I, Gunawardena J. Laplacian Dynamics on General Graphs. Bull. Math. Biol. 2013;75(11):2118–2149. doi:10.1007/s11538-013-9884-8.

[4] Phillips R, Milo R. Rates and Duration. *Cell Biology by the Numbers*, chapter 4. 2015;.

[5] Transtrum MK, Machta BB, Brown KS, Daniels BC, Myers CR, Sethna JP. Perspective: Sloppiness and Emergent Theories in Physics, Biology, and Beyond. Journal of Chemical Physics. 2015;143(1):010901. doi:10.1063/1.4923066.

[6] Feller G, Bussy OL, Houssier C, Gerday C. Structural and Functional Aspects of Chloride Binding to Alteromonas Haloplanctis Alpha-Amylase. J. Biol. Chem. 1996;271(39):23836–23841. doi:10.1074/jbc.271.39.23836.

[7] Li C, Begum A, Numao S, Park KH, Withers SG, Brayer GD. Acarbose Rearrangement Mechanism Implied by the Kinetic and Structural Analysis of Human Pancreatic Alpha-Amylase in Complex with Analogues and Their Elongated Counterparts. Biochem. 2005;44(9):3347–57. doi:10.1021/bi048334e.

[8] Changeux JP. Responses of Acetylcholinesterase from Torpedo Marmorata to Salts and Curarizing Drugs. Mol. Pharmacol. 1966;2(5):369–92.

[9] Wales ME, Madison LL, Glaser SS, Wild JR. Divergent Allosteric Patterns Verify the Regulatory Paradigm for Aspartate Transcarbamylase. J. Mol. Biol. 1999;294(5):1387–1400. doi:10.1006/jmbi.1999.3315.

[10] Howlett GJ, Blackburn MN, Compton JG, Schachman HK. Allosteric Regulation of Aspartate Transcarbamoylase. Analysis of the Structural and Functional Behavior in Terms of a Two-State Model. Biochem. 1977;16(23):5091–5099. doi:10.1021/bi00642a023.

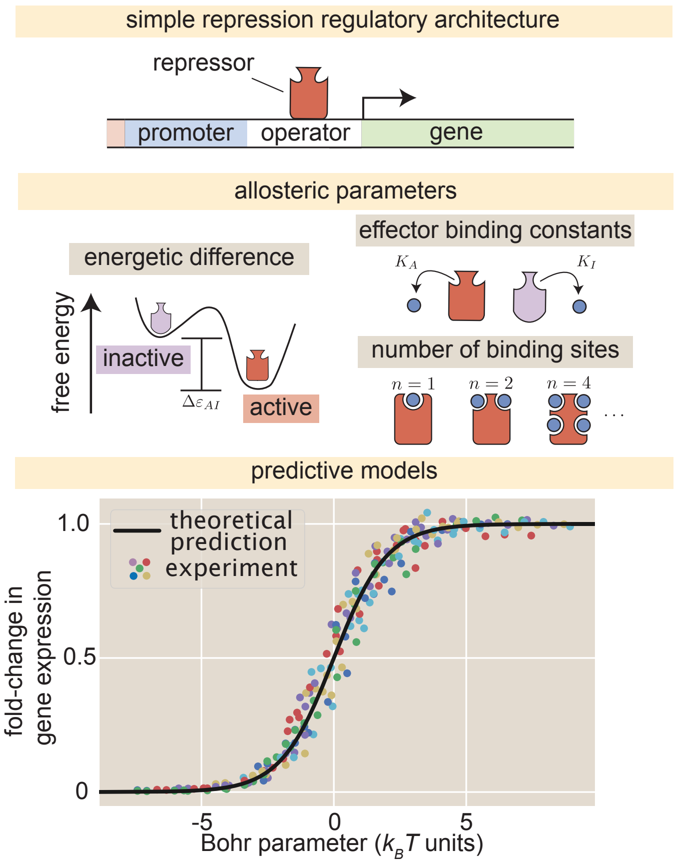
[11] Cockrell GM, Zheng Y, Guo W, Peterson AW, Truong JK, Kantrowitz ER. New Paradigm for Allosteric Regulation of *Escherichia coli* Aspartate Transcarbamoylase. Biochem. 2013;52(45):8036–8047. doi:10.1021/bi401205n.

[12] Fetler L, Kantrowitz ER, Vachette P. Direct Observation in Solution of a Preexisting Structural Equilibrium for a Mutant of the Allosteric Aspartate Transcarbamoylase. Proc. Natl. Acad. Sci. USA. 2007;104(2):495–500. doi: 10.1073/Proc.Natl.Acad.Sci.USA.0607641104.

[13] Mendes KR, Kantrowitz ER. The Pathway of Product Release from the R State of Aspartate Transcarbamoylase. J. Mol. Biol. 2010;401(5):940–8. doi: 10.1016/j.jmb.2010.07.003.

Chapter 2

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



Razo-Mejia M, Barnes S, Belliveau N, Chure G, Einav T, Mitchell L, Phillips R. Tuning Transcriptional Regulation through Signaling: A Predictive Theory of Allosteric Induction. Cell Systems. 2018;6(4):456-469. doi:10.1016/j.cels.2018.02.004

In June 2016, Rob and I prepared a beautiful paper analyzing data from our collaborator Mitch Lewis (University of Pennsylvania). The paper was ready to go and slated for PNAS, but just as I was about to submit, Rob came into the room and said, "don't submit!" It turned out that some parameters might be different from what we had assumed, but I pointed out that if these values were not more than an order of magnitude off (which Mitch believed was the case), our results would remain unchanged. Rob agreed, but he suggested that we *could quickly redo these experiments in our lab where we precisely know each parameter. Moreover, we could write a prequel where we verify that the system follows our model. It seemed that one paper had just become two papers, and the prequel paper is discussed in this chapter while the ready-to-go paper is discussed in Chapter 3.*

Little did I realize that this project would turn into one of the most rigorous, difficult, time-consuming, awe-inspiring, and high-quality papers of my PhD. Four experimentalists from my lab joined this project, and they developed an exquisitely robust data-analysis pipeline (as attested by the 45 pages of Methods and Supplementary Information in our manuscript). By sitting in on their meetings, I got a taste of what it means to be an experimentalist – the positive mindset needed to tackle setbacks and to redo experiments time and again until things work – as well as the joys and hardships of working in a large group. It took another year for the paper to be published, but looking back, the most important lesson I learned was that the final manuscript was far better than anything I could have written myself. Science is always best when done in a community or, better yet, with friends.

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 [1]. 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 for its DNA target site. *In vitro* studies have rigorously quantified the equilibria of different conformational states for allosteric transcription factors and measured the affinities of these states to the ligand [2, 3]. 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— that is, across both broad ranges of parameters and regulatory architectures —is important for multiple reasons. First, in the context of a specific gene, accurate prediction demonstrates that all components relevant to the 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 accurately extrapolate a system's behavior in other conditions. Moreover, there is an essential distinction between a predictive model, which is used to predict a system's behavior given a set of input variables, and a retroactive model, which is used to describe the behavior of data that has already been obtained. We note that even some of the most careful and rigorous analysis of transcriptional regulation often entails only a retroactive reflection on a single experiment. This raises the fear that each regulatory architecture may require a unique analysis that cannot carry over to other systems, a worry that is exacerbated by the prevalent use of phenomenological functions (e.g. Hill functions or ratios of polynomials) that can analyze a single data set but cannot be used to extrapolate a system's behavior in other conditions [4–8].

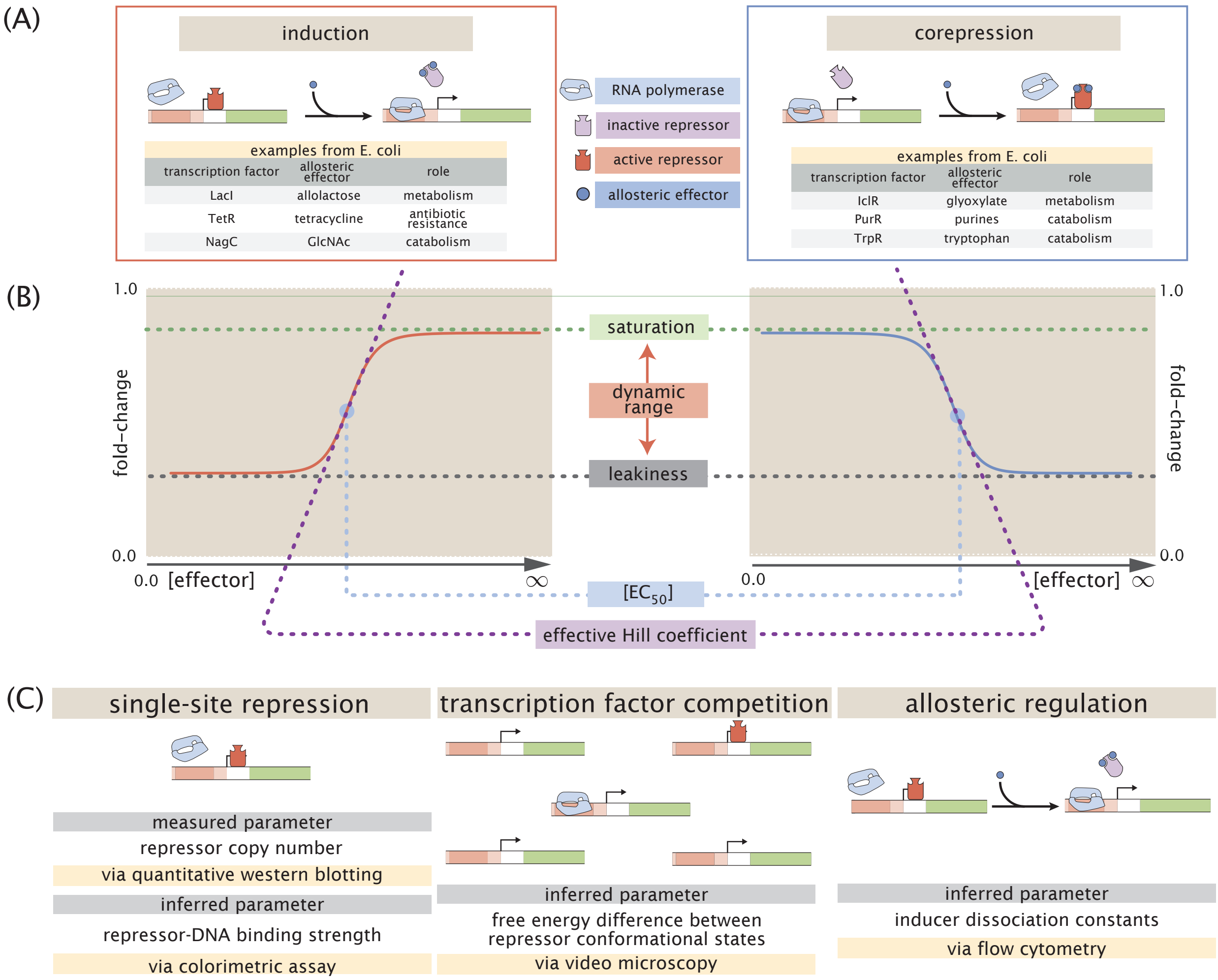
This work explores what happens when theory takes center stage, namely, we first write down the equations governing a system and describe its expected behavior across a wide array of experimental conditions, and only then do we set out to experimentally confirm these results. Building upon previous work [9–11] and the work of Monod, Wyman, and Changeux [12], we present a statistical mechanical rendering of allostery in the context of induction and corepression (shown schematically in Fig. 2.1(A) 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 [12]. 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 [9, 10, 13–16]. Thus, rather than viewing the behavior of each circuit as giving rise to its own unique input-output response, the MWC model provides a means to characterize these seemingly diverse behaviors using a single unified framework governed by a small set of parameters.

2.3 Results

2.3.1 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 [19, 20]. 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 [12]. This causes



examples from E. coli

transcription factor	allosteric effector	role
LacI	allolactose	metabolism
TetR	tetracycline	antibiotic resistance
NagC	GlcNAc	catabolism

examples from E. coli

transcription factor	allosteric effector	role
IclR	glyoxylate	metabolism
PurR	purines	catabolism
TrpR	tryptophan	catabolism

single-site repression: measured parameter: repressor copy number via quantitative western blotting; inferred parameter: repressor-DNA binding strength via colorimetric assay

transcription factor competition: inferred parameter: free energy difference between repressor conformational states via video microscopy

allosteric regulation: inferred parameter: inducer dissociation constants via flow cytometry

Figure 2.1: Transcription regulation architectures involving an allosteric repressor. (A) We consider a promoter regulated by an allosteric repressor (left panel), where the addition of an effector binds to the repressor and stabilizes the inactive state (the state with low affinity for DNA), thereby increasing gene expression. Corepression (right panel) is characterized by the same statistical mechanical model we develop. (B) A schematic response plotting 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, together with four key phenotypic properties of the response. (C) The simple repression architectures has been characterized using multiple experimental methods including colorimetric assays/quantitative Western blots and video microscopy, and add to this list the additional method of flow cytometry.

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 [9, 10, 19–22] (we note that non-equilibrium models of simple repression have been shown to have the same functional form that we derive below [23]). 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. 2.2A, the promoter can either be empty, occupied by RNAP, or occupied by either an active or inactive repressor. The probability of binding to the promoter will be affected by the protein copy number, which we denote as P for RNAP, R_A for active repressor, and R_I for inactive repressor. We note that repressors fluctuate between the active and inactive conformation in thermodynamic equilibrium, such that R_A and R_I will remain constant for a given inducer concentration [12]. 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 [9–11, 19–22, 24–26] posit that gene expression is proportional to the probability that the RNAP is bound to the promoter p_{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_{bound} directly is fraught with experimental difficulties, as determining the exact proportionality between expression and p_{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,

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

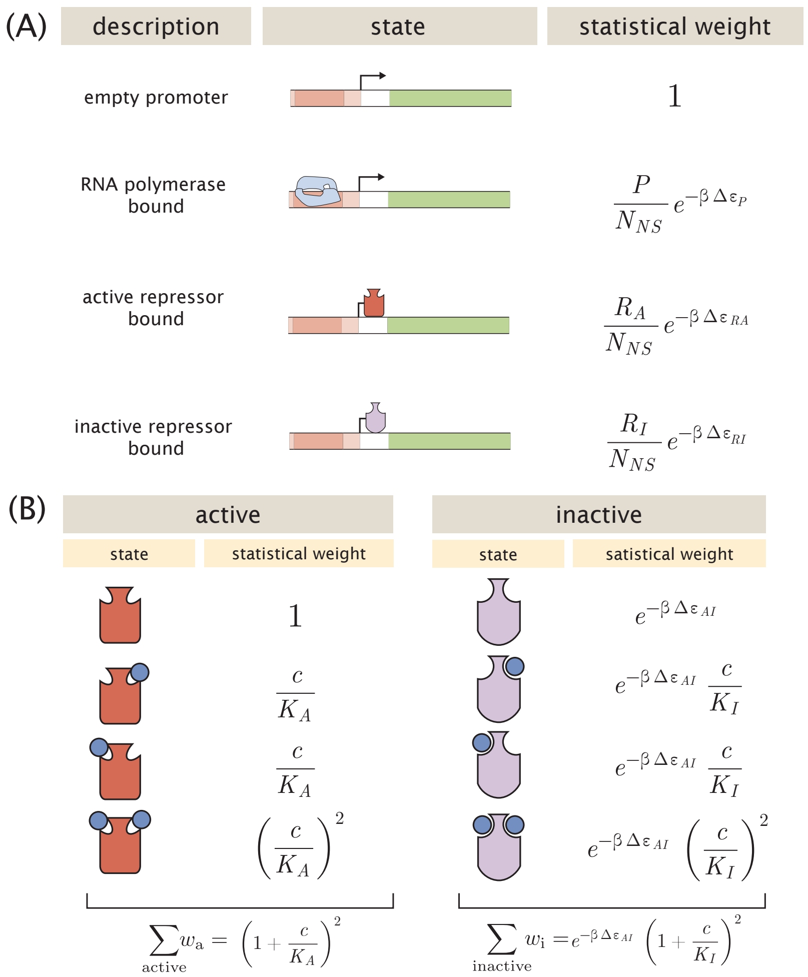

(A)
description
state
statistical weight
empty promoter
1
RNA polymerase bound
$\frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_P}$
active repressor bound
$\frac{R_A}{N_{NS}} e^{-\beta \Delta\varepsilon_{RA}}$
inactive repressor bound
$\frac{R_I}{N_{NS}} e^{-\beta \Delta\varepsilon_{RI}}$
(B)
active
state
statistical weight
1
$\frac{c}{K_A}$
$\frac{c}{K_A}$
$\left(\frac{c}{K_A}\right)^2$
inactive
state
satistical weight
$e^{-\beta \Delta\varepsilon_{AI}}$
$e^{-\beta \Delta\varepsilon_{AI}} \frac{c}{K_I}$
$e^{-\beta \Delta\varepsilon_{AI}} \frac{c}{K_I}$
$e^{-\beta \Delta\varepsilon_{AI}} \left(\frac{c}{K_I}\right)^2$
$\sum_{\text{active}} w_a = \left(1+\frac{c}{K_A}\right)^2$
$\sum_{\text{inactive}} w_i = e^{-\beta \Delta\varepsilon_{AI}} \left(1+\frac{c}{K_I}\right)^2$

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

We can simplify this expression using two well-justified approximations: (1) the weak promoter approximation $\frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_P} \ll 1$ implies that the promoter is most often not bound to RNAP ($N_{NS} = 4.6 \times 10^6$, $P \approx 10^3$ [27], $\Delta\varepsilon_P \approx -2$ to $-5\ k_B T$ [14], 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

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

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 [28].

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. 2.2B), namely,

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

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 $e^{-\Delta\varepsilon_{AI}}$ is sometimes denoted by L [12, 28] or $K_{\mathrm{RR}*}$ [26]). 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 inactive state in the absence of inducer. Thus, a repressor which favors the active state in the absence of inducer ($\Delta\varepsilon_{AI} > 0$) will be driven towards the inactive state upon inducer binding when $K_I < K_A$. The specific case of a repressor dimer with $n = 2$ inducer binding sites is shown in Fig. 2.2B.

Substituting $p_A(c)$ from Eq. 2.4 into Eq. 2.3 yields the general formula for induction of a simple repression regulatory architecture [23], namely,

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

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. 2.1B). 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 [10, 11, 24].

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 $L = e^{-\beta\Delta\varepsilon_{AI}}$ [26, 29], 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 Appendix S2.1 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 experimental conditions. For example, this framework can predict the response of the system at different repressor copy numbers R, repressor-operator affinities $\Delta\varepsilon_{RA}$, inducer concentrations c, and gene copy numbers.

2.3.2 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 [26, 30–33]. 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 [9]. Ref. [34] extended the theory used in that work to the case of multiple promoters competing for a given transcription factor, which was validated experimentally by Ref. [10], 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 Ref. [9] and Ref. [10]. Although the current work focuses on systems with a single site of repression, in Appendix S2.1 we utilize data from Ref. [10], 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 Fig. S2.1). After fixing $\Delta\varepsilon_{AI}$, 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 [35]. 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 Ref. [9] to generate strains with mean repressor copy number per cell of $R = 22 \pm 4$, 60 ± 20, 124 ± 30, 260 ± 40, 1220 ± 160, and 1740 ± 340, where the error denotes standard deviation of at least three replicates as measured by Ref. [9]. We note that R refers to the number of repressor dimers in the cell, which is twice the number of repressor tetramers reported by Ref. [9]; since both heads of the repressor are assumed to always be either specifically or non-specifically bound to the genome, the two repressor dimers in each LacI tetramer can be considered independently. Gene expression was measured using a Yellow Fluorescent Protein (YFP) gene, driven by a *lacUV5* promoter. Each of the six repressor copy number variants were paired with the native O1, O2, or O3 *lac* operator [36] 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 [9]. Additionally, we were able to obtain the value $\Delta\varepsilon_{AI} = 4.5\ k_BT$ by fitting to previous data as discussed in Appendix S2.1. 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 [37], 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. 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. 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.

2.3.3 Determination of the *in vivo* MWC Parameters

The three parameters that we tune experimentally are shown in Fig. 2.4A, 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 [10] to infer that $\Delta\varepsilon_{AI} = 4.5\ k_BT$ (see Appendix S2.1). 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 ($\Delta\varepsilon_{RA} = -13.9\ k_BT$ [9]) shown in Fig. 2.4D (white circles). Using Markov Chain Monte Carlo, we determine the most likely parameter values to be $K_A = 139^{+29}_{-22} \times 10^{-6}$ M and $K_I = 0.53^{+0.04}_{-0.04} \times 10^{-6}$ M, which are the modes of their respective distributions, where the superscripts and subscripts represent the upper and lower bounds of the 95th percentile of the parameter value distributions (see Fig. 2.4B). Unfortunately, we are not able to make a meaningful value-for-value comparison of our parameters to those of earlier studies [26, 31] because of uncertainties in both 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. 2.4C-E) together with the specific phenotypic properties described in Fig. 2.1 and discussed in detail below (see

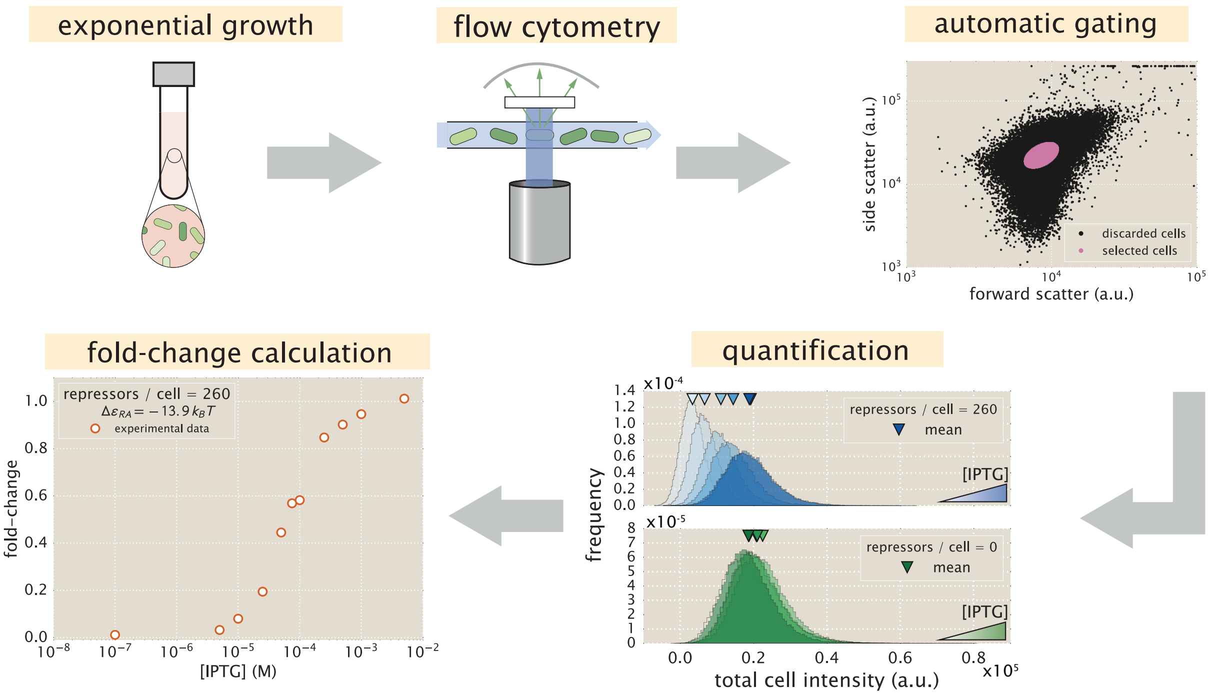


Figure 2.3: An experimental pipeline for high-throughput fold-change measurements. Cells are grown to exponential steady state and their fluorescence is measured using flow cytometry. Automatic gating methods using forward- and side-scattering are used to ensure that all measurements come from single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

Fig. 2.4F-J). The shaded regions in Fig. 2.4C-J denote the 95% credible regions.

We stress that the entire suite of predictions in Fig. 2.4 is based upon the induction profile of a single strain. Our ability to make such a broad range of predictions stems from the fact that our parameters of interest – such as the repressor copy number and DNA binding energy – appear as distinct physical parameters within our model. While the single data set in Fig. 2.4D could also be fit using a Hill function, such an analysis would be unable to predict any of the other curves in the figure. Phenomenological expressions such as the Hill function can describe data, but lack predictive power and are thus unable to build our intuition, help us design *de novo* input-output functions, or guide future experiments [25, 30].

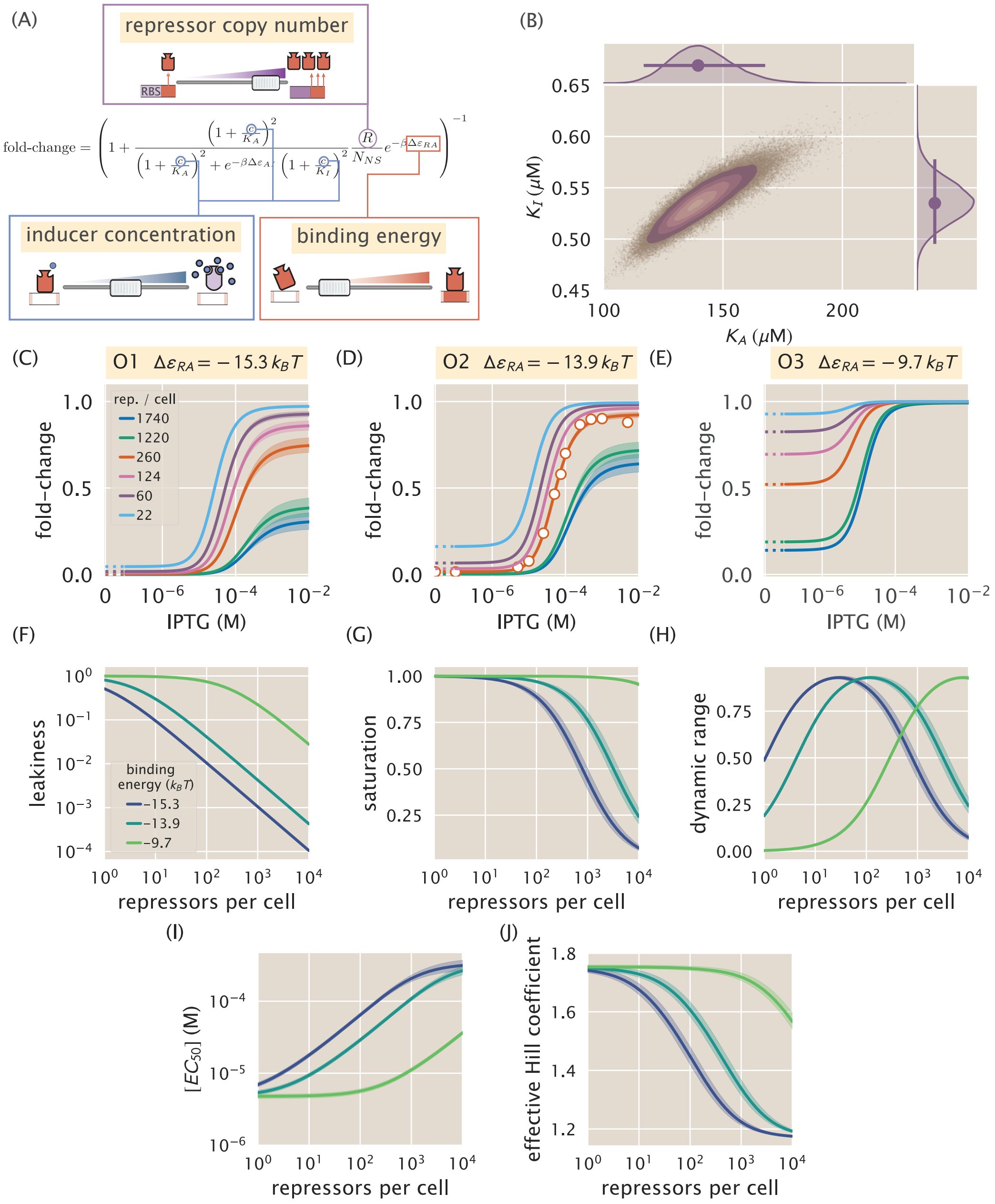


Figure 2.4: **Predicting induction profiles for different biological control parameters.** (A) We can quantitatively tune R via ribosomal binding site (RBS) modifications, $\Delta\varepsilon_{RA}$ by mutating the operator sequence, and c by adding different amounts of IPTG to the growth medium. (B) We infer the value of the dissociation constants K_A and K_I between the inducer and the repressor in the active and inactive states, respectively, using Bayesian parameter estimation from a single induction curve. (C-J) Predicted IPTG titration curves and key phenotypic parameters for different repressor copy numbers and operator strengths. Titration data for the O2 strain (white circles in Panel D) with $R = 260$, $\Delta\varepsilon_{RA} = -13.9\ k_BT$, $n = 2$, and $\Delta\varepsilon_{AI} = 4.5\ k_BT$ can be used to determine the thermodynamic parameters $K_A = 139^{+29}_{-22} \times 10^{-6}$ M and $K_I = 0.53^{+0.04}_{-0.04} \times 10^{-6}$ M (orange line). The remaining solid lines predict the fold-change Eq. 2.5 for all other combinations of repressor copy numbers (shown in the legend) and repressor-DNA binding energies corresponding to the O1 operator ($-15.3\ k_BT$), O2 operator ($-13.9\ k_BT$), and O3 operator ($-9.7\ k_BT$). Error bars of experimental data show the standard error of the mean (eight or more replicates) when this error is not smaller than the diameter of the data point. The shaded regions denote the 95% credible region.

2.3.4 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 Ref. [9]. 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. 2.5C) and all of the $R = 1220$ and $R = 1740$ strains. This may be partially due to imprecise previous determinations of their $\Delta\varepsilon_{RA}$ and R values. By performing a global fit where we infer all parameters including the repressor copy number R and the binding energy $\Delta\varepsilon_{RA}$, we found better agreement for these strains, although a discrepancy in the steepness of the response for all O3 strains remains (see Appendix S2.2). 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 $\Delta\varepsilon_{RA} = -17\ k_BT$ [9]. 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 Appendix S2.3).

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 and found that the inferred values of K_A and K_I depend minimally upon which strain is chosen, indicating that these parameter values are highly robust. We also performed a global fit using the data from all eighteen strains in which we fitted for the inducer dissociation constants K_A and K_I, the repressor copy number R, and the repressor DNA binding energy $\Delta\varepsilon_{RA}$ (see Appendix S2.2). The resulting parameter values were nearly identical to those fitted from any single strain. For the remainder of the text we continue using parameters fitted from the strain with $R = 260$ repressors and an O2 operator.

2.3.5 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

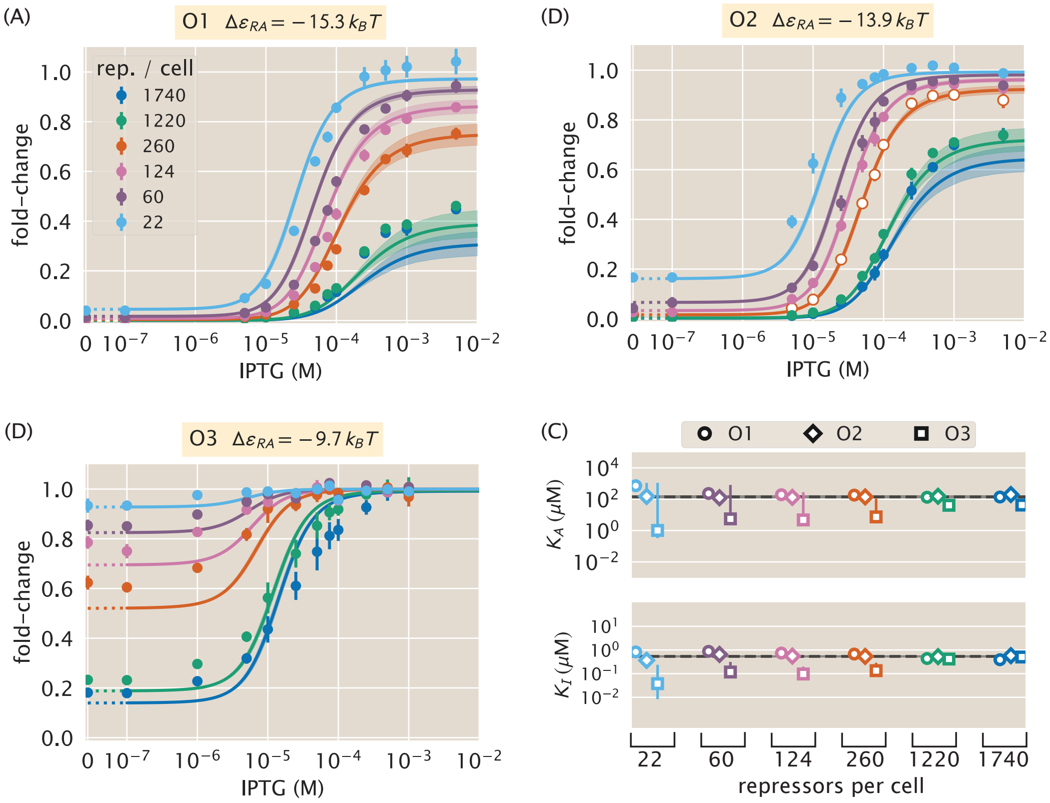


Figure 2.5: Comparison of predictions against measured and inferred data. Flow cytometry measurements of fold-change over a range of IPTG concentrations for (A) O1, (B) O2, and (C) O3 strains at varying repressor copy numbers, overlaid on the predicted responses. Error bars for the experimental data show the standard error of the mean (eight or more replicates). As discussed in Fig. 2.4, all of the predicted induction curves were generated prior to measurement by inferring the MWC parameters using a single data set (O2 R = 260, shown by white circles in Panel B). The predictions may therefore depend upon which strain is used to infer the parameters. (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.

responses (i.e. a large dynamic range) or finding a strong binding partner (i.e. a small [EC_{50}]) [38, 39]. While these properties are all individually informative, when taken together they capture the essential features of the induction response. We reiterate that a Hill function approach cannot predict these features *a priori* and furthermore requires fitting each curve individually. The MWC model, on the other hand, enables us to quantify how each trait depends upon a single set of physical parameters as shown by Fig. 2.4F-J.

We define these five phenotypic traits using expressions derived from the model, Eq. 2.5. These results build upon extensive work by Ref. [40], who computed many such properties for ligand-receptor binding within the MWC model. We begin by analyzing the leakiness, which is the minimum fold-change observed in the absence of ligand, given by

$$\begin{aligned}\text{leakiness} &= \text{fold-change}(c = 0)\\ &= \left(1 + \frac{1}{1 + e^{-\beta \Delta \varepsilon_{AI}}} \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_{RA}}\right)^{-1},\end{aligned} \tag{2.6}$$

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

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

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

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

These three properties are shown in Fig. 2.4F-H. Fig. 2.6A-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 Eq. 2.5 halfway between its minimum and maximum value,

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

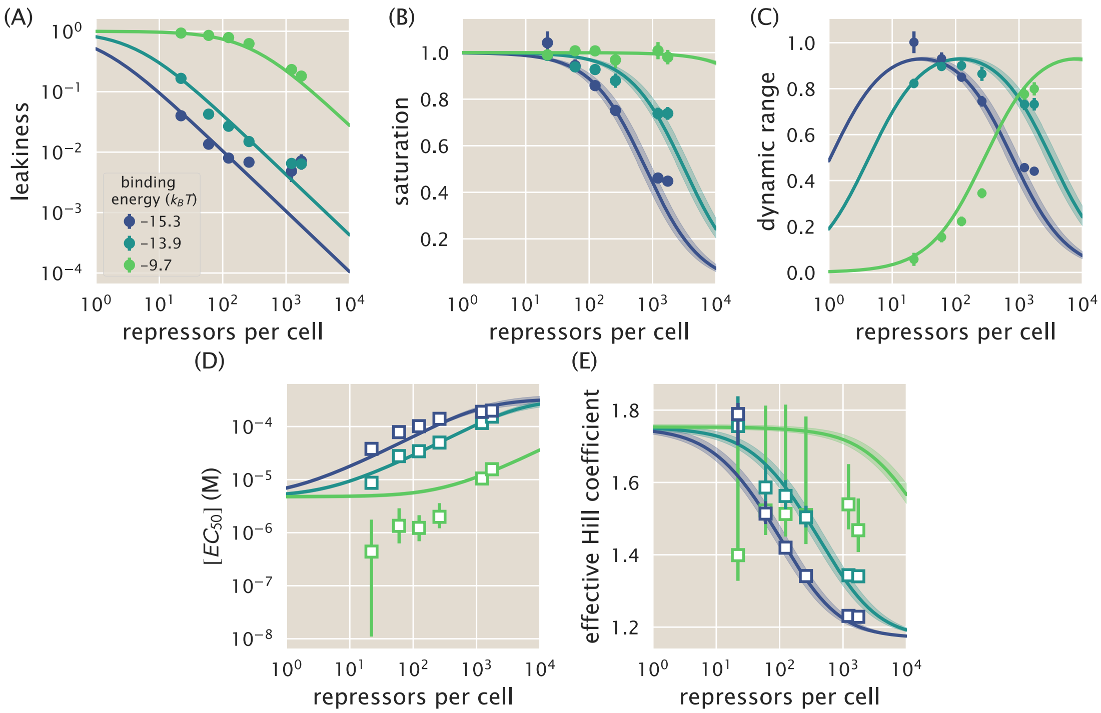


Figure 2.6: Predictions and experimental measurements of key properties of induction profiles. Data for the (A) leakiness, (B) saturation, and (C) dynamic range are obtained from fold-change measurements in Fig. 2.5 in the absence of IPTG and at saturating concentrations of IPTG. The three repressor-operator 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 (D) $[EC_{50}]$ and (E) effective Hill coefficient are inferred by individually fitting each operator-repressor pairing in Fig. 2.5A-C separately to Eq. 2.5 in order to smoothly interpolate between the data points. Error bars for A-C represent the standard error of the mean for eight or more replicates; error bars for D-E represent the 95% credible region for the parameter found by propagating the credible region of our estimates of K_A and K_I into Eqs. 2.9 and 2.10.

The effective Hill coefficient h, which quantifies the steepness of the curve at the $[EC_{50}]$ [28], 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. 2.4I-J shows how the $[EC_{50}]$ and effective Hill coefficient depend on the repressor copy number.

Fig. 2.6D-E shows the estimated values of the $[EC_{50}]$ and the effective Hill coefficient overlaid on the theoretical predictions. Both properties were obtained by fitting 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 operator are markedly off. The uncertainty with O3 arises from its nearly flat response, where the lack of dynamic range makes it impossible to determine the value of the inducer dissociation constants K_A and K_I, as can be seen in the uncertainty of both the $[EC_{50}]$ and effective Hill coefficient. Discrepancies between theory and data for O3 are improved, but not fully resolved, by performing a global fit or fitting the MWC model individually to each curve (see Appendix S2.2). It remains an open question how to account for discrepancies in O3, in particular regarding the significant mismatch between the predicted and fitted effective Hill coefficients.

2.3.6 Data Collapse of Induction Profiles

Our primary interest heretofore was to determine the system response at a specific inducer concentration, repressor copy number, and repressor-DNA binding energy. However, the cell does not necessarily "care about" the precise number of repressors in the system or the binding energy of an individual operator. The relevant quantity for cellular function is the fold-change enacted by the regulatory system. This raises the question: given a specific value of the fold-change, what combination of parameters will give rise to this desired response? In other words, what trade-offs between the parameters of the system will give rise to the same mean cellular output? These are key questions both for understanding how the system is governed and 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 [41–43], 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_BT units [23, 42, 43], which is given by

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

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 [23, 44, 45].

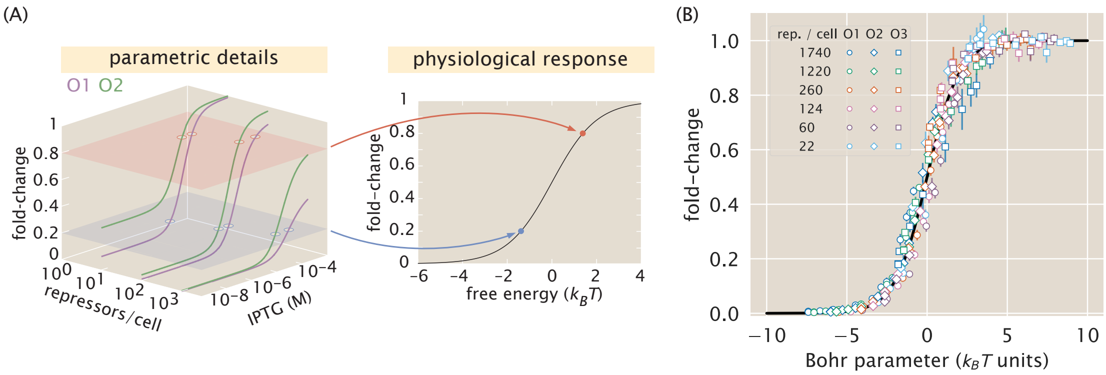


Figure 2.7: Fold-change data from a broad collection of different strains collapse onto a single master curve. (A) Any combination of parameters can be mapped to a single physiological response (i.e. fold-change) via the free energy, which encompasses the parametric details of the model. (B) Experimental data from Fig. 2.5 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 $n = 2$, $\Delta\varepsilon_{AI} = 4.5\ k_BT$, $K_A = 139 \times 10^{-6}$ 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.

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. 2.7A demonstrates how the various induction curves from Fig. 2.4C-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. 2.7B shows this data collapse for the 216 data points in Fig. 2.5A-C, demonstrating the close match between the theoretical predictions and experimental measurements across all eighteen strains.

There are many different combinations of parameter values that can result in the same free energy as defined in Eq. 2.12. For example, suppose a system originally has a fold-change of 0.2 at a specific inducer concentration, and then operator mutations increase the $\Delta\varepsilon_{RA}$ binding energy [46]. While this serves to initially increase both the free energy and the fold-change, a subsequent increase in the repressor copy number could bring the cell back to the original fold-change level. Such trade-offs hint that there need not be a single set of parameters that evoke a specific cellular response, but rather that the cell explores a large but degenerate space of parameters with multiple, equally valid paths.

2.4 Discussion

Since the early work by Monod, Wyman, and Changeux [12, 47], 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 [48], enzymatic reactions [45, 49], chemotaxis [42], quorum sensing [43], G-protein coupled receptors [50], physiologically important proteins [51, 52], 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 [17, 18]. A number of studies have focused on developing a quantitative understanding of allosteric regulatory systems. Ref. [28, 40] analytically derived fundamental properties of the MWC model, including the leakiness and dynamic range described in this work, noting the inherent trade-offs in these properties when tuning the model's parameters. Work in the Church and Voigt labs, among others, has expanded on the availability of allosteric circuits for synthetic biology [7, 8, 53, 54]. Recently, Daber *et al.* theoretically explored the induction of simple repression within the MWC model [31] and experimentally measured how mutations alter the induction profiles of transcription factors [26]. 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 [6]. Other work has attempted to use the *lac* system to reconcile *in vitro* and *in vivo* measurements [33, 55].

Although this body of work has done much to improve our understanding of allosteric transcription factors, there have been few attempts to explicitly connect quantitative models to experiments. Here, we generate a predictive model of allosteric transcriptional regulation and then test the model against a thorough set of experiments using well-characterized regulatory components. Specifically, we used the MWC model to build upon a well-established thermodynamic model of transcriptional regulation [9, 24], allowing us to compose the model from a minimal set of biologically meaningful parameters. This model combines both theoretical and experimental insights; for example, rather than considering gene expression directly we analyze the fold-change in expression, where the weak promoter approximation (see Eq. 2.3) circumvents uncertainty in the RNAP copy number. The resulting model depended upon experimentally accessible parameters, namely, the repressor copy number, the repressor-DNA binding energy, and the concentration of inducer. We tested these predictions on a range of strains whose repressor copy number spanned two orders of magnitude and whose DNA binding affinity spanned 6 k_BT. We argue that one would not be able to generate such a wide array of predictions by using a Hill function, which abstracts away the biophysical meaning of the parameters into phenomenological parameters [56].

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 $\Delta\varepsilon_{RA} = -13.9\ k_BT$ 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 (see Fig. 2.4). 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. 2.5A-C. Importantly, we find that fitting K_A and K_I to data from any other strain would have resulted in nearly identical predictions (see Fig. 2.5D). 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 stands 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 explicitly predict the values of the induction curves' key parameters, namely, the leakiness, saturation, dynamic range, $[EC_{50}]$, and the effective Hill coefficient (see Fig. 2.6). 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 Appendix S2.2), we are still unable to accurately predict the effective Hill coefficient or the $[EC_{50}]$. We further tried including additional states (such as allowing the inactive repressor to bind to the operator), relaxing the weak promoter approximation, accounting for changes in gene and repressor copy number throughout the cell cycle [57], and refitting the original binding energies from Ref. [13], 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 [7, 8].

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 [41–43]. Although it is experimentally straightforward to observe system responses to changes in effector concentration c, framing the input-output function in terms of c can give the misleading impression that changes in system parameters lead to fundamentally altered system responses. Alternatively, if one can find the “natural variable" that enables the output to collapse onto a single curve, it becomes clear that the system’s output is not governed by individual system parameters, but rather the contributions of multiple parameters that define the natural variable. When our fold-change data are plotted against the respective free energies for each construct, they collapse cleanly onto a single curve (see Fig. 2.7). This enables us to analyze how parameters can compensate each other. For example, rather than viewing strong repression as a consequence of low IPTG concentration c or high repressor copy number R, we can now observe that strong repression is achieved when the free energy $F(c) \leq -5k_BT$, 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 [24]. We can further explore different regulatory configurations such as corepression, activation, and coactivation, each of which are found in *E. coli* (see Appendix S2.4). 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 [58]. 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 [23, 26]. Finally, we expect that the kind of 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 [59–61].

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 Acknowledgements

This work has been a wonderful exercise in scientific collaboration. We thank Hernan Garcia for information and advice for working with these bacterial strains, Pamela Björkman and Rachel Galimidi for access and training for use of the Miltenyi Biotec MACSQuant flow cytometer, and Colin deBakker of Milteny Biotec for useful advice and instruction in flow cytometry. The experimental front of this work began at the Physiology summer course at the Marine Biological Laboratory in Woods Hole, MA operated by the University of Chicago. We thank Simon Alamos, Nalin Ratnayeke, and Shane McInally for their work on the project during the course. We also thank Suzannah Beeler, Justin Bois, Robert Brewster, Ido Golding, Soichi Hirokawa, Jané Kondev, Tom Kuhlman, Heun Jin Lee, Muir Morrison, Nigel Orme, Alvaro Sanchez, and Julie Theriot for useful advice and discussion. This work was supported by La Fondation Pierre-Gilles de Gennes, the Rosen Center at Caltech, and the National Institutes of Health DP1 OD000217 (Director's Pioneer Award), R01 GM085286, and 1R35 GM118043-01 (MIRA). Nathan Belliveau is a Howard Hughes Medical Institute International Student Research fellow.

References

[1] Lindsley JE, Rutter J. Whence Cometh the Allosterome? Proc. Natl. Acad. Sci. USA. 2006;103(28):10533–5. doi:10.1073/Proc.Natl.Acad.Sci.USA.0604452103.

[2] Harman JG. Allosteric Regulation of the cAMP Receptor Protein. Biochim. Biophys. Acta. 2001;1547(1):1–17. doi:10.1016/S0167-4838(01)00187-X.

[3] Lanfranco MF, Gárate F, Engdahl AJ, Maillard RA. Asymmetric Configurations in a Reengineered Homodimer Reveal Multiple Subunit Communication Pathways in Protein Allostery. J. Biol. Chem. 2017;292(15):6086–6093. doi: 10.1074/jbc.M117.776047.

[4] Setty Y, Mayo AE, Surette MG, Alon U. Detailed Map of a Cis-Regulatory Input Function. Proc. Natl. Acad. Sci. USA. 2003;100(13):7702–7707. doi: 10.1073/Proc.Natl.Acad.Sci.USA.1230759100.

[5] Poelwijk FJ, de Vos MG, Tans SJ. Tradeoffs and Optimality in the Evolution of Gene Regulation. Cell. 2011;146(3):462–470. doi:10.1016/j.cell.2011.06.035.

[6] Vilar JMG, Saiz L. Reliable Prediction of Complex Phenotypes from a Modular Design in Free Energy Space: An Extensive Exploration of the *lac* Operon. ACS Synth. Biol. 2013;2(10):576–586. doi:10.1021/sb400013w.

[7] Rogers JK, Guzman CD, Taylor ND, Raman S, Anderson K, Church GM. Synthetic Biosensors for Precise Gene Control and Real-Time Monitoring of Metabolites. Nucleic. Acids. Res. 2015;43(15):7648–7660. doi:10.1093/nar/gkv616.

[8] Rohlhill J, Sandoval NR, Papoutsakis ET. Sort-Seq Approach to Engineering a Formaldehyde-Inducible Promoter for Dynamically Regulated *Escherichia coli* Growth on Methanol. ACS Synth. Biol. 2017;6(8):1584–1595. doi: 10.1021/acssynbio.7b00114.

[9] Garcia HG, Phillips R. Quantitative Dissection of the Simple Repression Input-Output Function. Proc. Natl. Acad. Sci. USA. 2011;108(29):12173–8. doi:10.1073/Proc.Natl.Acad.Sci.USA.1015616108.

[10] Brewster RC, Weinert FM, Garcia HG, Song D, Rydenfelt M, Phillips R. The Transcription Factor Titration Effect Dictates Level of Gene Expression. Cell. 2014;156(6):1312–23. doi:10.1016/j.cell.2014.02.022.

[11] Weinert FM, Brewster RC, Rydenfelt M, Phillips R, Kegel WK. Scaling of Gene Expression with Transcription-Factor Fugacity. Phys. Rev. Lett. 2014; 113(25):1–5. doi:10.1103/PhysRevLett.113.258101.

[12] Monod J, Wyman J, Changeux JP. On the Nature of Allosteric Transitions: A Plausible Model. J. Mol. Biol. 1965;12:88–118. doi:10.1016/S0022-2836(65)80285-6.

[13] Garcia HG, Lee HJ, Boedicker JQ, Phillips R. Comparison and Calibration of Different Reporters for Quantitative Analysis of Gene Expression. Biophys. J. 2011;101(3):535–544. doi:10.1016/j.bpj.2011.06.026.

[14] Brewster RC, Jones DL, Phillips R. Tuning Promoter Strength Through RNA Polymerase Binding Site Design in *Escherichia coli*. PLoS Comput. Biol. 2012;8(12). doi:10.1371/journal.pcbi.1002811.

[15] Boedicker JQ, Garcia HG, Johnson S, Phillips R. DNA Sequence-Dependent Mechanics and Protein-Assisted Bending in Repressor-Mediated Loop Formation. Phys. Biol. 2013;10(6):066005. doi:10.1088/1478-3975/10/6/066005.

[16] Boedicker JQ, Garcia HG, Phillips R. Theoretical and Experimental Dissection of DNA Loop-Mediated Repression. Phys. Rev. Lett. 2013;110(1):018101. doi:10.1103/PhysRevLett.110.018101.

[17] Huang Z, Zhu L, Cao Y, Wu G, Liu X, Chen Y, et al. ASD: A Comprehensive Database of Allosteric Proteins and Modulators. Nucleic. Acids. Res. 2011; 39(Database):D663–D669. doi:10.1093/nar/gkq1022.

[18] Li M, Petukh M, Alexov E, Panchenko AR. Predicting the Impact of Missense Mutations on Protein-Protein Binding Affinity. J Chem. Theory. Comput. 2014;10(4):1770–1780. doi:10.1021/ct401022c.

[19] Ackers GK, Johnson AD, Shea MA. Quantitative Model for Gene Regulation by Lambda Phage Repressor. Proc. Natl. Acad. Sci. USA. 1982;79(4):1129–33.

[20] Buchler NE, Gerland U, Hwa T. On Schemes of Combinatorial Transcription Logic. Proc. Natl. Acad. Sci. USA. 2003;100(9):5136–41. doi:10.1073/Proc.Natl.Acad.Sci.USA.0930314100.

[21] Vilar JM, Leibler S. DNA Looping and Physical Constraints on Transcription Regulation. J. Mol. Biol. 2003;331(5):981–989. doi:10.1016/S0022-2836(03)00764-2.

[22] Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, et al. Transcriptional Regulation by the Numbers: Applications. Curr Opin Genetics Dev. 2005;15(2):125–135. doi:10.1016/j.gde.2005.02.006.

[23] Phillips R. Napoleon Is in Equilibrium. Annu. Rev. Condens. Matter Phys. 2015;6(1):85–111. doi:10.1146/annurev-conmatphys-031214-014558.

[24] Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, et al. Transcriptional Regulation by the Numbers: Models. Curr Opin Genetics Dev. 2005; 15(2):116–124. doi:10.1016/j.gde.2005.02.007.

[25] Kuhlman T, Zhang Z, Saier MH, Hwa T. Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. Proc. Natl. Acad. Sci. USA. 2007; 104(14):6043–6048. doi:10.1073/Proc.Natl.Acad.Sci.USA.0606717104.

[26] Daber R, Sochor MA, Lewis M. Thermodynamic Analysis of Mutant Lac Repressors. J. Mol. Biol. 2011;409(1):76–87. doi:10.1016/j.jmb.2011.03.057.

[27] Klumpp S, Hwa T. Growth-Rate-dependent Partitioning of RNA Polymerases in Bacteria. Proc. Natl. Acad. Sci. USA. 2008;105(51):20245–20250. doi:10.1073/Proc.Natl.Acad.Sci.USA.0804953105.

[28] Marzen S, Garcia HG, Phillips R. Statistical Mechanics of Monod-Wyman-Changeux (MWC) Models. J. Mol. Biol. 2013;425(9):1433–1460. doi:10.1016/j.jmb.2013.03.013.

[29] O'Gorman RB, Rosenberg JM, Kallai OB, Dickerson RE, Itakura K, Riggs AD, et al. Equilibrium Binding of Inducer to Lac Repressor.operator DNA Complex. J. Biol. Chem. 1980;255(21):10107–10114.

[30] Murphy KF, Balazsi G, Collins JJ. Combinatorial Promoter Design for Engineering Noisy Gene Expression. Proc. Natl. Acad. Sci. USA. 2007; 104(31):12726–12731. doi:10.1073/Proc.Natl.Acad.Sci.USA.0608451104.

[31] Daber R, Sharp K, Lewis M. One Is Not Enough. J. Mol. Biol. 2009; 392(5):1133–1144.

[32] Murphy KF, Adams RM, Wang X, Balázsi G, Collins JJ. Tuning and Controlling Gene Expression Noise in Synthetic Gene Networks. Nucleic. Acids. Res. 2010;38(8):2712–2726. doi:10.1093/nar/gkq091.

[33] Sochor MA. In Vitro Transcription Accurately Predicts Lac Repressor Phenotype *in Vivo* in *Escherichia coli*. PeerJ. 2014;2:e498. doi:10.7717/PeerJ.498.

[34] Rydenfelt M, Cox RS, Garcia H, Phillips R. Statistical Mechanical Model of Coupled Transcription from Multiple Promoters Due to Transcription Factor Titration. Phys. Rev. E. 2014;89(1):012702. doi:10.1103/PhysRevE.89.012702.

[35] Sivia D, Skilling J. *Data Analysis : A Bayesian Tutorial*. Oxford University Press, 2006.

[36] Oehler S, Amouyal M, Kolkhof P, von Wilcken-Bergmann B, Müller-Hill B. Quality and Position of the Three Lac Operators of *E. coli* Define Efficiency of Repression. EMBO J. 1994;13(14):3348–55.

[37] Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T. Interdependence of Cell Growth and Gene Expression: Origins and Consequences. Science. 2010;330(6007):1099–1102. doi:10.1126/science.1192588.

[38] Brophy JAN, Voigt CA. Principles of Genetic Circuit Design. Nat. Methods. 2014;11(5):508–520. doi:10.1038/nmeth.2926.

[39] Shis DL, Hussain F, Meinhardt S, Swint-Kruse L, Bennett MR. Modular, Multi-Input Transcriptional Logic Gating with Orthogonal LacI/GalR Family Chimeras. ACS Synth. Biol. 2014;3(9):645–651. doi:10.1021/sb500262f.

[40] Martins BMC, Swain PS. Trade-Offs and Constraints in Allosteric Sensing. PLoS Comput. Biol. 2011;7(11):1–13. doi:10.1371/journal.pcbi.1002261.

[41] Sourjik V, Berg HC. Receptor Sensitivity in Bacterial Chemotaxis. Proc. Natl. Acad. Sci. USA. 2002;99(1):123–127. doi:10.1073/Proc.Natl.Acad.Sci.USA.011589998.

[42] Keymer JE, Endres RG, Skoge M, Meir Y, Wingreen NS. Chemosensing in *Escherichia coli*: Two regimes of two-state receptors. Proc. Natl. Acad. Sci. USA. 2006;103(6):1786–1791. doi:10.1073/Proc.Natl.Acad.Sci.USA.0507438103.

[43] Swem LR, Swem DL, Wingreen NS, Bassler BL. Deducing Receptor Signaling Parameters from In Vivo Analysis: LuxN/AI-1 Quorum Sensing in Vibrio Harveyi. Cell. 2008;134(3):461–473. doi:10.1016/j.cell.2008.06.023.

[44] Mirny La. Nucleosome-Mediated Cooperativity Between Transcription Factors. Proc. Natl. Acad. Sci. USA. 2010;doi:10.1073/Proc.Natl.Acad.Sci.USA.0913805107.

[45] Einav T, Mazutis L, Phillips R. Statistical Mechanics of Allosteric Enzymes. J Phys. Chem. B. 2016;120(26):6021–6037. doi:10.1021/acs.jpcb.6b01911.

[46] Garcia HG, Sanchez A, Boedicker JQ, Osborne M, Gelles J, Kondev J, et al. Operator Sequence Alters Gene Expression Independently of Transcription Factor Occupancy in Bacteria. Cell. Rep. 2012;2(1):150–161. doi:10.1016/j.celrep.2012.06.004.

[47] Monod J, Changeux JP, Jacob F. Allosteric Proteins and Cellular Control Systems. J. Mol. Biol. 1963;6:306–329.

[48] Auerbach A. Thinking in Cycles: MWC Is a Good Model for Acetylcholine Receptor-Channels. J. Physiol. 2012;590(1):93–98. doi:10.1113/jphysiol.2011.214684.

[49] Velyvis A, Yang YR, Schachman HK, Kay LE. A Solution NMR Study Showing That Active Site Ligands and Nucleotides Directly Perturb the Allosteric Equilibrium in Aspartate Transcarbamoylase. Proc. Natl. Acad. Sci. USA. 2007;104(21):8815–20. doi:10.1073/Proc.Natl.Acad.Sci.USA.0703347104.

[50] Canals M, Lane JR, Wen A, Scammells PJ, Sexton PM, Christopoulos A. A Monod-Wyman-Changeux Mechanism Can Explain G Protein-Coupled Receptor (GPCR) Allosteric Modulation. J. Biol. Chem. 2012;287(1):650–659. doi:10.1074/jbc.M111.314278.

[51] Milo R, Hou JH, Springer M, Brenner MP, Kirschner MW. The Relationship Between Evolutionary and Physiological Variation in Hemoglobin. Proc. Natl. Acad. Sci. USA. 2007;104(43):16998–17003. doi:10.1073/Proc.Natl.Acad.Sci.USA.0707673104.

[52] Levantino M, Spilotros A, Cammarata M, Schirò G, Ardiccioni C, Vallone B, et al. The Monod-Wyman-Changeux Allosteric Model Accounts for the Quaternary Transition Dynamics in Wild Type and a Recombinant Mutant Human Hemoglobin. Proc. Natl. Acad. Sci. USA. 2012;109(37):14894–9. doi:10.1073/Proc.Natl.Acad.Sci.USA.1205809109.

[53] Lutz R, Bujard H. 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. Res. 1997;25(6):1203–10. doi:10.1093/NAR/25.6.1203.

[54] Moon TS, Lou C, Tamsir A, Stanton BC, Voigt CA. Genetic Programs Constructed from Layered Logic Gates in Single Cells. Nature. 2012;491(7423):249–253. doi:10.1038/nature11516.

[55] Tungtur S, Skinner H, Zhan H, Swint-Kruse L, Beckett D. In Vivo Tests of Thermodynamic Models of Transcription Repressor Function. Biophys. Chem. 2011;159(1):142–51. doi:10.1016/j.bpc.2011.06.005.

[56] Forsén S, Linse S. Cooperativity: Over the Hill. Trends. Biochem. Sci. 1995;20(12):495–497. doi:10.1016/S0968-0004(00)89115-X.

[57] Jones DL, Brewster RC, Phillips R. Promoter Architecture Dictates Cell-To-cell Variability in Gene Expression. Science. 2014;346(6216):1533–1536. doi:10.1126/science.1255301.

[58] Eldar A, Elowitz MB. Functional Roles for Noise in Genetic Circuits. Nature. 2010;467(7312):167–173. doi:10.1038/nature09326.

[59] Gerland U, Hwa T. On the Selection and Evolution of Regulatory DNA Motifs. J. Mol. Evol. 2002;55(4):386–400. doi:10.1007/s00239-002-2335-z.

[60] Berg J, Willmann S, Lässig M. Adaptive Evolution of Transcription Factor Binding Sites. BMC Evol. Biol. 2004;4(1):42. doi:10.1186/1471-2148-4-42.

[61] Zeldovich KB, Shakhnovich EI. Understanding Protein Evolution: From Protein Physics to Darwinian Selection. Annu. Rev. Phys. Chem. 2008; 59(1):105–127. doi:10.1146/annurev.physchem.58.032806.104449.

Chapter S2

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

A detailed description of the experimental methodology and the computational notebooks used in this work can be found in the published manuscript.

S2.1 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 Ref. [1] to determine the parameter $\Delta\varepsilon_{AI} = 4.5\ k_BT$, after which the remaining parameters K_A and K_I can be extracted from any induction profile with no further degeneracy.

S2.1.1 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 in Fig. 2.4C to the fold-change Eq. 2.5 and extract the three unknown parameters (K_A, K_I, and $\Delta\varepsilon_{AI}$), then multiple degenerate parameter sets would yield equally good fits. In other words, this data set alone is insufficient to uniquely determine the actual physical parameter values of the system. This problem persists even when fitting multiple data sets simultaneously as in Appendix S2.2.

In Fig. S2.1A, 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 Eq. 2.5 but with $\beta\Delta\varepsilon_{RA}$ modified to the form $\beta\Delta\tilde{\varepsilon}_{RA}$ in Eq. S2.4 to account for the underlying assumptions used when fitting previous data (see Section S2.1.2 for a full explanation of why this modification is needed).

The best-fit curves for several different values of $\Delta\varepsilon_{AI}$ are shown in Fig. S2.1B.

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. S2.1A 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 [2, 3].

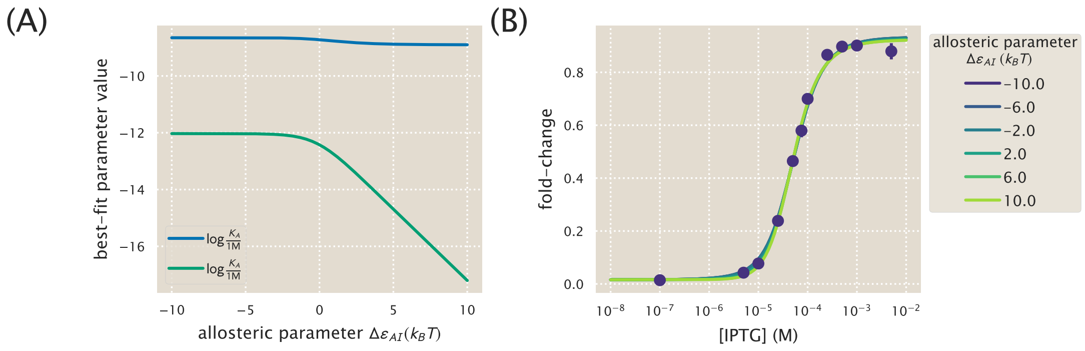


Figure S2.1: Multiple sets of parameters yield identical fold-change responses. (A) The data for the O2 strain ($\Delta\varepsilon_{RA} = -13.9\ k_BT$) with $R = 260$ in Fig. 2.4C was fit using Eq. 2.5 with $n = 2$. $\Delta\varepsilon_{AI}$ is forced to take on the value shown on the x-axis, while the 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.

S2.1.2 Computing $\Delta\varepsilon_{AI}$

As shown in the previous section, the fold-change response of a single strain is not sufficient to determine the three MWC parameters (K_A, K_I, and $\Delta\varepsilon_{AI}$), since degenerate sets of parameters yield nearly identical fold-change responses. To circumvent this degeneracy, we now turn to some previous data from the *lac* system in order to determine the value of $\Delta\varepsilon_{AI}$ in Eq. 2.5 for the induction of the Lac repressor. Specifically, we consider two previous sets of work from: (1) Ref. [4] and (2) Ref. [1], 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,

$$\text{fold-change}(c = 0) = \left(1 + \frac{R}{N_{NS}} e^{-\beta\Delta\tilde{\varepsilon}_{RA}}\right)^{-1}. \tag{S2.1}$$

Note that this functional form does not exactly match our fold-change Eq. 2.5 in the limit $c = 0$,

$$\text{fold-change}(c = 0) = \left(1 + \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} \frac{R}{N_{NS}} e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1}, \tag{S2.2}$$

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,

$$\frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} = p_A(0). \tag{S2.3}$$

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

$$\beta\Delta\tilde{\varepsilon}_{RA} = \beta\Delta\varepsilon_{RA} - \log\left(\frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}}\right). \tag{S2.4}$$

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. S2.2 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 Ref. [4] are effective parameters $\Delta\tilde{\varepsilon}_{RA}$. In this case, we are faced with an undetermined system in which we have more variables than equations, and we are thus unable to determine the value of $\Delta\varepsilon_{AI}$. In order to obtain this parameter, we must turn to a more complex regulatory scenario which provides additional constraints that allow us to fit for $\Delta\varepsilon_{AI}$.

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

To obtain an expression for a system with multiple promoters N, we follow Ref. [5], writing the fold-change in terms of the grand canonical ensemble as

$$\text{fold-change} = \frac{1}{1 + \lambda_r e^{-\beta\Delta\varepsilon_{RA}}}, \tag{S2.5}$$

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

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

$$R_{\text{tot}} = R_S + R_{NS}, \tag{S2.6}$$

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

$$R_S = N \frac{\lambda_r e^{-\beta\Delta\varepsilon_{RA}}}{1 + \lambda_r e^{-\beta\Delta\varepsilon_{RA}}}, \tag{S2.7}$$

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

$$R_{NS} = N_{NS} \frac{\lambda_r}{1 + \lambda_r}, \tag{S2.8}$$

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 Eqs. S2.7 and S2.8 into the modified Eq. S2.6 yields the form

$$p_A(0)R_{\text{tot}} = \frac{1}{1 + e^{-\beta\Delta\varepsilon_{AI}}} \left(N \frac{\lambda_r e^{-\beta\Delta\varepsilon_{RA}}}{1 + \lambda_r e^{-\beta\Delta\varepsilon_{RA}}} + N_{NS} \frac{\lambda_r}{1 + \lambda_r} \right), \tag{S2.9}$$

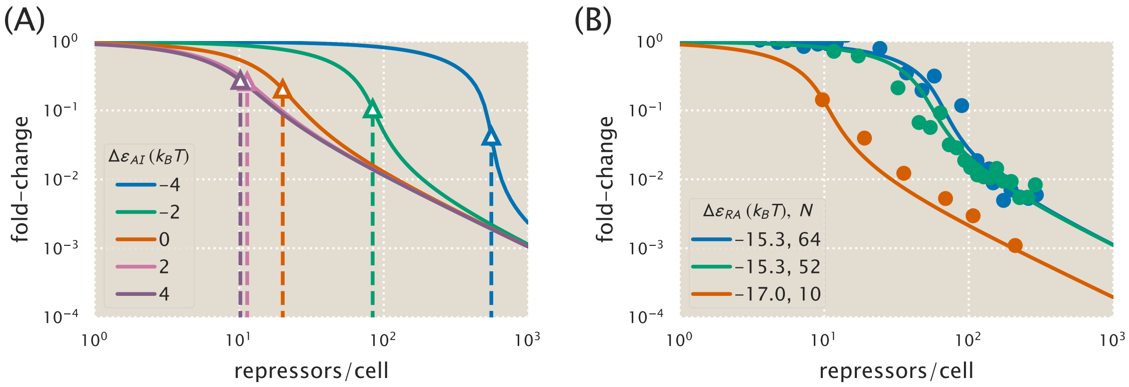


Figure S2.2: Fold-change of multiple identical genes. (A) In the presence of $N = 10$ identical promoters, the fold-change Eq. S2.5 depends strongly on the allosteric energy difference $\Delta\varepsilon_{AI}$ between the Lac repressor's active and inactive states. The vertical dotted lines represent the number of repressors at which $R_A = N$ for each value of $\Delta\varepsilon_{AI}$. (B) Using previous fold-change measurements for the operators and gene copy numbers shown, we can determine the most likely value $\Delta\varepsilon_{AI} = 4.5\ k_BT$ for LacI.

where we recall from Eq. S2.4 that $\beta\Delta\varepsilon_{RA} = \beta\Delta\tilde{\varepsilon}_{RA} + \log\left(\frac{1}{1+e^{-\beta\Delta\varepsilon_{AI}}}\right)$. Numerically solving for λ_r and plugging the value back into Eq. S2.5 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. S2.2A 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. S2.9 occurs near $p_A(0)R_{\text{tot}} = N$ (as shown by the triangles in Fig. S2.2A), 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 $\Delta\varepsilon_{AI} \gtrsim 5\ k_BT$, 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.

We now analyze experimental induction data for different strains with known plasmid copy numbers to determine $\Delta\varepsilon_{AI}$. Fig. S2.2B 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 Ref. [1]. By fitting these data to Eq. S2.5, we extracted the parameter value $\Delta\varepsilon_{AI} = 4.5\ k_BT$. Substituting this value into Eq. S2.3 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 Ref. [1, 4] were accurate.

S2.2 Global Fit of All Parameters

In the main text, we used the repressor copy numbers R and repressor-DNA binding energies $\Delta\varepsilon_{RA}$ as reported by Ref. [4]. However, any error in these previous measurements of R and $\Delta\varepsilon_{RA}$ will necessarily propagate into our own fold-change predictions. In this section, we take an alternative approach to fitting the physical parameters of the system to that used in the main text. First, rather than fitting only a single strain, we fit the entire data set in Fig. 2.5 along with microscopy data for the synthetic operator Oid (see Appendix S2.3). 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 S2.1.

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

Fig. S2.3 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 improves the agreement between the theory and the data. Table S2.2 summarizes the values of the parameters as obtained with this MCMC parameter inference. We note that even though we allowed the repressor copy numbers and repressor-DNA binding energies to vary, the resulting fit values were very close to the previously reported values. The fit values of the repressor copy numbers were all within one standard deviation of the previous reported values provided in Ref. [4]. 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_BT, which represents a small change in the biological scales we are considering. The biggest discrepancy between our fit values and the previous measurements arose for the synthetic Oid

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 repressor will be in the active state
p_{bound}	Probability RNAP is bound to promoter of interest, assumed proportional to gene expression
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
$\beta = \frac{1}{k_BT}$	The inverse product of the Boltzmann constant k_B and the temperature T of the system

Table S2.1: Key model parameters for induction of an allosteric repressor.

operator, which we discuss in more detail in Appendix S2.3.

S2.3 Applicability of Theory to the Oid Operator Sequence

In addition to the native operator sequences (O1, O2, and O3) considered in the main text, we were also interested in testing our model predictions against the synthetic Oid operator. In contrast to the other operators, Oid is one base pair shorter in length (20 bp), is fully symmetric, and is known to provide stronger repression than the native operator sequences considered so far. While the theory should be similarly applicable, measuring the lower fold-changes associated with this YFP construct was expected to be near the sensitivity limit for our flow cytometer, due to the especially strong binding energy of Oid ($\Delta\varepsilon_{RA} = -17.0\ k_BT$) [6]. Accordingly, fluorescence data for Oid were obtained using microscopy, which is more sensitive than flow cytometry.

We follow the approach of the main text and make fold-change predictions based on the parameter estimates from our strain with $R = 260$ and an O2 operator. These predictions are shown in Fig. S2.4A, 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 Appendix S2.2) to see whether we might find better agreement with the observed data. Interestingly, we find that the majority

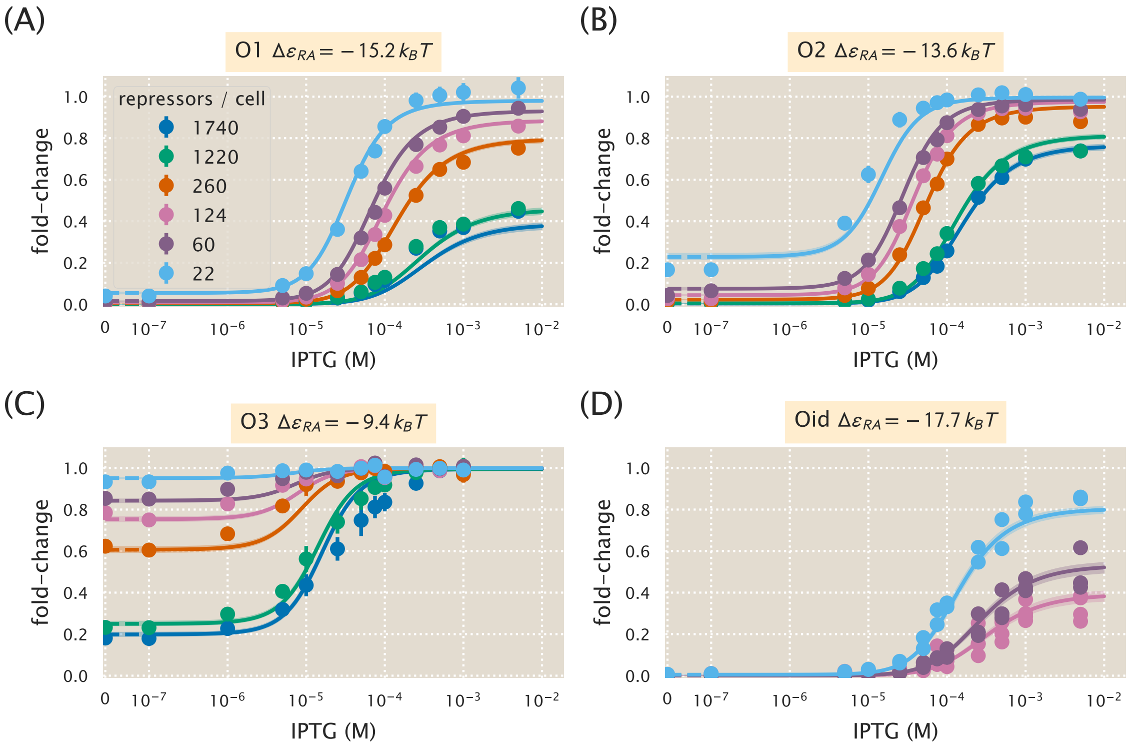


Figure S2.3: Global fit of dissociation constants, repressor copy numbers and binding energies. Theoretical predictions resulting from simultaneously fitting the dissociation constants K_A and K_I, the six repressor copy numbers R, and the four repressor-DNA binding energies $\Delta\varepsilon_{RA}$ 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. For the Oid operator, all of the data points are shown 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 indicates a linear scale, whereas solid lines represent a log scale.

of the parameters remain largely 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 Ref. [4]. In Fig. S2.4B we again plot the Oid fold-change data but with theoretical predictions using the new estimate for the Oid binding energy from our global fit and find substantially better agreement.

Fig. S2.5 shows the cumulative data from Ref. [4] and Ref. [1], 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

	Reported Values [4]	**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} μM
K_I	–	$0.73^{+0.04}_{-0.04}$ μM
R_{22}	22 ± 4	20^{+1}_{-1}
R_{60}	60 ± 20	74^{+4}_{-3}
R_{124}	124 ± 30	130^{+6}_{-6}
R_{260}	260 ± 40	257^{+9}_{-11}
R_{1220}	1220 ± 160	1191^{+32}_{-55}
R_{1740}	1740 ± 340	1599^{+75}_{-87}
O1 $\Delta\varepsilon_{RA}$	-15.3 ± 0.2 k_BT	$-15.2^{+0.1}_{-0.1}$ k_BT
O2 $\Delta\varepsilon_{RA}$	-13.9 ± 0.2 k_BT	$-13.6^{+0.1}_{-0.1}$ k_BT
O3 $\Delta\varepsilon_{RA}$	-9.7 ± 0.1 k_BT	$-9.4^{+0.1}_{-0.1}$ k_BT
Oid $\Delta\varepsilon_{RA}$	-17.0 ± 0.2 k_BT	$-17.7^{+0.2}_{-0.1}$ k_BT

Table S2.2: 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 ± 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_BT. The logarithms $\tilde{k}_A = -\log \frac{K_A}{1\,\text{M}}$ and $\tilde{k}_I = -\log \frac{K_I}{1\,\text{M}}$ of the dissociation constants were fit for numerical stability.

the binding energies from the global fit, including $\Delta\varepsilon_{RA} = -17.7$ k_BT, compare reasonably well with all previous measurements.

S2.4 Applications to Other Regulatory Architectures

In this section, we discuss how the theoretical framework presented in this work is sufficiently general to include a variety of regulatory architectures outside of simple repression by LacI. We begin by noting that the exact same formula for fold-change given in 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

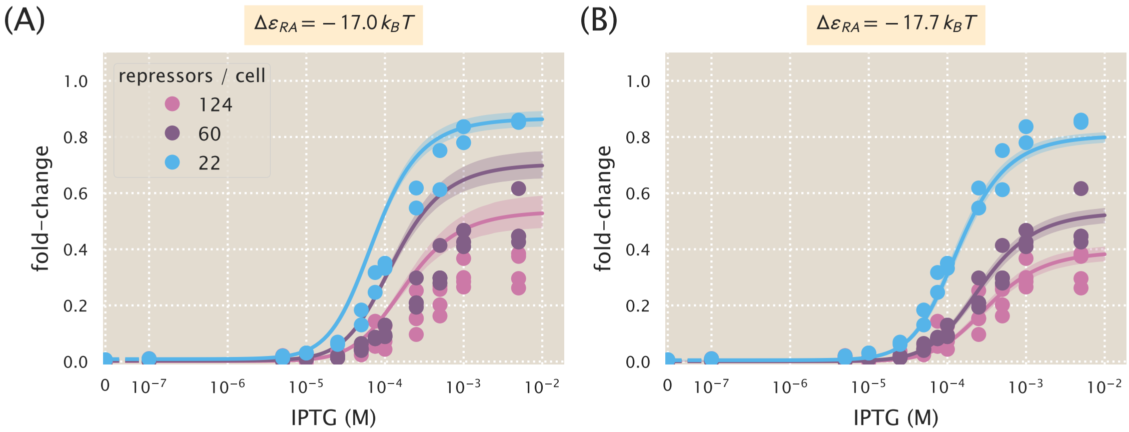


Figure S2.4: Predictions of fold-change for strains with an Oid binding sequence versus experimental measurements with different repressor copy numbers. (A) Experimental data is plotted against the parameter-free predictions that are based on our fit to the O2 strain with $R = 260$. Here we use the previously measured binding energy $\Delta\varepsilon_{RA} = -17.0\ k_BT$. (B) 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 Appendix S2.2). 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 B 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.

the interested reader to read Ref. [7, 8].

S2.4.1 Corepression

Consider a regulatory architecture where binding of a transcriptional repressor occludes the binding of RNAP to the DNA. A corepressor molecule binds to the repressor and shifts its allosteric equilibrium towards the active state in which it binds more tightly to the DNA, thereby decreasing gene expression (in contrast, an inducer shifts the allosteric equilibrium towards the inactive state where the repressor binds more weakly to the DNA). As in the main text, we can enumerate the states and statistical weights of the promoter and the allosteric states of the repressor. We note that these states and weights exactly match Fig. 2.2 and yield the same fold-change

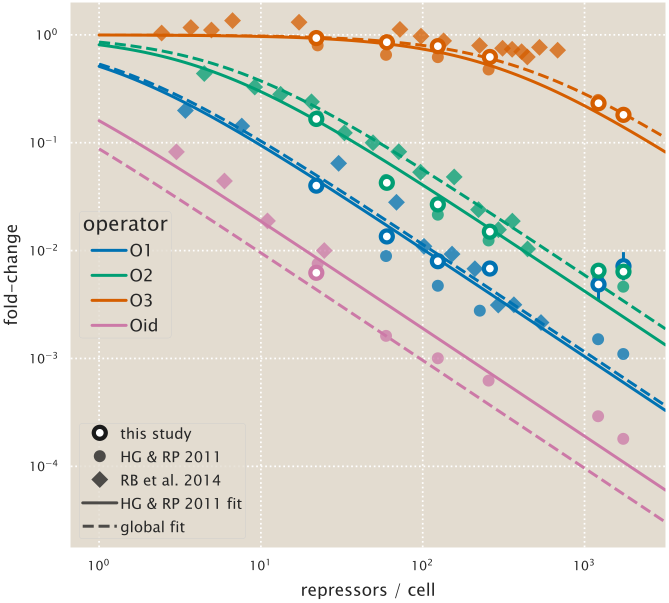


Figure S2.5: Comparison of fold-change predictions based on binding energies from Garcia and Phillips and those inferred from this work. Fold-change curves for the different repressor-DNA binding energies $\Delta\varepsilon_{RA}$ are plotted as a function of repressor copy number when IPTG concentration $c = 0$. Solid curves use previously determined binding energies, 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 this work and from previous measurements 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. S2.4 had extremely low fold-change in the absence of inducer ($c = 0$) so as to be indistinguishable from autofluorescence, and in fact their fold-change values in this limit were negative and hence do not appear on this plot.

equation as Eq. 2.5,

$$\text{fold-change} \approx \left(1 + \frac{\left(1+\frac{c}{K_A}\right)^n}{\left(1+\frac{c}{K_A}\right)^n + e^{\beta\Delta\varepsilon_{AI}}\left(1+\frac{c}{K_I}\right)^n}\frac{R}{N_{NS}}e^{-\beta\Delta\varepsilon_{RA}}\right)^{-1}, \tag{S2.10}$$

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 the main text from the induction data (see Fig. 2.4). Much as was performed in the main text, we can make some predictions about the how the response of a corepressor. In Fig. S2.6A, we show how varying the repressor copy number R and the repressor-DNA binding energy $\Delta\varepsilon_{RA}$ influence the response. We draw the reader's attention to the decrease in fold-change as the concentration of effector is increased.

S2.4.2 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 ε_{AP}. Note that in this architecture, binding of the activator does not occlude binding of the polymerase. Binding of a coactivator molecule shifts its allosteric equilibrium towards the active state ($K_A < K_I$), where the activator is more likely to be bound to the DNA and promote expression. Enumerating all of the states and statistical weights of this architecture and making the approximation that the promoter is weak generates a fold-change equation of the form

$$\text{fold-change} = \frac{1 + \frac{\left(1+\frac{c}{K_A}\right)^n}{\left(1+\frac{c}{K_A}\right)^n + e^{\beta\Delta\varepsilon_{AI}}\left(1+\frac{c}{K_I}\right)^n}\frac{A}{N_{NS}}e^{-\beta\Delta\varepsilon_{AA}}e^{-\beta\varepsilon_{AP}}}{1 + \frac{\left(1+\frac{c}{K_A}\right)^n}{\left(1+\frac{c}{K_A}\right)^n + e^{\beta\Delta\varepsilon_{AI}}\left(1+\frac{c}{K_I}\right)^n}\frac{A}{N_{NS}}e^{-\beta\Delta\varepsilon_{AA}}}, \tag{S2.11}$$

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 ε_{AP} is the interaction energy between the activator and the RNAP. Unlike in the cases of induction and corepression, the fold-change formula for activation includes terms from when the RNAP is bound by itself on the DNA as well as when both RNAP and the activator are simultaneously bound to the DNA. Fig. S2.6B 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 result in dramatic increase in fold-change.

As in the case of induction, the Eq. S2.11 is straightforward to generalize. For example, the relative values of K_I and K_A can be switched such that $K_I < K_A$ in which the secondary molecule drives the activator to assume the inactive state represents induction of an activator. While these cases might be viewed as separate biological phenomena, mathematically they can all be described by the same underlying formalism.

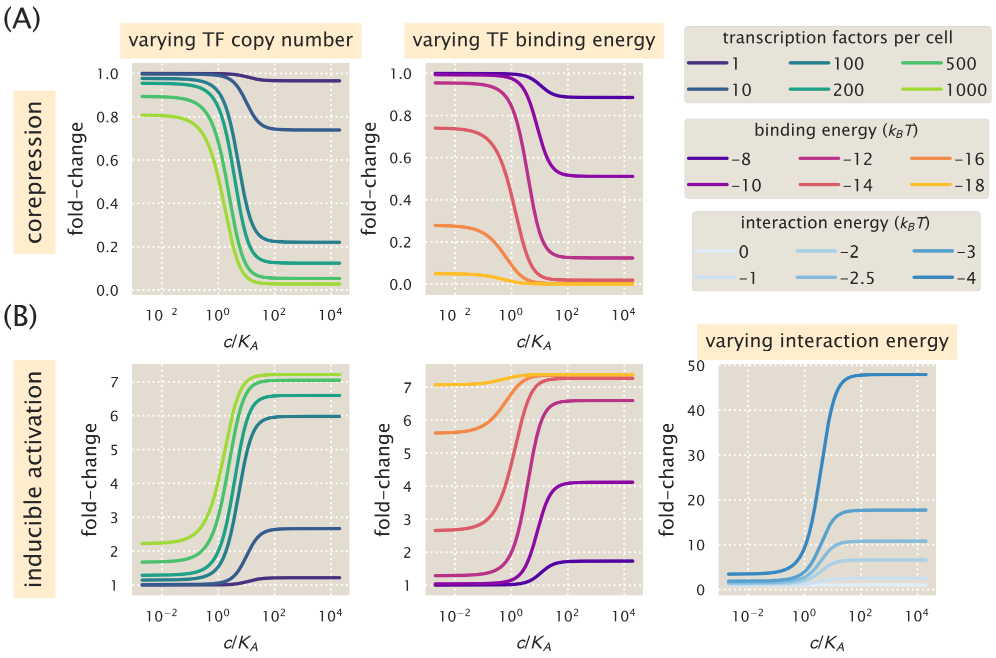


Figure S2.6: Representative fold-change predictions for allosteric corepression and activation. (A) Contrary to the case of induction described in the main text, addition of a corepressor decreases fold-change in gene expression. The left and right panels demonstrate how varying the values of the repressor 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 response, respectively. The right panel shows how varying the polymerase-activator interaction energy ε_{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.

References

[1] Brewster RC, Weinert FM, Garcia HG, Song D, Rydenfelt M, Phillips R. The Transcription Factor Titration Effect Dictates Level of Gene Expression. Cell. 2014;156(6):1312–23. doi:10.1016/j.cell.2014.02.022.

[2] Gardino AK, Volkman BF, Cho HS, Lee SY, Wemmer DE, Kern D. The NMR Solution Structure of BeF3-Activated Spo0F Reveals the Conformational Switch in a Phosphorelay System. J. Mol. Biol. 2003;331(1):245–254. doi:10.1016/S0022-2836(03)00733-2.

[3] Boulton S, Melacini G. Advances in NMR Methods To Map Allosteric Sites: From Models to Translation. Chem. Rev. 2016;116(11):6267–6304. doi:10.1021/acs.chemrev.5b00718.

[4] Garcia HG, Phillips R. Quantitative Dissection of the Simple Repression Input-Output Function. Proc. Natl. Acad. Sci. USA. 2011;108(29):12173–8. doi:10.1073/Proc.Natl.Acad.Sci.USA.1015616108.

[5] Weinert FM, Brewster RC, Rydenfelt M, Phillips R, Kegel WK. Scaling of Gene Expression with Transcription-Factor Fugacity. Phys. Rev. Lett. 2014;113(25):1–5. doi:10.1103/PhysRevLett.113.258101.

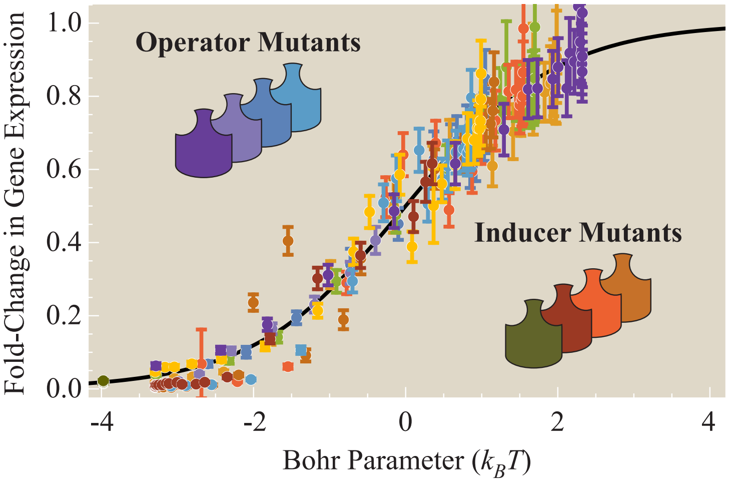
[6] Garcia HG, Lee HJ, Boedicker JQ, Phillips R. Comparison and Calibration of Different Reporters for Quantitative Analysis of Gene Expression. Biophys. J. 2011;101(3):535–544. doi:10.1016/j.bpj.2011.06.026.

[7] Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, et al. Transcriptional Regulation by the Numbers: Models. Curr Opin Genetics Dev. 2005;15(2):116–124. doi:10.1016/j.gde.2005.02.007.

[8] Marzen S, Garcia HG, Phillips R. Statistical Mechanics of Monod-Wyman-Changeux (MWC) Models. J. Mol. Biol. 2013;425(9):1433–1460. doi:10.1016/j.jmb.2013.03.013.

Chapter 3

MUTATIONS FROM A THERMODYNAMIC PERSPECTIVE



This short paper remains one of the most important works of my PhD, marking my first sojourn of modeling mutations in allosteric proteins. Shown below is the publication-ready draft written during the third year of my PhD, before four graduate students in the Phillips lab joined the project to see how well the model predictions matched experiments. Careful, rigorous experiments take time, and this project is ongoing in the lab, although we now have sufficient data to convincingly demonstrate that the key prediction in this work – that single mutants combine in an epistasis-free manner – was correct. In the interim, I have extended this idea of decoupling mutational epistasis through statistical mechanical modeling to other transcription factors (Chapter 4) and to ion channels (Chapter 5), giving further support for the model. Because the field of mutations is both central to biological evolution and steeped in complexity, I am particularly proud of the inroads I made with this work!

3.1 Abstract

Predicting the effects of mutations on protein function has been a long-standing goal in biology, and simple quantitative models to respond to this challenge are hard to come by. Here, we show that a thermodynamic model of transcriptional regulation coupled with a statistical mechanical treatment of the induction of allosteric proteins can be used to characterize large classes of mutants. Specifically, we link the physiological role of mutations within a transcription factor to physical parameters present in the Monod-Wyman-Changeux model of allostery. For example, a

mutation in a repressor's DNA or inducer binding domains only affects its DNA or inducer dissociation constants, respectively, leaving all other parameters unchanged. This model suggests a unique perspective to understand and quantify the roles of mutations and enables us, perhaps surprisingly, to collapse the plethora of data on different mutants into a single family of curves. It further provides tight constraints on important protein characteristics, such as the dynamic range and leakiness of inducible transcription factors, which bolsters our ability to theoretically predict the effects of mutations.

3.2 Introduction

Mutations pervade every aspect of biology from evolution to disease, providing researchers with a critical resource to understand and manipulate biological systems. Although significant strides have been made to characterize mutations, systematic efforts are hampered by the overwhelmingly large space of possible mutants. Given an average protein comprised of 300 amino acids, substituting each of the 20 amino acids in every position would yield more proteins than there are atoms in the universe.

The consequences of mutations are often extremely difficult to predict – even a single amino acid mutation may cause a protein to stop functioning, as seen in the case of hemoglobin and sickle cell anemia [1]. Consequently, it is tempting to treat each mutant as a new intellectual adventure with no visible quantitative link to its unmutated (i.e. wild type) form. The task of quantifying this sequence-structure relationship is a central one within diverse fields of biology ranging from molecular evolution to structural biology, and many quantitative techniques have been developed towards this goal [2–4]. This paper investigates a quantitative and mechanistic framework to predict the functional roles of mutations. More specifically, we explore the hypothesis that many point mutations have a local effect, which in the context of the Monod-Wyman-Changeux (MWC) model of allostery implies that point mutations only change a subset of the parameters characterizing that model, an interesting twist since the very notion of allostery is predicated upon the "communication" between different parts of a protein [5].

This paper builds upon substantial earlier work in the context of both chemotaxis and quorum sensing where different mutants in membrane receptors led to activity curves which could be classified as one-parameter families within the MWC framework [6, 7]. This approach enables us to characterize large classes of transcription factor mutants based upon the location and the physiological role of their underlying mutations. For example, we test the notion that any mutation inside of a binding motif only alters that binding region and leaves the coarse-grained description of the rest of the protein, through its associated parameters, unaffected and identical to the wild type form [8]. Such a treatment limits the phenotypes of binding domain mutants and makes it possible to theoretically predict the effects of mutations on important characteristics such as leakiness and dynamic range.

Specifically, we focus on the induction of Lac repressor in *Escherichia coli* and its effects on transcriptional regulation, a core cellular process that allows an organism to sense and respond to its environment by altering its level of gene expression. The Lac repressor regulates the transcription of the lactose (*lac*) operon whose gene products allow *E. coli* to digest the sugar lactose if glucose is not available while lactose is. The Lac repressor is an allosteric protein which exists in two conformations: an active state which tightly binds to the Lac operator and an inactive state with a much lower affinity for the DNA. When the natural inducer allolactose or the gratuitous inducer isopropyl β-D-1-thiogalactopyranoside (IPTG) bind to the Lac repressor, they make the repressor more likely to assume the inactive state, thereby hampering the repressor's ability to bind to the DNA and increasing gene expression.

Using the MWC model, we first characterize the thermodynamic parameters of the wild type Lac repressor (namely, its DNA binding affinity, inducer binding affinity, and the energy difference between the repressor's allosteric states). We then analyze three Lac repressors with mutations in their operator binding regions. We find that these mutants are well characterized by only varying their repressor-DNA binding affinity, leaving all other parameters identical to those describing the wild type Lac repressor. A key outcome of this analysis is that the data from the wild type Lac repressor and all mutants can be collapsed onto a single master curve. In this way, the MWC model provides a unifying framework that allows us to mathematically characterize different mutants as members of a one-parameter family.

We next analyze ten mutants with mutations in their inducer binding sites and find that these mutants also obey the simple model where only their repressor-inducer binding affinity changes. This notion may appear to contradict the principle of allostery, since when an inducer binds to a repressor it hampers that repressor's ability to bind to DNA. Yet our results demonstrate that although the Lac repressor's DNA and inducer binding domains are allosterically linked, they can still be independently modified as exhibited by the insulation of the corresponding thermodynamic parameters that govern these two distinct classes of mutants.

In addition to categorizing existing data, this framework provides new insights into important metrics such as the leakiness and dynamic range of a transcription factor. For example, the MWC model correctly predicts that mutating the inducer binding region has no effect on the leakiness of the Lac repressor. We also find that the wild type Lac repressor provides an excellent compromise between exhibiting a large dynamic range and small leakiness. The leakiness and dynamic range of all thirteen mutants studied were well characterized within the MWC framework, indicating that the DNA and inducer binding affinities are the correct set of variables, both complete and complementary, to study these Lac repressor mutants.

Lastly, it is important to realize that while we focus on the model system of the Lac repressor, this framework can be readily extended not only to other transcription factors [9], but beyond that to many other biological systems (such as enzymes [10, 11], transport proteins [12, 13], the cytoskeleton [14], and antibodies [15], to name a few) that are well characterized by models where each parameter is tied to a physical trait of a protein, thus raising the possibility of classifying mutants in a much wider setting.

3.3 Results

3.3.1 The Monod-Wyman-Changeux (MWC) Model

This paper builds upon an extensive dialogue between theory and experiment in transcriptional regulation. A first round of experiments measured the dependence of gene expression on repressor copy number and binding strength [16]. A second round of experiments pushed beyond the “independent promoter approximation” to acknowledge the fact that different genes compete for the same regulatory apparatus. Here too we were able to show that theoretical predictions of this subtle effect were consistent with their measured counterparts, even permitting the collapse of data from multiple experiments onto master curves [17, 18]. In the current paper, we consider yet another layer of complexity having to do with how signaling works in the context of transcription.

It has been shown that removing the tetramerization region in wild type Lac repressor creates a functional dimeric repressor that: (1) can bind to DNA; (2) exists in both an active and inactive allosteric conformation; and (3) has two binding sites for the inducer IPTG – we shall refer to this truncated dimeric protein as “the Lac repressor” for the remainder of this paper [5, 19].

As discussed previously, the behavior of the Lac repressor as a regulatory protein is well characterized by an equilibrium model where the probability of each state of repressor and RNA polymerase occupancy is proportional to its Boltzmann weight [20–22]. We begin with a summary of this model. Suppose there are P RNA polymerase (RNAP) and R repressor molecules in a cell. R_A repressors will be in the active state (the favored state when repressor is not bound to inducer; in this state the repressor binds tightly to DNA) and the remaining R_I repressors will be in the inactive state (the predominant state when repressor is bound to inducer; in this state the repressor binds weakly to DNA) so that $R_A + R_I = R$.

We first model the interaction between the Lac repressor and DNA by enumerating all possible states and their corresponding weights. As shown in Fig. 3.1, the Lac promoter can either be empty, occupied by RNAP, or occupied by either an active or inactive repressor molecule. Assume that there are N_{NS} non-specific sites on the DNA outside the Lac operator where RNAP or the Lac repressor can bind. $\Delta\epsilon_{PD}$ represents the energy difference between RNAP bound to the Lac operator or bound elsewhere on the DNA; $\Delta\epsilon_{RD,A}$ and $\Delta\epsilon_{RD,I}$ equal the difference in energy when the Lac repressor is bound to the Lac operator compared to when it is bound non-specifically elsewhere on the DNA in the active and inactive state, respectively. $\beta = \frac{1}{k_B T}$ where k_B is Boltzmann's constant and T is the temperature of the system.

In thermodynamic models of transcription, gene expression is proportional to the probability p_{bound} that RNAP is bound to the Lac operator which is given by

$$p_{\text{bound}}(R) = \frac{p}{1 + r_A + r_I + p}, \tag{3.1}$$

where

$$p = \frac{P}{N_{\text{NS}}} e^{-\beta\Delta\epsilon_{PD}} \tag{3.2}$$

$$r_A = \frac{R_A}{N_{\text{NS}}} e^{-\beta\Delta\epsilon_{RD,A}} \tag{3.3}$$

$$r_I = \frac{R_I}{N_{\text{NS}}} e^{-\beta\Delta\epsilon_{RD,I}}. \tag{3.4}$$

Gene expression can be readily measured experimentally by exploiting the fold-change,

$$\text{fold-change} \equiv \frac{p_{\text{bound}}(R)}{p_{\text{bound}}(0)} = \frac{1 + p}{1 + r_A + r_I + p}. \tag{3.5}$$

We can simplify this expression using two well-justified approximations: $p \ll 1$ and $r_I \ll r_A$. The first approximation is called the weak promoter approximation

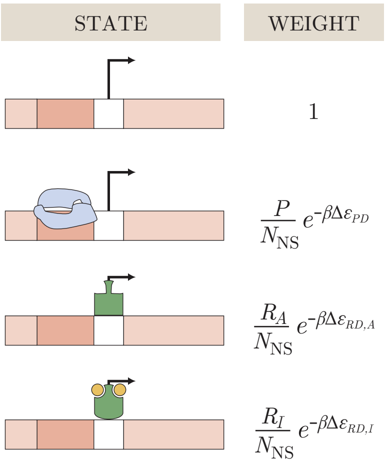

STATE
WEIGHT
1
$\frac{P}{N_{\text{NS}}} e^{-\beta\Delta\varepsilon_{PD}}$
$\frac{R_A}{N_{\text{NS}}} e^{-\beta\Delta\varepsilon_{RD,A}}$
$\frac{R_I}{N_{\text{NS}}} e^{-\beta\Delta\varepsilon_{RD,I}}$

Figure 3.1: States and weights for simple repression. Both RNAP (light blue) and repressor compete for DNA binding. There are R_A repressors in the active state (green, sharp) and R_I repressors in the inactive state (green, rounded), with the latter type typically bound to inducer (gold). The difference in energy between a repressor bound to the Lac operator and to another non-specific site on the DNA equals $\Delta\epsilon_{RD,A}$ in the active state and $\Delta\epsilon_{RD,I}$ in the inactive state; the P RNAP have a corresponding energy difference $\Delta\epsilon_{PD}$. The number of active repressors R_A includes repressors that are unbound, singly bound, or doubly bound to inducer, although the majority of active state repressors will not be bound to inducer (which pushes them into the inactive state). Similarly, the R_I term includes all inactive state repressors bound to any number of inducer molecules, with the most prevalent state shown in the figure.

and is valid for the wild type Lac promoter [23]. The second approximation follows because $e^{-\beta\Delta\epsilon_{RD,I}}$ is approximately 1000 times smaller than $e^{-\beta\Delta\epsilon_{RD,A}}$ for the Lac repressor [5]. Using these approximations, the fold-change reduces to the form

$$\text{fold-change} \approx \frac{1}{1 + r_A} = \left(1 + \frac{R_A}{N_{\text{NS}}} e^{-\beta\Delta\epsilon_{RD,A}}\right)^{-1}. \tag{3.6}$$

We now introduce the role of inducer binding, which changes the number of repressors in the active and inactive allosteric states. We define $p_A(c) \equiv \frac{R_A(c)}{R}$ to be the fraction of repressors in the active state given a concentration c of the inducer IPTG. We define V as the volume of an *E. coli* cell, $[R] = \frac{R}{V}$ as the concentration of repressors, and $K_{\text{DNA}} = \frac{N_{\text{NS}}}{V} e^{\beta\Delta\epsilon_{RD,A}}$ as the dissociation constant of the active repressor binding to the Lac operator DNA. This last expression, which links the physical energies of the system with the language of dissociation constants and chemical rates, is discussed in detail in the Supplementary Information. With these definitions, Eq. 3.6 becomes

$$\text{fold-change} = \left(1 + \frac{p_A(c)[R]V}{N_{\text{NS}}} e^{-\beta\Delta\epsilon_{RD,A}}\right)^{-1} = \left(1 + \frac{p_A(c)[R]}{K_{\text{DNA}}}\right)^{-1}. \tag{3.7}$$

As shown in Fig. 3.2, we can enumerate the relative likelihood of the eight possible conformations of the repressor (the repressor can be in an active or inactive state, and each of its two inducer binding sites can be empty or occupied), using the energy difference ϵ between a Lac repressor in the active and inactive state. From these eight states, we can compute the probability $p_A(c)$ that a repressor will be in the active state as the sum of the weights of the active states divided by the sum of the weights of every possible state, namely,

$$p_A(c) = \frac{\left(1 + \frac{c}{K_A}\right)^2}{\left(1 + \frac{c}{K_A}\right)^2 + e^{-\beta\epsilon}\left(1 + \frac{c}{K_I}\right)^2}. \tag{3.8}$$

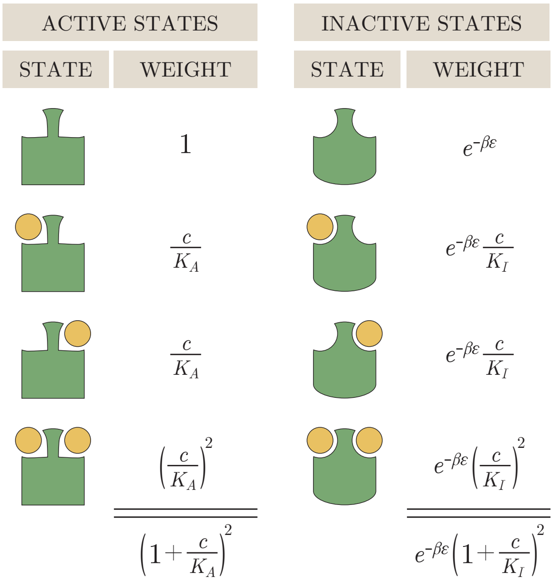

ACTIVE STATES
INACTIVE STATES
STATE
WEIGHT
STATE
WEIGHT
1
$e^{-\beta\varepsilon}$
$\frac{c}{K_A}$
$e^{-\beta\varepsilon}\frac{c}{K_I}$
$\frac{c}{K_A}$
$e^{-\beta\varepsilon}\frac{c}{K_I}$
$\left(\frac{c}{K_A}\right)^2$
$e^{-\beta\varepsilon}\left(\frac{c}{K_I}\right)^2$
$\left(1+\frac{c}{K_A}\right)^2$
$e^{-\beta\varepsilon}\left(1+\frac{c}{K_I}\right)^2$

Figure 3.2: The eight states of the Lac repressor. The Lac repressor (green) has an active conformation (left column) and inactive conformation (right column), with the energy difference between these two states given by ϵ. In each conformation, the repressor can bind an inducer (gold) at two sites. Each state is shown with its corresponding Boltzmann weight. If the sum of the active state weights shown (bottom left) is greater than the sum of the inactive state weights (bottom right), the repressor is more likely to be in the active state.

Substituting this result into Eq. 3.7 yields the complete formula

$$\text{fold-change} = \left(1 + \frac{\left(1 + \frac{c}{K_A}\right)^2}{\left(1 + \frac{c}{K_A}\right)^2 + e^{-\beta\epsilon}\left(1 + \frac{c}{K_I}\right)^2}\frac{[R]}{K_{\text{DNA}}}\right)^{-1}, \tag{3.9}$$

which predicts that given a concentration $[R]$ of Lac repressor and a concentration c of the inducer IPTG, the fold-change in gene expression will depend solely on 4 parameters: the DNA binding affinity of the repressor (K_{DNA}), the inducer binding affinities for the repressor in the active state (K_A) and inactive state (K_I), and the difference in free energy between the active and inactive states of the repressor (ϵ).

3.3.2 Lac Repressor Mutants

A protein with a single point mutation can often behave so differently from its unmutated form that it is difficult to draw a connection between the two. In terms of the thermodynamic model for fold-change described in Eq. 3.9, the standard practice for any mutant Lac repressor would be to measure fold-change at multiple concentrations of inducer and refit all four parameters K_{DNA}, K_A, K_I, and ϵ [13, 24, 25].

It has long been known both experimentally and through simulations that mutations deep within a binding pocket effect ligand binding much more than mutations far outside the binding domain [6, 7, 26–28]. For example, consider a Lac repressor with a single amino acid mutation in its DNA binding domain. We shall call such a mutant an *operator mutant*. We first show that operator mutants are well characterized by the simple model where only their affinity K_{DNA} between the Lac repressor and DNA changes while the other parameters K_A, K_I, and ϵ retain the same values as the wild type Lac repressor. We next test the hypothesis that a mutation within the Lac repressor's inducer binding domain (where allolactose or IPTG bind) only changes the thermodynamic parameters K_A and K_I. We shall call this latter mutant an *inducer mutant*.

The concept that mutations can be linked to a small subset of thermodynamic parameters has been shown to work well in the context of bacterial chemotaxis [29] and quorum sensing [7]. Extending it beyond these model systems would drastically improve our ability to characterize mutants [30] and help guide future experiments aimed at designing specific gene expression profiles. Here we show that, like in the cases of quorum sensing and chemotaxis, the use of allosteric models allows us to unify broad classes of mutants and make quantitative predictions about their leakiness and dynamic range. We now turn to an analysis of operator and inducer mutants within the Lac system.

The Bohr Parameter It is instructive to isolate the effects of the thermodynamic parameters by rewriting the fold-change Eq. 3.9 as

$$\text{fold-change} = \frac{1}{1 + e^{-\beta F(c)}} \tag{3.10}$$

where we have introduced the *Bohr parameter*,

$$F(c) \equiv -k_BT \log\left(\frac{\left(1+\frac{c}{K_A}\right)^2}{\left(1+\frac{c}{K_A}\right)^2 + e^{-\beta\epsilon}\left(1+\frac{c}{K_I}\right)^2}\frac{[R]}{K_{\text{DNA}}}\right). \tag{3.11}$$

The Bohr parameter captures many of the important properties of the system including whether DNA is more likely to be bound to repressor ($F(c) < 0$) or unbound ($F(c) > 0$) [6, 7]. Our reasoning for calling this the Bohr parameter is suggested by work from Mirny that hypothesizes an analogy between allosteric transitions in chromatin and in the binding of oxygen to hemoglobin [31]. Specifically, the Bohr effect refers to the change in oxygen's binding affinity to hemoglobin as a function of pH, which results in families of binding curves analogous to those found for different transcription factor mutants.

Consider the minimum and maximum values that $\frac{F(c)}{k_BT}$ can attain over all concentrations c of inducer. Assuming $K_A \geq K_I$, the minimum and maximum values are given by

$$\frac{F_{\text{min}}}{k_BT} = \lim_{c\to 0}\frac{F(c)}{k_BT} = \log\left(1+e^{-\beta\epsilon}\right) - \log\left(\frac{[R]}{K_{\text{DNA}}}\right) \tag{3.12}$$

and

$$\frac{F_{\text{max}}}{k_BT} = \lim_{c\to\infty}\frac{F(c)}{k_BT} = \log\left(1+e^{-\beta\epsilon}\left(\frac{K_A}{K_I}\right)^2\right) - \log\left(\frac{[R]}{K_{\text{DNA}}}\right), \tag{3.13}$$

respectively (if $K_A < K_I$, then F_{min} is achieved as $c \to \infty$ and F_{max} as $c \to 0$).

The Bohr parameter is a natural variable with which we can understand quantities such as the leakiness or dynamic range of a repressor. Leakiness is defined as the fold-change in the absence of inducer, $c = 0$, which relates how much a gene is expressed even when the repressor is working at full strength,

$$\text{leakiness} = \frac{1}{1+e^{-\beta F_{\text{min}}}}. \tag{3.14}$$

The dynamic range equals the difference between the maximum and minimum fold-change. Assuming $K_A \geq K_I$, maximum fold-change occurs when the system is saturated with inducer ($c \to \infty$) while the minimum fold-change occurs with no inducer ($c = 0$), resulting in a dynamic range given by

$$\text{dynamic range} = \frac{1}{1+e^{-\beta F_{\text{max}}}} - \frac{1}{1+e^{-\beta F_{\text{min}}}}. \tag{3.15}$$

The maximum and minimum values of the Bohr parameter, Eqs. 3.12 and 3.13, also provide a simple way to test the theory of operator mutants and inducer mutants discussed above. For operator mutants where only K_{DNA} changes, the difference $F_{\text{max}} - F_{\text{min}}$ is predicted to stay the same,

$$\frac{F_{\text{max}} - F_{\text{min}}}{k_B T} = \log\left(\frac{1 + e^{-\beta\epsilon}\left(\frac{K_A}{K_I}\right)^2}{1 + e^{-\beta\epsilon}}\right). \tag{3.16}$$

Experimentally, this invariant quantity can be determined by making two measurements of fold-change, one in the absence of inducer ($c = 0$, $F(c) \to F_{\text{min}}$) and the other when the system is saturated with inducer ($c \to \infty$, $F(c) \to F_{\text{max}}$), and then using Eq. 3.10 to transform from fold-change to the Bohr parameter. On the other hand, for inducer mutants where only K_A and K_I vary, F_{min} is predicted to exactly equal the wild type value (or equivalently, the leakiness of the wild type and any inducer mutant should be identical). We will see both of these results in the next two sections.

We now consider experimental measurements of a broad class of Lac repressor mutants that control the expression of a promoter driving the production of a GFP reporter. By altering the amount of inducer, the fold-change in gene-expression was measured using the GFP fluorescence in the presence and absence of Lac repressor as per Eq. 3.5. The calculations described here aim to discover to what extent the apparent complexity of these different induction curves can be tamed by the appropriate underlying theoretical models.

Operator Mutants Fig. 3.3A shows inducer titration curves for wild type Lac repressor and three mutants with point mutations in the DNA binding region [19]. The wild type curve was used to fit all four parameters (K_A, K_I, ϵ, and K_{DNA}). Then the three operator mutants were fit assuming that only their K_{DNA} parameter was different from the wild type. These fits are discussed in detail in the Supplementary Information. Fig. 3.3B plots the fold-change of each mutant collapsed onto a single master curve as a function of the Bohr parameter. The horizontal line segments stretch from F_{min} to F_{max} for each operator mutant. As implied by Eq. 3.16, $F_{\text{max}} - F_{\text{min}}$ will be the same for all three operator mutants and the wild type, which can be seen by the identical size of all four horizontal bars.

As a technical note, in fitting the variables K_A, K_I, ϵ, and K_{DNA} from Eq. 3.9, we make the further assumption that $[R]$ remains constant among all mutants (with $[R] \approx 11 \pm 2\,\text{nM}$ for the promoter used in the experiment) [16, 32]. However, point mutations may cause proteins to improperly fold and hence be degraded at significantly faster rates [33, 34], and if such an effect exists we implicitly absorb it into our K_{DNA} value.

Using Eqs. 3.14 and 3.15, we can plot the leakiness and dynamic range of a general operator mutant as a function of K_{DNA}, using the values of K_A, K_I, and ϵ from the wild type titration curve. Fig. 3.4 shows the leakiness and dynamic range for any K_{DNA} value along with the theoretical best fit values from the four titration curves in Fig. 3.3(A).

As shown in Fig. 3.4A, increasing K_{DNA} increases the leakiness in a sigmoidal fashion. From Fig. 3.4B, we see that the dynamic range has a symmetric peak (on a logarithmic plot); from Eq. 3.15 it is straightforward to see that the peak occurs at

$$\left(\frac{[R]}{K_{\text{DNA}}}\right)_{\substack{\text{max}\\ \text{dynamic}\\ \text{range}}} = \sqrt{\left(1+e^{-\beta\epsilon}\right)\left(1+e^{-\beta\epsilon}\left(\frac{K_A}{K_I}\right)^2\right)}. \tag{3.17}$$

Interestingly, the wild type values for Lac repressor appear to be an excellent com-

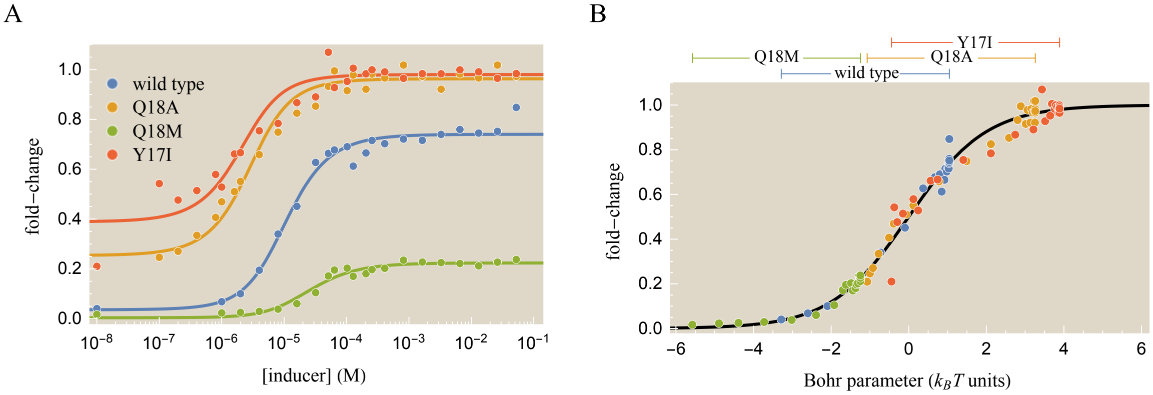


Figure 3.3: Operator mutants only vary their DNA binding affinity. (A) Fold-change in gene expression of the promoter controlled by Lac repressor as a function of inducer (IPTG) concentration. Data is shown for wild type Lac repressor and three mutants with point mutations in their DNA binding region. First, the wild type Lac repressor data was fit to the theoretical fold-change expression, Eq. 3.9. Following this, the fold-change profiles of the three mutants were fit by only varying their DNA binding affinity, K_{DNA}, while keeping their remaining thermodynamic parameters (K_A, K_I, ϵ) equal to the wild type values. (B) Each mutant can be collapsed onto the same curve using the Bohr parameter $F(c)$ given by Eq. 3.11. A larger Bohr parameter indicates that the repressor is less likely to be bound to the Lac operator. The horizontal bars stretching from F_{min} to F_{max} have the same length for all four repressors as determined by Eq. 3.16. Standard deviation from triplicate measurements and the best fit parameters are shown in the Supplementary Information.

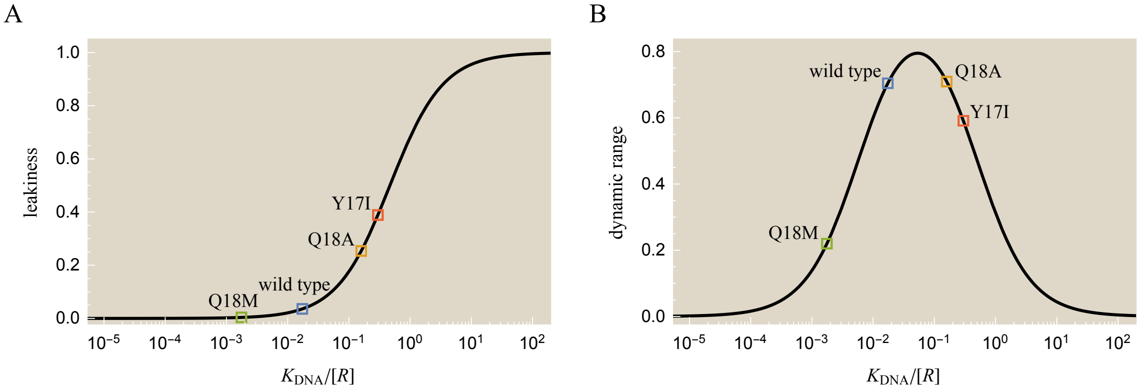


Figure 3.4: Theoretical values of leakiness and dynamic range for operator mutants. (A) Assuming a fixed repressor concentration $[R]$ for all operator mutants, leakiness monotonically increases with increasing K_{DNA} while (B) the dynamic range has a peak. The wild type repressor exhibits a good compromise between having a large dynamic range and a small leakiness. Leakiness and dynamic range values for the four repressors shown are taken from the best fit curves in Fig. 3.3 and are not measured directly from data. Thus, we represent these values as empty squares, and they fall exactly upon the black theoretical curve.

promise between small leakiness and large dynamic range.

Inducer Mutants Fig. 3.5A shows inducer titration curves for wild type Lac repressor and ten mutants with point mutations in the inducer binding region [19]. Fig. 3.5B collapses all of these titration profiles onto a single master curve using the Bohr parameter. As in the case of operator mutants, the data matches well to the simple model where only the inducer binding constants K_A and K_I vary within the mutants, while the K_{DNA} and ϵ parameters are equal to those of the wild type repressor. Note from Eq. 3.5 that fold-change values greater than 1 are an indication of the noise in the measurements. The fits and the error are discussed further in the Supplementary Information.

Of these ten inducer mutants, only one (Q291R) was rendered completely unable to bind to inducer (or alternatively its K_A and K_I increased to the point that an IPTG concentration larger than 0.1 M was required to induce it). Yet even if Q291R causes the Lac repressor to cease functioning, the fact that all three operator mutants and nine out of the ten inducer mutants we considered are well categorized by this simple allosteric model gives confidence in the possible broader applicability of such thinking.

From Eqs. 3.12 and 3.14, the leakiness of all inducer mutants should be exactly

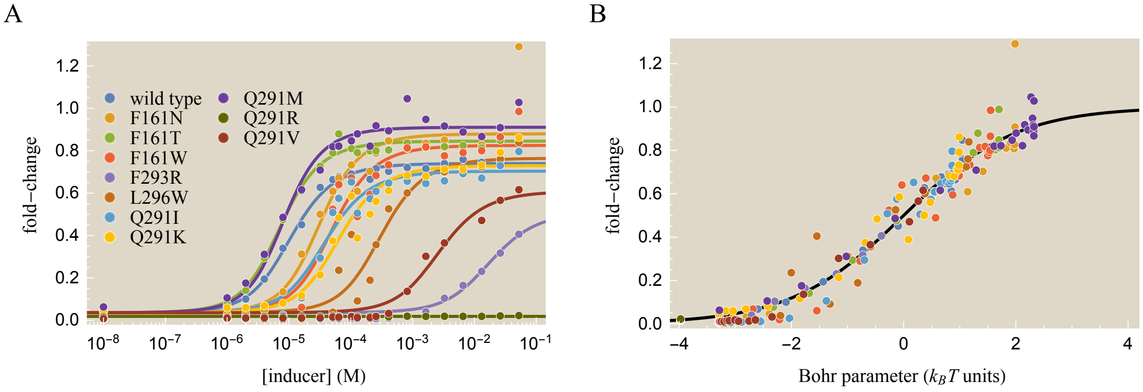


Figure 3.5: Inducer mutants only vary their inducer binding affinities. (A) Ten different point mutations in the Lac repressor's inducer binding region can be well characterized by only varying the mutant repressor's binding affinity to IPTG in the active (K_A) and inactive (K_I) state, while keeping the other thermodynamic parameters (K_{DNA} and ϵ) in Eq. 3.9 equal to the wild type values. (B) Data collapse using the Bohr parameter from Eq. 3.11. Standard deviation from triplicate measurements and the best fit parameters are shown in the Supplementary Information.

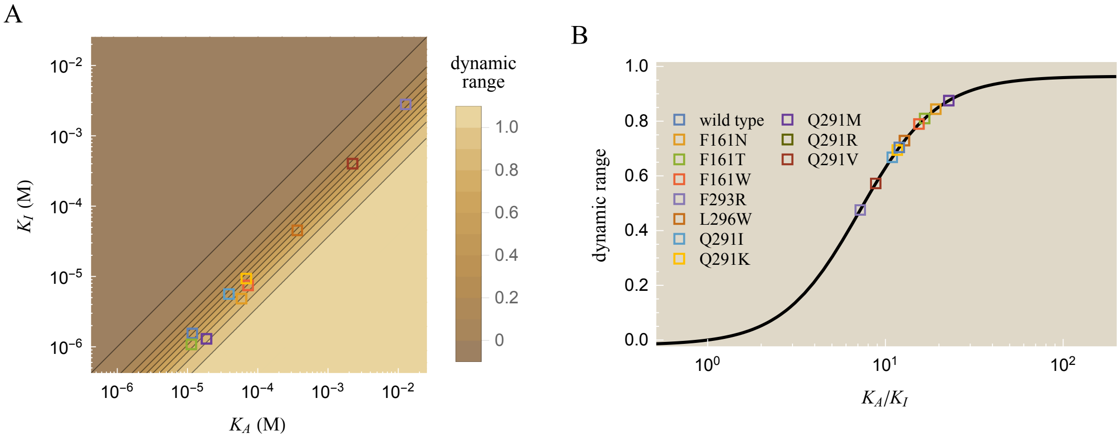


Figure 3.6: Theoretical values of the dynamic range for inducer mutants. (A) The dynamic range, Eq. 3.15, as a function of K_A and K_I, using the wild type values for ϵ and K_{DNA}. (B) Dynamic range as a function of K_A/K_I overlaid with the dynamic range values of the ten titration curves from Fig. 3.5A. Note that K_A/K_I is much more tightly constrained about the wild type value than the individual K_A and K_I values. The dynamic range values shown are from the theoretical best fit curves in Fig. 3.5 and are not measured directly from data.

equal to the leakiness of the wild type Lac repressor, which is indeed confirmed by the titration curves in Fig. 3.5A. Turning to the dynamic range, Fig. 3.6A shows the K_A and K_I values for the inducer mutants which collectively span 2-3 decades. However, Fig. 3.6B demonstrates that the K_A/K_I values of all inducer mutants is very tightly constrained (less than 1 decade) around the wild type value. In other words,

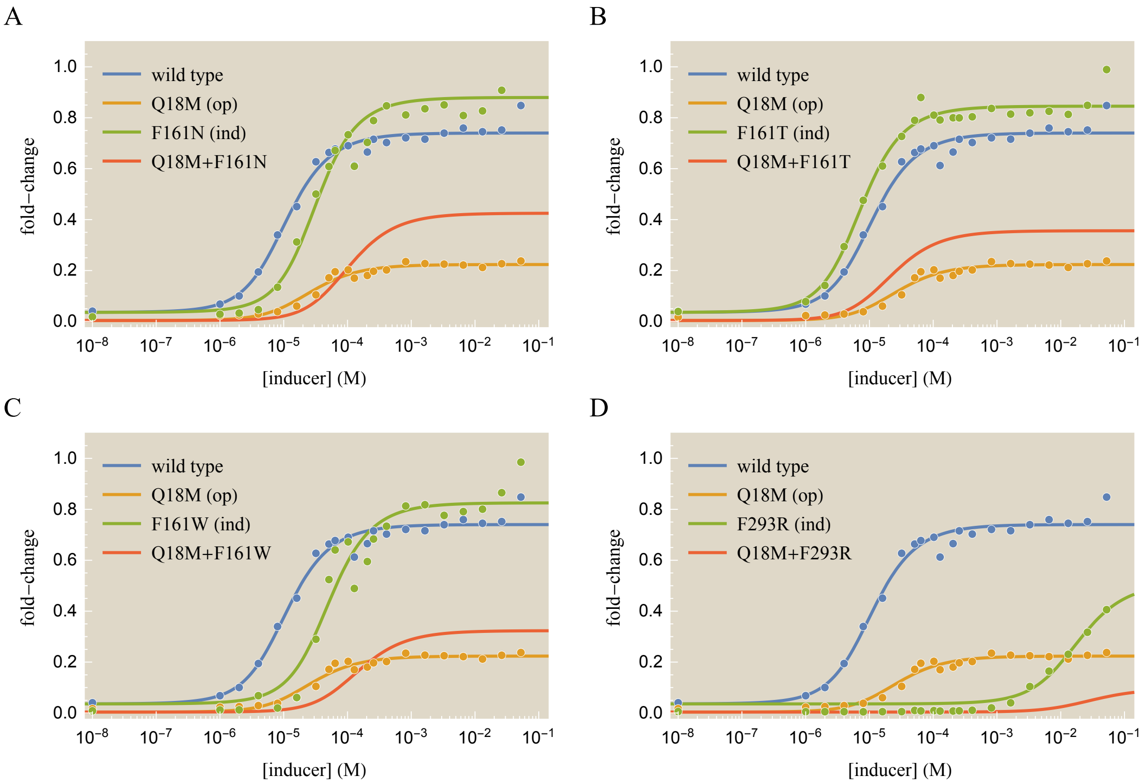


Figure 3.7: Predicted behavior of double mutants. The anticipated fold-change of a double mutant, with a point mutation in the operator and inducer binding domains, using the K_{DNA} parameter from the operator mutant and the K_A and K_I parameters from the inducer mutant. Predictions are shown for the fold-change of the double mutation of Q18M in the DNA binding domain and (A) F161N, (B) F161T, (C) F161W, and (D) F293R in the inducer binding domain.

any mutation in the inducer binding domain that increases the affinity of the inducer to active-state repressor (e.g. by altering avidity or electrostatic interactions) also seems to increase the affinity of inducer to inactive-state repressor by approximately the same amount. Note that in our discussion of leakiness and dynamic range above, the ratio K_A/K_I emerged as the critical parameter in the MWC model, rather than the individual values of K_A and K_I.

3.4 Discussion

We have shown that within the Monod-Wyman-Changeux statistical mechanical model of allostery, the physiological role of mutations within the Lac repressor have a deep, intuitive connection with the thermodynamic parameters characterizing that protein. Said another way, the Lac repressor is amenable to a treatment of mutations as local perturbations to protein structure. Compared to the wild type Lac repressor, mutations inside the operator binding domain will only noticeably affect the DNA binding affinity (K_{DNA}) while mutations within the inducer binding domain only alter the inducer binding affinities (K_A and K_I) in the fold-change Eq. 3.9.

While general relationships are well-known between quantities such as leakiness and dynamic range for allosteric proteins [30], the framework developed here has made it possible to make mechanistically specific predictions about the consequences of different classes of mutations. For example, we predict that mutating the inducer binding domain should not change the leakiness of repressor mutants, which has been confirmed in all ten inducer mutants tested. In addition, we have been able to more tightly constrain the relationship between leakiness and dynamic range for our different classes of mutants. For instance, significant effort has been invested to find a Lac repressor mutant with a larger dynamic range but comparable leakiness to the wild type [5]. Our analysis has shown that operator mutants are unlikely to yield these characteristics and that inducer mutants can at best yield only a slightly improved dynamic range.

We have also shown that a subset of the possible mutations of Lac repressor seem to act locally, only changing the MWC parameters corresponding to the physical region of the mutation. While we acknowledge that not all mutations may be amenable to a similar treatment, we are intrigued by the success of the simple model proposed here and are curious to explore the limits of this mindset as the mutational landscape becomes more complicated. For example, it would be interesting to test whether the effects of double mutants, with a point mutation in the DNA binding region and another point mutation in the inducer binding region, will dictate the K_{DNA} parameter from the former mutation while K_A and K_I will be fixed by the latter mutation [7, 29]. Fig. 3.7 shows four such double mutants, although any combination of an operator and inducer mutant analyzed in this paper is possible. These results assume that the two mutations are completely independent and do not interact with each other, a trait which appears supported by the clear classes of operator and inducer mutants presented in the text, but which ultimately must be verified experimentally. In addition, it remains an open question whether one can access the ϵ parameter, the energy difference between the active and inactive allosteric conformations, in Lac repressor independently of the DNA and inducer binding affinities and create an ϵ mutant. Answering these questions will surely lead to new insights into the ways in which evolution has shaped proteins by giving a sense of which parameters are evolutionarily accessible.

3.5 Methods

Detailed notes about how the induction profiles of the Lac repressor were measured can be found in [5]. Both the Lac repressor and Lac promoter were on low copy number plasmids. Provided that the number of repressors is significantly larger than the number of plasmids, the form of the fold-change equation remains unchanged from Eq. 3.9 [18].

References

[1] Hoban MD, Cost GJ, Mendel MC, Romero Z, Kaufman ML, Joglekar AV, et al. Correction of the Sickle Cell Disease Mutation in Human Hematopoietic Stem/progenitor Cells. Blood. 2015;125(17):2597–604. doi:10.1182/Blood-2014-12-615948.

[2] Rodrigues JV, Bershtein S, Li A, Lozovsky ER, Hartl DL, Shakhnovich EI. Biophysical Principles Predict Fitness Landscapes of Drug Resistance. Proc. Natl. Acad. Sci. USA. 2016;113(11):E1470–E1478. doi:10.1073/Proc.Natl.Acad.Sci.USA.1601441113.

[3] Sadowski MI, Jones DT. The Sequence-Structure Relationship and Protein Function Prediction. Curr. Opin. Struct. Biol. 2009;19(3):357–62. doi:10.1016/j.sbi.2009.03.008.

[4] Lunzer M, Miller SP, Felsheim R, Dean AM. The Biochemical Architecture of an Ancient Adaptive Landscape. Science (New York, N.Y.). 2005; 310(5747):499–501. doi:10.1126/science.1115649.

[5] Daber R, Sochor MA, Lewis M. Thermodynamic Analysis of Mutant Lac Repressors. J. Mol. Biol. 2011;409(1):76–87. doi:10.1016/j.jmb.2011.03.057.

[6] Endres RG, Wingreen NS. Precise Adaptation in Bacterial Chemotaxis Through Assistance Neighborhoods. Proc. Natl. Acad. Sci. USA. 2006; 103(35):13040–4. doi:10.1073/Proc.Natl.Acad.Sci.USA.0603101103.

[7] Swem LR, Swem DL, Wingreen NS, Bassler BL. Deducing Receptor Signaling Parameters from In Vivo Analysis: LuxN/AI-1 Quorum Sensing in Vibrio Harveyi. Cell. 2008;134(3):461–473. doi:10.1016/j.cell.2008.06.023.

[8] Markiewicz P, Kleina L, Cruz C, Ehret S, Miller J. Genetic Studies of the Lac Repressor. XIV. Analysis of 4000 Altered *Escherichia coli* Lac Repressors Reveals Essential and Non-Essential Residues, as Well as "Spacers" Which Do Not Require a Specific Sequence. J. Mol. Biol. 1994;240(5):421–433. doi:10.1006/jmbi.1994.1458.

[9] Li GW, Burkhardt D, Gross C, Weissman JS. Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. Cell. 2014;157(3):624–635. doi:10.1016/j.cell.2014.02.033.

[10] Einav T, Mazutis L, Phillips R. Statistical Mechanics of Allosteric Enzymes. J Phys. Chem. B. 2016;120(26):6021–6037. doi:10.1021/acs.jpcb.6b01911.

[11] Velyvis A, Yang YR, Schachman HK, Kay LE. A Solution NMR Study Showing That Active Site Ligands and Nucleotides Directly Perturb the Allosteric Equilibrium in Aspartate Transcarbamoylase. Proc. Natl. Acad. Sci. USA. 2007;104(21):8815–20. doi:10.1073/Proc.Natl.Acad.Sci.USA.0703347104.

[12] Marzen S, Garcia HG, Phillips R. Statistical Mechanics of Monod-Wyman-Changeux (MWC) Models. J. Mol. Biol. 2013;425(9):1433–1460. doi:10.1016/j.jmb.2013.03.013.

[13] Levantino M, Spilotros A, Cammarata M, Schirò G, Ardiccioni C, Vallone B, et al. The Monod-Wyman-Changeux Allosteric Model Accounts for the Quaternary Transition Dynamics in Wild Type and a Recombinant Mutant Human Hemoglobin. Proc. Natl. Acad. Sci. USA. 2012;109(37):14894–9. doi:10.1073/Proc.Natl.Acad.Sci.USA.1205809109.

[14] Mohapatra L, Goode BL, Jelenkovic P, Phillips R, Kondev J. Design Principles of Length Control of Cytoskeletal Structures. Annu. Rev. Biophys. 2015; 45:85–116.

[15] Harris SL, Fernsten P. Thermodynamics and Density of Binding of a Panel of Antibodies to High-Molecular-weight Capsular Polysaccharides. Clinical and vaccine immunology: CVI. 2009;16(1):37–42. doi:10.1128/CVI.00290-08.

[16] Garcia HG, Phillips R. Quantitative Dissection of the Simple Repression Input-Output Function. Proc. Natl. Acad. Sci. USA. 2011;108(29):12173–8. doi:10.1073/Proc.Natl.Acad.Sci.USA.1015616108.

[17] Brewster RC, Weinert FM, Garcia HG, Song D, Rydenfelt M, Phillips R. The Transcription Factor Titration Effect Dictates Level of Gene Expression. Cell. 2014;156(6):1312–23. doi:10.1016/j.cell.2014.02.022.

[18] Weinert FM, Brewster RC, Rydenfelt M, Phillips R, Kegel WK. Scaling of Gene Expression with Transcription-Factor Fugacity. Phys. Rev. Lett. 2014; 113(25):1–5. doi:10.1103/PhysRevLett.113.258101.

[19] Daber R, Sharp K, Lewis M. One Is Not Enough. J. Mol. Biol. 2009; 392(5):1133–1144.

[20] Ackers GK, Johnson AD, Shea MA. Quantitative Model for Gene Regulation by Lambda Phage Repressor. Proc. Natl. Acad. Sci. USA. 1982;79(4):1129–33.

[21] Buchler NE, Gerland U, Hwa T. On Schemes of Combinatorial Transcription Logic. Proc. Natl. Acad. Sci. USA. 2003;100(9):5136–41. doi:10.1073/Proc.Natl.Acad.Sci.USA.0930314100.

[22] Phillips R. Napoleon Is in Equilibrium. Annu. Rev. Condens. Matter Phys. 2015;6(1):85–111. doi:10.1146/annurev-conmatphys-031214-014558.

[23] Brewster RC, Jones DL, Phillips R. Tuning Promoter Strength Through RNA Polymerase Binding Site Design in *Escherichia coli*. PLoS Comput. Biol. 2012;8(12). doi:10.1371/journal.pcbi.1002811.

[24] Yonetani T, Park SI, Tsuneshige A, Imai K, Kanaori K. Global Allostery Model of Hemoglobin. Modulation of O(2) Affinity, Cooperativity, and Bohr Effect by Heterotropic Allosteric Effectors. J. Biol. Chem. 2002;277(37):34508–20. doi:10.1074/jbc.M203135200.

[25] Grosman C, Auerbach A. The Dissociation of Acetylcholine from Open Nicotinic Receptor Channels. Proc. Natl. Acad. Sci. USA. 2001;98(24):14102–14107. doi:10.1073/Proc.Natl.Acad.Sci.USA.251402498.

[26] Levy ED. A Simple Definition of Structural Regions in Proteins and Its Use in Analyzing Interface Evolution. J. Mol. Biol. 2010;403(4):660–70. doi: 10.1016/j.jmb.2010.09.028.

[27] Li M, Petukh M, Alexov E, Panchenko AR. Predicting the Impact of Missense Mutations on Protein-Protein Binding Affinity. J Chem. Theory. Comput. 2014;10(4):1770–1780. doi:10.1021/ct401022c.

[28] Mishra P, Flynn JM, Starr TN, Bolon DNA. Systematic Mutant Analyses Elucidate General and Client-Specific Aspects of Hsp90 Function. Cell. Rep. 2016;15(3):588–98. doi:10.1016/j.celrep.2016.03.046.

[29] Keymer JE, Endres RG, Skoge M, Meir Y, Wingreen NS. Chemosensing in *Escherichia coli*: Two regimes of two-state receptors. Proc. Natl. Acad. Sci. USA. 2006;103(6):1786–1791. doi:10.1073/Proc.Natl.Acad.Sci.USA.0507438103.

[30] Martins BMC, Swain PS. Trade-Offs and Constraints in Allosteric Sensing. PLoS Comput. Biol. 2011;7(11):1–13. doi:10.1371/journal.pcbi.1002261.

[31] Mirny La. Nucleosome-Mediated Cooperativity Between Transcription Factors. Proc. Natl. Acad. Sci. USA. 2010;doi:10.1073/Proc.Natl.Acad.Sci.USA.0913805107.

[32] Razo-Mejia M, Boedicker JQ, Jones D, DeLuna A, Kinney JB, Phillips R. Comparison of the Theoretical and Real-World Evolutionary Potential of a Genetic Circuit. Phys. Biol. 2014;11(2):026005. doi:10.1088/1478-3975/11/2/026005.

[33] Pakula AA, Young VB, Sauer RT. Bacteriophage Lambda Cro Mutations: Effects on Activity and Intracellular Degradation. Proc. Natl. Acad. Sci. USA. 1986;83(23):8829–8833. doi:10.1073/Proc.Natl.Acad.Sci.USA.83.23.8829.

[34] Tungtur S, Skinner H, Zhan H, Swint-Kruse L, Beckett D. In Vivo Tests of Thermodynamic Models of Transcription Repressor Function. Biophys. Chem. 2011;159(1):142–51. doi:10.1016/j.bpc.2011.06.005.

[35] Caprio M. LevelScheme: A Level Scheme Drawing and Scientific Figure Preparation System for Mathematica. Comput. Phys. Commun. 2005;171(2):107–118. doi:10.1016/j.cpc.2005.04.010.

Chapter S3

SUPPLEMENTARY INFORMATION FOR MUTATIONS FROM A THERMODYNAMIC PERSPECTIVE

S3.1 Linking Thermodynamics and Statistical Mechanics

In this section, we discuss how to pass naturally between the language of thermodynamics, which utilizes rate constants and dissociation constants, and the language of statistical mechanics, which focuses upon energies of different states. We begin by rewriting Eq. 3.6, the fold-change of the Lac operator in the language of statistical mechanics,

$$\text{fold-change} = \frac{1}{1 + \frac{R_A}{N_{\text{NS}}} e^{-\beta \Delta \epsilon_{RD,A}}}. \tag{S3.1}$$

Using Fig. 3.1, the probability of active Lac repressor binding to the Lac operator versus the probability that the Lac operator is unoccupied equals $\frac{R_A}{N_{\text{NS}}} e^{-\beta \Delta \epsilon_{RD,A}}$. However, this ratio must also be given by the thermodynamic form $\frac{[R_A]}{K_{\text{DNA}}}$ where $[R_A]$ is the concentration of active Lac repressors and K_{DNA} is the dissociation constant between the repressor and operator. To link these two quantities, we consider the volume V of our system (the volume of an *E. coli*) so that $[R_A] = \frac{R_A}{V}$. Thus, we find that $K_{\text{DNA}} = \frac{N_{\text{NS}}}{V} e^{\beta \Delta \epsilon_{RD,A}}$ from which Eq. 3.7 follows.

S3.2 Data Fitting

In this section, we describe the fitting procedure used to match the experimental fold-change data at various concentrations of inducer (Fig. 3.3A and Fig. 3.5A) with the theoretically predicted curves of the form

$$\text{fold-change} = \left(1 + \frac{\left(1 + \frac{c}{K_A}\right)^2}{\left(1 + \frac{c}{K_A}\right)^2 + e^{-\beta\epsilon} \left(1 + \frac{c}{K_I}\right)^2} \frac{[R]}{K_{\text{DNA}}} \right)^{-1}. \tag{S3.2}$$

All fitting was done using nonlinear regression (NonlinearModelFit in *Mathematica*). Initial guesses were randomly chosen from the realistic parameter space $K_A, K_I \in [10^{-9}\,\text{M}, 10^{-3}\,\text{M}]$, $[R]/K_{\text{DNA}} \in [10^{-3}, 10^{3}]$, and $\beta\epsilon \in [-5, 5]$ until a sufficiently good fit ($R^2 > 0.95$) was found.

It must be noted that, as with nearly all models, there are serious ambiguities in the best fit values since multiple values may yield nearly identical curves. In point of

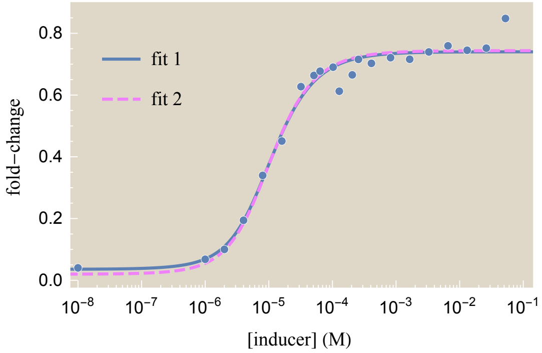


Figure S3.1: Multiple sets of thermodynamic parameters can yield nearly identical functional forms. The blue data points represent the wild type values of fold-change from Fig. 3.3A together with the best-fit line shown as *fit 1*. Fold-change for this fit was given by Eq. S3.2 with the parameters $K_A = 1.5 \times 10^{-5}$ M, $K_I = 1.2 \times 10^{-6}$ M, $e^{-\beta\epsilon} = 1.1$, and $[R]/K_{\text{DNA}} = 60$. Overlaid on top is the dashed curve *fit 2* with identical K_A and K_I parameters but with the unrealistic parameters $e^{-\beta\epsilon} = 2.2 \times 10^4$, and $[R]/K_{\text{DNA}} = 1.1 \times 10^6$. Note that the ratio $e^{\beta\epsilon}[R]/K_{\text{DNA}} = 0.02$ is the same in both cases, but the individual parameters can vary enormously.

fact, consider the wild type best fit parameters (shown in Table S3.1). Because K_I is sufficiently smaller than K_A and $e^{-\beta\epsilon} \approx 1$, the term $\left(1 + \frac{c}{K_A}\right)^2$ in the denominator of Eq. S3.2 is much less than $e^{-\beta\epsilon}\left(1 + \frac{c}{K_I}\right)^2$ for nearly the entire range of inducer concentrations measured in Fig. 3.3. Thus the fold-change is very well approximated by the form

$$\text{fold-change} \approx \left(1 + \frac{\left(1 + \frac{c}{K_A}\right)^2}{\left(1 + \frac{c}{K_I}\right)^2} e^{\beta\epsilon} \frac{[R]}{K_{\text{DNA}}}\right)^{-1}. \tag{S3.3}$$

Therefore, it is really the ratio $e^{\beta\epsilon}\frac{[R]}{K_{\text{DNA}}}$ rather than the individual parameters $e^{-\beta\epsilon}$ and $\frac{[R]}{K_{\text{DNA}}}$ that are tightly constrained by the fitting. Fig. S3.1 demonstrates this phenomenon by plotting the best fit curve for the wild type repressor overlaid on another curve with unrealistically large parameters.

This attribute of models, sometimes dubbed "sloppiness," is well known [1]. With this in mind, our results below demonstrate that our framework is *sufficient* to describe the induction profile of Lac repressor, but that the individual parameter values (i.e. K_A, K_I, K_{DNA}, and $e^{-\beta\epsilon}$) are *not tightly determined* by these fits. Finally, we point out that while fitting the wild type Lac repressor has a certain amount of sloppiness, consequently fitting the operator mutants (with a 1 parameter fit) and inducer mutants (with a 2 parameter fit) has much less sloppiness.

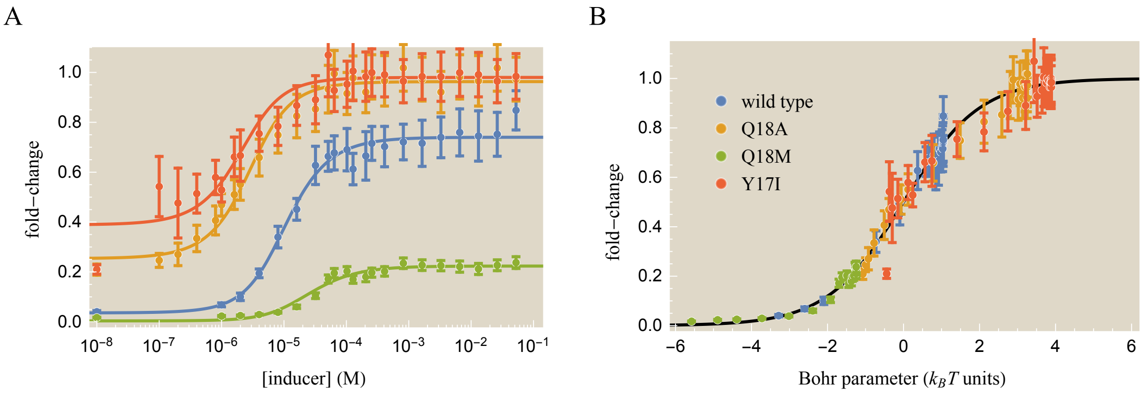


Figure S3.2: Fold-change of Lac repressor operator mutants. The curves show fold-change as a function of (A) effector (IPTG) concentration and (B) the Bohr parameter given in Eq. 3.11. Error bars represent the standard deviation from triplicate measurements.

S3.2.1 Operator Mutants

Consider the operator mutant titration curves, Fig. 3.3, from the main text. To obtain the theoretical curves, we first fit the wild type data to the fold change Eq. S3.2 by letting all four parameters K_A, K_I, ϵ, and $[R]/K_{\text{DNA}}$ vary. The resulting best fit values are shown in the top row of Table S3.1. Then, the three parameters K_A, K_I, and ϵ were held fixed while only $[R]/K_{\text{DNA}}$ was allowed to vary for the operator mutant data. The resulting $[R]/K_{\text{DNA}}$ values are shown in Table S3.1.

Fig. S3.2 shows the operator mutant data together with the standard deviation obtained through triplicate measurements within the same experiment. However, these error bars *were not* used when fitting the data, as we strongly suspect that there was systematic error unaccounted for, especially at low IPTG concentrations where it is difficult to distinguish signal from background. In particular, the standard deviation of the data points at 10^{-8} M inducer, which are a factor of 5-10 times smaller than the standard deviation at larger inducer concentrations, would disproportionately skew any fitting that assumes these standard deviations accurately represent the true measurement error. Without a systematic measurement of error, we opted to weigh each data point evenly.

S3.2.2 Inducer Mutants

We now turn to the inducer mutant titration curves, Fig. 3.5, from the main text. We used the same wild type parameters as in the case of the operator mutants, as shown in the top row of Table S3.2. We then fixed the value of ϵ and $[R]/K_{\text{DNA}}$ for all inducer mutants using the wild type values and fit the two inducer binding

Mutant	K_A (M)	K_I (M)	$e^{-\beta\epsilon}$	$[R]/K_{\text{DNA}}$
wild type	1.5×10^{-5}	1.2×10^{-6}	1.1	60
Q18A	—	—	—	6.5
Q18M	—	—	—	570
Y17I	—	—	—	3.5

Table S3.1: Best fit parameters for the wild type Lac repressor and three mutants with a single amino acid mutation in their operator binding site. The three mutants only fit their K_{DNA} parameter. Empty spaces indicate that the wild type parameter value was used.

affinities K_A and K_I for the inducer mutant data, with the resulting values shown in Table S3.2.

Fig. S3.3 shows the operator mutant data together with the standard deviation obtained through triplicate measurements. As in the case of the operator mutants discussed above, these error bars were not used when fitting the data, but instead all data points were weighted equally.

All of the best fit values for these inducer mutants seem reasonable except for Q291R. This particular mutation either made Lac repressor incapable of binding to inducer or raised its K_A and K_I so much that IPTG concentrations larger than 0.1 M are required to induce it. In either case, fitting data points with zero fold-change will inevitably lead to near-infinite K_A and K_I values. Hence, we ignored this mutant in our analysis within the main text.

In general, while mutants such as Q291R with flat-line fold-change technically qualify to be inducer mutants (i.e. their data can be well fit by the theory), no real information can be gained by studying them within our framework. Hence, we recommend to first try inducing such repressors at maximum IPTG concentrations (around 1 M) and if no fold-change can be detected to ignore these mutants.

References

[1] Transtrum MK, Machta BB, Brown KS, Daniels BC, Myers CR, Sethna JP. Perspective: Sloppiness and Emergent Theories in Physics, Biology, and Beyond. J. Chem. Phys. 2015;143(1):010901. doi:10.1063/1.4923066.

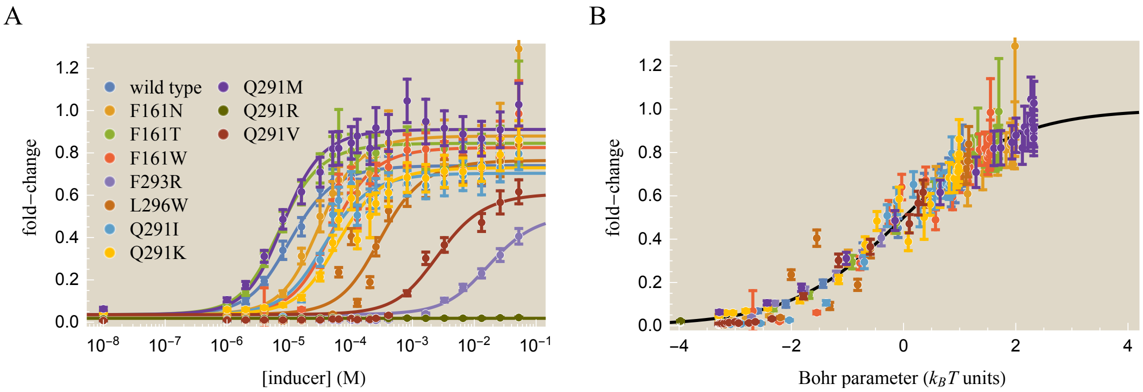


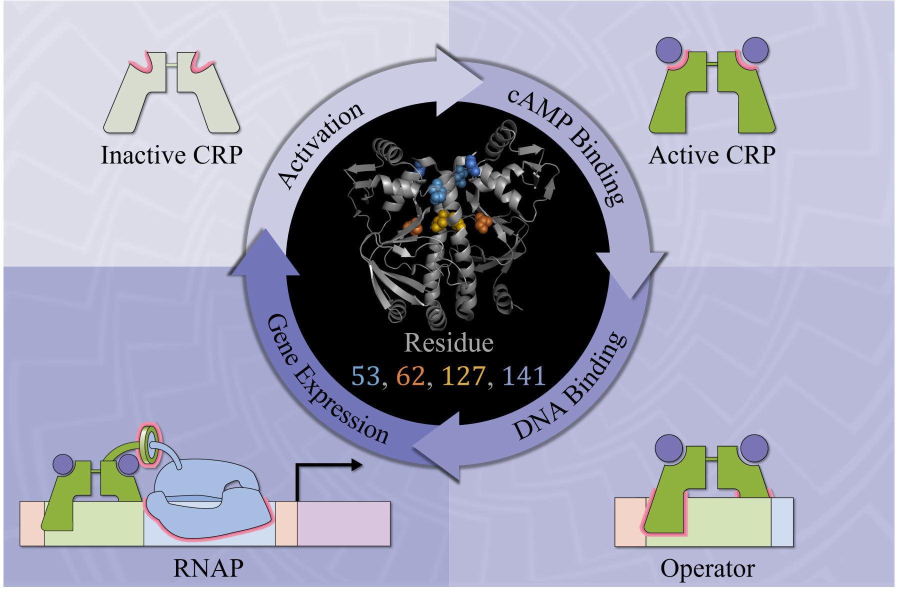
Figure S3.3: Fold-change of Lac repressor inducer mutants. The curves show fold-change as a function of (A) effector (IPTG) concentration and (B) the Bohr parameter given in Eq. 3.11. Error bars represent the standard deviation from triplicate measurements.

Mutant	K_A (M)	K_I (M)	$e^{-\beta\epsilon}$	$[R]/K_{\text{DNA}}$
wild type	1.5×10^{-5}	1.2×10^{-6}	1.1	60
F161N	7.4×10^{-5}	3.8×10^{-6}	—	—
F161T	1.4×10^{-5}	8.6×10^{-7}	—	—
F161W	9.2×10^{-5}	5.9×10^{-6}	—	—
F293R	1.6×10^{-2}	2.2×10^{-3}	—	—
L296W	4.6×10^{-4}	3.6×10^{-5}	—	—
Q291I	4.9×10^{-5}	4.5×10^{-6}	—	—
Q291K	8.7×10^{-5}	7.4×10^{-6}	—	—
Q291M	2.3×10^{-5}	1.0×10^{-6}	—	—
Q291R*	6.8×10^{-25}	2.5×10^{-24}	—	—
Q291V	2.8×10^{-3}	3.2×10^{-4}	—	—

Table S3.2: Best fit parameters for the wild type Lac repressor and ten mutants with a single amino acid mutation in their inducer binding site. The ten mutants only fit their K_A and K_I parameters which represent the repressor's affinity to the inducer in the active and inactive states, respectively. Dashes indicate that the wild type parameter values were used. * Q291R had zero fold-change for all IPTG concentrations measured, to which a wide variety of parameters can be fit. Hence it was discarded from our analysis of inducer mutants.

Chapter 4

THEORETICAL ANALYSIS OF INDUCER AND OPERATOR BINDING FOR CYCLIC-AMP RECEPTOR PROTEIN MUTANTS



Einav T, Duque J, Phillips R. Theoretical Analysis of Inducer and Operator Binding for Cyclic-AMP Receptor Protein Mutants. PLOS ONE (Public Library of Science). 2018;13(9):e0204275. doi:10.1371/journal.pone.0204275

This was the first project that I had developed from the ground up. Rob and I had met Julia Duque the previous summer at MBL, and we enjoyed interacting with her so much that Rob invited her to visit Caltech for three months and conduct post-course research with us. As we were looking for projects, I read a beautiful paper by Rodrigo Maillard and immediately knew that it was perfect. Together with my Lac repressor paper (Chapter 3) and ion channel paper (Chapter 5), these projects make a compelling case that the MWC model of allostery can refine how we think about mutations.

4.1 Abstract

Allosteric transcription factors undergo binding events at inducer binding sites as well as at distinct DNA binding domains, and it is difficult to disentangle the structural and functional consequences of these two classes of interactions. We compare the ability of two statistical mechanical models – the Monod-Wyman-Changeux (MWC) and the Koshland-Némethy-Filmer (KNF) models of protein conformational change – to characterize the multi-step activation mechanism of the broadly acting cyclic-AMP receptor protein (CRP). We first consider the allosteric transition resulting from cyclic-AMP binding to CRP, then analyze how CRP binds to its operator, and finally investigate the ability of CRP to activate gene expression. We use these models to examine a beautiful recent experiment that created a single-chain version of the CRP homodimer, creating six mutants using all possible combinations of the wild type, D53H, and S62F subunits. We demonstrate that the MWC model can explain the behavior of all six mutants using a small, self-consistent set of parameters whose complexity scales with the number of subunits, providing a significant benefit over previous models. In comparison, the KNF model not only leads to a poorer characterization of the available data but also fails to generate parameter values in line with the available structural knowledge of CRP. In addition, we discuss how the conceptual framework developed here for CRP enables us to not merely analyze data retrospectively, but has the predictive power to determine how combinations of mutations will interact, how double mutants will behave, and how each construct would regulate gene expression.

4.2 Introduction

Transcriptional regulation lies at the heart of cellular decision making, and understanding how cells modify the myriad of players involved in this process remains challenging. The cyclic-AMP receptor protein (CRP; also known as the catabolite receptor protein, CAP) is an allosteric transcription factor that regulates over 150 genes in *Escherichia coli* [1–4]. Upon binding to cyclic-AMP (cAMP), the homodimeric CRP undergoes a conformational change whereby two alpha helices reorient to open a DNA binding domain [5], allowing CRP to bind to DNA and affect transcription [6–8]. While much is known about the molecular details of CRP and how different mutations modify its functionality [9, 10], each new CRP mutant is routinely analyzed in isolation using phenomenological models. We argue that given the hard-won structural insights into the conformational changes of proteins like CRP, it is important to test how well mechanistically motivated models of such proteins can characterize the wealth of available data.

The picture that has emerged from various domains of biology is that allostery involves the interplay of a spectrum of dynamically linked states [11–17]. In some systems, it is straightforward to partition these states into the physiologically relevant categories; for example, CRP naturally divides into the cAMP unbound, singly bound, and doubly bound states as well as the DNA bound and unbound states. Nuclear magnetic resonance (NMR) and isothermal titration calorimetry (ITC) have begun to tease out the precise thermodynamics of the underlying interactions between these states [18, 19]. These methods have demonstrated that allosteric regulation in CRP includes both large structural changes as well as entropic modifications that make the protein more rigid [20, 21]. In this work, we ask whether we can capitalize upon this detailed knowledge of the system to construct a coarse-grained model of the multi-step activation cycle of CRP shown in Fig. 4.1A using a compact set of parameters. Specifically, we investigate variants of the Monod-Wyman-Changeux (MWC) model, which posits that both CRP subunits fluctuate concurrently between an active and inactive conformational state [22], and the Koshland-Némethy-Filmer (KNF) model, which proposes that each subunit must independently transition from an inactive to active state upon ligand binding [23], adapted for the CRP system. These two models have been investigated in a wide variety of allosteric systems, and evidence for both models as well as their shortcomings have been extensively analyzed [24–28]. Nevertheless, the simple thermodynamic view provided by the MWC and KNF models provides fertile ground to both verify how well we understand the critical factors governing CRP behavior as well as to explore hypotheses about mutational perturbations to the system.

Our paper is inspired by a recent *in vitro* study of CRP performed by Lanfranco *et al.* who engineered a single-chain CRP molecule whose two subunits are tethered together by an unstructured polypeptide linker [29]. This construct enabled them to mutate each subunit independently, providing a novel setting within which to analyze the combinatorial effects of mutations. Specifically, they took three distinct CRP subunits – the wild type (WT) subunit and the well characterized mutations D53H and S62F (denoted D and S, respectively) originally chosen to perturb the transcription factor's cAMP binding domain [30, 31] – and linked them together in every possible combination to create six CRP mutants as shown in Fig. 4.1B (black and pink boxes). Lanfranco *et al.* measured the cAMP-binding and DNA-binding capabilities of these mutants, separating these two key components of transcription

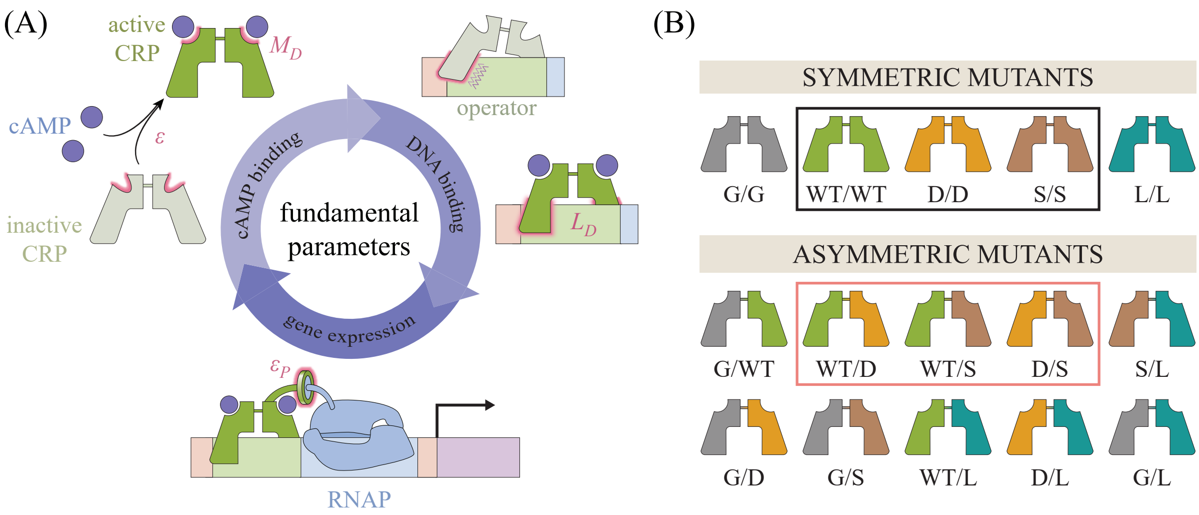


Figure 4.1: Key parameters governing CRP function. (A) Within the MWC and KNF models, each CRP subunit can assume either an active or an inactive conformation with a free energy difference ϵ between the two states. cAMP can bind to CRP (with a dissociation constant M_D^A in the active state and M_D^I in the inactive state) and promotes the active state ($M_D^A < M_D^I$ in the MWC model; $M_D^I \to \infty$ in the KNF model). Active CRP has a higher affinity for the operator (L_D^A) than the inactive state (L_D^I). When CRP is bound to DNA, it promotes RNA polymerase binding through an interaction energy ϵ_P, thereby enhancing gene expression. (B) Lanfranco *et al.* constructed a single-chain CRP molecule whose two subunits could be mutated independently. All possible dimers are shown using five mutant subunits: wild type (WT), D (D53H), S (S62F), G (G141Q), and L (L148R). Lanfranco *et al.* constructed the six mutants comprised of WT, D, and S (black and pink boxes) and analyzed each mutant independently.

factor activation. In this work, we present an analysis of these CRP mutants that demonstrates how their diverse phenotypes are related by their subunit compositions.

More specifically, the effects of mutations are often difficult to interpret, and indeed the results from Lanfranco *et al.* showed no clear pattern. The behavior of each mutant was analyzed *independently* by fitting its binding curve to a second order polynomial [29]. In this work, we propose an alternative framework that bolsters our understanding of the system in two significant ways: (1) we link the response functions of each CRP construct to its subunit composition, closing the gap between structure and function and (2) the number of parameters in our model scales linearly with the number of subunits whereas the number of parameters in the original analysis scaled with the number of CRP mutants (i.e. the square of the number of subunits). The advantage of this scaling behavior grows with the number of subunits. For example, this work focuses on the CRP mutants made by Lanfranco *et al.* using three subunits (black and pink boxes in Fig. 4.1B). If we include two additional well-characterized mutants – such as G141Q (G) and L148R (L) [32] – for a total of $N = 5$ subunits, our model would only require $2N = 10$ parameters to describe the $\frac{N(N+1)}{2} = 15$ mutants whereas a model analyzing each mutant independently would require 30 parameters (2 per mutant). With $N = 10$ subunits, we would require 20 parameters to understand 55 mutants while a model characterizing individual mutants would require 110 parameters.

In addition to analyzing the available *in vitro* data for these mutants, we consider how each construct would promote gene expression *in vivo*. Because CRP is a global activator, its activity within the cell is tightly regulated by enzymes that produce, degrade, and actively transport cAMP [7]. We discuss how these processes can either be modeled theoretically or excised experimentally and calibrate our resulting framework for transcription using gene expression measurements for wild type CRP. In this manner, we find a small, self-consistent set of parameters able to characterize each step of CRP activation shown in Fig. 4.1A.

The remainder of this paper is organized as follows. First, we characterize the interaction between cAMP and CRP for the six CRP mutants created by Lanfranco *et al.* and quantify the key parameters governing this behavior. Next, we analyze the interaction between CRP and DNA and discuss how the inferred parameters align with structural knowledge of the system. Finally, we consider how CRP enhances gene expression and extend the results from Lanfranco *et al.* to predict the activation profiles of the CRP mutants within a cellular environment.

4.3 Results

4.3.1 The Interaction between CRP and cAMP

In this section, we examine the cAMP-CRP binding process through the lenses of generalized MWC and KNF models which tie each mutant's behavior to its subunit composition. We find that both frameworks can characterize data from a suite of CRP mutants using a compact set of parameters, but only the interpretation of the MWC parameters is consistent with structural knowledge of CRP.

4.3.1.1 MWC Model

We first formulate a description of cAMP-CRP binding using a generalized form of the MWC model, where the two subunits of each CRP molecule fluctuate concurrently between an active and inactive state. The different conformations of

CRP binding to cAMP and their corresponding Boltzmann weights are shown in Fig. 4.2A. We define the free energy difference between inactive CRP and active CRP as 2ϵ (or ϵ per subunit). ϵ will be large and negative since the activator is preferentially inactive in the absence of ligand, which will allow us to simplify the description of the system. $\beta = \frac{1}{k_B T}$ where k_B is Boltzmann's constant and T represents temperature. The two cAMP binding events are known to be cooperative [26, 33, 34], where the magnitude and the sign of this cooperativity (whether it is favorable or unfavorable) strongly depends upon the conditions of the buffer, mutational perturbations to the system, and whether the full or partial CRP protein is considered [29, 35, 36]. To that end, we introduce two types of cooperativity. First, the classic MWC model is inherently cooperative, as the binding of each ligand alters the probable conformation and hence binding affinity of the other binding site; however, this mode of cooperativity can only be favorable [37]. Because CRP may also exhibit negative cooperativity, we introduce explicit interaction energies ϵ_{int}^A and ϵ_{int}^I between two ligands in the active and inactive CRP states, respectively. For simplicity, and because it will enable us to characterize the CRP collectively rather than requiring a unique parameter for each mutant, we assume that these explicit cooperative interactions are the same across all constructs (see Supporting Information Section S4.1 where we relax such assumptions).

For each cAMP-CRP dissociation constant M_X^Y, the subscript denotes which CRP subunit it describes – either the left (L) or right (R) subunit – while the superscript denotes the active (A) or inactive (I) state of CRP. Note that the left and right subunits may be different (see Fig. 4.1B). Given a cAMP concentration $[M]$, the fraction of occupied cAMP binding sites is given by

$$\text{fractional CRP occupancy}([M]) =$$
$$\frac{\frac{1}{2}\left(\frac{[M]}{M_L^A}+\frac{[M]}{M_R^A}\right)+e^{-\beta\epsilon_{\text{int}}^A}\frac{[M]}{M_L^A}\frac{[M]}{M_R^A}+e^{-2\beta\epsilon}\left(\frac{1}{2}\left(\frac{[M]}{M_L^I}+\frac{[M]}{M_R^I}\right)+e^{-\beta\epsilon_{\text{int}}^I}\frac{[M]}{M_L^I}\frac{[M]}{M_R^I}\right)}{1+\frac{[M]}{M_L^A}+\frac{[M]}{M_R^A}+e^{-\beta\epsilon_{\text{int}}^A}\frac{[M]}{M_L^A}\frac{[M]}{M_R^A}+e^{-2\beta\epsilon}\left(1+\frac{[M]}{M_L^I}+\frac{[M]}{M_R^I}+e^{-\beta\epsilon_{\text{int}}^I}\frac{[M]}{M_L^I}\frac{[M]}{M_R^I}\right)}. \quad (4.1)$$

Here, the fractional occupancy of CRP bound to zero, one, or two cAMP equals 0, 1/2, and 1, respectively. Experimentally, the fractional occupancy was measured *in vitro* in the absence of DNA using ANS fluorescence which utilizes a fluorescent probe triggered by the conformational change of cAMP binding to CRP [29].

Lanfranco *et al.* considered CRP subunits with either the D53H or S62F point mutations (hereafter denoted by D and S, respectively), with the D subunit binding more strongly to cAMP than the wild type while the S subunit binds more weakly

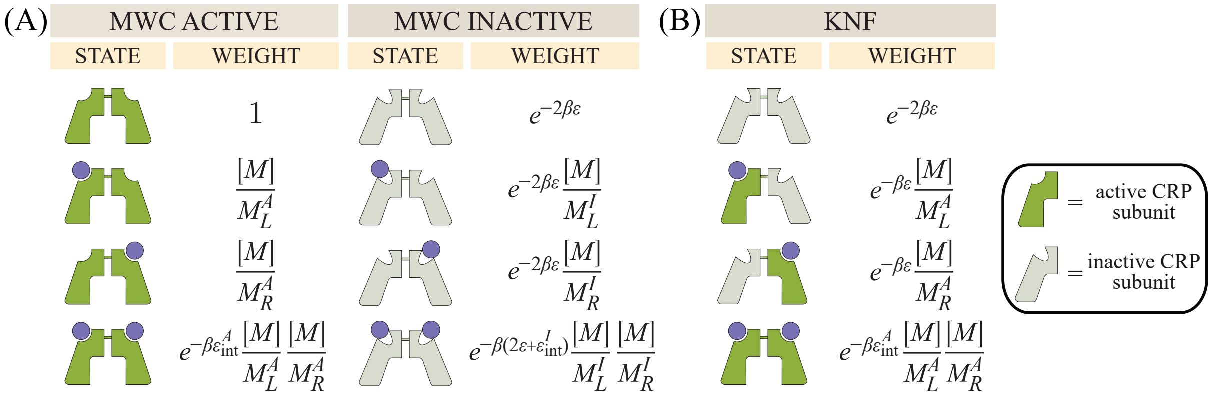

(A)
MWC ACTIVE
STATE
WEIGHT
1
$\frac{[M]}{M_L^A}$
$\frac{[M]}{M_R^A}$
$e^{-\beta\epsilon_{\text{int}}^A}\frac{[M]}{M_L^A}\frac{[M]}{M_R^A}$
MWC INACTIVE
STATE
WEIGHT
$e^{-2\beta\varepsilon}$
$e^{-2\beta\varepsilon}\frac{[M]}{M_L^I}$
$e^{-2\beta\varepsilon}\frac{[M]}{M_R^I}$
$e^{-\beta(2\varepsilon+\epsilon_{\text{int}}^I)}\frac{[M]}{M_L^I}\frac{[M]}{M_R^I}$
(B)
KNF
STATE
WEIGHT
$e^{-2\beta\varepsilon}$
$e^{-\beta\varepsilon}\frac{[M]}{M_L^A}$
$e^{-\beta\varepsilon}\frac{[M]}{M_R^A}$
$e^{-\beta\epsilon_{\text{int}}^A}\frac{[M]}{M_L^A}\frac{[M]}{M_R^A}$
= active CRP subunit
= inactive CRP subunit

Figure 4.2: Macroscopic states and Boltzmann weights for cAMP binding to CRP. (A) Within the MWC model, cAMP (purple circles) may bind to a CRP subunit in either the active (dark green) or inactive (light green) state. M_L^A and M_L^I represent the dissociation constants of the left subunit in the active and inactive states, respectively, while M_R^A and M_R^I represent the analogous dissociation constants for the right subunit. $[M]$ denotes the concentration of cAMP and ϵ represents the free energy difference between each subunit's inactive and active states. ϵ_{int}^A and ϵ_{int}^I represent a cooperative energy when two cAMP are bound to CRP in the active and inactive states, respectively. (B) The KNF model assumes that the two CRP subunits are inactive when unbound to cAMP and transition to the active state immediately upon binding to cAMP. The parameters have the same meaning as in the MWC model, but states where one subunit is active while the other is inactive are allowed.

as shown in Fig. 4.3A. While we could characterize the dose-response curves of each CRP mutant independently – for example, by using Eq. 4.1 to extract a set of parameters for each mutant – such an analysis lacks a direct connection between the subunit composition and the corresponding binding behavior. Instead, we assume that the cAMP binding affinity for each subunit should be uniquely dictated by that subunit's identity as either the WT, D, or S subunit. To that end, we represent the fractional occupancy of $\text{CRP}_{\text{D/WT}}$ using Eq. 4.1 with one D subunit ($M_L^A = M_{\text{D}}^A$, $M_L^I = M_{\text{D}}^I$) and one WT subunit ($M_R^A = M_{\text{WT}}^A$, $M_R^I = M_{\text{WT}}^I$). The equations for the remaining CRP mutants follow analogously, tying the behavior of each mutant to its subunit composition. For simplicity, we will assume that the D and S mutations do not alter the cAMP interaction energies ϵ_{int}^A and ϵ_{int}^I.

One difficulty in inferring parameter values from Eq. 4.1 is that degenerate sets of parameters may produce equivalent binding curves. For example, in S1 Text section A, we demonstrate how the same cAMP-CRP binding curves can arise from an arbitrarily large and negative free energy difference ($\epsilon \to -\infty$) provided that the dissociation constants scale appropriately. In that same supporting information

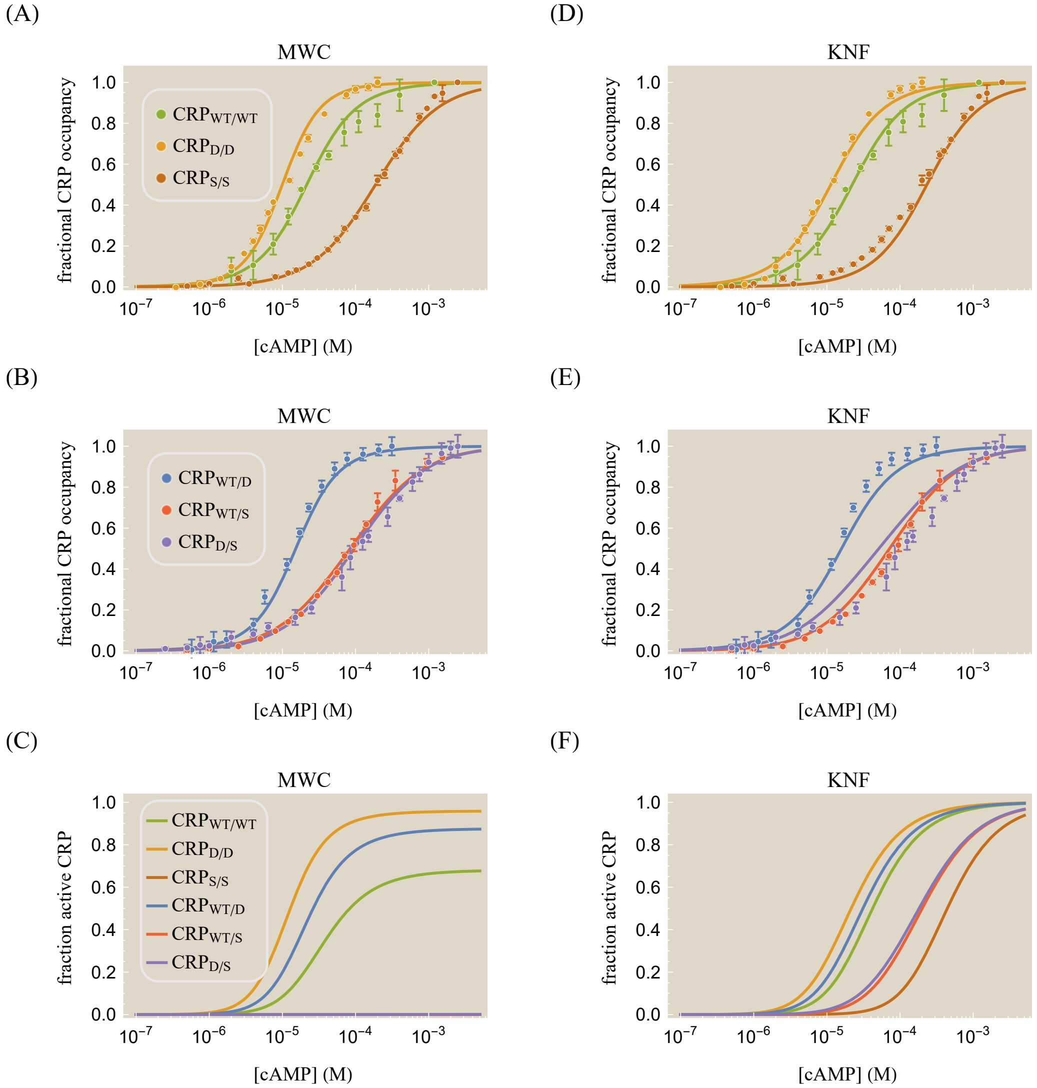


Figure 4.3: cAMP binding curves for different CRP mutants. In addition to the wild type CRP subunit (denoted WT), the mutation D53H (denoted D) and the mutation S62F (denoted S) can be applied to either subunit as indicated by the subscripts in the legend. Curves were characterized using the (A-C) MWC or (D-F) KNF model. The response of the asymmetric mutants (Panels B,E) lie between those of the symmetric mutants (Panels A,D). The fraction of CRP in the active state (Panels C,F) is markedly different between the two models; in the MWC model the mutants with an S subunit will be inactive even in the limit of saturating cAMP. Error bars represent the (corrected) sample standard deviation.

section, we demonstrate how this degeneracy can be excised so that Eq. 4.1 is well approximated by the following form,

MWC Parameter	Best-Fit Value	KNF Parameter	Best-Fit Value
$\tilde{M}_{\text{WT}}^{A}, M_{\text{WT}}^{I}$	$\{25 \pm 1, 40 \pm 3\} \times 10^{-6}$ M	$\bar{M}_{\text{WT}}^{A}$	$(30 \pm 2) \times 10^{-6}$ M
$\tilde{M}_{\text{D}}^{A}, M_{\text{D}}^{I}$	$\{10 \pm 1, 50 \pm 5\} \times 10^{-6}$ M	$\bar{M}_{\text{D}}^{A}$	$(20 \pm 1) \times 10^{-6}$ M
$\tilde{M}_{\text{S}}^{A}, M_{\text{S}}^{I}$	$\{\geq 1000, 200 \pm 10\} \times 10^{-6}$ M	$\bar{M}_{\text{S}}^{A}$	$(350 \pm 10) \times 10^{-6}$ M
$\epsilon_{\text{int}}^{I}$	$0.0 \pm 0.2\, k_B T$	$\epsilon_{\text{int}}^{A}$	$-0.8 \pm 0.2\, k_B T$

Table 4.1: Parameters for cAMP binding to CRP. The data in Fig. 4.3 can be characterized using a single set of dissociation constants for the WT, D, and S subunits whose values and standard errors are shown. To excise parameter degeneracy, the active-inactive free energy difference ϵ and the cAMP interaction energy in the active state $\epsilon_{\text{int}}^{A}$ are absorbed into the active state dissociation constants in the MWC model (Eqs. 4.2 and 4.3). Similarly, ϵ is absorbed into the KNF dissociation constants (Eqs. 4.6 and 4.7).

$$\text{fractional CRP occupancy}([M]) \approx \frac{\frac{[M]}{\tilde{M}_L^A}\frac{[M]}{\tilde{M}_R^A} + \frac{1}{2}\left(\frac{[M]}{M_L^I} + \frac{[M]}{M_R^I}\right) + e^{-\beta\epsilon_{\text{int}}^{I}}\frac{[M]}{M_L^I}\frac{[M]}{M_R^I}}{\frac{[M]}{\tilde{M}_L^A}\frac{[M]}{\tilde{M}_R^A} + \left(1 + \frac{[M]}{M_L^I} + \frac{[M]}{M_R^I} + e^{-\beta\epsilon_{\text{int}}^{I}}\frac{[M]}{M_L^I}\frac{[M]}{M_R^I}\right)}, \tag{4.2}$$

where we have neglected the unbound and singly-cAMP-bound active CRP states and defined the effective dissociation constants

$$\tilde{M}_L^A = e^{-\beta\epsilon} e^{\beta\epsilon_{\text{int}}^{A}/2} M_L^A \tag{4.3}$$

and

$$\tilde{M}_R^A = e^{-\beta\epsilon} e^{\beta\epsilon_{\text{int}}^{A}/2} M_R^A. \tag{4.4}$$

Using Eq. 4.2, we can extract the set of effective dissociation constants for the WT, D, and S subunits that determine the behavior of all six CRP mutants. The resulting parameters (shown in Table 4.1) give rise to the cAMP-CRP binding curves in Fig. 4.3A and B. Note that in removing the parameter degeneracy using Eqs. 4.3 and 4.4, we can no longer determine the individual values of ϵ, $\epsilon_{\text{int}}^{A}$, and the active state dissociation constants M_X^A, but rather only the parameter combinations $\tilde{M}_X^A$. On the other hand, the inactive state cooperativity energy $\epsilon_{\text{int}}^{I}$ can be unambiguously determined to be negligible. The effective dissociation constant of the S subunit in the MWC model can only be bounded from below as $\tilde{M}_{\text{S}}^{A} \geq 1000 \times 10^{-6}$ M. However, NMR measurements reported that in the limit of saturating cAMP, the S/S mutant will be inactive state 98% of the time (see Fig. 4.3C and Supporting Information Section S4.1) which corresponds to a value of $\tilde{M}_{\text{S}}^{A} \approx 1300 \times 10^{-6}$ M [20].

In Supporting Information Section S4.1, we demonstrate that the symmetric CRP mutants in Fig. 4.3A provide sufficient information to approximate the behavior of the asymmetric mutants in Fig. 4.3B. We further show that fitting each CRP data set individually to the MWC or KNF models without constraining the WT, D, and S subunits to a single unified set of dissociation constants results in only a marginal improvement over the constrained fitting. Finally, we analyze the slope of each cAMP binding response and explain why they are nearly identical for the six CRP mutants. In Supporting Information Section S4.2, we investigate the effects of the double mutation D+S on a single subunit by comparing its CRP occupancy data supposing that the change in free energy from both mutations is additive and independent. Within this epistasis-free model, we can similarly predict the behavior of other double mutants including $\text{CRP}_{\text{D/D+S}}$, $\text{CRP}_{\text{S/D+S}}$, and $\text{CRP}_{\text{D+S/D+S}}$.

Lastly, we reiterate that the MWC model presented here provides a coarse-grained model of the system. For example, experiments have revealed that the first cAMP binding does not alter the conformation of the second subunit, although it does drastically diminish its protein motions [34]. In the MWC model, these effects are captured both by the inherent cooperativity [37] as well as by the explicit interaction energies $\epsilon_{\text{int}}^{A}$ and $\epsilon_{\text{int}}^{I}$, since within this model the binding of one cAMP can induce the other CRP subunit to change (e.g. changing the unbound inactive state into the active singly-bound state). In light of these results, we next consider an alternative model of the system which explicitly assumes that each subunit only becomes active upon ligand binding.

4.3.1.2 KNF Model

We now turn to a KNF analysis of CRP, where the two subunits are individually inactive when not bound to cAMP and become active upon binding as shown in Fig. 4.2B. Some studies have claimed that cAMP binding to one CRP subunit does not affect the state of the other subunit, in support of the KNF model [34]. Other studies, meanwhile, have reported that a fraction of CRP molecules are active even in the absence of cAMP, thereby favoring an MWC interpretation [9, 38]. To determine whether either model can accurately represent the system, we explore some of the consequences of a KNF interpretation of CRP.

Using the statistical mechanical states of the system in Fig. 4.2B, the occupancy of

CRP is given by

$$\text{fractional CRP occupancy}([M]) = \frac{\frac{e^{-\beta\epsilon}}{2}\left(\frac{[M]}{M_L^A}+\frac{[M]}{M_R^A}\right)+e^{-\beta\epsilon_{\text{int}}^A}\frac{[M]}{M_L^A}\frac{[M]}{M_R^A}}{e^{-2\beta\epsilon}+e^{-\beta\epsilon}\left(\frac{[M]}{M_L^A}+\frac{[M]}{M_R^A}\right)+e^{-\beta\epsilon_{\text{int}}^A}\frac{[M]}{M_L^A}\frac{[M]}{M_R^A}}. \tag{4.5}$$

where the parameters have the same meaning as in the MWC model. Multiplying the numerator and denominator by $e^{2\beta\epsilon}$, we obtain the form

$$\text{fractional CRP occupancy}([M]) = \frac{\frac{1}{2}\left(\frac{[M]}{\bar{M}_L^A}+\frac{[M]}{\bar{M}_R^A}\right)+e^{-\beta\epsilon_{\text{int}}^A}\frac{[M]}{\bar{M}_L^A}\frac{[M]}{\bar{M}_R^A}}{1+\frac{[M]}{\bar{M}_L^A}+\frac{[M]}{\bar{M}_R^A}+e^{-\beta\epsilon_{\text{int}}^A}\frac{[M]}{\bar{M}_L^A}\frac{[M]}{\bar{M}_R^A}} \tag{4.6}$$

where, similar to the MWC model effective dissociation constants Eqs. 4.3 and 4.4, we have defined

$$\bar{M}_L^A = e^{-\beta\epsilon}M_L^A \tag{4.7}$$

and

$$\bar{M}_R^A = e^{-\beta\epsilon}M_R^A. \tag{4.8}$$

This simplification occurs because within the KNF model, a CRP monomer only switches from the inactive to active state upon cAMP binding. As a result, the free energy of cAMP binding to CRP and the free energy of the CRP undergoing its inactive-to-active state conformational always occur concurrently and may be combined into the effective dissociation constants $\bar{M}_L^A$ and $\bar{M}_R^A$.

As shown in Fig. 4.3D and Fig. 4.3E, the KNF model can approximately characterize the six mutant CRP binding curves, although the S/S and WT/D responses lie slightly below the data while the D/S curve deviates above the data. These discrepancies could potentially be alleviated by letting the interaction energy ϵ_{int}^A vary with each mutant, although doing so would significantly increase the number of parameters in the model (which would then scale with the number of mutants rather than the number of subunits). However, a greater failing of the KNF model is that it predicts that at saturating cAMP concentrations the protein will always be completely active, even though the S/S mutant is 98% inactive in this limit (Fig. 4.3F) [20]. These results suggest that a more complex variant of the KNF model should be used to quantitatively dissect the CRP system.

4.3.2 The Interaction between CRP and DNA

We now turn to the second binding interaction experienced by CRP, namely, that between CRP and DNA. Since the preceding analysis demonstrated that the KNF

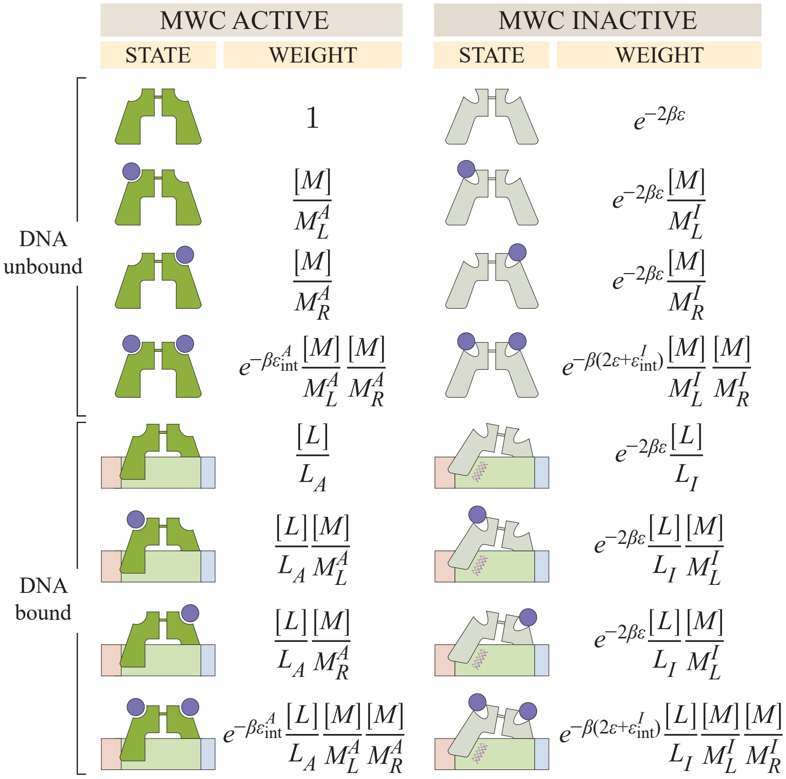

MWC ACTIVE
STATE
WEIGHT
MWC INACTIVE
STATE
WEIGHT
DNA unbound
DNA bound
1
$e^{-2\beta\varepsilon}$
$\frac{[M]}{M_L^A}$
$e^{-2\beta\varepsilon}\frac{[M]}{M_L^I}$
$\frac{[M]}{M_R^A}$
$e^{-2\beta\varepsilon}\frac{[M]}{M_R^I}$
$e^{-\beta\varepsilon_{\text{int}}^A}\frac{[M]}{M_L^A}\frac{[M]}{M_R^A}$
$e^{-\beta(2\varepsilon+\varepsilon_{\text{int}}^I)}\frac{[M]}{M_L^I}\frac{[M]}{M_R^I}$
$\frac{[L]}{L_A}$
$e^{-2\beta\varepsilon}\frac{[L]}{L_I}$
$\frac{[L]}{L_A}\frac{[M]}{M_L^A}$
$e^{-2\beta\varepsilon}\frac{[L]}{L_I}\frac{[M]}{M_L^I}$
$\frac{[L]}{L_A}\frac{[M]}{M_R^A}$
$e^{-2\beta\varepsilon}\frac{[L]}{L_I}\frac{[M]}{M_L^I}$
$e^{-\beta\varepsilon_{\text{int}}^A}\frac{[L]}{L_A}\frac{[M]}{M_L^A}\frac{[M]}{M_R^A}$
$e^{-\beta(2\varepsilon+\varepsilon_{\text{int}}^I)}\frac{[L]}{L_I}\frac{[M]}{M_L^I}\frac{[M]}{M_R^I}$

Figure 4.4: States and weights for CRP binding to DNA. The DNA unbound states from Fig. 4.2 are shown together with the DNA bound states. The Boltzmann weight of each DNA bound state is proportional to the concentration $[L]$ of CRP and inversely proportional to the CRP-DNA dissociation constants L_A or L_I for the active and inactive states, respectively.

model considered here cannot characterize the existing data, we proceed by only analyzing the MWC model.

Consider a concentration $[L]$ of CRP whose subunits either assume an active state (where they tightly bind to DNA with a dissociation constant L_A) or in an inactive state (characterized by weaker DNA binding with dissociation constant L_I satisfying $L_I > L_A$). The states and weights of this system within the generalized MWC model are shown in Fig. 4.4.

Lanfranco *et al.* fluorescently labeled a short, 32 bp DNA sequence that binds to CRP. Using a spectrometer, they measured the anisotropy of this fluorescence when different concentrations of CRP and cAMP were added *in vitro* [29]. The data are shown in Fig. 4.5A for $\text{CRP}_{\text{D/S}}$ for various concentrations of the receptor and effector. When CRP binds, it slows the random tumbling of the DNA so that over very short time scales the fluorescence is oriented along a particular axis, resulting in a larger anisotropy readout. Unbound DNA is defined as having anisotropy equal to 1 while DNA-bound CRP with 0, 1, or 2 bound cAMP have higher anisotropy values of $1 + r_0$, $1 + r_1$, and $1 + r_2$, respectively. Thus, the total anisotropy within the model is given by the weighted sum of each species [39], namely,

$$\text{anisotropy} = 1 + r_0 p_0 + r_1 p_1 + r_2 p_2. \tag{4.9}$$

Here, p_0, p_1, and p_2 represent the probabilities that DNA-bound CRP will be bound to 0, 1, and 2 cAMP molecules, respectively. Using the effective dissociation constants (Eqs. 4.3 and 4.4) and neglecting all terms proportional to the small quantity $e^{\beta\epsilon}$, we can write these probabilities as

$$p_0 = \frac{e^{2\beta\epsilon}\frac{[L]}{L_A} + \frac{[L]}{L_I}}{Z} \approx \frac{\frac{[L]}{L_I}}{Z}, \tag{4.10}$$

$$p_1 = \frac{e^{2\beta\epsilon}\frac{[L]}{L_A}\left(\frac{[M]}{M_L^A} + \frac{[M]}{M_R^A}\right) + \frac{[L]}{L_I}\left(\frac{[M]}{M_L^I} + \frac{[M]}{M_R^I}\right)}{Z} \approx \frac{\frac{[L]}{L_I}\left(\frac{[M]}{M_L^I} + \frac{[M]}{M_R^I}\right)}{Z}, \tag{4.11}$$

and

$$p_2 = \frac{e^{2\beta\epsilon}e^{-\beta\epsilon_{\text{int}}^A}\frac{[L]}{L_A}\frac{[M]}{M_L^A}\frac{[M]}{M_R^A} + e^{-\beta\epsilon_{\text{int}}^I}\frac{[L]}{L_I}\frac{[M]}{M_L^I}\frac{[M]}{M_R^I}}{Z} \approx \frac{\frac{[L]}{L_A}\frac{[M]}{\tilde{M}_L^A}\frac{[M]}{\tilde{M}_R^A} + e^{-\beta\epsilon_{\text{int}}^I}\frac{[L]}{L_I}\frac{[M]}{M_L^I}\frac{[M]}{M_R^I}}{Z} \tag{4.12}$$

with

$$\begin{aligned} Z &= e^{2\beta\epsilon}\left(1 + \frac{[L]}{L_A}\right)\left(1 + \frac{[M]}{M_L^A} + \frac{[M]}{M_R^A} + e^{-\beta\epsilon_{\text{int}}^A}\frac{[M]}{M_L^A}\frac{[M]}{M_R^A}\right) \\ &\quad + \left(1+\frac{[L]}{L_I}\right)\left(1 + \frac{[M]}{M_L^I} + \frac{[M]}{M_R^I} + e^{-\beta\epsilon_{\text{int}}^I}\frac{[M]}{M_L^I}\frac{[M]}{M_R^I}\right) \\ &\approx \left(1 + \frac{[L]}{L_A}\right)\frac{[M]}{\tilde{M}_L^A}\frac{[M]}{\tilde{M}_R^A} + \left(1 + \frac{[L]}{L_I}\right)\left(1 + \frac{[M]}{M_L^I} + \frac{[M]}{M_R^I} + e^{-\beta\epsilon_{\text{int}}^I}\frac{[M]}{M_L^I}\frac{[M]}{M_R^I}\right). \end{aligned} \tag{4.13}$$

In making these approximations, we have assumed the stricter conditions $e^{2\beta\epsilon}\frac{L_I}{L_A} \ll 1$ and $e^{2\beta\epsilon}\frac{L_I}{L_A}\frac{M_X^I}{\tilde{M}_X^A} \ll 1$ for the WT, D, and S subunits, all of which are valid assumptions for this system.

Fig. 4.5 shows the resulting best-fit curves for the anisotropy data, with the corresponding $\text{CRP}_{\text{D/S}}$ DNA dissociation constants given in Table 4.2. Since $1 + r_0 \approx 1$, cAMP-unbound CRP binds poorly to DNA, in accordance with the inactive state crystal structure whose DNA recognition helices are buried inside the protein [10].

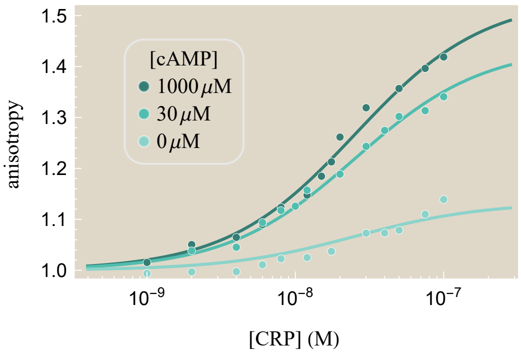


Figure 4.5: The interaction between CRP and DNA. Anisotropy of 32-bp fluorescein-labeled *lac* promoter binding to $\text{CRP}_{\text{D/S}}$ at different concentrations of cAMP. An anisotropy of 1 corresponds to unbound DNA while higher values imply that DNA is bound to CRP. In the presence of cAMP, more CRP subunits will be active, and hence there will be greater anisotropy for any given concentration of CRP. The sample standard deviation $\sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(y_{\text{theory}}^{(j)}-y_{\text{data}}^{(j)})^2}$ is 0.01, with the corresponding parameters given in Tables 4.1 and 4.2.

Additionally, the anisotropy $1 + r_1 = 1.7$ of the DNA-CRP-cAMP complex is larger than that of both the cAMP-unbound state and the doubly bound state DNA-CRP-$(\text{cAMP})_2$ with $1 + r_2 = 1.4$; this suggests that CRP-$(\text{cAMP})_2$ binds more weakly to DNA than CRP-cAMP. However, we note that these results depend upon the anisotropy values for the three CRP states (r_j in Table 4.2); Lanfranco *et al.* assumed that difference between the singly-cAMP-bound CRP state and the unbound CRP state should be the same as the difference between the doubly- and singly-cAMP-bound states and subsequently determined that the singly- and doubly-cAMP bound CRP states bind with roughly the same affinity to DNA. That said, previous studies have supported the claim that the singly-cAMP bound state binds tightest to DNA using multiple experimental methods including proteolytic digestion by subtilisin, chemical modification of Cys-178, and fluorescence measurements [40–42]. Given the ability of the MWC model to characterize the cAMP-binding and DNA-binding data of Lanfranco *et al.*, we next consider the final step in the CRP activation cycle, namely, how well CRP can enhance gene expression.

4.3.3 Implications of Mutations for *in vivo* Systems

Since CRP is a global transcriptional activator that governs many metabolic genes in *E. coli* [8], introducing mutations *in vivo* may vastly change cell behavior. Nevertheless, because the framework introduced above is very generic, it can be readily applied to other transcriptional activators that regulate a more limited number of

MWC Parameter	Best-Fit Value
r_0, r_1, r_2	$\{0.1, 0.8, 0.5\} \pm 0.1$
L_A, L_I	$\{\leq 30, 30 \pm 10\} \times 10^{-9}$ M

Table 4.2: Parameters for CRP binding to DNA. The anisotropy data for CRP$_{\text{D/S}}$ characterized using Eq. 4.9, as shown in Fig. 4.5. Each value is given as a mean ± standard error. The uncertainty in the $\tilde{M}_{\text{S}}^{A}$ parameter (shown in Table 4.1) leads to a corresponding uncertainty in the active CRP dissociation constant L_A.

genes. In that spirit, we briefly explore how the CRP mutants characterized in the Lanfranco *et al.* experiments would behave *in vivo* assuming that they only affect a single gene.

4.3.3.1 Simple Activation

Consider a cell with cAMP concentration $[M]$ and CRP concentration $[L]$ where the population of CRP is split between an active $[L_A]$ and an inactive $[L_I]$ conformation. Suppose the cell has a concentration $[P]$ of RNA polymerase (RNAP) which have a dissociation constant P_D with a promoter of interest. The thermodynamic states of the system are shown in Fig. 4.6, where the activator can bind to and recruit RNAP via an interaction energy ϵ_{P,L_A} between active CRP and RNAP with a weaker interaction ϵ_{P,L_I} between inactive CRP and RNAP. Without these two interaction energies ($\epsilon_{P,L_A} = \epsilon_{P,L_I} = 0$), the RNAP and CRP binding events would be independent and there would be no activation. Moreover, if the two activation energies were the same ($\epsilon_{P,L_A} = \epsilon_{P,L_I}$), the system could not exhibit the level of activation seen in the data (see Supporting Information Section S4.1).

We assume that gene expression is equal to the product of the RNAP transcription rate r_{trans} and the probability that RNAP is bound to the promoter of interest, namely,

$$\text{activity} = r_{\text{trans}} \frac{\frac{[P]}{P_D}\left(1 + \frac{[L_I]}{L_I} e^{-\beta\epsilon_{P,L_I}} + \frac{[L_A]}{L_A} e^{-\beta\epsilon_{P,L_A}}\right)}{\frac{[P]}{P_D}\left(1 + \frac{[L_I]}{L_I} e^{-\beta\epsilon_{P,L_I}} + \frac{[L_A]}{L_A} e^{-\beta\epsilon_{P,L_A}}\right) + 1 + \frac{[L_I]}{L_I} + \frac{[L_A]}{L_A}}. \tag{4.14}$$

Several additional factors influence gene expression *in vivo*. First, cAMP is synthesized endogenously by *cyaA* and degraded by *cpdA*, although both of these genes have been knocked out for the data set shown in Fig. 4.7A (see Methods and Ref. [7]). Furthermore, cAMP is actively transported out of a cell leading to a smaller concentration of intracellular cAMP. Following Kuhlman *et al.*, we will assume that the intracellular cAMP concentration is proportional to the extracellular concentra-

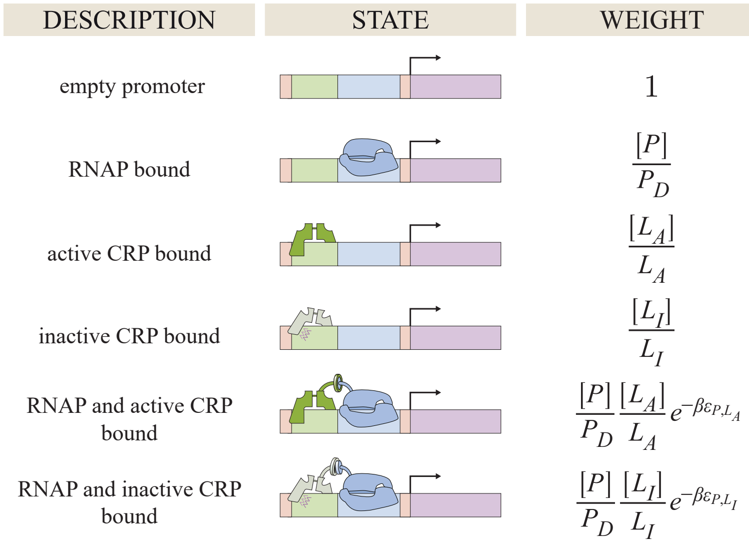

DESCRIPTION
STATE
WEIGHT
empty promoter
1
RNAP bound
$\frac{[P]}{P_D}$
active CRP bound
$\frac{[L_A]}{L_A}$
inactive CRP bound
$\frac{[L_I]}{L_I}$
RNAP and active CRP bound
$\frac{[P]}{P_D}\frac{[L_A]}{L_A}e^{-\beta\varepsilon_{P,L_A}}$
RNAP and inactive CRP bound
$\frac{[P]}{P_D}\frac{[L_I]}{L_I}e^{-\beta\varepsilon_{P,L_I}}$

Figure 4.6: States and weights for a simple activation motif. Binding of RNAP (blue) to a promoter is facilitated by the binding of the activator CRP. Simultaneous binding of RNAP and CRP is facilitated by an interaction energy ϵ_{P,L_A} for active CRP (dark green) and ϵ_{P,L_I} for inactive CRP (light green). cAMP (not drawn) influences the concentration of active and inactive CRP as shown in Fig. 4.4.

tion, namely, $\gamma[M]$ (with $0 < \gamma < 1$) [43, 44]. Hence, the concentration of active CRP satisfies $\frac{[L_A]}{[L]} = p^L_{\text{act}}(\gamma[M])$ where the fraction of active CRP p^L_{act} is given by Fig. 4.2A as

$$\begin{aligned} p^L_{\text{act}}([M]) &= \frac{1 + \frac{[M]}{M^A_L} + \frac{[M]}{M^A_R} + e^{-\beta\epsilon^A_{\text{int}}}\frac{[M]}{M^A_L}\frac{[M]}{M^A_R}}{1 + \frac{[M]}{M^A_L} + \frac{[M]}{M^A_R} + e^{-\beta\epsilon^A_{\text{int}}}\frac{[M]}{M^A_L}\frac{[M]}{M^A_R} + e^{-2\beta\epsilon}\left(1 + \frac{[M]}{M^I_L} + \frac{[M]}{M^I_R} + e^{-\beta\epsilon^I_{\text{int}}}\frac{[M]}{M^I_L}\frac{[M]}{M^I_R}\right)} \\ &\approx \frac{e^{-\beta\epsilon^A_{\text{int}}}\frac{[M]}{\tilde{M}^A_L}\frac{[M]}{\tilde{M}^A_R}}{e^{-\beta\epsilon^A_{\text{int}}}\frac{[M]}{\tilde{M}^A_L}\frac{[M]}{\tilde{M}^A_R} + \left(1 + \frac{[M]}{M^I_L} + \frac{[M]}{M^I_R} + e^{-\beta\epsilon^I_{\text{int}}}\frac{[M]}{M^I_L}\frac{[M]}{M^I_R}\right)}. \end{aligned} \tag{4.15}$$

In the last step, we have again introduced the effective dissociation constants from Eqs. 4.3 and 4.4 and dropped any terms proportional to $e^{\beta\epsilon}$. In addition to these considerations, proteins *in vivo* may experience crowding, additional forms of modification, and competition by other promoters. However, since our primary goal is to understand how CRP mutations will affect gene expression, we proceed with the simplest model and neglect the effects of crowding, modification, and competition.

Because of the uncertainty in the dissociation constant L_A between active CRP and DNA (see Table 4.2), it is impossible to unambiguously determine the transcription parameters from the single data set for wild type CRP shown in Fig. 4.7A. Instead, we select one possible set of parameters ($\frac{[P]}{P_D} = 130 \times 10^{-6}$, $r_{\text{trans}} = 5 \times 10^5 \frac{\text{MU}}{\text{hr}}$, $\gamma = 0.1$,

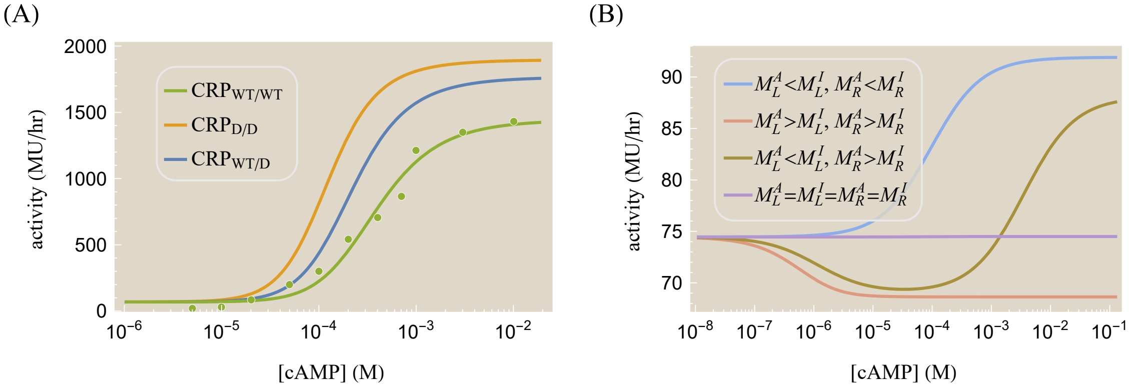


Figure 4.7: Predicted gene expression profiles for a simple activation architecture. (A) Gene expression for wild type CRP, where 1 Miller Unit (MU) represents a standardized amount of β-galactosidase activity. This data was used to determine the relevant parameters in Eq. 4.14 for the promoter in the presence of $[L] = 1.5\,\mu$M of CRP. The predicted behavior of the CRP mutants is shown using their corresponding cAMP dissociation constants. (B) The spectrum of possible gene expression profiles can be categorized based upon the cAMP-CRP binding affinity in each subunit. In all cases, we assumed $M_L^A = M_R^A = 3 \times 10^{-6}$ M and $e^{-\beta\epsilon_{\text{int}}^A} = 0$. The activation response (blue) was generated using $M_L^I = M_R^I = 6 \times 10^{-6}$ M. The repression response (orange) used $M_L^I = M_R^I = 10^{-7}$ M. The peaked response (gold) used $M_L^I = 10^{-7}$ M and $M_R^I = 300 \times 10^{-6}$ M. The flat response used $M_L^I = M_R^I = 3 \times 10^{-6}$ M. The remaining parameters in both plots were $\frac{[P]}{P_D} = 130 \times 10^{-6}$, $r_{\text{trans}} = 5 \times 10^5 \frac{\text{MU}}{\text{hr}}$, $\gamma = 0.1$, $\epsilon_{P,L_A} = -3\,k_BT$, $\epsilon_{P,L_I} = 0\,k_BT$, $\epsilon = -3\,k_BT$, and those shown in Tables 4.1 and 4.2.

$\epsilon_{P,L_A} = -3\,k_BT$, and $\epsilon_{P,L_I} = 0\,k_BT$) that is consistent with the wild type data. Next, we inserted the other cAMP-CRP dissociation constants (given in Table 4.1) into Eq. 4.14 to predict the gene expression profiles of the CRP mutants. Fig. 4.7A show the possible behavior of the CRP$_{\mathrm{D/D}}$ and CRP$_{\mathrm{WT/D}}$ mutants. As expected, replacing a WT subunit with a D subunit shifts the gene expression profile leftwards since the D subunit has a higher cAMP affinity (see Fig. 4.3A). Interestingly, the substitution of WT with D subunits comes with a concomitant increase in the maximum gene expression because at saturating cAMP concentrations, a larger fraction of CRP$_{\mathrm{D/D}}$ is active compared to CRP$_{\mathrm{WT/WT}}$ (96% and 68%, respectively) as seen by using Eq. 4.15 and the parameters in Table 4.1. Note that we cannot predict the behavior of any of the CRP mutants with S subunits due to the large uncertainty in $\tilde{M}_{\mathrm{S}}^A$.

Lastly, we probe the full spectrum of phenotypes that could arise from the activity function provided in Eq. 4.14 for any CRP mutant by considering all possible values of the cAMP-CRP dissociation constants M_L^A, M_L^I, M_R^A, and M_R^I in Eq. 4.15. In particular, we relax our assumption that cAMP binding promotes the CRP's active state, as a CRP mutation may exist whose inactive state binds more tightly to cAMP than its active state. Fig. 4.7B demonstrates that given such a mutation, a variety of novel phenotypes may arise. The standard sigmoidal activation response is achieved when cAMP binding promotes the active state in both CRP subunits ($M_L^A < M_L^I$, $M_R^A < M_R^I$). A repression phenotype is achieved in the opposite extreme when cAMP binding favors the inactive CRP state ($M_L^A > M_L^I$, $M_R^A > M_R^I$); we note that the ability to switch between a repressing and activating phenotype was achieved in the Lac repressor with as few as three mutations (see the R^c phenotypes in Ref. [46]). When one subunit is activated and the other is repressed by cAMP ($M_L^A < M_L^I$, $M_R^A > M_R^I$ or $M_L^A > M_L^I$, $M_R^A < M_R^I$), a peaked response can form. If the CRP subunits have the same affinity for cAMP in the active and inactive states ($M_L^A = M_L^I = M_R^A = M_R^I$), then CRP will behave identically for all concentrations of CRP, generating a flat-line response. It will be interesting to see whether these phenotypes can be achieved experimentally.

4.4 Discussion

The recent work of Lanfranco *et al.* provides a window into the different facets of gene regulation through activation [29]. Using insights from their *in vitro* experiments, we can break down the process of activation into its key steps, namely: (1) the binding of cAMP to make the activator CRP competent to bind DNA (Fig. 4.3); (2) the binding of CRP to DNA (Fig. 4.5); and (3) the recruitment of RNAP to promote gene expression (Fig. 4.7A). In this work, we generalized the classic MWC and KNF models to include a cAMP interaction energy as well as different DNA-binding affinities for the various cAMP-CRP bound states, allowing us to globally analyze the CRP binding data. Whereas biological research relishes the unique nuances in each system, the physical sciences suggest that common motifs – such as the prevalence of systems adopting an MWC-like description – lead to equally profound insights into the underlying principles governing systems.

By concurrently modeling the multi-step process of activation, we begin to unravel relationships and set strict limits for the binding energies and dissociation constants governing these systems. One hurdle to precisely fixing these values for CRP has been that many different sets of parameters produce the same degenerate responses. This parameter degeneracy is surprisingly common when modeling biological systems [47, 48], and we discuss how to account for it within the MWC and KNF models of CRP. A key feature of our analysis is that it permits us to identify the relevant parameter combinations for the system, quantify how well we can infer their values, and suggest which future experiments should be pursued to best constrain the behavior of the system.

Lanfranco *et al.* further explored how mutations in one or both subunits of CRP would influence its behavior. Specifically, they used three distinct subunits (WT, D, and S) to create the six CRP mutants shown in Fig. 4.1B (black and pink boxes). In this work, we showed that the effects of these mutations can be naturally understood through simple thermodynamic models so that each mutation need not be analyzed individually as if it had no relation to any other mutant. Instead, a compact set of parameters characterizing each subunit (see Table 4.1) could self-consistently characterize the cAMP-binding of all six mutants. The MWC model was shown to successfully describe the CRP activation data for all mutants whereas the KNF model led to a poorer characterization of the data and moreover incorrectly predicted the inhomogeneous population of CRP in the absence and presence of saturating cAMP. Even though an MWC description of the system was sufficient for the data set considered here, the full CRP system exhibits richer behavior that may require more generalized models that include the ensemble of different states seen by NMR [34, 49]. Nevertheless, it remains a useful exercise to understand how much of a system's behavior can be successfully captured by such simple models [50].

The models presented here suggest several avenues to further our understanding of CRP. First, we note that both the MWC and KNF models can serve as a springboard for more complex descriptions of CRP or other regulatory architectures [51]. However, a key advantage of simple frameworks lies in their ability to *predict* how different CRP subunits combine. For example, in Supporting Information Section S4.1 we demonstrate how the data from the three symmetric CRP mutants in Fig. 4.3A can be used to coarsely predict the asymmetric mutant responses in Fig. 4.3B. It would be interesting to see whether such predictions continue to hold as more mutant subunits are characterized, such as for the expanded suite of mutants shown in Fig. 4.1B. This framework has the potential to harness the combinatorial complexity of oligomeric proteins and presents a possible step towards systematically probing the space of mutations. In addition, any deviations in these predictions will provide further information on how allostery propagates in this system.

Second, several groups have proposed that multiple CRP mutations (K52N, T127, S128, G141K, G141Q, A144T, L148K, H159L from Refs. [9, 32, 52]) only affect the free energy difference ϵ between the CRP subunit's active and inactive states while leaving the cAMP-CRP dissociation constants unchanged. Our model predicts a narrow spectrum of phenotypes for such mutants, since the dependence of the ϵ parameter is solely confined to the effective dissociation constants (see Eqs. 4.3 and 4.4).

Finally, the framework considered here can be used to predict how the CRP mutants generated by Lanfranco *et al.* would behave *in vivo*. We calibrated the $\text{CRP}_{\text{WT/WT}}$ gene expression profile using data from Ref. [7] and suggested how the remaining CRP mutants may function within a simple activation regulatory architecture given the currently available data (see Fig. 4.7). It would be interesting to measure such constructs – or better yet, similar activators that regulate very few genes – within the cell and test the intersection of our *in vivo* and *in vitro* understanding both in the realm of the multi-step binding events of transcription factors as well as in quantifying the effects of mutations.

4.5 Methods

As described in Ref. [29], the fractional CRP occupancy data in Fig. 4.3 was measured *in vitro* using 8-anilino-1-naphthalenesulfonic acid (ANS) fluorescence which is triggered by the conformational change of cAMP binding to CRP. Experiments were conducted in 20 mM Tris, 50 mM NaCl, 1 mM EDTA, pH 7.8, and at 25°C. The CRP-DNA anisotropy data in Fig. 4.5 was measured *in vitro* by tagging the end of a 32 bp *lac* promoter with a fluorescein molecule and measuring its anisotropy with a spectrometer. When CRP is bound to DNA, anisotropy arises from two sources: the fast bending of the flanking DNA sequence and the slower rotation of the CRP-DNA complex. Sources of error include oligomerization of CRP, the bending of the flanking DNA, and nonspecific binding of CRP to the DNA.

The *in vivo* gene expression data was taken from Kuhlman *et al.* using the *lac* operon *E. coli* strain TK310 [7]. This strain had two genes knocked out: *cyaA* (a gene encoding adenylate cyclase, which endogenously synthesizes cAMP) and *cpdA* (encoding cAMP-phosphodiesterase, which degrades cAMP within the cell). Experiments were done at saturating concentrations of inducer ([IPTG] = 1 mM) so that Lac repressor negligibly binds to the operator [53]. In this limit, the only transcription factor affecting gene expression is the activator CRP. Gene expression was measured using β-galactosidase activity.

A Mathematica notebook that contains all of the data, reproduces the fitting (using both nonlinear regression and MCMC), and generates the plots in this paper can be found in the supplement of the online publication.

4.6 Acknowledgements

We thank Lacramioara Bintu for bringing the recent developments on CRP to our attention as well as Terry Hwa, Tom Kuhlman, and Michael Manhart for helpful discussions. All plots were made with data obtained from the authors or using WebPlotDigitizer [54]. This work was supported in the RP group by La Fondation Pierre-Gilles de Gennes, the Rosen Center at Caltech, and the National Institutes of Health through DP1 OD000217 (Director's Pioneer Award), R01 GM085286, and 1R35 GM118043-01 (MIRA). We are grateful to the Burroughs-Wellcome Fund for its support of the Physiology Course at the Marine Biological Laboratory, where part of the work on this work was done, and for a post-course research grant (JD).

References

[1] Martínez-Antonio A, Collado-Vides J. Identifying Global Regulators in Transcriptional Regulatory Networks in Bacteria. Curr. Opin. Microbiol. 2003; 6(5):482–489. doi:10.1016/j.mib.2003.09.002.

[2] You C, Okano H, Hui S, Zhang Z, Kim M, Gunderson CW, et al. Coordination of Bacterial Proteome with Metabolism by Cyclic AMP Signalling. Nature. 2013;500(7462):301–306. doi:10.1038/nature12446.

[3] Vilar JMG, Saiz L. Reliable Prediction of Complex Phenotypes from a Modular Design in Free Energy Space: An Extensive Exploration of the *lac* Operon. ACS Synth. Biol. 2013;2(10):576–586. doi:10.1021/sb400013w.

[4] Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muniz-Rascado L, Garcia-Sotelo JS, et al. RegulonDB Version 9.0: High-Level Integration of Gene Regulation, Coexpression, Motif Clustering and Beyond. Nucleic. Acids. Res. 2016;44(D1):D133–D143. doi:10.1093/nar/gkv1156.

[5] Popovych N, Tzeng SR, Tonelli M, Ebright RH, Kalodimos CG. Structural Basis for CAMP-Mediated Allosteric Control of the Catabolite Activator Protein. Proc. Natl. Acad. Sci. USA. 2009;106(17):6927–6932. doi: 10.1073/Proc.Natl.Acad.Sci.USA.0900595106.

[6] Hudson JM, Fried MG. Co-Operative Interactions Between the Catabolite Gene Activator Protein and the Lac Repressor at the Lactose Promoter. J. Mol. Biol. 1990;214(2):381–396. doi:10.1016/0022-2836(90)90188-R.

[7] Kuhlman T, Zhang Z, Saier MH, Hwa T. Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. Proc. Natl. Acad. Sci. USA. 2007; 104(14):6043–6048. doi:10.1073/Proc.Natl.Acad.Sci.USA.0606717104.

[8] Kochanowski K, Gerosa L, Brunner SF, Christodoulou D, Nikolaev YV, Sauer U. Few Regulatory Metabolites Coordinate Expression of Central Metabolic Genes in *Escherichia coli*. Mol. Syst. Biol. 2017;13(1):903. doi:10.15252/msb.20167402.

[9] Youn H, Koh J, Roberts GP. Two-State Allosteric Modeling Suggests Protein Equilibrium as an Integral Component for Cyclic AMP (cAMP) Specificity in the cAMP Receptor Protein of *Escherichia coli*. J. Bacteriol. 2008; 190(13):4532–4540. doi:10.1128/JB.00074-08.

[10] Sharma H, Yu S, Kong J, Wang J, Steitz TA. Structure of Apo-CAP Reveals That Large Conformational Changes Are Necessary for DNA Binding. Proc. Natl. Acad. Sci. USA. 2009;106(39):16604–9. doi:10.1073/Proc.Natl.Acad.Sci.USA.0908380106.

[11] Gunasekaran K, Ma B, Nussinov R. Is Allostery an Intrinsic Property of All Dynamic Proteins? Proteins: Structure, Function and Genetics. 2004; 57(3):433–443. doi:10.1002/prot.20232.

[12] Won HS, Lee YS, Lee SH, Lee BJ. Structural Overview on the Allosteric Activation of Cyclic AMP Receptor Protein. Biochim. Biophys. Acta. 2009; 1794(9):1299–308. doi:10.1016/j.bbapap.2009.04.015.

[13] Changeux JP. Allostery and the Monod-Wyman-Changeux Model After 50 Years. Annu. Rev. Biophys. 2012;41:103–33. doi:10.1146/annurev-biophys-050511-102222.

[14] Motlagh HN, Wrabl JO, Li J, Hilser VJ. The Ensemble Nature of Allostery. Nature. 2014;508(7496):331–339. doi:10.1038/nature13001.

[15] Tsai CJ, Nussinov R. A Unified View of “How Allostery Works”. PLoS Comput. Biol. 2014;10(2):e1003394. doi:10.1371/journal.pcbi.1003394.

[16] Liu J, Nussinov R. Energetic Redistribution in Allostery to Execute Protein Function. Proc. Natl. Acad. Sci. USA. 2017;114(29):7480–7482. doi:10.1073/Proc.Natl.Acad.Sci.USA.1709071114.

[17] White JT, Li J, Grasso E, Wrabl JO, Hilser VJ. Ensemble Allosteric Model: Energetic Frustration Within the Intrinsically Disordered Glucocorticoid Receptor. Philosophical transactions of the Royal Society of London. Series B, Biological sciences. 2018;373(1749):20170175. doi:10.1098/rstb.2017.0175.

[18] Baldwin AJ, Kay LE. NMR Spectroscopy Brings Invisible Protein States into Focus. Nat. Chem. Biol. 2009;5(11):808–814. doi:10.1038/nchembio.238.

[19] Freiburger L, Auclair K, Mittermaier A. Global ITC Fitting Methods in Studies of Protein Allostery. Methods. 2015;76:149–161. doi:10.1016/J.YMETH.2014.12.018.

[20] Tzeng SR, Kalodimos CG. Dynamic Activation of an Allosteric Regulatory Protein. Nature. 2009;462(7271):368–372. doi:10.1038/nature08560.

[21] Grutsch S, Brüschweiler S, Tollinger M. NMR Methods to Study Dynamic Allostery. PLoS Comput. Biol. 2016;12(3):e1004620. doi:10.1371/journal.pcbi.1004620.

[22] Monod J, Wyman J, Changeux JP. On the Nature of Allosteric Transitions: A Plausible Model. J. Mol. Biol. 1965;12:88–118. doi:10.1016/S0022-2836(65)80285-6.

[23] Koshland DE, Némethy G, Filmer D. Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits. Biochem. 1966;5(1):365–385. doi:10.1021/bi00865a047.

[24] James LC, Roversi P, Tawfik DS. Antibody Multispecificity Mediated by Conformational Diversity. Science. 2003;299(5611):1362–1367. doi:10.1126/science.1079731.

[25] Bahar I, Chennubhotla C, Tobi D. Intrinsic Enzyme Dynamics in the Unbound State and Relation to Allosteric Regulation. Curr. Opin. Struct. Biol. 2007;17(6):633–640. doi:10.1016/j.sbi.2007.09.011.

[26] Cui Q, Karplus M. Allostery and Cooperativity Revisited. Protein Science. 2008;17(8):1295–1307. doi:10.1110/ps.03259908.

[27] Park PSH, Lodowski DT, Palczewski K. Activation of G Protein–Coupled Receptors: Beyond Two-State Models and Tertiary Conformational Changes. Annual Review of Pharmacology and Toxicology. 2008;48(1):107–141. doi:10.1146/annurev.pharmtox.48.113006.094630.

[28] del Sol A, Tsai CJ, Ma B, Nussinov R. The Origin of Allosteric Functional Modulation: Multiple Pre-Existing Pathways. Structure. 2009;17(8):1042–1050. doi:10.1016/j.str.2009.06.008.

[29] Lanfranco MF, Gárate F, Engdahl AJ, Maillard RA. Asymmetric Configurations in a Reengineered Homodimer Reveal Multiple Subunit Communication Pathways in Protein Allostery. J. Biol. Chem. 2017;292(15):6086–6093. doi:10.1074/jbc.M117.776047.

[30] Lin SH, Lee JC. Communications Between the High-Affinity Cyclic Nucleotide Binding Sites in *E. coli* Cyclic AMP Receptor Protein. Biochem. 2002;41(39):11857–11867. doi:10.1021/BI026099Z.

[31] Dai J, Lin SH, Kemmis C, Chin AJ, Lee JC. Interplay Between Site-Specific Mutations and Cyclic Nucleotides in Modulating DNA Recognition by *Escherichia coli* Cyclic AMP Receptor Protein. Biochem. 2004;43(28):8901–8910. doi:10.1021/bi0499359.

[32] Lin SH, Kovac L, Chin AJ, Chin CCQ, Lee JC. Ability of *E. coli* Cyclic AMP Receptor Protein To Differentiate Cyclic Nucelotides: Effects of Single Site Mutations. Biochem. 2002;41(9):2946–2955. doi:10.1021/BI0119215.

[33] Heyduk E, Heyduk T, Lee JC. Intersubunit Communications in *Escherichia coli* Cyclic AMP Receptor Protein: Studies of the Ligand Binding Domain. Biochem. 1992;31(14):3682–3688. doi:10.1021/bi00129a017.

[34] Popovych N, Sun S, Ebright RH, Kalodimos CG. Dynamically Driven Protein Allostery. Nature Structural and Molecular Biology. 2006;13(9):831–838. doi:10.1038/nsmb1132.

[35] Takahashi M, Blazy B, Baudras A, Hillen W. Ligand-Modulated Binding of a Gene Regulatory Protein to DNA. Quantitative Analysis of Cyclic-AMP Induced Binding of CRP from *Escherichia coli* to Non-Specific and Specific DNA Targets. J. Mol. Biol. 1989;207(4):783–796. doi:10.1016/0022-2836(89) 90244-1.

[36] Yu S, Maillard RA, Gribenko AV, Lee JC. The N-Terminal Capping Propensities of the D-Helix Modulate the Allosteric Activation of the *Escherichia coli* cAMP Receptor Protein. J. Biol. Chem. 2012;287(47):39402–39411. doi:10.1074/jbc.M112.404806.

[37] Marzen S, Garcia HG, Phillips R. Statistical Mechanics of Monod-Wyman-Changeux (MWC) Models. J. Mol. Biol. 2013;425(9):1433–1460. doi:10.1016/j.jmb.2013.03.013.

[38] Tzeng SR, Kalodimos CG. Protein Activity Regulation by Conformational Entropy. Nature. 2012;488:236–240. doi:10.1038/nature11271.

[39] Heyduk T, Lee JC. Application of Fluorescence Energy Transfer and Polarization to Monitor *Escherichia coli* cAMP Receptor Protein and *lac* Promoter Interaction. Proc. Natl. Acad. Sci. USA. 1990;87(5):1744–1748.

[40] Heyduk T, Lee JC. *Escherichia coli* cAMP Receptor Protein: Evidence for Three Protein Conformational States with Different Promoter Binding Affinities. Biochem. 1989;28(17):6914–6924. doi:10.1021/bi00443a021.

[41] Pyles EA, Lee JC. Mode of Selectivity in Cyclic AMP Receptor Protein-Dependent Promoters in *Escherichia coli*. Biochem. 1996;35(4):1162–1172. doi:10.1021/bi952187q.

[42] Harman JG. Allosteric Regulation of the cAMP Receptor Protein. Biochim. Biophys. Acta. 2001;1547(1):1–17. doi:10.1016/S0167-4838(01)00187-X.

[43] Li G, Young KD. A CAMP-Independent Carbohydrate-Driven Mechanism Inhibits *tnaA* Expression and TnaA Enzyme Activity in *Escherichia coli*. Microbiology (United Kingdom). 2014;160(PART 9):2079–2088. doi:10.1099/mic.0.080705-0.

[44] Goldenbaum PE, Hall GA. Transport of Cyclic Adenosine 3’,5’-Monophosphate Across *Escherichia coli* Vesicle Membranes. J. Bacteriol. 1979;140(2):459–467.

[45] Cossart P, Gicquel-Sanzey B. Regulation of Expression of the *crp* Gene of *Escherichia coli* K-12: *in Vivo* Study. J. Bacteriol. 1985;161(1):454–457.

[46] Daber R, Sochor MA, Lewis M. Thermodynamic Analysis of Mutant Lac Repressors. J. Mol. Biol. 2011;409(1):76–87. doi:10.1016/j.jmb.2011.03.057.

[47] Hines KE, Middendorf TR, Aldrich RW. Determination of Parameter Identifiability in Nonlinear Biophysical Models: A Bayesian Approach. J. Gen. Physiol. 2014;143(3):401–416. doi:10.1085/jgp.201311116.

[48] Transtrum MK, Machta BB, Brown KS, Daniels BC, Myers CR, Sethna JP. Perspective: Sloppiness and Emergent Theories in Physics, Biology, and Beyond. J. Chem. Phys. 2015;143(1):010901. doi:10.1063/1.4923066.

[49] Hilser VJ, García-Moreno EB, Oas TG, Kapp G, Whitten ST. A Statistical Thermodynamic Model of the Protein Ensemble. Chem. Rev. 2006; 106(5):1545–1558. doi:10.1021/cr040423+.

[50] Gunawardena J. Models in Biology: Accurate Descriptions of Our Pathetic Thinking. BMC Biol. 2014;12(1):29. doi:10.1186/1741-7007-12-29.

[51] Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, et al. Transcriptional Regulation by the Numbers: Models. Curr Opin Genetics Dev. 2005; 15(2):116–124. doi:10.1016/j.gde.2005.02.007.

[52] Youn H, Kerby RL, Conrad M, Roberts GP. Study of Highly Constitutively Active Mutants Suggests How cAMP Activates cAMP Receptor Protein. J. Biol. Chem. 2006;281(2):1119–1127. doi:10.1074/jbc.M509421200.

[53] Razo-Mejia M, Barnes SL, Belliveau NM, Chure G, Einav T, Lewis M, et al. Tuning Transcriptional Regulation Through Signaling: A Predictive Theory of Allosteric Induction. Cell Systems. 2018;6(4):456–469. doi:10.1016/j.cels.2018.02.004.

[54] Rohatgi A. WebPlotDigitizer, 2017.

Chapter S4

SUPPLEMENTARY INFORMATION FOR THEORETICAL ANALYSIS OF INDUCER AND OPERATOR BINDING FOR CYCLIC-AMP RECEPTOR PROTEIN MUTANTS

S4.1 Additional Characterizations of the cAMP-CRP Binding Data

In this section, we probe more deeply into the cAMP-CPR data presented in Fig. 4.3 of the main text. We first explore how both the MWC and KNF descriptions for this binding process enable us to use a subset of the available data to predict the remaining data. We then take the opposite approach and analyze each CRP mutant independently within each model and determine how well each data set can conform to an MWC or KNF description. Finally, we touch upon how NMR data allows us to make direct contact between experiment and theory by measuring the fraction of active CRP molecules at saturating cAMP.

S4.1.1 Predicting the Asymmetric Mutants from the Symmetric Mutants

In the main text, our goal was to characterize the full suite of data generated by Lanfranco *et al.* in order to determine the best-possible description of the CRP system. In this section, our goal is not to analyze a system retrospectively, but rather to test how well knowledge of a subset of the CRP mutants can predict the behavior of the remaining mutants. Suppose that Lanfranco *et al.* had only measured cAMP binding for the symmetric CRP mutants (black box in Fig. 4.1B). Can we use this data to predict the behavior of the asymmetric mutants (pink box in Fig. 4.1B)?

Fig. S4.1A shows the calibration of the MWC model Eq. 4.1 to the three symmetric CRP mutants which, being the only data fit, are very well characterized. The corresponding parameters are shown in Table S4.1. Aside from $\tilde{M}_{\text{WT}}^{S}$ and the interaction energy $\epsilon_{\text{int}}^{I}$, the remaining parameters are all within a factor of 2 of their more precise values obtained by fitting the entire data set (Table 4.1 in the main text).

Since the three symmetric CRP mutants contain all three types of subunits (WT, D, and S), we can now predict the behavior of the asymmetric CRP mutants without recourse to any further fitting. Using Eq. 4.2 and the parameters in Table S4.1, the resulting predictions are plotted in Fig. S4.1B using dashed lines to emphasize

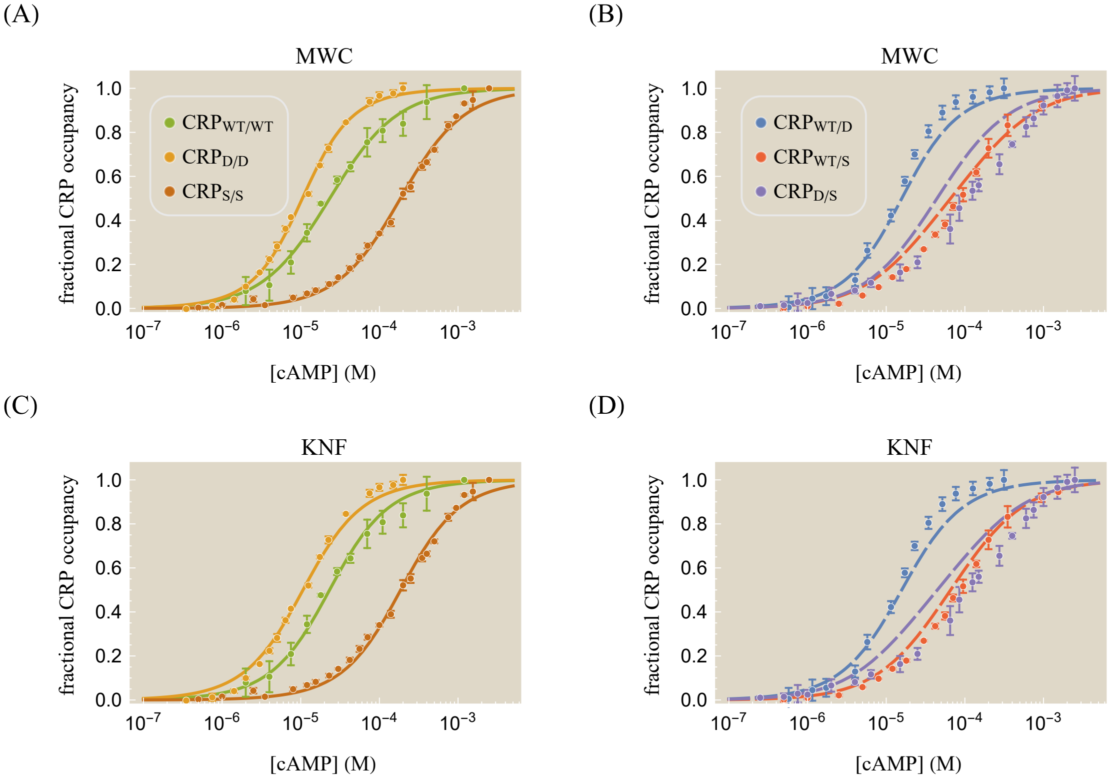


Figure S4.1: Predicting the behavior of the asymmetric CRP mutants. (A) We infer the cAMP-CRP dissociation constants for the WT, D, and S subunits using the best-fit characterizations (solid curves) of the three symmetric CRP mutants. (B) Using these values, we can predict the behavior of the asymmetric mutants (dashed curves) with no further fitting. (C) In an analogous manner, the KNF model can be calibrated using the symmetric CRP mutants. (D) The corresponding predictions of the KNF model for the asymmetric CRP mutants. The (corrected) sample standard deviations for the asymmetric mutant predictions are 0.08 for both the MWC and KNF models, and the resulting best-fit parameters are shown in Table S4.1.

that the data in those plots were not used to fit these curves. Fig. S4.1C shows the analogous calibration of the KNF model using the symmetric CRP mutants and Fig. S4.1D displays the corresponding predictions of the asymmetric mutants.

As expected, the predicted cAMP-CRP binding curves for both models are worse than the main text fits, since in the main text *all* of the cAMP-CRP data was used to infer the parameters. While these predictions only capture the approximate shape of the data, it must be noted that they provide an answer to the otherwise impossible question: how can you predict the behavior of an asymmetric CRP mutant without measuring it? Note the predictive power scales with the number of subunits. For example, the five symmetric mutants shown in Fig. 4.1B can be used to predict the

MWC Parameter	Best-Fit Value	KNF Parameter	Best-Fit Value
$\tilde{M}_{\text{WT}}^A, M_{\text{WT}}^I$	$\{20 \pm 2, 20 \pm 3\} \times 10^{-6}$ M	$\bar{M}_{\text{WT}}^A$	$(30 \pm 3) \times 10^{-6}$ M
$\tilde{M}_{\text{D}}^A, M_{\text{D}}^I$	$\{10 \pm 1, 20 \pm 3\} \times 10^{-6}$ M	$\bar{M}_{\text{D}}^A$	$(10 \pm 2) \times 10^{-6}$ M
$\tilde{M}_{\text{S}}^A, M_{\text{S}}^I$	$\{200 \pm 10, 220 \pm 20\} \times 10^{-6}$ M	$\bar{M}_{\text{S}}^A$	$(250 \pm 20) \times 10^{-6}$ M
ϵ_{int}^I	$2.2 \pm 0.2\, k_B T$	ϵ_{int}^A	$-0.6 \pm 0.2\, k_B T$

Table S4.1: Best-fit parameters for the symmetric CRP mutant fitting. Using only the symmetric CRP mutant data in Fig. S4.1 allows us to infer the following cAMP-CRP binding parameters from which we can predict the behavior of the asymmetric mutants. These values should be compared with the corresponding best-fit parameters when fitting the entire data set (Table 4.1).

behavior of the ten asymmetric mutants. More generally, given N subunits we could carry out N experiments on the symmetric mutants and predict the responses of the $\frac{N(N-1)}{2}$ asymmetric mutants. Since the number of mutants scales quadratically with N, the number of predictions grows much faster than the number of experiments that must be carried out, enabling us to harness the combinatorial complexity of oligomeric proteins. The expediency of checking the space of mutants theoretically may be well worth the decrease in the quality of this characterization.

S4.1.2 Individual Characterizations of each CRP Mutant

In this section, we relax our assumption that each CRP subunit behaves identically regardless of the composition of its partner subunit and instead analyze a broader question: how well can the MWC or KNF models presented in the text characterize each individual cAMP-CRP binding curve? In other words, suppose there is a complex interaction between CRP subunits, so that the free energy change $\frac{c}{M_{\text{WT}}^A}$ of cAMP binding to either subunit in CRP$_{\text{WT/WT}}$ is different from the corresponding free energy change of the WT subunit in CRP$_{\text{WT/D}}$. To this end, we fit each CRP mutant's binding data to either the MWC model Eq. 4.1 or the KNF model Eq. 4.5 with no constraint between the parameters of the individual fittings. We note that the original analysis of the six mutants conducted by Lanfranco *et al.* was carried out in an analogous manner by fitting each mutant individually [6].

To simplify this analysis, we restrict ourselves to the case where the two interaction energies are negligible ($\epsilon_{\text{int}}^A = \epsilon_{\text{int}}^I = 0$); even with this restriction, we find that both the MWC and KNF models can fit the individual data sets remarkably well. In the case where the interaction energies are allowed to be non-zero (not shown), the MWC model negligibly improves while the KNF model becomes indistinguishable from the MWC model. However, we now proceed with the assumption that both

Mutant Fit	MWC Parameter	Best-Fit Value	KNF Parameter	Best-Fit Value
$\text{CRP}_{\text{WT/WT}}$	$\tilde{M}^A_{\text{WT}}$, M^I_{WT}	$\{210, 20\} \times 10^{-6}$ M	$\bar{M}^A_{\text{WT}}$	20×10^{-6} M
$\text{CRP}_{\text{D/D}}$	$\tilde{M}^A_{\text{D}}$, M^I_{D}	$\{10, 20\} \times 10^{-6}$ M	$\bar{M}^A_{\text{D}}$	10×10^{-6} M
$\text{CRP}_{\text{S/S}}$	$\tilde{M}^A_{\text{S}}$, M^I_{S}	$\{370, 220\} \times 10^{-6}$ M	$\bar{M}^A_{\text{S}}$	180×10^{-6} M
$\text{CRP}_{\text{WT/D}}$	$\tilde{M}^A_{\text{WT}}$, M^I_{WT}, $\tilde{M}^A_{\text{D}}$, M^I_{D}	$\{20, 40, 10, 40\} \times 10^{-6}$ M	$\bar{M}^A_{\text{WT}}$, $\bar{M}^A_{\text{D}}$	$\{10, 10\} \times 10^{-6}$ M
$\text{CRP}_{\text{WT/S}}$	$\tilde{M}^A_{\text{WT}}$, M^I_{WT}, $\tilde{M}^A_{\text{S}}$, M^I_{S}	$\{50, 40, 160, 2000\} \times 10^{-6}$ M	$\bar{M}^A_{\text{WT}}$, $\bar{M}^A_{\text{S}}$	$\{80, 80\} \times 10^{-6}$ M
$\text{CRP}_{\text{D/S}}$	$\tilde{M}^A_{\text{D}}$, M^I_{D}, $\tilde{M}^A_{\text{S}}$, M^I_{S}	$\{90, 40, 160, 1200\} \times 10^{-6}$ M	$\bar{M}^A_{\text{D}}$, $\bar{M}^A_{\text{S}}$	$\{50, 220\} \times 10^{-6}$ M

Table S4.2: Best-fit parameters for the individual CRP mutant fitting. The following parameters were determined by fitting each cAMP-CRP data set separately. Thus, each CRP mutant yield slightly different values for the same fit parameters. In both models, the interactions energies are assumed to be zero, $\epsilon^A_{\text{int}} = \epsilon^I_{\text{int}} = 0$.

interaction energies are vanishingly small.

Fig. S4.2 shows the resulting individual fits for the CRP mutants within the MWC and KNF models, with the corresponding parameters given in Table S4.2. We find that both models can characterize all of the data sets very well; aside from small errors in the KNF description of $\text{CRP}_{\text{WT/D}}$, nearly every data point is within one standard deviation of the predicted value. Thus, both models are capable of characterizing the cAMP-CRP binding behavior, and any discrepancies between the theory and data in Fig. 4.3 may ultimately be attributed to the assumption posited in the text that the WT, D, and S subunits must function identically regardless of the identity of the other CRP subunit. We note that fitting each curve individually results in a sample standard deviation that is a factor of 2 smaller than fitting the six mutants using a single set of parameters (Fig. 4.3); we stress again that this improvement in fit quality comes at the cost of losing both a unified description of the system as well as the predictive power discussed in the previous section.

That said, analyzing each curve individually has the merit in providing the best possible characterization of the data. For example, the individual fitting provides a smooth interpolation between the data points of each CRP mutant, enabling us to compute additional properties of the binding such as the slope at the half-maximal effective concentration (also known as the effective Hill coefficient [7]), which is given by

$$h = \left(2 \frac{d}{d \log[M]} \log\left(\text{fractional CRP occupancy}\right) \right)_{[M]=[EC_{50}^{\text{cAMP}}]}. \tag{S4.1}$$

Table S4.3 shows that the effective Hill coefficients of each data set is roughly one. To understand this result, note that the fractional CRP occupancy Eq. 4.6 within the

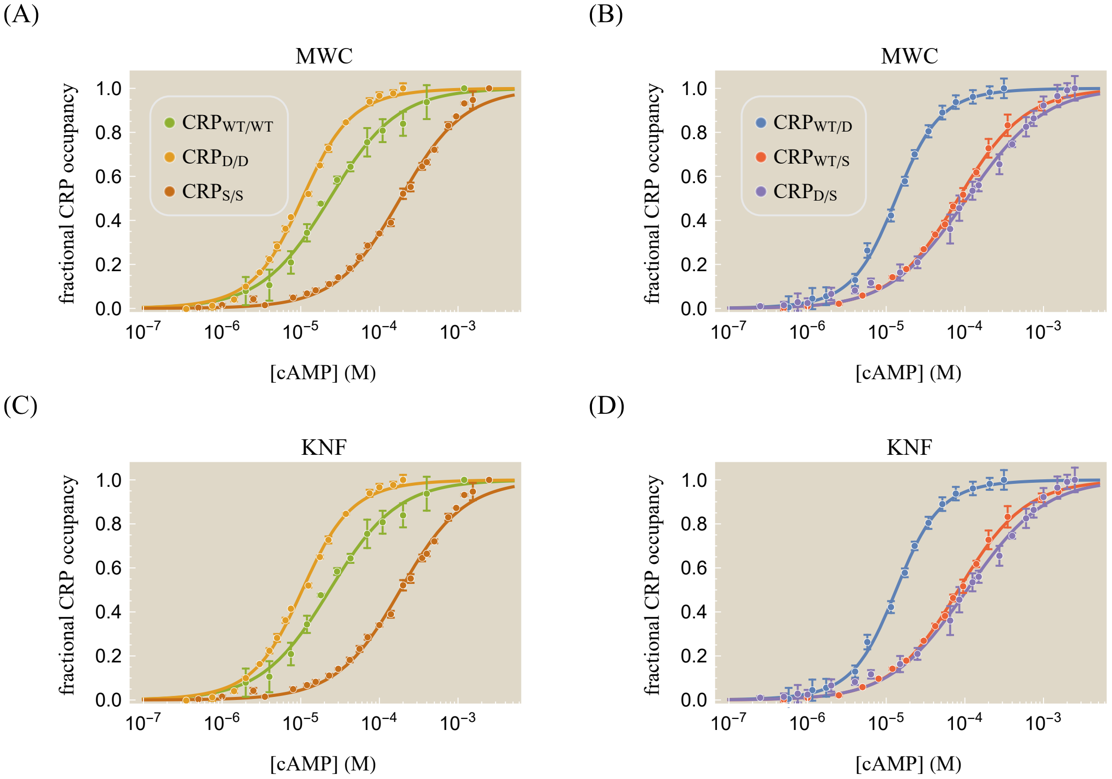


Figure S4.2: Individual characterization of each CRP mutant. Each of the (A) symmetric and (B) asymmetric CRP mutants are characterized individually using the MWC model Eq. 4.1, showing how closely the model could match the data if the assumption that each subunit behaves identically and independently is relaxed. Similarly, each of the (C) symmetric and (D) asymmetric CRP mutants are characterized separately using the KNF model Eq. 4.5. The sample standard deviation equals 0.02 for the MWC model and 0.04 for the KNF model, and the best-fit parameters for both models are given in Table S4.2.

KNF model when both subunits have the same cAMP affinity ($\bar{M}_L^A = \bar{M}_R^A$) is given by the Hill equation

$$\text{fractional CRP occupancy}([M]) = \frac{\frac{[M]}{\bar{M}_L^A}}{1 + \frac{[M]}{\bar{M}_L^A}} \tag{S4.2}$$

which has an effective Hill coefficient of 1. In other words, if we linearly approximate the fractional occupancy of any curve in Fig. S4.2C at its midpoint on the log-linear plot, any symmetric CRP mutant will transition from being unbound to cAMP (fractional CRP occupancy ≈ 0) to mostly bound (fractional CRP occupancy ≈ 1) over approximately one order of magnitude in cAMP concentration.

From Table S4.2, the three asymmetric CRP mutants also have nearly identical cAMP binding affinities in their two subunits within the KNF model, thereby leading

Mutant Fit	MWC Effective Hill Coefficient	KNF Effective Hill Coefficient
CRP$_{\text{WT/WT}}$	1.0	1.0
CRP$_{\text{D/D}}$	1.4	1.0
CRP$_{\text{S/S}}$	1.1	1.0
CRP$_{\text{WT/D}}$	1.5	1.0
CRP$_{\text{WT/S}}$	1.0	1.0
CRP$_{\text{D/S}}$	0.9	0.9

Table S4.3: Effective Hill coefficients for the CRP mutants. The effective Hill coefficient Eq. S4.1 is approximately one for all of the CRP mutants in both the MWC and KNF models.

to a value of approximately one for each of their effective Hill coefficients as well. Note that the slopes at the midpoints of the CRP$_{\text{D/D}}$ and CRP$_{\text{WT/D}}$ binding curves are slightly steeper than predicted by the KNF model. On the other hand, the effective Hill coefficient for the MWC model, which can be derived in an analogous manner, is more complex than the KNF expression but captures the slopes of all six CRP mutants more accurately, as is seen in Fig. S4.2A and B.

We end with the cautionary note that a Hill coefficient of order unity does not imply that there is little-to-no cooperativity in the system. Indeed, Lanfranco *et al.* determined that CRP$_{\text{WT/WT}}$ is 15x more cooperative than CRP$_{\text{S/S}}$ and 5x less cooperative than CRP$_{\text{D/D}}$, features that are completely masked by only considering the effective Hill coefficient (see the cooperativity (c) column in Table 1 as well as Eq (1) of Ref. [6]). Although the precise definition of cooperativity depends upon the model used, the slope at the half-way point of a sigmoidal response may not be a good indicator of the energies and dissociation constants governing a system.

S4.1.3 Comparing the Fraction of CRP in the Active State

In this section, we derive the fraction of active CRP, enabling us to use NMR measurements to compare the predictions of the MWC and KNF models to experiment. Using Fig. 4.2, the probability that CRP will be in the active state is given by the sum of active weights divided by the sum of all weights, namely,

$$\text{fraction active CRP}([M]) = \frac{1 + \frac{[M]}{M_L^A} + \frac{[M]}{M_R^A} + e^{-\beta\epsilon_{\text{int}}^A}\frac{[M]}{M_L^A}\frac{[M]}{M_R^A}}{1 + \frac{[M]}{M_L^A} + \frac{[M]}{M_R^A} + e^{-\beta\epsilon_{\text{int}}^A}\frac{[M]}{M_L^A}\frac{[M]}{M_R^A} + e^{-2\beta\epsilon}\left(1 + \frac{[M]}{M_L^I} + \frac{[M]}{M_R^I} + e^{-\beta\epsilon_{\text{int}}^I}\frac{[M]}{M_L^I}\frac{[M]}{M_R^I}\right)} \tag{S4.3}$$

$$\approx \frac{\frac{[M]}{\tilde{M}_L^A}\frac{[M]}{\tilde{M}_R^A}}{\frac{[M]}{\tilde{M}_L^A}\frac{[M]}{\tilde{M}_R^A} + \left(1 + \frac{[M]}{M_L^I} + \frac{[M]}{M_R^I} + e^{-\beta\epsilon_{\text{int}}^I}\frac{[M]}{M_L^I}\frac{[M]}{M_R^I}\right)}, \tag{S4.4}$$

where we have applied the same approximations as in Eqs. 4.2-4.4. Similarly, the fraction of active CRP in the KNF model is given by

$$\text{fraction active CRP}([M]) = \frac{e^{-\beta\epsilon_{\text{int}}^A}\frac{[M]}{M_L^A}\frac{[M]}{M_R^A}}{e^{-2\beta\epsilon} + e^{-\beta\epsilon}\left(\frac{[M]}{M_L^A} + \frac{[M]}{M_R^A}\right) + e^{-\beta\epsilon_{\text{int}}^A}\frac{[M]}{M_L^A}\frac{[M]}{M_R^A}} \tag{S4.5}$$

$$= \frac{e^{-\beta\epsilon_{\text{int}}^A}\frac{[M]}{\bar{M}_L^A}\frac{[M]}{\bar{M}_R^A}}{1 + \frac{[M]}{\bar{M}_L^A} + \frac{[M]}{\bar{M}_R^A} + e^{-\beta\epsilon_{\text{int}}^A}\frac{[M]}{\bar{M}_L^A}\frac{[M]}{\bar{M}_R^A}}, \tag{S4.6}$$

where we have used Eqs. 4.7 and 4.8.

Fig. 4.3C,F show the resulting predictions for the fraction of active CRP in both models, with the two models starkly disagreeing in the limit of saturating cAMP. Whereas the KNF model predicts that CRP must ultimately be doubly bound and hence active for sufficiently large concentrations of cAMP, the MWC model purports that some CRP may be inactive regardless of how high the cAMP concentration is raised. In particular, the MWC model predicts that the S subunit is highly biased towards the inactive state, so that only a minute fraction of the S/S mutant will be active at saturating cAMP. This MWC viewpoint agrees with NMR data which shows that only 2% of the S/S mutant is active at saturating cAMP [8].

S4.1.4 CRP Activation with the Same RNAP Affinity in the Active and Inactive CRP States

In this section, we demonstrate why the activation energies between RNAP and active-CRP (ϵ_{P,L_A}) as well as between RNAP and inactive-CRP (ϵ_{P,L_I}) must be different in order to characterize the activation data in Fig. 4.7A.

To orient ourselves, the maximum possible cAMP activation (see (Fig. 4.6 and Eq. 4.14) is achieved if in the absence of cAMP all CRP was inactive ($[L_A] = 0$, $[L_I] = [L]$) and in the limit of saturating cAMP all CRP would be active ($[L_A] = [L]$,

$[L_I] = 0$). In such a case, fold-change in activity (i.e. the activity in the presence of saturating cAMP divided by the activity in the absence of cAMP) that we would like to maximize would be given by

$$\text{fold-change in activity} = \frac{\left(1 + e^{-\beta\epsilon_{P,L_A}}\frac{[L]}{L_A}\right)\left(1 + \frac{[L]}{L_I} + \left(1 + e^{-\beta\epsilon_{P,L_I}}\frac{[L]}{L_I}\right)\frac{[P]^2}{P_D^2}\right)}{\left(1 + e^{-\beta\epsilon_{P,L_I}}\frac{[L]}{L_I}\right)\left(1 + \frac{[L]}{L_A} + \left(1 + e^{-\beta\epsilon_{P,L_A}}\frac{[L]}{L_A}\right)\frac{[P]^2}{P_D^2}\right)}. \tag{S4.7}$$

In the limit where the interactions energies are the same and either really small ($\epsilon_{P,L_A} = \epsilon_{P,L_I} = 0$) or really large ($\epsilon_{P,L_A} = \epsilon_{P,L_I} \to -\infty$), the fold change goes to 1 (the only assumption necessary is that $\frac{[P]}{P_D} \ll 1$, as is true in the data set we consider where the best-fit value is $\frac{[P]}{P_D} \approx 10^{-6}$). A in fold-change emerges at intermediate values of the activation energy, but such a model is unable to fit the 30-fold increase in activation of the experimental data in Fig. 4.7A.

In the opposite limit where the activation energies are infinitely far apart from each other ($\epsilon_{P,L_A} \to -\infty$ and $\epsilon_{P,L_I} = 0$), the fold-change in activation reduces to the large value of $\left(\frac{[P]}{P_D}\right)^{-2} \approx 10^{12}$, which is more than capable of matching the 30-fold increase in activation seen in the data. As described in the main text, the much more modest energy difference of $\epsilon_{P,L_A} = -3k_BT$ and $\epsilon_{P,L_I} = 0$ can characterize the data.

S4.2 Multiple Mutations within a Subunit

In addition to their symmetric mutants (WT/WT, D/D, S/S) and asymmetric mutants (WT/D, WT/S, D/S), Lanfranco *et al.* constructed one additional WT/D+S mutant that had both the D and S mutation within one subunit. The purpose of this mutant was to study the difference between intersubunit communication (D/S) and intrasubunit communication (WT/D+S) in an allosteric protein. In this section, we consider the simplest possible model for this double mutant, namely, that the change in free energy incurred by both the D and S mutations is additive and independent. While mutations are often epistatic (i.e. either not independent or nonlinear; for example see Ref. [9]), this simple model provides a null hypothesis to test whether there are any interactions between such mutations.

Since the WT subunit has already been characterized (see Table 4.1), we only need to determine how the D+S subunit behaves. To proceed, we will assume that the effects of the D and S mutations on cAMP-CRP affinity are independent as, for example, has been shown to be the case for mutations in the pore region of the nAChR ion channel [10]. More precisely, if the D mutation increases the cAMP-CRP binding energy by $2\,k_BT$ and the S mutation changes it by $-3\,k_BT$, then we assume the D+S mutant changes the cAMP-CRP binding energy by their sum, $-1\,k_BT$. Because dissociation constants are proportional to the exponential of the cAMP-CRP binding energy, $K_D \propto e^{\beta\epsilon_{\text{bound}}}$, the D+S mutation translates into a multiplicative effect on the dissociation constant. For example, the inactive state dissociation constants in the MWC model will obey

$$M^I_{\text{D+S}} = M^I_{\text{WT}} \frac{M^I_{\text{D}}}{M^I_{\text{WT}}} \frac{M^I_{\text{S}}}{M^I_{\text{WT}}}. \tag{S4.8}$$

In both the MWC and KNF models, we can obtain an analogous expression for the active state dissociation constants by changing $I \to A$ in the superscripts. Multiplying both sides by $e^{-\beta\epsilon}$ leads to the equivalent statement for the effective dissociation constants,

$$\tilde{M}^A_{\text{D+S}} = \tilde{M}^A_{\text{WT}} \frac{\tilde{M}^A_{\text{D}}}{\tilde{M}^A_{\text{WT}}} \frac{\tilde{M}^A_{\text{S}}}{\tilde{M}^A_{\text{WT}}}. \tag{S4.9}$$

Using these relations, we can predict the behavior of $\text{CRP}_{\text{WT/D+S}}$ without recourse to fitting. Fig. S4.3A shows the predictions for the MWC model (dashed black curve) together with the experimental measurements (black data points). For reference, the data and best-fit curves for the three symmetric mutants (green, gold, brown) are also shown. Fig. S4.3B demonstrates that a very similar prediction is achieved by the KNF model. In both cases, the experimental measurements roughly follow the theoretical predictions, suggesting that there may be some epistatic interaction in the system, but that assuming linearity and independence for the D and S mutations provides a reasonable zeroth order approximation to the behavior of the system. While we find it interesting that the KNF prediction better matches the $\text{CRP}_{\text{WT/D+S}}$ data, it does not exonerate the KNF model's shortcomings described in the main text (e.g. predicting that $\text{CRP}_{\text{S/S}}$ will be 100% active in the limit of saturating cAMP even though experimental measurements show it to be 2% active in this limit). It would be interesting to compare the ability of the MWC and KNF models to predict the behavior of other double mutants, as would be possible by introducing additional subunit mutations such as G141Q and L148R [11].

With this characterization of the D+S subunit, we can similarly predict how the other possible CRP mutants (D/D+S, S/D+S, D+S/D+S) would behave, as shown in Fig. S4.4. Since Lanfranco *et al.* did not construct any of these mutants, they provide a unique testbed to probe how well the notion of independent mutations holds up within the context of CRP.

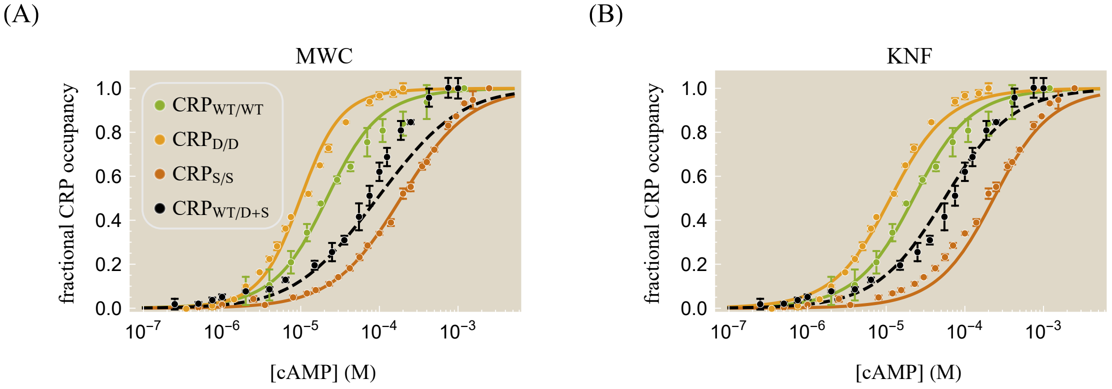


Figure S4.3: Effect of a D+S double mutation. With both the (A) MWC and (B) KNF models, the D and S mutations are assumed to be independent and additive, leading to the modified dissociation constants given by Eqs. S4.8 and S4.9. The predicted behavior of the D+S subunit (black line, drawn dashed to emphasize that it was not fit to the data) loosely follows the experimental data (black points) for both models. For reference, the symmetric mutants (WT/WT, D/D, S/S) from from Fig. 4.3 are also shown. Parameters used were the same as in Table 4.1 with no recourse to fitting.

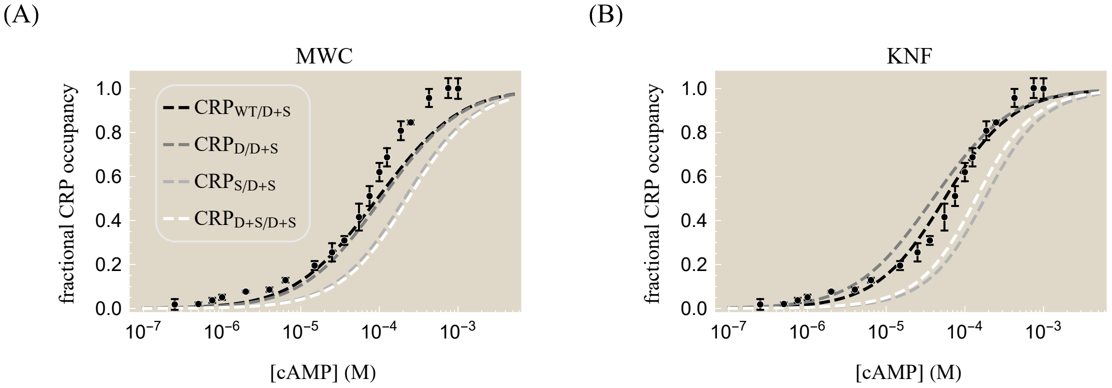


Figure S4.4: Predicting the behavior of other D+S mutants. Using Eqs. S4.8 and S4.9, the behavior of any CRP mutant with a D+S subunit can be modeled. The four possibilities are shown for the (A) MWC and (B) KNF models together with the data on the WT/D+S mutant. Parameters used were the same as in Table 4.1 with no recourse to fitting.

References

[1] Gunasekara SM, Hicks MN, Park J, Brooks CL, Serate J, Saunders CV, et al. Directed Evolution of the *Escherichia coli* cAMP Receptor Protein at the cAMP Pocket. J. Biol. Chem. 2015;290(44):26587–26596. doi:10.1074/jbc.M115.678474.

[2] Cheng X, Lee JC. Differential Perturbation of Intersubunit and Interdomain

Communications by Glycine 141 Mutation in *Escherichia coli* CRP. Biochem. 1998;37(1):51–60. doi:10.1021/bi9719455.

[3] Takahashi M, Blazy B, Baudras A, Hillen W. Ligand-Modulated Binding of a Gene Regulatory Protein to DNA. Quantitative Analysis of Cyclic-AMP Induced Binding of CRP from *Escherichia coli* to Non-Specific and Specific DNA Targets. J. Mol. Biol. 1989;207(4):783–796. doi:10.1016/0022-2836(89) 90244-1.

[4] Popovych N, Sun S, Ebright RH, Kalodimos CG. Dynamically Driven Protein Allostery. Nature Structural and Molecular Biology. 2006;13(9):831–838. doi:10.1038/nsmb1132.

[5] Yu S, Maillard RA, Gribenko AV, Lee JC. The N-Terminal Capping Propensities of the D-Helix Modulate the Allosteric Activation of the *Escherichia coli* cAMP Receptor Protein. J. Biol. Chem. 2012;287(47):39402–39411. doi:10.1074/jbc.M112.404806.

[6] Lanfranco MF, Gárate F, Engdahl AJ, Maillard RA. Asymmetric Configurations in a Reengineered Homodimer Reveal Multiple Subunit Communication Pathways in Protein Allostery. J. Biol. Chem. 2017;292(15):6086–6093. doi: 10.1074/jbc.M117.776047.

[7] Marzen S, Garcia HG, Phillips R. Statistical Mechanics of Monod-Wyman-Changeux (MWC) Models. J. Mol. Biol. 2013;425(9):1433–1460. doi:10. 1016/j.jmb.2013.03.013.

[8] Tzeng SR, Kalodimos CG. Dynamic Activation of an Allosteric Regulatory Protein. Nature. 2009;462(7271):368–372. doi:10.1038/nature08560.

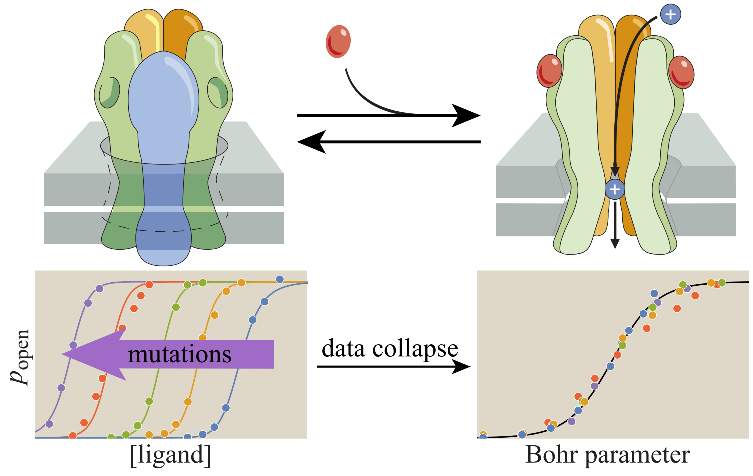
[9] Rodrigues JV, Bershtein S, Li A, Lozovsky ER, Hartl DL, Shakhnovich EI. Biophysical Principles Predict Fitness Landscapes of Drug Resistance. Proc. Natl. Acad. Sci. USA. 2016;113(11):E1470–E1478. doi:10.1073/Proc.Natl. Acad.Sci.USA.1601441113.

[10] Auerbach A. Thinking in Cycles: MWC Is a Good Model for Acetylcholine Receptor-Channels. J. Physiol. 2012;590(1):93–98. doi:10.1113/jphysiol. 2011.214684.

[11] Lin SH, Kovac L, Chin AJ, Chin CCQ, Lee JC. Ability of *E. coli* Cyclic AMP Receptor Protein To Differentiate Cyclic Nucelotides: Effects of Single Site Mutations. Biochem. 2002;41(9):2946–2955. doi:10.1021/BI0119215.

Chapter 5

MONOD-WYMAN-CHANGEUX ANALYSIS OF LIGAND-GATED ION CHANNEL MUTANTS



Einav T, Phillips R. Monod-Wyman-Changeux Analysis of Ligand-Gated Ion Channel Mutants. The Journal of Physical Chemistry B. 2017;121(15):3813-3824. doi:10.1021/acs.jpcb.6b12672

In March 2016, Rob emailed me a manuscript with ion channel data together with the single comment: "The plot is very sneaky (in a bad way)." I read the article and created a model to understand the data, sending the results to Rob at 4am and immediately falling asleep. I awoke to see that Rob had responded ten minutes after my email asking me if I can talk, and then emailing me again two hours later telling me that he was quite excited about these results. When I awoke, we discussed the calculations and Rob gave me many excellent suggestions that I implemented that day and night, sending him the results at 6am. I got out of bed at noon to find more feedback on my latest notes which I immediately implemented. And with that, the first half of this paper was written in three days. Finding the data for the second ion channel took another week. It was the most intense and enjoyable burst of activity of my PhD, and there is no better feeling then to be perfectly in sync with your advisor. The icing on the cake came when we posted the paper on the bioRxiv and got an invitation the next day to submit to the Biophysical Journal (to which we politely informed them that we had already submitted to JPC B). One day, I hope to surpass the speed of even this project!

5.1 Abstract

We present a framework for computing the gating properties of ligand-gated ion channel mutants using the Monod-Wyman-Changeux (MWC) model of allostery. We derive simple analytic formulas for key functional properties such as the leakiness, dynamic range, half-maximal effective concentration ($[EC_{50}]$), and effective Hill coefficient, and explore the full spectrum of phenotypes that are accessible through mutations. Specifically, we consider mutations in the channel pore of nicotinic acetylcholine receptor (nAChR) and the ligand binding domain of a cyclic nucleotide-gated (CNG) ion channel, demonstrating how each mutation can be characterized as only affecting a subset of the biophysical parameters. In addition, we show how the unifying perspective offered by the MWC model allows us, perhaps surprisingly, to collapse the plethora of dose-response data from different classes of ion channels into a universal family of curves.

5.2 Introduction

Ion channels are signaling proteins responsible for a huge variety of physiological functions ranging from responding to membrane voltage, tension, and temperature to serving as the primary players in the signal transduction we experience as vision [1]. Broadly speaking, these channels are classified on the basis of the driving forces that gate them. In this work, we explore one such classification for ligand-gated ion channel mutants based on the Monod-Wyman-Changeux (MWC) model of allostery. In particular, we focus on mutants in two of the arguably best studied ligand-gated ion channels: the nicotinic acetylcholine receptor (nAChR) and the cyclic nucleotide-gated (CNG) ion channel shown schematically in Fig. 5.1 [2, 3].

The MWC model has long been used in the contexts of both nAChR and CNG ion channels [4–6]. Although careful analysis of these systems has revealed that some details of ligand-gated ion channel dynamics are not captured by this model (e.g. the existence and interaction of multiple conducting states [7, 8]), the MWC framework nevertheless captures many critical features of a channel's response and offers one of the simplest settings to explore its underlying mechanisms. For example, knowledge of both the nAChR and CNG systems' molecular architecture and our ability to measure their complex kinetics has only recently become sufficiently advanced to tie the effects of mutations to key biophysical parameters. Purohit and Auerbach used combinations of mutations to infer the nAChR gating energy, finding that unliganded nAChR channels open up for a remarkably brief 80 μs every 15 minutes [9]. Statistical mechanics has been used to show how changes to the

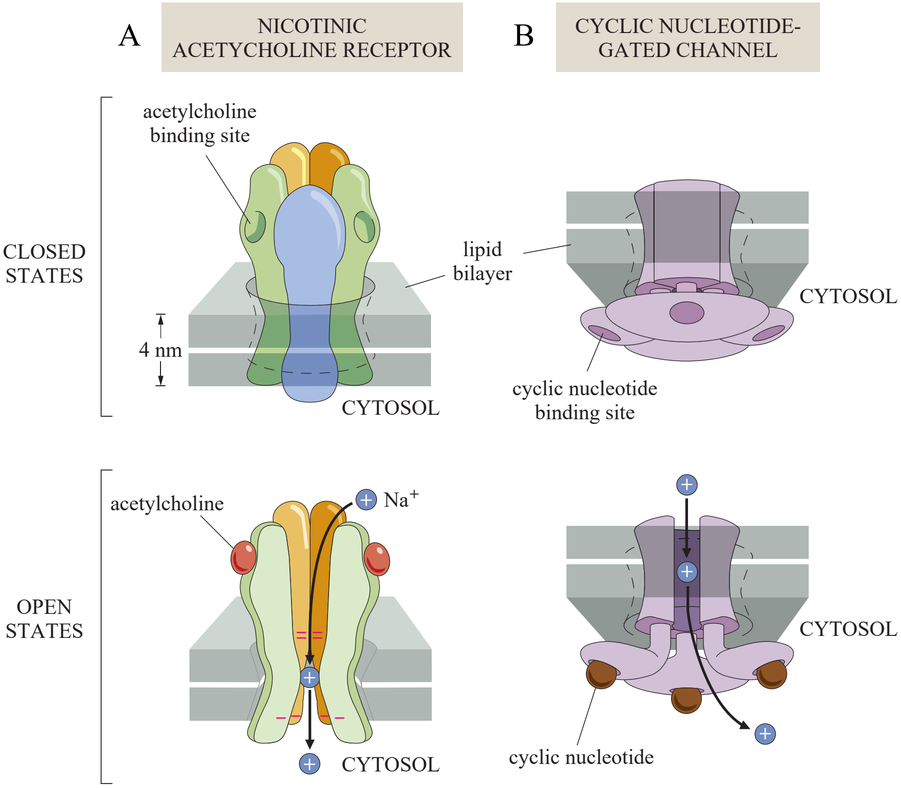

A
NICOTINIC ACETYCHOLINE RECEPTOR
B
CYCLIC NUCLEOTIDE-GATED CHANNEL
CLOSED STATES
acetylcholine binding site
lipid bilayer
4 nm
CYTOSOL
CYTOSOL
cyclic nucleotide binding site
OPEN STATES
acetylcholine
Na^+
CYTOSOL
CYTOSOL
cyclic nucleotide

Figure 5.1: Schematic of nAChR and CNGA2 ion channels. (A) The heteropentameric nicotinic acetylcholine receptor (nAChR) has two ligand binding sites for acetylcholine outside the cytosol. (B) The homotetrameric cyclic nucleotide-gated (CNGA2) has four ligand binding sites, one on each subunit, for cAMP or cGMP located inside the cytosol. Both ion channels have a higher probability of being closed in the absence of ligand and open when bound to ligand.

energy difference between conformations in allosteric proteins translate to different functional behavior (i.e. how it modifies the leakiness, dynamic range, $[EC_{50}]$ and the effective Hill coefficient) [10, 11], and we extend this work to find simple analytic approximations that are valid within the context of ion channels. Using this methodology, we systematically explore the full range of behaviors that may be induced by different types of mutations. This analysis enables us to quantify the inherent trade-offs between key properties of ion channel dose-response curves and potentially paves the way for future biophysical models of evolutionary fitness in which the genotype (i.e. amino acid sequence) of allosteric molecules is directly connected to phenotype (i.e. properties of a channel's response).

To this end, we consider two distinct classes of mutants which tune different sets of MWC parameters – either the ion channel gating energy or the ligand-channel dissociation constants. Previous work by Auerbach *et al.* demonstrated that these two sets of physical parameters can be independently tuned within the nAChR ion channel; pore mutations only alter the channel gating energy whereas mutations within the ligand binding domain only affect the ligand-channel dissociation constants [12]. Utilizing this parameter independence, we determine the full spectrum of nAChR phenotypes given an arbitrary set of channel pore mutations and show why a linear increase in the channel gating energy leads to a logarithmic shift in the nAChR dose-response curve. Next, we consider recent data from CNGA2 ion channels with mutations in their ligand binding pocket [13]. We hypothesize that modifying the ligand binding domain should not alter the channel gating energy and demonstrate how the entire class of CNGA2 mutants can be simultaneously characterized with this constraint. This class of mutants sheds light on the fundamental differences between homooligomeric channels composed of a single type of subunit and heterooligomeric channels whose distinct subunits can have different ligand binding affinities.

By viewing mutant data through its effects on the underlying biophysical parameters of the system, we go well beyond simply fitting individual dose-response data, instead creating a framework with which we can explore the full expanse of ion channel phenotypes available through mutations. Using this methodology, we: (1) analytically compute important ion channel characteristics, namely the leakiness, dynamic range, $[EC_{50}]$, and effective Hill coefficient, (2) link the role of mutations with thermodynamic parameters, (3) show how the behavior of an entire family of mutants can be predicted using only a subset of the members of that family, (4) quantify the pleiotropic effect of point mutations on multiple phenotypic traits and characterize the correlations between these diverse effects, and (5) collapse the data from multiple ion channels onto a single master curve, revealing that such mutants form a one-parameter family. In doing so, we present a unified framework to collate the plethora of data known about such channels.

5.3 Model

Electrophysiological techniques can measure currents across a single cell's membrane. The current flowing through a ligand-gated ion channel is proportional to the average probability $p_{\text{open}}(c)$ that the channel will be open at a ligand concentration c. For an ion channel with m identical ligand binding sites (see Fig. 5.2), this probability is given by the MWC model as

$$p_{\text{open}}(c) = \frac{\left(1 + \frac{c}{K_{\text{O}}}\right)^m}{\left(1 + \frac{c}{K_{\text{O}}}\right)^m + e^{-\beta\epsilon}\left(1 + \frac{c}{K_{\text{C}}}\right)^m}, \tag{5.1}$$

where K_{O} and K_{C} represent the dissociation constants between the ligand and the open and closed ion channel, respectively, c denotes the concentration of the ligand, ϵ (called the gating energy) denotes the free energy difference between the closed and open conformations of the ion channel in the absence of ligand, and $\beta = \frac{1}{k_{\text{B}}T}$ where k_{B} is Boltzmann's constant and T is the temperature of the system. Wild type ion channels are typically closed in the absence of ligand ($\epsilon < 0$) and open when bound to ligand ($K_{\text{O}} < K_{\text{C}}$). Fig. 5.2 shows the possible conformations of the nAChR ($m = 2$) and CNGA2 ($m = 4$) ion channels together with their Boltzmann weights. $p_{\text{open}}(c)$ is given by the sum of the open state weights divided by the sum of all weights. Note that the MWC parameters K_{O}, K_{C}, and ϵ may be expressed as ratios of the experimentally measured rates of ligand binding and unbinding as well as the transition rates between the open and closed channel conformations (see Supporting Information section S5.1.1).

Current measurements are often reported as *normalized* current, implying that the current has been stretched vertically to run from 0 to 1, as given by

$$\text{normalized current} = \frac{p_{\text{open}}(c) - p_{\text{open}}^{\text{min}}}{p_{\text{open}}^{\text{max}} - p_{\text{open}}^{\text{min}}}. \tag{5.2}$$

$p_{\text{open}}(c)$ increases monotonically as a function of ligand concentration c, with a minimum value in the absence of ligand given by

$$p_{\text{open}}^{\text{min}} = p_{\text{open}}(0) = \frac{1}{1 + e^{-\beta\epsilon}}, \tag{5.3}$$

and a maximum value in the presence of saturating levels of ligand given as

$$p_{\text{open}}^{\text{max}} = \lim_{c\to\infty} p_{\text{open}}(c) = \frac{1}{1 + e^{-\beta\epsilon}\left(\frac{K_{\text{O}}}{K_{\text{C}}}\right)^m}. \tag{5.4}$$

Using the above two limits, we can investigate four important characteristics of ion channels [10, 11]. First, we examine the *leakiness* of an ion channel, or the fraction of time a channel is open in the absence of ligand, namely,

$$\text{leakiness} = p_{\text{open}}^{\text{min}}. \tag{5.5}$$

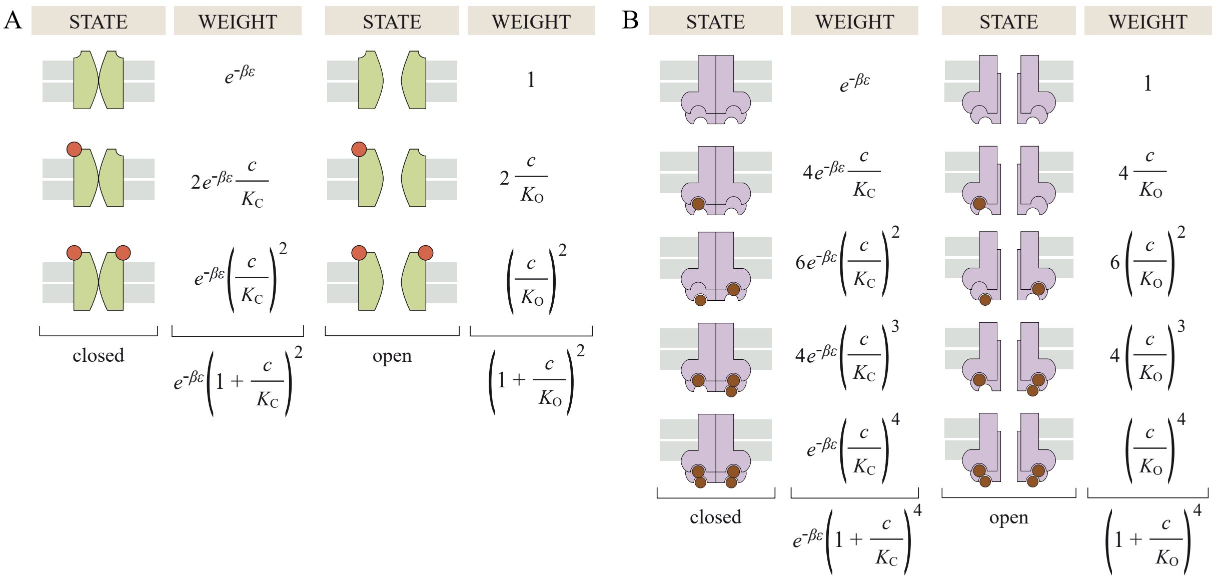

A
STATE
WEIGHT
STATE
WEIGHT
$e^{-\beta\varepsilon}$
1
$2e^{-\beta\varepsilon}\frac{c}{K_C}$
$2\frac{c}{K_O}$
$e^{-\beta\varepsilon}\left(\frac{c}{K_C}\right)^2$
$\left(\frac{c}{K_O}\right)^2$
closed
$e^{-\beta\varepsilon}\left(1+\frac{c}{K_C}\right)^2$
open
$\left(1+\frac{c}{K_O}\right)^2$
B
STATE
WEIGHT
STATE
WEIGHT
$e^{-\beta\varepsilon}$
1
$4e^{-\beta\varepsilon}\frac{c}{K_C}$
$4\frac{c}{K_O}$
$6e^{-\beta\varepsilon}\left(\frac{c}{K_C}\right)^2$
$6\left(\frac{c}{K_O}\right)^2$
$4e^{-\beta\varepsilon}\left(\frac{c}{K_C}\right)^3$
$4\left(\frac{c}{K_O}\right)^3$
$e^{-\beta\varepsilon}\left(\frac{c}{K_C}\right)^4$
$\left(\frac{c}{K_O}\right)^4$
closed
$e^{-\beta\varepsilon}\left(1+\frac{c}{K_C}\right)^4$
open
$\left(1+\frac{c}{K_O}\right)^4$

Figure 5.2: Probability that a ligand-gated ion channel is open as given by the MWC model. (A) Microscopic states and Boltzmann weights of the nAChR ion channel (green) binding to acetylcholine (orange). (B) Corresponding states for the CNGA2 ion channel (purple) binding to cGMP (brown). The behavior of these channels is determined by three physical parameters: the affinity between the receptor and ligand in the open (K_O) and closed (K_C) states and the free energy difference ϵ between the closed and open conformations of the ion channel.

Next we determine the *dynamic range*, or the difference between the probability of the maximally open and maximally closed states of the ion channel, given by

$$\text{dynamic range} = p_{\text{open}}^{\text{max}} - p_{\text{open}}^{\text{min}}. \tag{5.6}$$

Ion channels that minimize leakiness only open upon ligand binding, and ion channels that maximize dynamic range have greater contrast between their open and closed states. Just like $p_{\text{open}}(c)$, leakiness and dynamic range lie within the interval $[0, 1]$.

Two other important characteristics are measured from the normalized current. The *half maximal effective concentration* $[EC_{50}]$ denotes the concentration of ligand at which the normalized current of the ion channel equals ½, namely,

$$p_{\text{open}}([EC_{50}]) = \frac{p_{\text{open}}^{\text{min}} + p_{\text{open}}^{\text{max}}}{2}. \tag{5.7}$$

The *effective Hill coefficient* h equals twice the log-log slope of the normalized current evaluated at $c = [EC_{50}]$,

$$h = 2\frac{d}{d\log c}\log\left(\frac{p_{\text{open}}(c) - p_{\text{open}}^{\text{min}}}{p_{\text{open}}^{\text{max}} - p_{\text{open}}^{\text{min}}}\right)_{c=[EC_{50}]}, \tag{5.8}$$

which reduces to the standard Hill coefficient for the Hill function [14]. The $[EC_{50}]$ determines how the normalized current shifts left and right, while the effective Hill coefficient corresponds to the slope at $[EC_{50}]$. Together, these two properties determine the approximate window of ligand concentrations for which the normalized current transitions from 0 to 1.

In the limit $1 \ll e^{-\beta\epsilon} \ll \left(\frac{K_C}{K_O}\right)^m$, which we show below is relevant for both the nAChR and CNGA2 ion channels, the various functional properties of the channel described above can be approximated to leading order as (see Supporting Information section S5.2):

$$\text{leakiness} \approx e^{\beta\epsilon} \tag{5.9}$$

$$\text{dynamic range} \approx 1 \tag{5.10}$$

$$[EC_{50}] \approx e^{-\beta\epsilon/m} K_O \tag{5.11}$$

$$h \approx m. \tag{5.12}$$

5.4 Results

5.4.1 nAChR Mutants can be Categorized using Free Energy

Muscle-type nAChR is a heteropentamer with subunit stoichiometry $\alpha_2\beta\gamma\delta$, containing two ligand binding sites for acetylcholine at the interface of the α-δ and α-γ subunits [15]. The five homologous subunits have M2 transmembrane domains which move symmetrically during nAChR gating to either occlude or open the ion channel [16]. By introducing a serine in place of the leucine at a key residue (L251S) within the M2 domain present within each subunit, the corresponding subunit is able to more easily transition from the closed to open configuration, shifting the dose-response curve to the left (see Fig. 5.3A) [17]. For example, wild type nAChR is maximally stimulated with 100 μM of acetylcholine, while a mutant ion channel with one L251S mutation is more sensitive and only requires 10 μM to saturate its dose-response curve.

Labarca *et al.* used L251S mutations to create ion channels with n mutated subunits [17]. Fig. 5.3A shows the resulting normalized current for several of these mutants; from right to left the curves represent $n = 0$ (wild type) to $n = 4$ (an ion channel with four of its five subunits mutated). One interesting trend in the data is that each additional mutation shifts the normalized current to the left by approximately one decade in concentration (see Supporting Information section S5.1.2). This constant shift in the dose-response curves motivated Labarca *et al.* to postulate that mutating each subunit increases the gating free energy ϵ by a fixed amount.

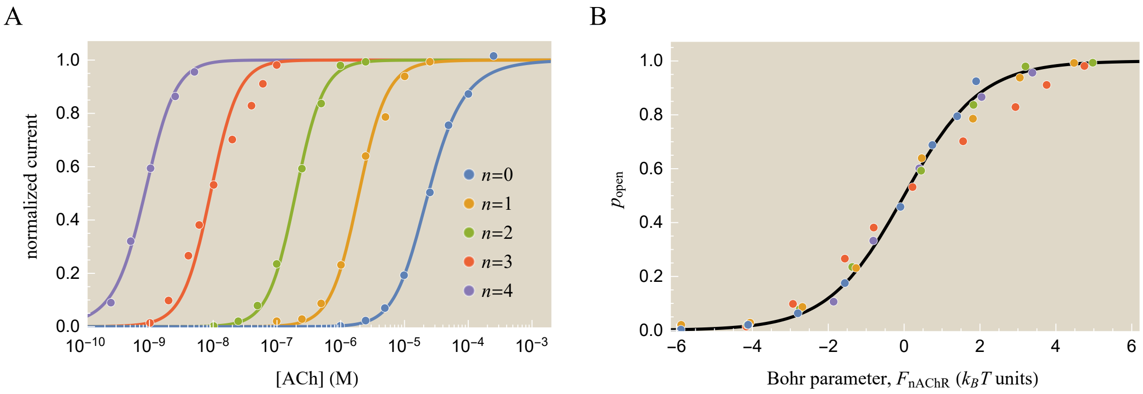


Figure 5.3: Characterizing nicotinic acetylcholine receptors with *n* subunits carrying the L251S mutation. (A) Normalized currents of mutant nAChR ion channels at different concentrations of the agonist acetylcholine (ACh). The curves from right to left show a receptor with $n = 0$ (wild type), $n = 1$ ($\alpha_2\beta\gamma^*\delta$), $n = 2$ ($\alpha_2^*\beta\gamma\delta$), $n = 3$ ($\alpha_2\beta^*\gamma^*\delta^*$), and $n = 4$ ($\alpha_2^*\beta\gamma^*\delta^*$) mutations, where asterisks ($*$) denote a mutated subunit. Fitting the data (solid lines) to Eqs. 5.1 and 5.2 with $m = 2$ ligand binding sites determines the three MWC parameters $K_{\text{O}} = 0.1 \times 10^{-9}$ M, $K_{\text{C}} = 60 \times 10^{-6}$ M, and $\beta\epsilon^{(n)} = [-4.0, -8.5, -14.6, -19.2, -23.7]$ from left ($n = 4$) to right ($n = 0$). With each additional mutation, the dose-response curve shifts to the left by roughly a decade in concentration while the ϵ parameter increases by roughly 5 $k_{\text{B}}T$. (B) The probability $p_{\text{open}}(c)$ that the five ion channels are open can be collapsed onto the same curve using the Bohr parameter $F_{\text{nAChR}}(c)$ given by Eq. 5.13. A positive Bohr parameter indicates that c is above the $[EC_{50}]$. See Supporting Information section S5.3 for details on the fitting procedure.

To test this idea, we analyze the nAChR data at various concentrations c of the ligand acetylcholine using the MWC model Eq. 5.1 with $m = 2$ ligand binding sites. Because the L251S mutation is approximately 4.5 nm from the ligand binding domain [18], we assume that the ligand binding affinities K_{O} and K_{C} are unchanged for the wild type and mutant ion channels, an assumption that has been repeatedly verified by Auerbach *et al.* for nAChR pore mutations [12]. Fig. 5.3A shows the best-fit theoretical curves assuming all five nAChR mutants have the same K_{O} and K_{C} values but that each channel has a distinct gating energy $\epsilon^{(n)}$ (where the superscript n denotes the number of mutated subunits). These gating energies were found to increase by roughly 5 $k_{\text{B}}T$ per n, as would be expected for a mutation that acts equivalently and independently on each subunit.

One beautiful illustration of the power of the MWC model lies in its ability to provide a unified perspective to view data from many different ion channels. Following earlier work in the context of both chemotaxis and quorum sensing [19, 20], we rewrite the probability that the nAChR receptor is open as

$$p_{\text{open}}(c) \equiv \frac{1}{1 + e^{-\beta F(c)}}, \tag{5.13}$$

where this last equation defines the *Bohr parameter* [21]

$$F(c) = -k_{\text{B}}T \log \left(e^{-\beta\epsilon} \frac{\left(1 + \frac{c}{K_{\text{C}}}\right)^m}{\left(1 + \frac{c}{K_{\text{O}}}\right)^m} \right). \tag{5.14}$$

The Bohr parameter quantifies the trade-offs between the physical parameters of the system (in the case of nAChR, between the entropy associated with the ligand concentration c and the gating free energy $\beta\epsilon$). When the Bohr parameters of two ion channels are equal, both channels will elicit the same physiological response. Using Eqs. 5.1 and 5.13 to convert the normalized current data into the probability p_{open} (see Supporting Information section S5.1.3), we can collapse the dose-response data of the five nAChR mutants onto a single master curve as a function of the Bohr parameter for nAChR, $F_{\text{nAChR}}(c)$, as shown in Fig. 5.3B. In this way, the Bohr parameter maps the full complexity of a generic ion channel response into a single combination of the relevant physical parameters of the system.

5.4.1.1 Full Spectrum of nAChR Gating Energy Mutants

We next consider the entire range of nAChR phenotypes achievable by only modifying the gating free energy ϵ of the wild type ion channel. For instance, any combination of nAChR pore mutations would be expected to not affect the ligand dissociation constants and thus yield an ion channel within this class (see Supporting Information section S5.1.4 for one such example). For concreteness, we focus on how the ϵ parameter tunes key features of the dose-response curves, namely the leakiness, dynamic range, $[EC_{50}]$, and effective Hill coefficient h (see Eqs. 5.5-5.12), although we note that other important phenotypic properties such as the intrinsic noise and capacity have also been characterized for the MWC model [10]. Fig. 5.4 shows these four characteristics, with the open squares representing the properties of the five best-fit dose-response curves from Fig. 5.3A.

Fig. 5.4A implies that all of the mutants considered here have negligible leakiness; the probability that the wild type channel ($\beta\epsilon^{(0)} = -23.7$) will be open is less than 10^{-10}. Experimental measurements have shown that such spontaneous openings occur extremely infrequently in nAChR [22], although direct measurement is difficult for such rare events. Other mutational analysis has predicted gating energies around

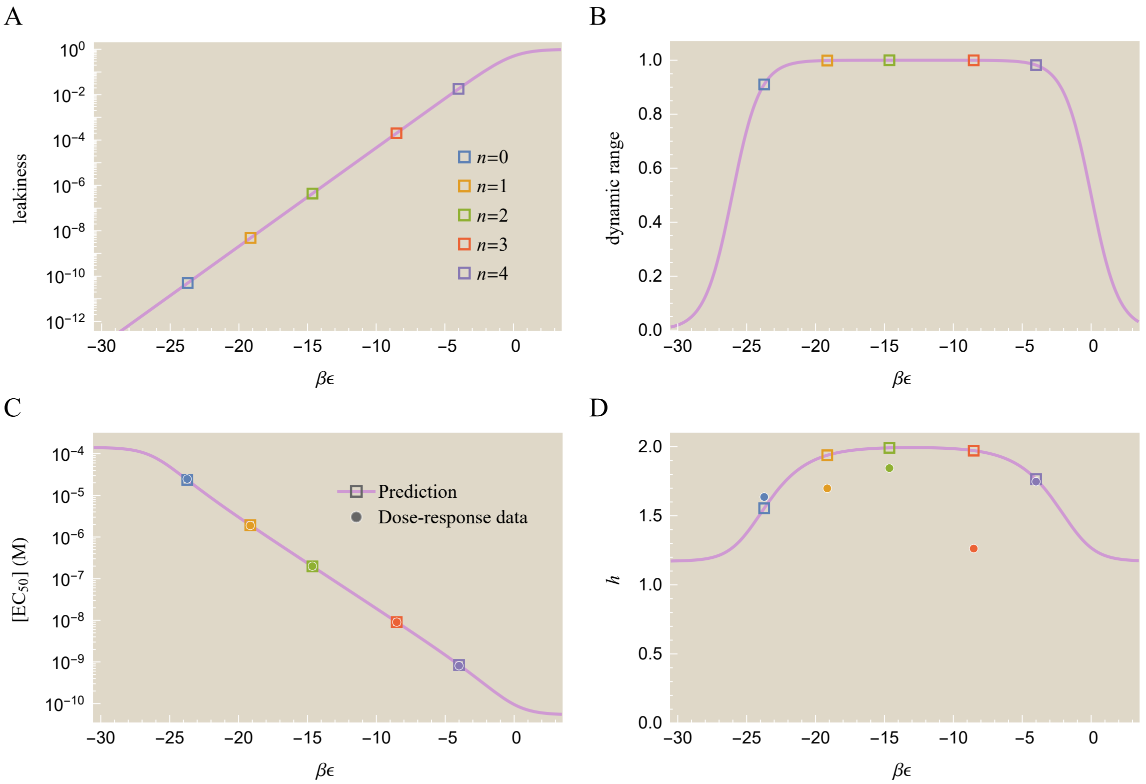


Figure 5.4: Theoretical prediction and experimental measurements for mutant nAChR ion channel characteristics. The open squares mark the $\beta\epsilon$ values of the five dose response curves from Fig. 5.3A. (A) The leakiness given by Eq. 5.5 increases exponentially with each mutation. (B) The dynamic range from Eq. 5.6 is nearly uniform for all mutants. (C) The $[EC_{50}]$ decreases exponentially with each mutation. (D) The effective Hill coefficient h is predicted to remain approximately constant. $[EC_{50}]$ and h offer a direct comparison between the best-fit model predictions (open squares) and the experimental measurements (solid circles) from Fig. 5.3A. While the $[EC_{50}]$ matches well between theory and experiment, the effective Hill coefficient h is significantly noisier.

$\beta\epsilon^{(0)} \approx -14$ (corresponding to a leakiness of 10^{-6}) [12], but we note that such a large wild type gating energy prohibits the five mutants in Fig. 5.3 from being fit as a single class of mutants with the same K_{O} and K_{C} values (see Supporting Information section S5.3.2). If this large wild type gating energy is correct, it may imply that the L251S mutation also affects the K_{O} and K_{C} parameters, though the absence of error bars on the original data make it hard to quantitatively assess the underlying origins of these discrepancies.

Fig. 5.4B asserts that all of the mutant ion channels should have full dynamic range except for the wild type channel, which has a dynamic range of 0.91. In comparison, the measured dynamic range of wild type nAChR is 0.95, close to our predicted value [12]. Note that only when the dynamic range approaches unity does the normalized current become identical to p_{open}; for lower values, information about the leakiness and dynamic range is lost by only measuring normalized currents.

We compare the $[EC_{50}]$ (Fig. 5.4C) and effective Hill coefficient h (Fig. 5.4D) with the nAChR data by interpolating the measurements (see Supporting Information section S5.3.3) in order to precisely determine the midpoint and slope of the response. The $[EC_{50}]$ predictions faithfully match the data over four orders of magnitude. Because each additional mutation lowers the $[EC_{50}]$ by approximately one decade, the analytic form Eq. 5.11 implies that ϵ increases by roughly 5 $k_B T$ per mutation, enabling the ion channel to open more easily. In addition to the L251S mutation considered here, another mutation (L251T) has also been found to shift $[EC_{50}]$ by a constant logarithmic amount (see Supporting Information section S5.1.4) [23]. We also note that many biological systems logarithmically tune their responses by altering the energy difference between two allosteric states, as seen through processes such as phosphorylation and calmodulin binding [24]. This may give rise to an interesting interplay between physiological time scales where such processes occur and evolutionary time scales where traits such as the $[EC_{50}]$ may be accessed via mutations like those considered here [25].

Lastly, the Hill coefficients of the entire class of mutants all lie between 1.5 and 2.0 except for the $n = 3$ mutant whose dose-response curve in Fig. 5.3A is seen to be flatter than the MWC prediction. We also note that if the L251S mutation moderately perturbs the K_O and K_C values, it would permit fits that more finely attune to the specific shape of each mutant's data set. That said, the dose-response curve for the $n = 3$ mutant could easily be shifted by small changes in the measured values, and hence, without recourse to error bars, it is difficult to make definitive statements about the value adopted for h for this mutant.

Note that the simplified expressions Eqs. 5.9-5.12 for the leakiness, dynamic range, $[EC_{50}]$, and effective Hill coefficient apply when $1 \ll e^{-\beta\epsilon} \ll \left(\frac{K_C}{K_O}\right)^m$, which given the values of K_C and K_C for the nAChR mutant class translates to $-22 \lesssim \beta\epsilon \lesssim -5$. The $n = 1$, 2, and 3 mutants all fall within this range, and hence each subsequent mutation exponentially increases their leakiness and exponentially decreases their $[EC_{50}]$, while their dynamic range and effective Hill coefficient remain indifferent to the L251S mutation. The $\beta\epsilon$ parameters of the $n = 0$ and $n = 4$ mutants lie at the edge of the region of validity, so higher order approximations can be used to more precisely fit their functional characteristics (see Supporting Information section S5.2).

5.4.2 Heterooligomeric CNGA2 Mutants can be Categorized using an Expanded MWC Model

The nAChR mutant class discussed above had two equivalent ligand binding sites, and only the gating free energy $\beta\epsilon$ varied for the mutants we considered. In this section, we use beautiful data for the olfactory CNGA2 ion channel to explore the unique phenotypes that emerge from a heterooligomeric ion channel whose subunits have different ligand binding strengths.

The wild type CNGA2 ion channel is made up of four identical subunits, each with one binding site for the cyclic nucleotide ligands cAMP or cGMP [26]. Within the MWC model, the probability that this channel is open is given by Eq. 5.1 with $m = 4$ ligand binding sites (see Fig. 5.2B). Wongsamitkul *et al.* constructed a mutated subunit with lower affinity for ligand and formed tetrameric CNGA2 channels from different combinations of mutated and wild type subunits (see Fig. 5.5) [13]. Since the mutation specifically targeted the ligand binding sites, these mutant subunits were postulated to have new ligand dissociation constants but the same free energy difference $\beta\epsilon$.

We can extend the MWC model to compute the probability p_{open} that these CNGA2 constructs will be open. The states and weights of an ion channel with n mutated subunits (with ligand affinities K_{O}^* and K_{C}^*) and $m-n$ wild type subunits (with ligand affinities K_{O} and K_{C}) is shown in Fig. 5.5, and its probability to be open is given by

$$p_{\text{open}}(c) = \frac{\left(1+\frac{c}{K_{\text{O}}}\right)^{m-n}\left(1+\frac{c}{K_{\text{O}}^*}\right)^{n}}{\left(1+\frac{c}{K_{\text{O}}}\right)^{m-n}\left(1+\frac{c}{K_{\text{O}}^*}\right)^{n}+e^{-\beta\epsilon}\left(1+\frac{c}{K_{\text{C}}}\right)^{m-n}\left(1+\frac{c}{K_{\text{C}}^*}\right)^{n}}. \qquad (5.15)$$

Measurements have confirmed that the dose-response curves of the mutant CNGA2 channels only depend on the total number of mutated subunits n and not on the positions of those subunits (for example both $n = 2$ with adjacent mutant subunits and $n = 2$ with mutant subunits on opposite corners have identical dose-response curves) [13].

Fig. 5.6A shows the normalized current of all five CNGA2 constructs fit to a single set of K_{O}, K_{C}, K_{O}^*, K_{C}^*, and ϵ parameters. Since the mutated subunits have weaker affinity to ligand (leading to the larger dissociation constants $K_{\text{O}}^* > K_{\text{O}}$ and $K_{\text{C}}^* > K_{\text{C}}$), the $[EC_{50}]$ shifts to the right as n increases. As in the case of nAChR, we

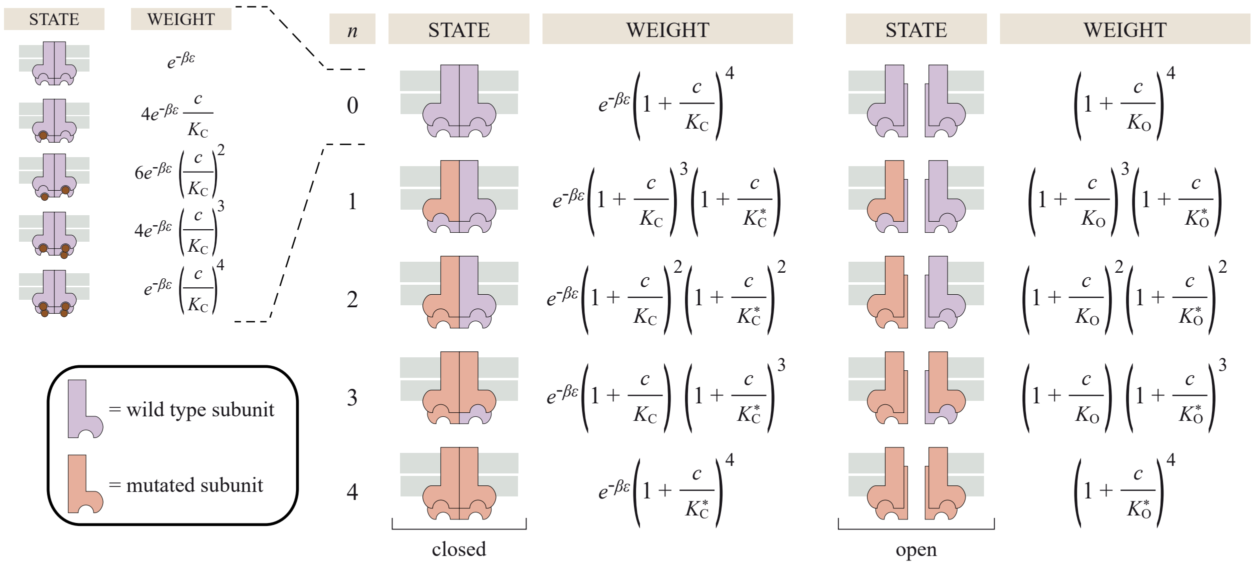

STATE
WEIGHT
$e^{-\beta\varepsilon}$
$4e^{-\beta\varepsilon}\frac{c}{K_C}$
$6e^{-\beta\varepsilon}\left(\frac{c}{K_C}\right)^2$
$4e^{-\beta\varepsilon}\left(\frac{c}{K_C}\right)^3$
$e^{-\beta\varepsilon}\left(\frac{c}{K_C}\right)^4$
= wild type subunit
= mutated subunit
n
STATE
WEIGHT
STATE
WEIGHT
0
$e^{-\beta\varepsilon}\left(1+\frac{c}{K_C}\right)^4$
$\left(1+\frac{c}{K_O}\right)^4$
1
$e^{-\beta\varepsilon}\left(1+\frac{c}{K_C}\right)^3\left(1+\frac{c}{K_C^*}\right)$
$\left(1+\frac{c}{K_O}\right)^3\left(1+\frac{c}{K_O^*}\right)$
2
$e^{-\beta\varepsilon}\left(1+\frac{c}{K_C}\right)^2\left(1+\frac{c}{K_C^*}\right)^2$
$\left(1+\frac{c}{K_O}\right)^2\left(1+\frac{c}{K_O^*}\right)^2$
3
$e^{-\beta\varepsilon}\left(1+\frac{c}{K_C}\right)\left(1+\frac{c}{K_C^*}\right)^3$
$\left(1+\frac{c}{K_O}\right)\left(1+\frac{c}{K_O^*}\right)^3$
4
$e^{-\beta\varepsilon}\left(1+\frac{c}{K_C^*}\right)^4$
$\left(1+\frac{c}{K_O^*}\right)^4$
closed
open

Figure 5.5: States and weights for mutant CNGA2 ion channels. CNGA2 mutants with $m = 4$ subunits were constructed using n mutated (light red) and $m - n$ wild type subunits (purple). The affinity between the wild type subunits to ligand in the open and closed states (K_O and K_C) is stronger than the affinity of the mutated subunits (K_O^* and K_C^*). The weights shown account for all possible ligand configurations, with the inset explicitly showing all of the closed states for the wild type ($n = 0$) ion channel from Fig. 5.2B. The probability that a receptor with n mutated subunits is open is given by its corresponding open state weight divided by the sum of open and closed weights in that same row.

can collapse the data from this family of mutants onto a single master curve using the Bohr parameter $F_{\text{CNGA2}}(c)$ from Eqs. 5.13 and 5.15, as shown in Fig. 5.6B.

Although we analyze the CNGA2 ion channels in equilibrium, we can glimpse the dynamic nature of the system by computing the probability of each channel conformation. Fig. 5.7A shows the ten possible states of the wild type ($n = 0$) channel, the five open states O_j and the five closed states C_j with $0 \leq j \leq 4$ ligands bound. Fig. 5.7B shows how the probabilities of these states are all significantly shifted to the right in the fully mutated ($n = 4$) channel since the mutation diminishes the channel-ligand affinity. The individual state probabilities help determine which of the intermediary states can be ignored when modeling. One extreme simplification that is often made is to consider the Hill limit, where all of the states are ignored save for the closed, unbound ion channel (C_0) and the open, fully bound ion channel (O_4). The drawbacks of such an approximation are two-fold: (1) at intermediate ligand concentrations ($c \in [10^{-7}, 10^{-5}]$ M for $n = 0$ and $c \in [10^{-4}, 10^{-2}]$ M for $n = 4$) the ion channel spends at least 10% of its time in the remaining states which results in

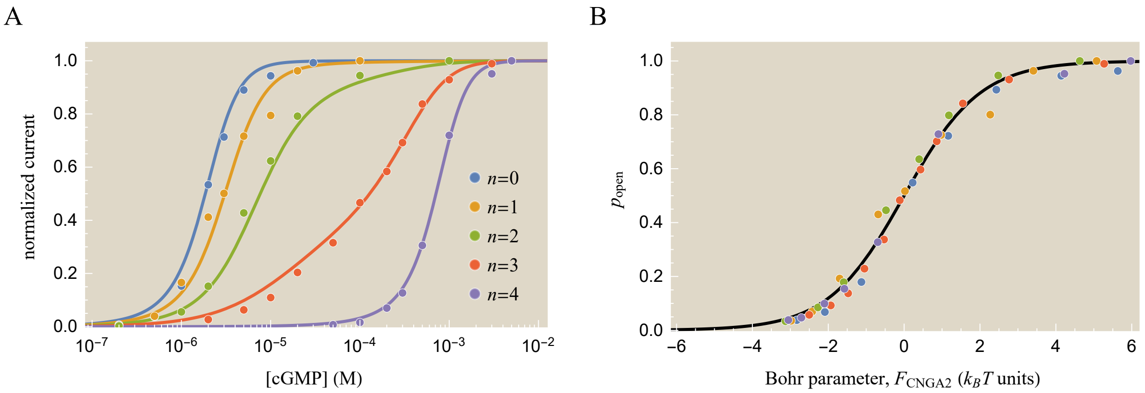


Figure 5.6: Normalized currents for CNGA2 ion channels with a varying number *n* of mutant subunits. (A) Dose-response curves for CNGA2 mutants composed of $4 - n$ wild type subunits and n mutated subunits with weaker affinity for the ligand cGMP. Once the free energy ϵ and the ligand dissociation constants of the wild type subunits (K_{O} and K_{C}) and mutated subunits (K_{O}^* and K_{C}^*) are fixed, each mutant is completely characterized by the number of mutated subunits n in Eq. 5.15. Theoretical best-fit curves are shown using the parameters $K_{\text{O}} = 1.2 \times 10^{-6}$ M, $K_{\text{C}} = 20 \times 10^{-6}$ M, $K_{\text{O}}^* = 500 \times 10^{-6}$ M, $K_{\text{C}}^* = 140 \times 10^{-3}$ M, and $\beta\epsilon = -3.4$. (B) Data from all five mutants collapses onto a single master curve when plotted as a function of the Bohr parameter given by Eq. 5.13. See Supporting Information section S5.3 for details on the fitting.

fundamentally different dynamics than what is predicted by the Hill response and (2) even in the limits such as $c = 0$ and $c \to \infty$ where the C_0 and O_4 states dominate the system, the Hill limit ignores the leakiness and dynamic range of the ion channel (requiring them to exactly equal 0 and 1, respectively), thereby glossing over these important properties of the system.

5.4.2.1 Characterizing CNGA2 Mutants based on Subunit Composition

We now turn to the leakiness, dynamic range, $[EC_{50}]$, and effective Hill coefficient h of a CNGA2 ion channel with n mutated and $m - n$ wild type subunits. Detailed derivations for the following results are provided in Supporting Information section

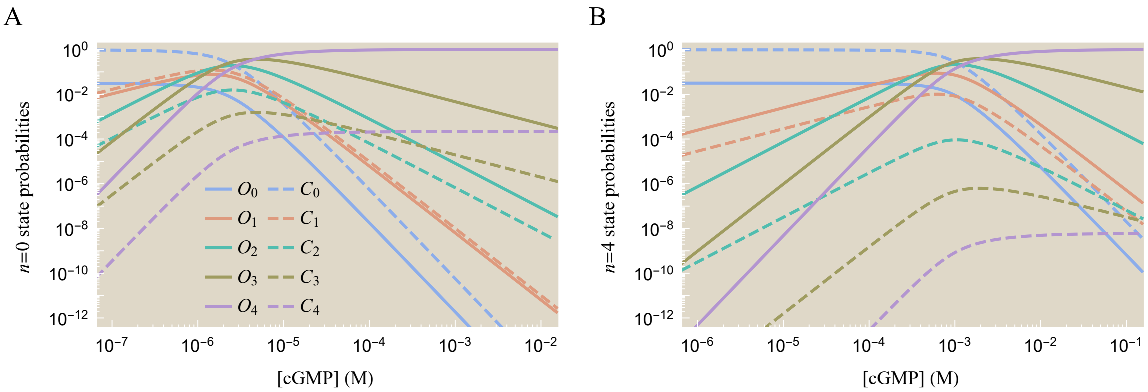


Figure 5.7: Individual state probabilities for the wild type and mutant CNGA2 ion channels. (A) The state probabilities for the wild type ($n = 0$) ion channel. The subscripts of the open (O_j) and closed (C_j) states represent the number of ligands bound to the channel. States with partial occupancy, $1 \leq j \leq 3$, are most likely to occur in a narrow range of ligand concentrations [cGMP] $\in [10^{-7}, 10^{-5}]$ M, outside of which either the completely empty C_0 or fully occupied O_4 states dominate the system. (B) The state probabilities for the $n = 4$ channel. Because the mutant subunits have a weaker affinity to ligand ($K_\text{O}^* > K_\text{O}$ and $K_\text{C}^* > K_\text{C}$), the state probabilities are all shifted to the right.

S5.2:

$$\text{leakiness} = \frac{1}{1 + e^{-\beta\epsilon}} \tag{5.16}$$

$$\text{dynamic range} = \frac{1}{1 + e^{-\beta\epsilon}\left(\frac{K_\text{O}}{K_\text{C}}\right)^{m-n}\left(\frac{K_\text{O}^*}{K_\text{C}^*}\right)^{n}} - \frac{1}{1 + e^{-\beta\epsilon}} \tag{5.17}$$

$$[EC_{50}] \approx \begin{cases} e^{-\beta\epsilon/m} K_\text{O} & n = 0 \\ e^{-2\beta\epsilon/m} \frac{K_\text{O} K_\text{O}^*}{K_\text{O} + K_\text{O}^*} & n = \frac{m}{2} \\ e^{-\beta\epsilon/m} K_\text{O}^* & n = m \end{cases} \tag{5.18}$$

$$h \approx \begin{cases} m & n = 0 \\ \frac{m}{2} & n = \frac{m}{2} \\ m & n = m. \end{cases} \tag{5.19}$$

Note that we recover the original MWC model results Eqs. 5.5-5.12 for the $n = 0$ wild type ion channel. Similarly, the homooligomeric $n = m$ channel is also governed by the MWC model with $K_\text{O} \to K_\text{O}^*$ and $K_\text{C} \to K_\text{C}^*$. We also show the $[EC_{50}]$ and h formulas for the $n = \frac{m}{2}$ case to demonstrate the fundamentally different scaling behavior that this heterooligomeric channel exhibits with the MWC parameters.

Fig. 5.8A shows that all of the CNGA2 mutants have small leakiness, which can be understood from their small ϵ value and Eq. 5.16. In addition, the first term in the dynamic range Eq. 5.17 is approximately 1 because the open state affinities are always smaller than the closed state affinities by at least a factor of ten, which is then raised to the fourth power. Thus, all of the mutants have a large dynamic range as shown in Fig. 5.8B. Experimentally, single channel measurements confirmed that the wild type $n = 0$ channel is nearly always closed in the absence of ligand; in the opposite limit of saturating cGMP, it was found that $p_{\text{open}}^{\text{max}} = 0.99$ for both the $n = 0$ and $n = m$ ion channels (see Supporting Information section S5.3.2) [13].

The $[EC_{50}]$ and effective Hill coefficient h are shown in Fig. 5.8C and D. In contrast to the nAChR case, where each additional mutation decreased $[EC_{50}]$, each CNGA2 mutation tends to increase $[EC_{50}]$, although not by a uniform amount. The effective Hill coefficient has a particularly complex behavior, first decreasing with each of the first three subunit mutations and then finally increasing back to the wild type level for the fully mutated ion channel. To explain this decrease in h for the heterooligomeric channels, we first note that the wild type $n = 0$ channel has a sharp response about its $[EC_{50}] \approx e^{-\beta\epsilon/m} K_{\text{O}}$ while the fully mutated $n = m$ channel has a sharp response about $[EC_{50}] \approx e^{-\beta\epsilon/m} K_{\text{O}}^*$. Roughly speaking, the response of the heterooligomeric channels with $1 \leq n \leq 3$ will occur throughout the full range between $e^{-\beta\epsilon/m} K_{\text{O}}$ and $e^{-\beta\epsilon/m} K_{\text{O}}^*$, which causes the dose-response curves to flatten out and leads to the smaller effective Hill coefficient. Such behavior could influence, for example, the response of the heterooligomeric nAChR ion channel if the two acetylcholine binding pockets diverged to have different ligand affinities.

Although we have focused on the particular mutants created by Wongsamitkul *et al.*, it is straightforward to apply this framework to other types of mutations. For example, in Supporting Information section S5.2.2 we consider the effect of modifying the K_{O} and K_{C} parameters of all four CNGA2 channels simultaneously. This question is relevant for physiological CNGA2 channels where a mutation in the gene would impact all of the subunits in the homooligomer, in contrast to the Wongsamitkul constructs where the four subunits were stitched together within a single gene. We find that when K_{O} and K_{C} vary for all subunits, the leakiness, dynamic range, and effective Hill coefficient remain nearly fixed for all parameter values, and that only the $[EC_{50}]$ scales linearly with K_{O} as per Eq. 5.11. In order to affect the other properties, either the gating energy $\beta\epsilon$ or the number of subunits m would need to be changed.

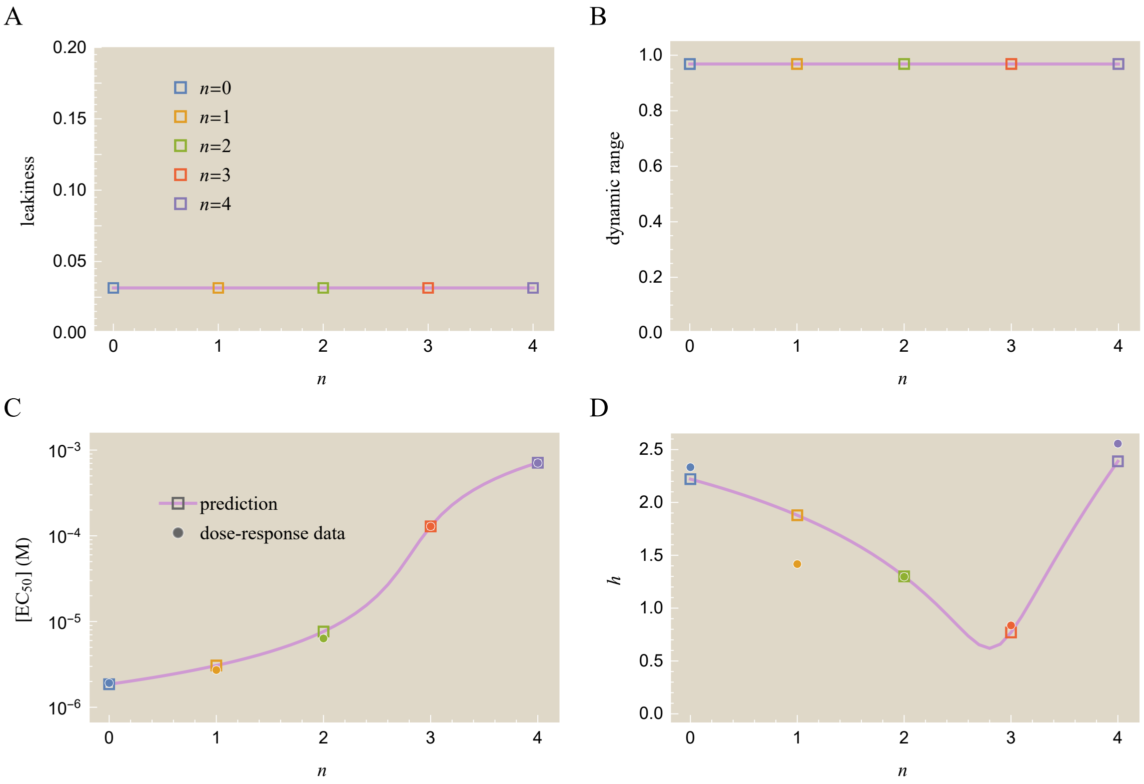


Figure 5.8: Theoretical prediction and experimental measurements for mutant CNGA2 ion channel characteristics. The open squares represent the five mutant ion channels in Fig. 5.6 with n mutated subunits. (A) All ion channels have small leakiness. (B) The dynamic range of all channels is near the maximum possible value of unity, indicating that they rarely open in the absence of ligand and are always open in the presence of saturating ligand concentrations. (C) The $[EC_{50}]$ increases non-uniformly with the number of mutant subunits. Also shown are the measured values (solid circles) interpolated from the data. (D) The effective Hill coefficient has a valley due to the competing influences of the wild type subunits (which respond at μM ligand concentrations) and the mutant subunits (which respond at mM concentrations). Although the homotetrameric channels ($n = 0$ and $n = 4$) both have sharp responses, the combined effect of having both types of subunits ($n = 1$, 2, and 3) leads to a flatter response.

5.4.3 Extrapolating the Behavior of a Class of Mutants

In this section, we explore how constant trends in both the nAChR and CNGA2 data presented above provide an opportunity to characterize the full class of mutants based on the dose-response curves from only a few of its members. Such trends may well be applicable to other ion channel systems, enabling us to theoretically probe a larger space of mutants than what is available from current data.

First, we note that because the $[EC_{50}]$ of the five nAChR mutants fell on a line

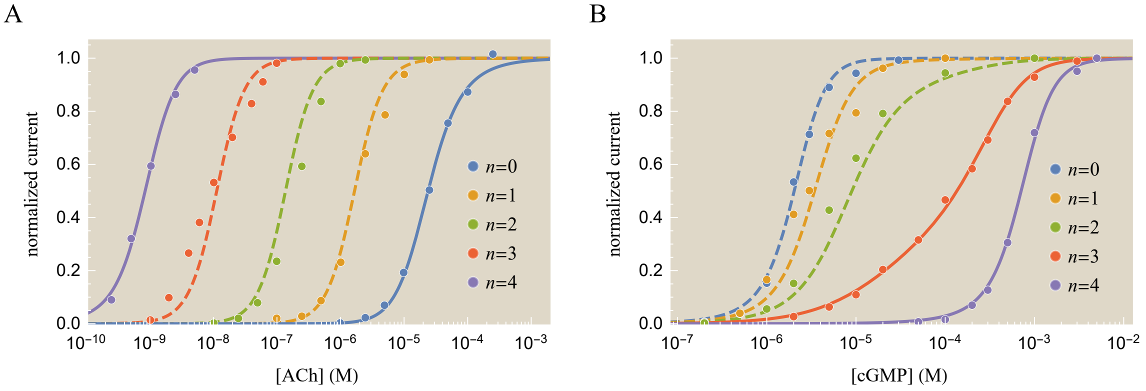


Figure 5.9: Predicting the dose-response of a class of mutants using a subset of its members. (A) The MWC parameters of the nAChR mutants can be fixed using only two data sets (solid lines), which together with Eq. 5.20 predict the dose-response curves of the remaining mutants (dashed lines). (B) For the CNGA2 channel, the properties of both the wild type and mutant subunits can also be fit using two data sets, accurately predicting the responses of the remaining three mutants. Supporting Information section S5.4 demonstrates the results of using alternative pairs of mutants to fix the thermodynamic parameters in both systems.

in Fig. 5.4C, we can predict the response of the entire class of mutants by only considering the dose-response curves of two of its members and extrapolating the behavior of the remaining mutants using linear regression. Experimentally, such a characterization arises because the L251S mutation acts nearly identically and independently across subunits to change the gating free energy of nAChR [12, 17, 23]. This implies that mutating n subunits would yield an ion channel with gating energy

$$\epsilon^{(n)} = \epsilon^{(0)} + n\Delta\epsilon, \tag{5.20}$$

where $\epsilon^{(0)}$ is the wild type gating energy and $\Delta\epsilon$ is the change in free energy per mutation. This functional form is identical to the mismatch model for transcription factor binding, where each additional mutation-induced mismatch adds a constant energy cost to binding [27]. Fig. 5.9A demonstrates how fitting only the $n = 0$ and $n = 4$ dose-response curves (solid lines) accurately predicts the behavior of the $n = 1$, 2, and 3 mutants (dashed lines). In Supporting Information section S5.4, we carry out such predictions using all possible input pairs.

We now turn to the CNGA2 ion channel where, once the K_{O}, K_{C}, K_{O}^*, K_{C}^*, and ϵ parameters are known, the dose-response curve of any mutant can be predicted by varying n in Eq. 5.15. Fig. 5.9B demonstrates that the wild type ion channel ($n = 0$) and the ion channel with only one mutated subunit ($n = 1$) can accurately predict

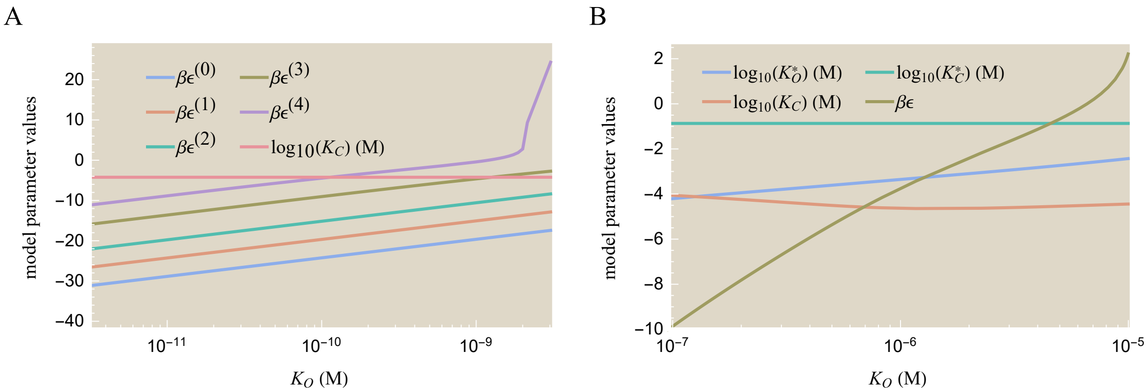


Figure 5.10: Degenerate parameter sets for nAChR and CNGA2 model fitting. Different sets of biophysical parameters can yield the same system response. (A) Data for the nAChR system in Fig. 5.3 is fit by constraining K_O to the value shown on the x-axis. The remaining parameters can compensate for this wide range of K_O values. (B) The CNGA2 system in Fig. 5.6 can similarly be fit by constraining the K_O value, although fit quality decreases markedly outside the narrow range shown. Any set of parameters shown for either system leads to responses with $R^2 > 0.96$.

the dose-response curves of the remaining mutants. Supporting Information section S5.4 explores the resulting predictions using all possible input pairs.

5.4.4 MWC Model allows for Degenerate Parameter Sets

One critical aspect of extracting model parameters from data is that degenerate sets of parameters may yield identical outputs, which suggests that there are fundamental limits to how well a single data set can fix parameter values [25, 28]. This phenomenon, sometimes dubbed "sloppiness," may even be present in models with very few parameters such as the MWC model considered here. Fig. 5.10 demonstrates the relationship between the best-fit parameters within the nAChR and CNGA2 systems. For concreteness, we focus solely on the nAChR system.

After the value of K_O is fixed (to that shown on the x-axis of Fig. 5.10A), the remaining parameters are allowed to freely vary in order to best fit the nAChR data. Although every value of $K_\text{O} \in \left[10^{-11} \text{-} 10^{-9}\right]$ M yields a nearly identical response curve in excellent agreement with the data (with a coefficient of determination $R^2 > 0.96$), we stress that dissociation constants are rarely found in the range $K_\text{O} \ll 10^{-10}$ M. In addition, a dissociation constant above the nM range, $K_\text{O} \gg 10^{-9}$ M, cannot fit the data well and is therefore invalidated by the model. Thus, we may suspect the true parameter values will fall around the interval $K_\text{O} \in \left[10^{-10} \text{-} 10^{-9}\right]$ M for the nAChR system. K_O could ultimately be fixed by measuring the leakiness

Eq. 5.5 (and thereby fixing $\beta\epsilon$) for any of the ion channel mutants.

Two clear patterns emerge from Fig. 5.10A: (1) The value of K_C is approximately constant for all values of K_O and (2) the five free energies all vary as $\beta\epsilon^{(n)} = 2\log\left(K_O\right) + n\,(\text{constant})$. This suggests that K_C and $e^{-\beta\epsilon/2}K_O$ are the fundamental parameter combinations of the system.

We end by noting that the notion of sloppiness, while detrimental to fixing the physical parameter values of a system, nevertheless suggests that multiple evolutionary paths may lead to optimal ion channel phenotypes, providing another mechanism by which allostery promotes a protein's capacity to adapt [29].

5.5 Discussion

There is a deep tension between the great diversity of biological systems and the search for unifying perspectives to help integrate the vast data that has built up around this diversity. Several years ago at the Institut Pasteur, a meeting was convened to celebrate the 50th anniversary of the allostery concept, which was pioneered in a number of wonderful papers in the 1960s and which since then has been applied to numerous biological settings [30–34]. Indeed, that meeting brought together researchers working in physiology, neuroscience, gene regulation, cell motility, and beyond, making it very clear that allostery has great reach as a conceptual framework for thinking about many of the key macromolecules that drive diverse biological processes.

In this paper, we have built upon this significant previous work and explored how the Monod-Wyman-Changeux model can serve as a unifying biophysical framework for whole suites of ion channel mutants (see Figs. 5.3 and 5.6). Specifically, we used two well-studied ligand-gated ion channels to explore the connection between mutations, the MWC parameters, and the full spectrum of dose-response curves which may be induced by those mutations. In addition, we have shown how earlier insights into the nature of "data collapse" in the context of bacterial chemotaxis and quorum sensing [19, 20] can be translated into the realm of ion channels. By introducing the Bohr parameter, we are able to capture the nonlinear combination of thermodynamic parameters which governs the system's response.

For both the nAChR and CNGA2 ion channels, we showed that precise predictions of dose-response curves can be made for an entire class of mutants by only using data from two members of this class (Fig. 5.9). In other words, the information contained in a single dose-response curve goes beyond merely providing data for that specific ion channel. Ultimately, because the total space of all possible mutants is too enormous for any significant fraction to be explored experimentally, we hope that a coupling of theory with experiment will provide a step toward mapping the relation between channel function (phenotype) and the vast space of protein mutations.

Moreover, we used the MWC model to determine analytic formulas for key properties such as the leakiness, dynamic range, $[EC_{50}]$, and the effective Hill coefficient, which together encapsulate much of the information in dose-response curves. These relationships tie into the extensive knowledge about phenotype-genotype maps [27, 29, 35], enabling us to quantify the trade-offs inherent in an ion channel's response. For example, when modifying the ion channel gating free energy, the changes in the leakiness and $[EC_{50}]$ are always negatively correlated (Fig. 5.4), whereas modifying the ligand binding domain will not affect the leakiness but may change the $[EC_{50}]$ (Fig. 5.8 and Supporting Information section S5.2.2). The ability to navigate between the genotype and phenotype of proteins is crucial in many bioengineering settings, where site-directed mutagenesis is routinely employed to find mutant proteins with specific characteristics (e.g. a low leakiness and large dynamic range) [36–38].

While general formulas for these phenotypic properties were elegantly derived in earlier work [10], we have shown that such relations can be significantly simplified in the context of ion channels where $1 \ll e^{-\beta\epsilon} \ll \left(\frac{K_C}{K_O}\right)^m$ (see Eqs. 5.9-5.12). This approximation is applicable for the range of parameters spanned by both the nAChR and CNGA2 systems, and we suspect it may hold for many other ion channels. These formulas provide a simple, intuitive framework to understand the effects of mutations. For example, they suggest the following: (1) Channel pore mutations that increase ϵ will exponentially increase the leakiness of the channel, although the constraint $1 \ll e^{-\beta\epsilon}$ ensures that this leakiness will still be small. Ligand domain mutations are not expected to affect leakiness at all. (2) Channel pore mutations will exponentially decrease the $[EC_{50}]$ with increasing ϵ, although this effect is diminished for ion channels with multiple subunits. For mutations in the ligand binding domain, the $[EC_{50}]$ will increase linearly with the dissociation constant K_O between the ligand and the open ion channel (see Supporting Information section S5.2.2). (3) Neither the dynamic range nor the effective Hill coefficient will be significantly perturbed by either type of mutation. (4) Transforming a homooligomeric ion channel into a heterooligomer can generate a significantly flatter response. For example, even though the CNGA2 channel composed of either all wild type or all mutant subunits had a very sharp response, a channel composed of both subunits had a smaller effective Hill coefficient (see Fig. 5.8D).

The framework presented here could be expanded in several exciting ways. First, it remains to be seen whether channel pore mutations and ligand binding domain mutations are completely independent, or whether there is some cross-talk between them. This question could be probed by creating a channel pore mutant (whose dose-response curves would fix its new $\tilde{\epsilon}$ values), a ligand domain mutant (whose new K_{O}^* and K_{C}^* values would be characterized from its response curve), and then creating the ion channel with both mutations. If these two mutations are independent, the response of the double mutant can be predicted *a priori* using $\tilde{\epsilon}$, K_{O}^*, and K_{C}^*.

We also note that the MWC model discussed here does not consider several important aspects relating to the dynamics of ion channel responses. Of particular importance is the phenomenon of desensitization which significantly modifies an ion channel's response in physiological settings [39, 40]. In addition, some ion channels have multiple open and closed conformations [7, 41, 42] while other channels exhibit slow switching between the channel states [43]. Exploring these additional complexities within generalizations of the MWC model would be of great interest.

Finally, we believe that the time is ripe to construct an explicit biophysical model of fitness to calculate the relative importance of mutation, selection, and drift in dictating the diversity of allosteric proteins such as the ion channels considered here. Such a model would follow in the conceptual footsteps laid in the context of fitness of transcription factors binding [27, 35, 44], protein folding stability [45–47], and influenza evolution [48]. This framework would enable us to make precise, quantitative statements about intriguing trends; for example, nearly all nAChR pore mutations appear to increase a channel's leakiness, suggesting that minimizing leakiness may increase fitness [12]. One could imagine that computing derivatives such as $\frac{d\text{leakiness}}{d\epsilon}$, a quantity analogous to the magnetic susceptibility in physics, would be correlated with how likely an ϵ mutation is to be fixed. The goal of such fitness functions is to map the complexity of the full evolutionary space (i.e. changes to a protein amino acid sequence) onto the MWC parameters and then determine how these parameters evolve in time. In this way, the complexity of sequence and structure would fall onto the very low dimensional space governed by ϵ, K_{O}, and K_{C}.

5.6 Conclusions

We have shown that the Monod-Wyman-Changeux (MWC) model of allostery can quantitatively account for the behavior of broad classes of mutant ligand-gated ion channels using the three biophysical parameters featured in that model: the free energy difference ϵ between the closed and open conformations of the ion channel in the absence of ligand and the dissociation constants K_O and K_C between the ligand and the open and closed ion channel, respectively. By examining nAChR and CNGA2 ion channel mutants, we demonstrated that pore mutations can be quantitatively understood as only affecting the ϵ parameter while mutations in the ligand binding site only alter K_O and K_C (Figs. 5.3 and 5.6).

Building upon these insights, we derived simple analytic approximations for several important properties of an ion channel's response, namely, the leakiness, dynamic range, $[EC_{50}]$, and effective Hill coefficient (Eqs. 5.9-5.12 for nAChR, Eqs. 5.16-5.19 for CNGA2). These formulae are not limited in scope to nAChR and CNGA2, but rather reflect more general properties of ligand-gated ion channels. Utilizing these analytic relations, we quantified the spectrum of possible phenotypes achievable by mutating an ion channel pore or ligand binding domain (Figs. 5.4 and 5.8).

In examining ion channels through the lens of the MWC model, we gained several intriguing insights into the dynamics of channel behavior. First, critical information about the leakiness and dynamic range of a channel's response gets lost when reporting normalized current rather than absolute current (or when only considering the Hill limit). Second, when the physical parameters induced by a series of mutations exhibit a simple trend, the behavior of an entire class of mutants may be extrapolated from only a few of its members (Fig. 5.9). Third, care must be taken when extracting the physical parameters that characterize the MWC model from data, as multiple degenerate parameter sets can often lead to nearly identical response curves (Fig. 5.10). Finally, data from disparate mutants collapse onto a single master curve as a function of the Bohr parameter Eq. 5.14, which represents the combination of physical parameters that completely characterizes an ion channel's response. In such cases, rather than thinking of each mutant as a story unto itself, we can instead categorize and understand mutants through simple one-parameter families.

5.7 Acknowledgements

We thank Richard Aldrich, Stephanie Barnes, Nathan Belliveau, Chico Camargo, Griffin Chure, Vahe Galstyan, Lea Goentoro, Michael Manhart, Chris Miller, Muir Morrison, Manuel Razo-Mejia, Noah Olsman, Allyson Sgro, and Jorge Zañudo for their sharp insights and valuable feedback on this work. We are also grateful to Henry Lester, Klaus Benndorf, and Vasilica Nache for helpful discussions as well as sharing their ion channel data.

References

[1] Hille B. *Ion Channels of Excitable Membranes*. Sinauer Associates, Sunderland, 2001.

[2] Changeux JP. Protein Dynamics and the Allosteric Transitions of Pentameric Receptor Channels. Biophys. Rev. 2014;6(3-4):311–321. doi:10.1007/s12551-014-0149-z.

[3] Kaupp UB, Seifert R. Cyclic Nucleotide-Gated Ion Channels. Physiol. Rev. 2002;82(3):769–824. doi:10.1152/physrev.00008.2002.

[4] Karlin A. On the Application of "A Plausible Model" of Allosteric Proteins to the Receptor for Acetylcholine. J. Theor. Biol. 1967;16(2):306–320. doi:10.1016/0022-5193(67)90011-2.

[5] Goulding EH, Tibbs GR, Siegelbaum SA. Molecular Mechanism of Cyclic-Nucleotide-Gated Channel Activation. Nature. 1994;372(6504):369–374. doi:10.1038/372369a0.

[6] Changeux JP, Edelstein SJ. *Nicotinic Acetylcholine Receptors: From Molecular Biology to Cognition.* Odile Jacob, New York, 2005.

[7] Ruiz M, Karpen JW. Single Cyclic Nucleotide-Gated Channels Locked in Different Ligand-Bound States. Nature. 1997;389(6649):389–392. doi:10.1038/38744.

[8] Biskup C, Kusch J, Schulz E, Nache V, Schwede F, Lehmann F, et al. Relating Ligand Binding to Activation Gating in CNGA2 Channels. Nature. 2007;446(7134):440–443. doi:10.1038/nature05596.

[9] Purohit P, Auerbach A. Unliganded Gating of Acetylcholine Receptor Channels. Proc. Natl. Acad. Sci. USA. 2009;106(1):115–120. doi:10.1073/Proc.Natl.Acad.Sci.USA.0809272106.

[10] Martins BMC, Swain PS. Trade-Offs and Constraints in Allosteric Sensing. PLoS Comput. Biol. 2011;7(11):1–13. doi:10.1371/journal.pcbi.1002261.

[11] Marzen S, Garcia HG, Phillips R. Statistical Mechanics of Monod-Wyman-Changeux (MWC) Models. J. Mol. Biol. 2013;425(9):1433–1460. doi:10.1016/j.jmb.2013.03.013.

[12] Auerbach A. Thinking in Cycles: MWC Is a Good Model for Acetylcholine Receptor-Channels. J. Physiol. 2012;590(1):93–98. doi:10.1113/jphysiol.2011.214684.

[13] Wongsamitkul N, Nache V, Eick T, Hummert S, Schulz E, Schmauder R, et al. Quantifying the Cooperative Subunit Action in a Multimeric Membrane Receptor. Sci. Rep. 2016;6(1):20974. doi:10.1038/srep20974.

[14] Hill TL. *Cooperativity Theory in Biochem: Steady-State and Equilibrium Systems*. Springer-Verlag, New York, 1985.

[15] Changeux JP, Edelstein SJ. Allosteric Receptors After 30 Years. Neuron. 1998;21(5):959–980. doi:10.1016/S0896-6273(00)80616-9.

[16] Unwin N. Acetylcholine Receptor Channel Imaged in the Open State. Nature. 1995;373(6509):37–43. doi:10.1038/373037a0.

[17] Labarca C, Nowak MW, Zhang H, Tang L, Deshpande P, Lester HA. Channel Gating Governed Symmetrically by Conserved Leucine Residues in the M2 Domain of Nicotinic Receptors. Nature. 1995;376(6540):514–516. doi:10.1038/376514a0.

[18] Unwin N. Nicotinic Acetylcholine Receptor at 9A Resolution. J. Mol. Biol. 1993;229(4):1101–24.

[19] Keymer JE, Endres RG, Skoge M, Meir Y, Wingreen NS. Chemosensing in *Escherichia coli*: Two regimes of two-state receptors. Proc. Natl. Acad. Sci. USA. 2006;103(6):1786–1791. doi:10.1073/Proc.Natl.Acad.Sci.USA.0507438103.

[20] Swem LR, Swem DL, Wingreen NS, Bassler BL. Deducing Receptor Signaling Parameters from In Vivo Analysis: LuxN/AI-1 Quorum Sensing in Vibrio Harveyi. Cell. 2008;134(3):461–473. doi:10.1016/j.cell.2008.06.023.

[21] Phillips R. Napoleon Is in Equilibrium. Annu. Rev. Condens. Matter Phys. 2015;6(1):85–111. doi:10.1146/annurev-conmatphys-031214-014558.

[22] Nayak TK, Purohit PG, Auerbach A. The Intrinsic Energy of the Gating Isomerization of a Neuromuscular Acetylcholine Receptor Channel. J. Gen. Physiol. 2012;139(5):349–58. doi:10.1085/jgp.201110752.

[23] Filatov GN, White MM. The Role of Conserved Leucines in the M2 Domain of the Acetylcholine Receptor in Channel Gating. Mol. Pharmacol. 1995;48(3):379–84.

[24] Olsman N, Goentoro L. Allosteric Proteins as Logarithmic Sensors. Proc. Natl. Acad. Sci. USA. 2016;113(30):E4423–30. doi:10.1073/Proc.Natl.Acad. Sci.USA.1601791113.

[25] Milo R, Hou JH, Springer M, Brenner MP, Kirschner MW. The Relationship Between Evolutionary and Physiological Variation in Hemoglobin. Proc. Natl. Acad. Sci. USA. 2007;104(43):16998–17003. doi:10.1073/Proc.Natl.Acad. Sci.USA.0707673104.

[26] Kusch J, Zifarelli G. Patch-Clamp Fluorometry: Electrophysiology Meets Fluorescence. Biophys. J. 2014;106(6):1250–7. doi:10.1016/j.bpj.2014.02.006.

[27] Berg J, Willmann S, Lässig M. Adaptive Evolution of Transcription Factor Binding Sites. BMC Evol. Biol. 2004;4(1):42. doi:10.1186/1471-2148-4-42.

[28] Transtrum MK, Machta BB, Brown KS, Daniels BC, Myers CR, Sethna JP. Perspective: Sloppiness and Emergent Theories in Physics, Biology, and Beyond. J. Chem. Phys. 2015;143(1):010901. doi:10.1063/1.4923066.

[29] Raman AS, White KI, Ranganathan R. Origins of Allostery and Evolvability in Proteins: A Case Study. Cell. 2016;166(2):468–80. doi:10.1016/j.cell.2016.05.047.

[30] Monod J, Changeux JP, Jacob F. Allosteric Proteins and Cellular Control Systems. J. Mol. Biol. 1963;6:306–329.

[31] Monod J, Wyman J, Changeux JP. On the Nature of Allosteric Transitions: A Plausible Model. J. Mol. Biol. 1965;12:88–118. doi:10.1016/S0022-2836(65)80285-6.

[32] Koshland DE, Némethy G, Filmer D. Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits. Biochem. 1966; 5(1):365–385. doi:10.1021/bi00865a047.

[33] Changeux JP. Allostery and the Monod-Wyman-Changeux Model After 50 Years. Annu. Rev. Biophys. 2012;41:103–33. doi:10.1146/annurev-biophys-050511-102222.

[34] Gerhart J. From Feedback Inhibition to Allostery: The Enduring Example of Aspartate Transcarbamoylase. FEBS J. 2014;281(2):612–620. doi:10.1111/febs.12483.

[35] Gerland U, Hwa T. On the Selection and Evolution of Regulatory DNA Motifs. J. Mol. Evol. 2002;55(4):386–400. doi:10.1007/s00239-002-2335-z.

[36] Flytzanis NC, Bedbrook CN, Chiu H, Engqvist MKM, Xiao C, Chan KY, et al. Archaerhodopsin Variants with Enhanced Voltage-Sensitive Fluorescence in Mammalian and Caenorhabditis Elegans Neurons. Nat. Commun. 2014; 5:4894. doi:10.1038/ncomms5894.

[37] Vogt A, Guo Y, Tsunoda SP, Kateriya S, Elstner M, Hegemann P. Conversion of a Light-Driven Proton Pump into a Light-Gated Ion Channel. Sci. Rep. 2015;5:16450. doi:10.1038/srep16450.

[38] Wietek J, Wiegert JS, Adeishvili N, Schneider F, Watanabe H, Tsunoda SP, et al. Conversion of Channelrhodopsin into a Light-Gated Chloride Channel. Science. 2014;344(6182):409–412.

[39] Edelstein SJ, Schaad O, Henry E, Bertrand D, Changeux JP. A Kinetic Mechanism for Nicotinic Acetylcholine Receptors Based on Multiple Allosteric Transitions. Biol. Cybern. 1996;75(5):361–79.

[40] Plested AJR. Structural Mechanisms of Activation and Desensitization in Neurotransmitter-Gated Ion Channels. Nature Structural & Molecular Biology. 2016;23(6):494–502. doi:10.1038/nsmb.3214.

[41] VanSchouwen B, Melacini G. Cracking the Allosteric Code of NMR Chemical Shifts. Proc. Natl. Acad. Sci. USA. 2016;113(34):9407–9409. doi:10.1073/Proc.Natl.Acad.Sci.USA.1611068113.

[42] Jackson MB. Kinetics of Unliganded Acetylcholine Receptor Channel Gating. Biophys. J. 1986;49(3):663–72. doi:10.1016/S0006-3495(86)83693-1.

[43] Bicknell BA, Goodhill GJ. Emergence of Ion Channel Modal Gating from Independent Subunit Kinetics. Proc. Natl. Acad. Sci. USA. 2016;113(36):E5288–97. doi:10.1073/Proc.Natl.Acad.Sci.USA.1604090113.

[44] Lynch M, Hagner K. Evolutionary Meandering of Intermolecular Interactions Along the Drift Barrier. Proc. Natl. Acad. Sci. USA. 2015;112(1):E30–8. doi:10.1073/Proc.Natl.Acad.Sci.USA.1421641112.

[45] Zeldovich KB, Shakhnovich EI. Understanding Protein Evolution: From Protein Physics to Darwinian Selection. Annu. Rev. Phys. Chem. 2008; 59(1):105–127. doi:10.1146/annurev.physchem.58.032806.104449.

[46] Serohijos AW, Shakhnovich EI. Merging Molecular Mechanism and Evolution: Theory and Computation at the Interface of Biophysics and Evolutionary Population Genetics. Curr. Opin. Struct. Biol. 2014;26:84–91. doi:10.1016/j.sbi.2014.05.005.

[47] Peleg O, Choi JM, Shakhnovich EI. Evolution of Specificity in Protein-Protein Interactions. Biophys. J. 2014;107(7):1686–1696. doi:10.1016/j.bpj.2014.08.004.

[48] Łuksza M, Lässig M. A Predictive Fitness Model for Influenza. Nature. 2014; 507(7490):57–61. doi:10.1038/nature13087.

Chapter S5

SUPPLEMENTARY INFORMATION FOR MONOD-WYMAN-CHANGEUX ANALYSIS OF LIGAND-GATED ION CHANNEL MUTANTS

S5.1 Additional Ion Channel Data

In this section, we explore some of the additional experimental measurements available for the nAChR and CNGA2 systems studied above and elaborate on several calculations mentioned in the text. In S5.1.1, we analyze the time scale required for an ion channel to reach equilibrium. In S5.1.2, we present data on additional L251S nAChR mutants. Using these mutants, we examine the approximation made in the text that only the *total* number of mutations, and not the identity of the subunits mutated, influences the resulting nAChR mutant behavior. In S5.1.3, we examine $p_{\text{open}}(c)$ for the classes of ion channels considered in the text and comment on how this probability differs from the normalized current. In S5.1.4, we examine data from a similar class of L251T mutations and show that their qualitative behavior is similar to the L251S mutants. In S5.1.5, we discuss measurements of combinations of CNGA2 ion channels.

S5.1.1 Dynamics Towards Equilibrium

$$\begin{tikzcd}
O \arrow[r, shift left, "k_{\text{on}}[L]"] \arrow[d, shift right, "k_{-}^{e}"'] & OL \arrow[l, shift left, "k_{\text{off}}^{(O)}"] \arrow[d, shift right, "k_{-}^{o}"'] \\
C \arrow[r, shift left, "k_{\text{on}}[L]"] \arrow[u, shift right, "k_{+}^{e}"'] & CL \arrow[l, shift left, "k_{\text{off}}^{(C)}"] \arrow[u, shift right, "k_{+}^{o}"']
\end{tikzcd}$$

Figure S5.1: Rates for an ion channel with one ligand binding site. The ion channel tends to transition from the closed (C) state to the open (O) state after binding to ligand (L). We assume both ion channel states have the same diffusion-limited on-rate $k_{\text{on}} = 10^9 \, \frac{1}{\text{M}\cdot\text{s}}$. The remaining rates of the bound states should satisfy $k_{\text{off}}^{(C)} > k_{\text{off}}^{(O)}$ and $k_{+}^{o} > k_{-}^{o}$ so that ligand binding drives the ion channel to the open state OL.

In this section, we derive an exact expression for the time constant for which an ion channel with one ligand binding site will come to equilibrium. This analysis can be readily extended numerically to include multiple ligand binding sites.

Fig. S5.1 shows the rates between the four possible ion channel states: the unbound open (O) and closed (C) states as well as the bound open (OL) and closed (CL) states. We assume that there is a sufficient ligand $[L]$ in the system so that when the ligand binds to the ion channels its concentration does not appreciably diminish. Hence the rate equations for the system can be written in matrix form (with bold denoting vectors and matrices) as

$$\frac{d\boldsymbol{E}}{dt} = \boldsymbol{K}\boldsymbol{E} \tag{S5.1}$$

where the right hand side represents the product of the transition matrix

$$\boldsymbol{K} = \begin{pmatrix} -(k_+^e + k_{\text{on}}[L]) & k_{\text{off}}^{(C)} & k_-^e & 0 \\ k_{\text{on}}[L] & -(k_+^o + k_{\text{off}}^{(C)}) & 0 & k_-^o \\ k_+^e & 0 & -(k_-^e + k_{\text{on}}[L]) & k_{\text{off}}^{(O)} \\ 0 & k_+^o & k_{\text{on}}[L] & -(k_-^o + k_{\text{off}}^{(O)}) \end{pmatrix} \tag{S5.2}$$

and the vector representing the occupancy of each ion channel state

$$\boldsymbol{E} = \begin{pmatrix} [C] \\ [CL] \\ [O] \\ [OL] \end{pmatrix}. \tag{S5.3}$$

The matrix $\boldsymbol{K}$ can be decomposed as

$$\boldsymbol{K} = \boldsymbol{V}^{-1}\boldsymbol{\Lambda}\boldsymbol{V} \tag{S5.4}$$

where $\boldsymbol{V}$'s columns are the eigenvectors of $\boldsymbol{K}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix whose entries are the eigenvalues of $\boldsymbol{K}$. In general, it is known that the eigenvalues of such a matrix $\boldsymbol{K}$ representing the dynamics of any graph such as Fig. S5.1 has one eigenvalue that is 0 while the remaining eigenvalues are non-zero and have negative real parts [1]. (Indeed, because all of the columns of $\boldsymbol{K}$ add up to zero, $\boldsymbol{K}$ is not full rank and hence one of its eigenvalues must be zero.) Defining the vector

$$\tilde{\boldsymbol{E}} \equiv \boldsymbol{V}\boldsymbol{E} = \begin{pmatrix} \tilde{E}_1 \\ \tilde{E}_2 \\ \tilde{E}_3 \\ \tilde{E}_4 \end{pmatrix}, \tag{S5.5}$$

Eq. S5.1 can be rewritten as

$$\frac{d\tilde{\boldsymbol{E}}}{dt} = \boldsymbol{\Lambda}\tilde{\boldsymbol{E}}. \tag{S5.6}$$

If the eigenvalues of Λ are λ_1, λ_2, λ_3, and 0, then $\tilde{E}_j = c_j e^{\lambda_j t}$ for $j = 1, 2, 3$ and $\tilde{E}_4 = c_4$ where the c_j's are constants determined by initial conditions. Since the $\tilde{E}_j$'s are linear combinations of $[C]$, $[CS]$, $[O]$, and $[OS]$, this implies that the $-\frac{1}{\lambda_1}$,$-\frac{1}{\lambda_2}$, and $-\frac{1}{\lambda_3}$ (or $-\frac{1}{\Re(\lambda_j)}$ if the eigenvalues are complex) are the time scales for the normal modes of the system to come to equilibrium, with the largest value representing the time scale τ for the entire system to reach equilibrium,

$$\tau = \max\left(-\frac{1}{\lambda_1}, -\frac{1}{\lambda_2}, -\frac{1}{\lambda_3}\right). \tag{S5.7}$$

Although the eigenvalues of this matrix can be calculated in closed form, as roots of a cubic function, the full expression is complicated. Instead, we write the Taylor expansion of λ_1, λ_2, and λ_3 in the limit $k_+^o \to \infty$, since we suspect that the transition from $CS \to OS$ is extremely fast. In this limit, the λ_js take the forms

$$\lambda_1 = -(k_{\text{off}}^{(O)} + k_{\text{on}}[L]) + O\left(\frac{1}{k_+^o}\right) \tag{S5.8}$$

$$\lambda_2 = -(k_{\text{off}}^{(C)} + k_-^o + k_+^o) + O\left(\frac{1}{k_+^o}\right) \tag{S5.9}$$

$$\lambda_3 = -(k_{\text{on}}[L] + k_-^e + k_+^e) + O\left(\frac{1}{k_+^o}\right). \tag{S5.10}$$

Fig. S5.2 shows an example of how the system attains its equilibrium starting from a random initial condition. The exact time scale Eq. S5.7 using the matrix eigenvalues leads to $\tau = 1.1 \times 10^{-3}$ s, which is very close to the approximation using Eqs. S5.8-S5.10 which yields $\tau^{(approx)} = 1.0 \times 10^{-3}$ s. The exact time scale is shown in Fig. S5.2 as a dashed line, and states achieve near total equilibrium by $t = 10^{-2}$ s.

As a point of reference for this time scale described above for the system to come to equilibrium, there are two other relevant times scales for an ion channel: (1) the time scale for an ion channel to switch between the open and closed conformations and (2) the time scale for an ion channel to stay in its open conformation before switching to the closed conformation. The former occurs on the microsecond scale for nAChR [2], while the latter occurs on the millisecond scale [3, 4]. Thus, the time to transition between the closed and open conformations can be ignored, and the system reaches equilibrium after only a few transitions between the open and closed states.

Lastly, we compute the fractional occupancy of the four states ion channel states in steady state, $\frac{dE}{dt} = \mathbf{0}$. We first make the standard assumption that the system is not

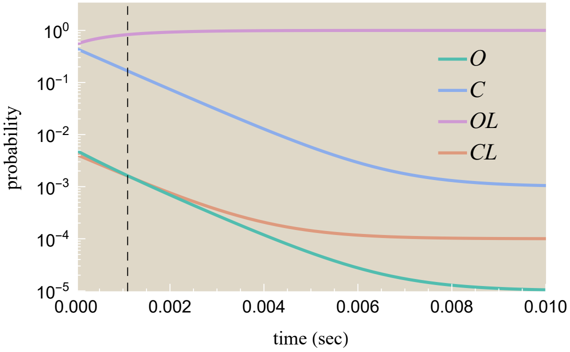


Figure S5.2: Kinetics of a system heading towards equilibrium. The relative probabilities of the four states are computed using Eqs. S5.1 and S5.2 and the rate constants $k_{\text{on}}[L] = 10^3 \frac{1}{\text{s}}$, $k_{\text{off}}^{(O)} = 10^{-2} \frac{1}{\text{s}}$, $k_{\text{off}}^{(C)} = 10^4 \frac{1}{\text{s}}$, $k_+^o = 10^4 \frac{1}{\text{s}}$, $k_-^o = 10 \frac{1}{\text{s}}$, $k_+^e = 10 \frac{1}{\text{s}}$, and $k_-^e = 10^3 \frac{1}{\text{s}}$. Note that the rate constants must satisfy the cycle condition: the product of rates moving clockwise equals the product of rates going counterclockwise. The dashed line indicates the exact time scale Eq. S5.7 for the system to reach equilibrium. Initial conditions were chosen randomly as $p_O = 0.005$, $p_C = 0.45$, $p_{OL} = 0.54$, and $p_{CL} = 0.005$.

expending energy to drive a cyclic flux in the system. Formally, this implies that the rate constants satisfy the cycle condition: the product of rates moving clockwise in Fig. S5.1 equals the product of rates going counterclockwise [5],

$$k_{\text{on}}[L] k_-^o k_{\text{off}}^{(C)} k_+^e = k_-^e k_{\text{off}}^{(O)} k_{\text{on}}[L] k_+^o. \tag{S5.11}$$

With this condition, the fractional occupancy of each state is given by

$$[C] = \frac{\frac{k_-^e}{k_+^e}}{\left(1 + \frac{k_{\text{on}}[L]}{k_{\text{off}}^{(O)}}\right) + \frac{k_-^e}{k_+^e}\left(1 + \frac{k_{\text{on}}[L]}{k_{\text{off}}^{(C)}}\right)} \tag{S5.12}$$

$$[CL] = \frac{\frac{k_-^e}{k_+^e} \frac{k_{\text{on}}[L]}{k_{\text{off}}^{(C)}}}{\left(1 + \frac{k_{\text{on}}[L]}{k_{\text{off}}^{(O)}}\right) + \frac{k_-^e}{k_+^e}\left(1 + \frac{k_{\text{on}}[L]}{k_{\text{off}}^{(C)}}\right)} \tag{S5.13}$$

$$[O] = \frac{1}{\left(1 + \frac{k_{\text{on}}[L]}{k_{\text{off}}^{(O)}}\right) + \frac{k_-^e}{k_+^e}\left(1 + \frac{k_{\text{on}}[L]}{k_{\text{off}}^{(C)}}\right)} \tag{S5.14}$$

$$[OL] = \frac{\frac{k_{\text{on}}[L]}{k_{\text{off}}^{(O)}}}{\left(1 + \frac{k_{\text{on}}[L]}{k_{\text{off}}^{(O)}}\right) + \frac{k_-^e}{k_+^e}\left(1 + \frac{k_{\text{on}}[L]}{k_{\text{off}}^{(C)}}\right)}, \tag{S5.15}$$

A system in steady state which satisfies the cycle condition must necessarily be in thermodynamic equilibrium [6], which implies that these fractional occupancies must be identical to the result derived from the Boltzmann distribution (see Fig. 5.2). And indeed, this correspondence is made explicit if we define

$$e^{-\beta\epsilon} = \frac{k_-^e}{k_+^e} \tag{S5.16}$$

$$K_{\text{O}} = \frac{k_{\text{off}}^{(O)}}{k_{\text{on}}} \tag{S5.17}$$

$$K_{\text{C}} = \frac{k_{\text{off}}^{(C)}}{k_{\text{on}}}. \tag{S5.18}$$

In this way, the MWC parameters can be defined through the ratios of the rate parameters of the system.

S5.1.2 Additional nAChR Mutants

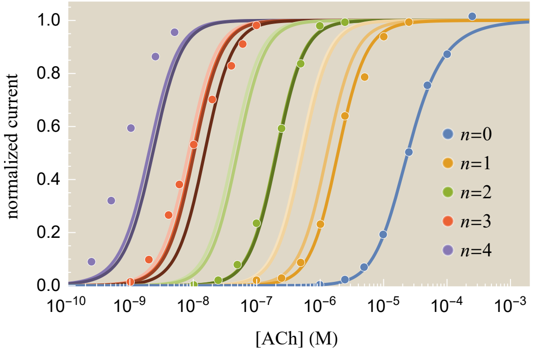


Figure S5.3: Categorizing the full set of ion channel mutants. Using the best-fit K_{O} and K_{C} values obtained from the five mutants in Fig. 5.3, we can use the measured $[EC_{50}]$ value for each mutant in Table S5.1 to determine its $\beta\epsilon$ parameter. Thus, a single data point for each mutant enables us to predict its complete dose-response curve. All mutants with the same total number n of mutations are plotted in shades of the same color, together with the complete dose-response curves from Fig. 5.3. Note that while each mutant family spans a range of $[EC_{50}]$ values, the classes are distinct and do not overlap.

In addition to the five constructs shown in Fig. 5.3, namely $n = 0$ (wild type), $n = 1$ ($\alpha_2\beta\gamma^*\delta$), $n = 2$ ($\alpha_2^*\beta\gamma\delta$), $n = 3$ ($\alpha_2\beta^*\gamma^*\delta^*$), and $n = 4$ ($\alpha_2^*\beta\gamma^*\delta^*$), Labarca *et al.* constructed multiple other ion channel mutants listed in Table S5.1 [3]. While complete dose-response curves are not available for these other constructs, their $[EC_{50}]$ values were measured. Using the K_{O} and K_{C} values for this entire class of mutants given in Fig. 5.3, we can use the $[EC_{50}]$ measurements to fit the $\beta\epsilon$ value of each mutant, thereby providing us with a complete description of each mutant.

In particular, we can predict the dose-response curves of each of these mutants, as shown in Fig. S5.3. We overlay the data from Fig. 5.3 on top of these theoretical curves, where mutants with the same total number n of mutations are drawn as shades of the same color. Note that there was some error in the original measurements, since the reported $[EC_{50}]$ value for the $n = 4$ ($\alpha_2^*\beta\gamma^*\delta^*$) mutant shown in purple dots in Fig. S5.3 should clearly be less than 10^{-9} M, even though it was given as $(2.0 \pm 0.6) \times 10^{-9}$ M in Ref. 3.

n	subunits	$[EC_{50}]$ (nM)
0	$\alpha_2\beta\gamma\delta$	24,010
1	$\alpha\alpha^*\beta\gamma\delta$	1,290
	$\alpha_2\beta^*\gamma\delta$	531
	$\alpha_2\beta\gamma^*\delta$	1,910
	$\alpha_2\beta\gamma\delta^*$	486
2	$\alpha_2^*\beta\gamma\delta$	202
	$\alpha_2\beta^*\gamma^*\delta$	49.7
	$\alpha_2\beta^*\gamma\delta^*$	208 ± 69
	$\alpha_2\beta\gamma^*\delta^*$	42.7
3	$\alpha_2^*\beta^*\gamma\delta$	10.3
	$\alpha_2^*\beta\gamma^*\delta$	15.1
	$\alpha_2^*\beta\gamma\delta^*$	8.4 ± 1.3
	$\alpha_2\beta^*\gamma^*\delta^*$	9.8 ± 1.3
4	$\alpha_2^*\beta^*\gamma\delta^*$	2.3
	$\alpha_2^*\beta\gamma^*\delta^*$	2.0 ± 0.6
5	$\alpha_2^*\beta^*\gamma^*\delta^*$	< 1

Table S5.1: Dose-response relations for mouse muscle ACh receptors containing various numbers of mutated L251S subunits (*n*). Mutated subunits are indicated by an asterisk ($*$). Standard error of the mean for $[EC_{50}]$ was less than 10% of the mean, except where given. Responses for the $\alpha_2^*\beta^*\gamma^*\delta^*$ mutant were too small for reliable measurements.

Fig. S5.3 demonstrates that not all subunit mutations cause a tenfold decrease in $[EC_{50}]$, but rather that there is a small spread in $[EC_{50}]$ depending on precisely which subunit was mutated. This variation is not unreasonable given that $\alpha_2\beta\gamma\delta$ nAChR is a heteropentamer. Indeed, such subunit-dependent spreading in $[EC_{50}]$ values has also been seen in other heteromeric ion channels [7, 8] but is absent within homomeric ion channels such as the CNGA2 ion channel explored in the text [4].

To explore this subunit-dependent shift in the dose-response curves, we now relax the assumption that mutating any of the four nAChR subunits results in an identical increase of roughly 5 k_BT to the allosteric gating energy ϵ. Instead, we allow each type of subunit to shift ϵ by a different amount upon mutation. We begin by writing the ϵ parameter of wild type nAChR as

$$\epsilon_{\alpha_2\beta\gamma\delta} = 2\epsilon_\alpha + \epsilon_\beta + \epsilon_\gamma + \epsilon_\delta, \tag{S5.19}$$

where ϵ_j denotes the gating energy contribution from subunit j and we have assumed that the five subunits independently contribute to channel gating. Upon mutation, we define the free energy differences of each type of subunit as

$$\Delta\epsilon_\alpha \equiv \epsilon_{\alpha^*} - \epsilon_\alpha \tag{S5.20}$$

$$\Delta\epsilon_\beta \equiv \epsilon_{\beta^*} - \epsilon_\beta \tag{S5.21}$$

$$\Delta\epsilon_\gamma \equiv \epsilon_{\gamma^*} - \epsilon_\gamma \tag{S5.22}$$

$$\Delta\epsilon_\delta \equiv \epsilon_{\delta^*} - \epsilon_\delta, \tag{S5.23}$$

where ϵ_{j^*} denotes the gating energy from the mutated subunit j.

The allosteric energy of any nAChR mutant can be found using the wild type energy $\epsilon_{\alpha_2\beta\gamma\delta} = -23.7\ k_BT$ from the main text together with $\Delta\epsilon_\alpha$, $\Delta\epsilon_\beta$, $\Delta\epsilon_\delta$, and $\Delta\epsilon_\gamma$. For example, the gating energy of $\alpha_2\beta^*\gamma\delta$ is given by $\epsilon_{\alpha_2\beta^*\gamma\delta} = \epsilon_{\alpha_2\beta\gamma\delta} + \Delta\epsilon_\beta$ while that of $\alpha_2^*\beta\gamma\delta^*$ is given by $\epsilon_{\alpha_2^*\beta\gamma\delta^*} = \epsilon_{\alpha_2\beta\gamma\delta} + 2\Delta\epsilon_\alpha + \Delta\epsilon_\delta$.

Using the measured $[EC_{50}]$ values of all the mutants in Table S5.1, we can fit the four $\Delta\epsilon_j$'s to determine how the different subunits increase the ion channel gating energy upon mutation. We find the values $\Delta\epsilon_\alpha = 4.4\ k_BT$, $\Delta\epsilon_\beta = 5.3\ k_BT$, $\Delta\epsilon_\gamma = 5.4\ k_BT$, and $\Delta\epsilon_\delta = 5.2\ k_BT$, which show a small spread about the value of roughly 5 k_BT found in the text by assuming that all four $\Delta\epsilon_j$'s are identical. To show the goodness of fit, we can compare the $[EC_{50}]$ values from this model to the experimental measurements in Table S5.1, as shown in Fig. S5.4.

S5.1.3 $p_{\text{open}}(c)$ Curves

Although the dose-response curves we analyze for nAChR were all presented using normalized current, the underlying physical process – namely, the opening and closing of the ion channel – is not required to go from 0 to 1. Fig. S5.5 shows the normalized dose-response curves from Fig. 5.3A together with the average probability that each ion channel mutant will be open, $p_{\text{open}}(c)$. Note that these $p_{\text{open}}(c)$ curves have exactly the same shape as the normalized current curves but are compressed in the vertical direction to have the leakiness and dynamic range specified by Fig. 5.4A and B.

From the viewpoint of these $p_{\text{open}}(c)$ curves, various nuances of this ion channel class stand out more starkly. For example, the four mutant channels have $p_{\text{open}}^{\text{max}} \approx 1$,

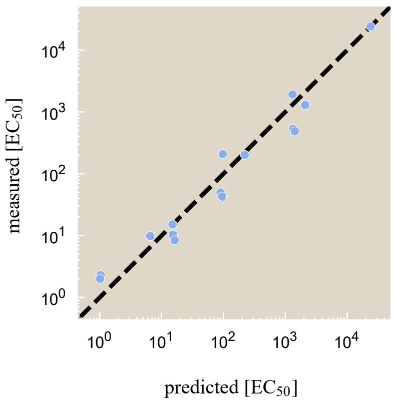


Figure S5.4: Mutating different nAChR subunits changes the gating energy ϵ by different amounts. Using a linear model where each subunit independently contributes to channel gating, we fit all of the $[EC_{50}]$ values in Table S5.1 to compute the increase of the gating energy ϵ when each subunit of $\alpha_2\beta\gamma\delta$ nAChR is mutated (see Eqs. S5.20-S5.23). Upon mutation, a subunit of type j increases the gating energy by $\Delta\epsilon_j$, where $\Delta\epsilon_\alpha = 4.4\ k_BT$, $\Delta\epsilon_\beta = 5.3\ k_BT$, $\Delta\epsilon_\gamma = 5.4\ k_BT$, and $\Delta\epsilon_\delta = 5.2\ k_BT$. For each mutant in Table S5.1, the $[EC_{50}]$ from the model can be compared to the corresponding experimental measurement, with the black dashed line denoting the line of equality $y = x$.

noticeably larger than the $p_{\text{open}}^{\text{max}} \approx 0.9$ value of the wild type channel. In addition, the $n = 4$ mutant is the only ion channel with non-negligible leakiness, and Fig. 5.4A suggests that an $n = 5$ mutant with all five subunits carrying the L251S mutation would have an even larger leakiness value greater than ½. In other words, the $n = 5$ ion channel is open more than half the time even in the absence of ligand, which could potentially cripple or kill the cell. This may explain why Labarca *et al.* made the $n = 5$ strain but were unable to measure its properties [3].

Fig. S5.6 repeats this same analysis for the CNGA2 dose-response curves from Fig. 5.6A. In this case, all of the ion channel mutants have uniformly small values of $p_{\text{open}}^{\text{min}} \approx 0.03$ and uniformly large $p_{\text{open}}^{\text{max}} \approx 1$, as indicated by Fig. 5.8A and B. Therefore, the $p_{\text{open}}(c)$ curves look very similar to the normalized currents.

S5.1.4 nAChR L251T Mutation

In this section, we consider a separate nAChR data from the one considered in the main paper. Filatov and White constructed nAChR ion channel mutants closely related to those of Labarca *et al.* but employing a L251T mutation [9]. They measured the $[EC_{50}]$ of multiple such constructs with the L251T mutation on different subsets of nAChR subunits, with the results shown in Fig. S5.7A as a function of the total number of mutated subunits n.

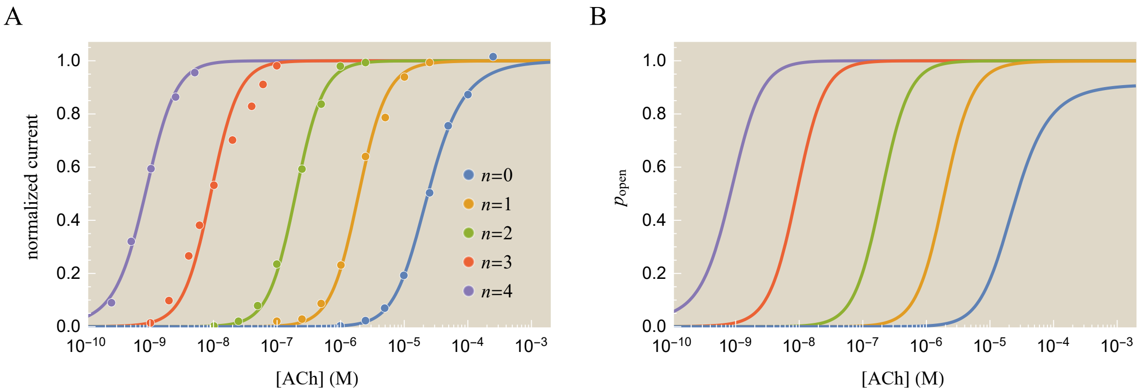


Figure S5.5: Probability that an nAChR mutant will be open. (A) Normalized current curves of the five nAChR mutants from Fig. 5.3A. (B) The probability that each ion channel will be open is given by Eq. 5.1. Note that the wild type ion channel has a smaller dynamic range and the $n = 4$ mutant has a noticeably larger leakiness than the other mutants.

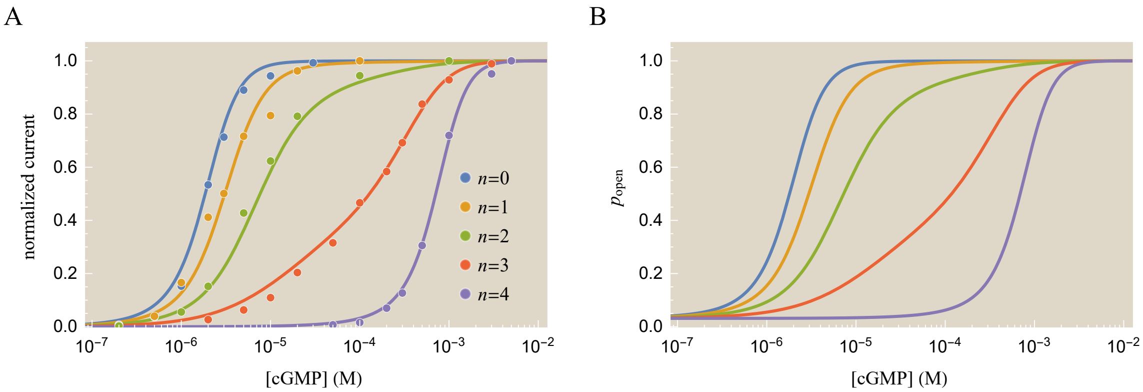


Figure S5.6: Probability that a CNGA2 mutant will be open. (A) CNGA2 dose-response curves from Fig. 5.6A. (B) The probability that each ion channel will be open is given by Eq. 5.1. Since all of the channels have small leakiness (≈ 0.03) and large dynamic range, the $p_{\text{open}}(c)$ curves are nearly identical to the normalized current curves.

As in the case of the L251S mutations from Labarca (see Fig. S5.3 and Table S5.1), there was some variation in $[EC_{50}]$ between different mutants with the same total number of mutations n, but the entire class of mutants is well approximated as having $[EC_{50}]$ exponentially decrease with each additional mutation. Utilizing our analytical formula for the $[EC_{50}]$ of nAChR, Eq. 5.11, and assuming that each mutation changes ϵ by a fixed amount $\Delta\epsilon$, the shift in $[EC_{50}]$ due to n mutations is given by

$$[EC_{50}] = e^{-\beta(\epsilon^{(0)}+n\Delta\epsilon)/2} K_{\text{O}}. \tag{S5.24}$$

We can fit the logarithm of the $[EC_{50}]$ values in Fig. S5.7A to a linear function going through the wild type ($n = 0$) data point to obtain $\Delta\epsilon = 3.64\ k_B T$ from the slope of this line. This value is comparable to that found for the L251S mutation (where $\Delta\epsilon = 5\ k_B T$).

With the gating energy now fully determined for any number of mutations n, and using the K_O and K_C parameters from Fig. 5.3, we now have a complete theoretical model of the L251T nAChR mutant class. For example, we can plot the predicted dose-response curves for all such mutants. Fig. S5.7B shows these predictions together with experimentally measured responses from the wild type channel and three mutant constructs. The dose-response predictions should match the data on average for the entire class of mutants, although individual channel responses may be slightly off. For example, Fig. S5.7A indicates that the $[EC_{50}]$ of the $n = 1$ and $n = 2$ mutants will be lower than predicted while that of the $n = 4$ and $n = 5$ mutants (whose dose-response data was not provided) will be higher than predicted.

S5.1.5 Combining Multiple Ion Channels

In this section, we consider the dose-response curve for the case in which the cell harbors both wild type and mutant ion channels. Given n_1 ion channels whose dose-response curves are governed by $p_{1,\text{open}}(c)$ and n_2 ion channels with a different response $p_{2,\text{open}}(c)$, the current produced by the combination of these two ion channels is given by

$$\text{current} \propto n_1 p_{1,\text{open}}(c) + n_2 p_{2,\text{open}}(c). \tag{S5.25}$$

Experimental measurements are computed on a relative scale so that the data runs from 0 to 1. Analytically, this amounts to subtracting the leakiness and dividing by the dynamic range,

$$(\text{normalized current})_{\text{tot}} = \frac{n_1 p_{1,\text{open}}(c) + n_2 p_{2,\text{open}}(c) - n_1 p_{1,\text{open}}^{\text{min}} - n_2 p_{2,\text{open}}^{\text{min}}}{n_1 p_{1,\text{open}}^{\text{max}} + n_2 p_{2,\text{open}}^{\text{max}} - n_1 p_{1,\text{open}}^{\text{min}} - n_2 p_{2,\text{open}}^{\text{min}}}. \tag{S5.26}$$

Wongsamitkul *et al.* constructed cells expressing both the $n = 0$ wild type ion channels and the $n = 4$ fully mutated ion channels in a ratio of 1:1 (i.e. $n_1 = n_2$) as shown in Fig. S5.8 [4].

Recall from Fig. 5.8 that these ion channels have very small leakiness ($p_{1,\text{open}}^{\text{min}} \approx p_{2,\text{open}}^{\text{min}} \approx 0$) and nearly full dynamic range ($p_{1,\text{open}}^{\text{max}} \approx p_{2,\text{open}}^{\text{max}} \approx 1$). This implies that $p_{1,\text{open}}(c) \approx (\text{normalized current})_1$ and $p_{2,\text{open}}(c) \approx (\text{normalized current})_2$, so that

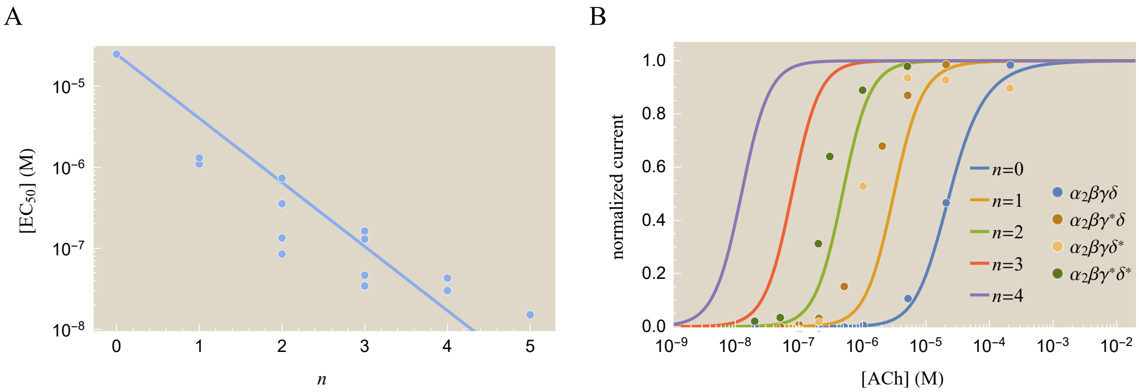


Figure S5.7: Effects of L251T mutations on nAChR. (A) $[EC_{50}]$ values for another class of L251T mutations introduced at different combinations of subunits. This data set is separate from the L251S mutation considered in the main text. The $[EC_{50}]$ mainly depends on the total number of mutations, $[EC_{50}] \propto e^{-1.82n}$, although there is slight variation depending upon which subunits are mutated. From Eq. S5.24, we find that each mutation imparts $\Delta\epsilon = 3.64\ k_B T$. (B) Once the MWC parameters have been fixed from the $[EC_{50}]$ measurements, we can predict the full dose-response curves for the entire class of L251T nAChR mutants. Overlaid on these theoretical prediction are four experimentally measured response curves for the wild type ($\alpha_2\beta\gamma\delta$), two $n = 1$ single mutants ($\alpha_2\beta\gamma^*\delta$ and $\alpha_2\beta\gamma\delta^*$), and the $n = 2$ double mutant ($\alpha_2\beta\gamma^*\delta^*$). We expect the predicted dose-response curves to match the data on average for the entire class of mutants, but Part A shows that the $[EC_{50}]$ of the $n = 1$ and $n = 2$ mutants will be overestimated while that of the $n = 4$ and $n = 5$ mutants will be underestimated. Asterisks ($*$) in the legend denote L251T mutations.

the total normalized current due to the combination of ion channels is given by

$$\text{(normalized current)}_{\text{tot}} = \frac{\text{(normalized current)}_1 + \text{(normalized current)}_2}{2}. \tag{S5.27}$$

Fig. S5.8 shows that this simple prediction compares well to the measured data.

S5.2 Computing nAChR and CNGA2 Characteristics

In S5.2.1, we derive Eqs. 5.9-5.12, the approximations for the leakiness, dynamic range, $[EC_{50}]$, and the effective Hill coefficient h for the general MWC model Eq. 5.1. We begin by Taylor expanding the well known exact expressions from Ref. 10 in the limit $1 \ll e^{-\beta\epsilon} \ll \left(\frac{K_C}{K_O}\right)^m$, which we found to be appropriate for both the nAChR and CNGA2 ion channels, and find the lowest order approximations.

Following that, in S5.2.2 we consider how mutations in the ligand dissociation constants K_O and K_C affect these four properties. We show that ion channel dose-

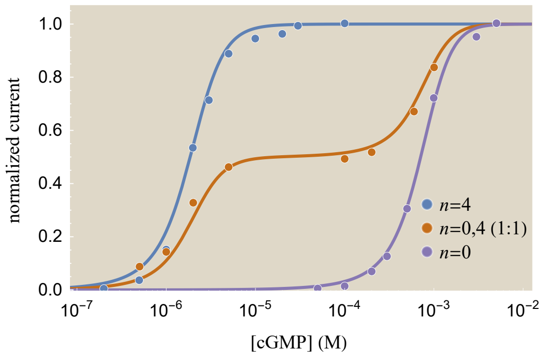


Figure S5.8: Normalized currents for combinations of CNGA2 ion channels. Channel currents of cells producing equal amounts of wild type $n = 0$ and the $n = 4$ mutant ion channels. As shown in Eq. S5.27, the resulting dose-response curve equals the average of the $n = 0$ and $n = 4$ individual response curves.

response curves are robust to changes in K_O and K_C aside from left-right shifts dictated by $[EC_{50}] = e^{-\beta\epsilon/m} K_O$. This discussion complements the nAChR section of the text where we considered mutations of the $\beta\epsilon$ parameter.

Lastly, in S5.2.3 we determine how ion channels comprised of a mix of wild type subunits (with ligand dissociation constants K_O and K_C) and mutant subunits (with dissociation constants K_O^* and K_C^*) influences the four properties. Specifically, we focus on the analytically tractable case where half of the subunits are wild type and the other half are mutated (see Eqs. 5.18 and 5.19 in the text).

S5.2.1 Characteristics of the MWC Model

Using $1 \ll e^{-\beta\epsilon}$, the leakiness Eq. 5.5 can be expanded as

$$\text{leakiness} = \frac{1}{1 + e^{-\beta\epsilon}} \approx e^{\beta\epsilon}. \tag{S5.28}$$

Therefore, ion channels have a very small leakiness which scales exponentially with $\beta\epsilon$. Fig. S5.9A shows that this is a good approximation across the entire range of parameters within the class of nAChR mutants, $-24 \leq \beta\epsilon \leq -4$.

The dynamic range Eq. 5.6 can be similarly expanded to obtain

$$\text{dynamic range} = \frac{1}{1 + e^{-\beta\epsilon}\left(\frac{K_O}{K_C}\right)^m} - \frac{1}{1 + e^{-\beta\epsilon}} \approx 1 - e^{-\beta\epsilon}\left(\frac{K_O}{K_C}\right)^m - e^{\beta\epsilon}. \tag{S5.29}$$

Keeping only the lowest order term yields the approximation Eq. 5.10 that ion channels have full dynamic range. Fig. S5.9B shows that keeping the first order terms in Eq. S5.29 also captures the behavior of the wild type channel ($\beta\epsilon^{(0)} = -23.7$) and the $n = 4$ mutant ($\beta\epsilon^{(4)} = -4.0$).

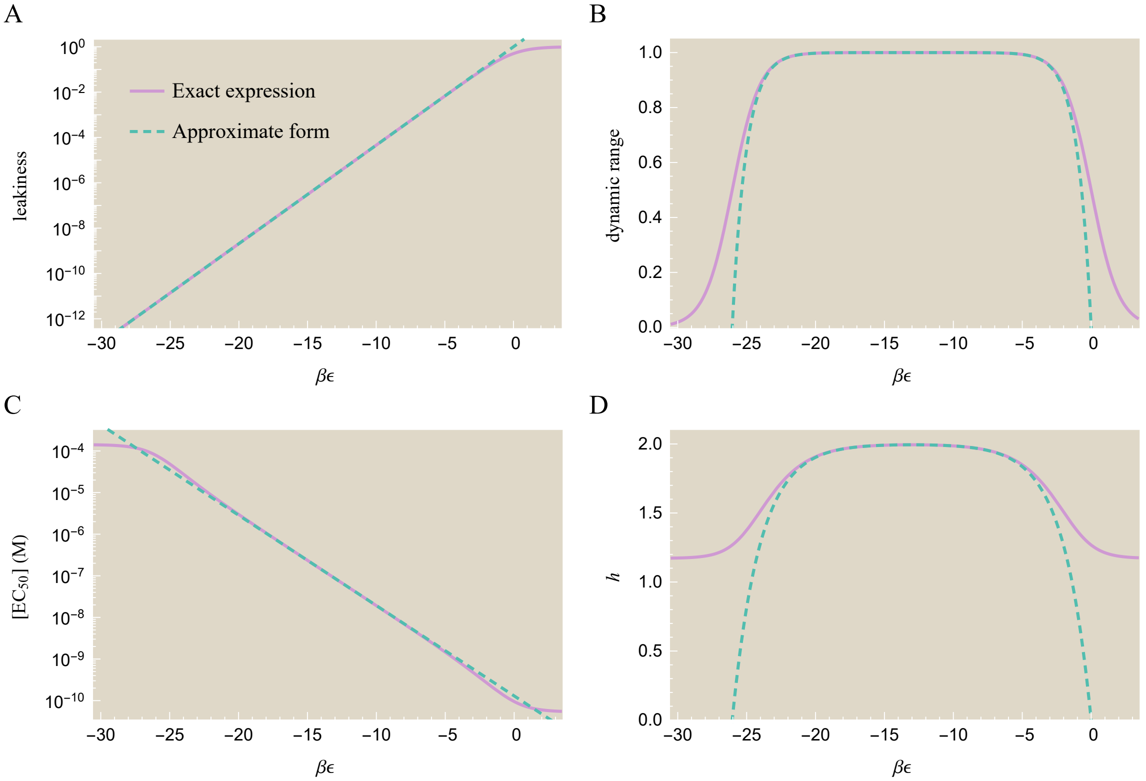


Figure S5.9: Exact and approximate expressions for nAChR characteristics. The approximations Eqs. S5.28-S5.34 (dashed, teal) are valid in the limit $1 \ll e^{-\beta\epsilon} \ll \left(\frac{K_C}{K_O}\right)^m$ where they closely match the exact expressions (purple). (A) Leakiness can be approximated as an exponentially increasing function of $\beta\epsilon$. (B) To lowest order, the dynamic range of an ion channel should approach unity, with deviations only at very large and very small $\beta\epsilon$ values. (C) The $[EC_{50}]$ is an exponentially decreasing function of $\beta\epsilon$. (D) The effective Hill coefficient is roughly constant for all mutants, but as with the dynamic range it decreases for very large and very small $\beta\epsilon$ values.

We next turn to the $[EC_{50}]$ Eq. 5.7, whose exact analytic formula is given by [11]

$$[EC_{50}] = K_O \frac{1 - \lambda^{\frac{1}{m}}}{\lambda^{\frac{1}{m}} - \frac{K_O}{K_C}} \tag{S5.30}$$

where

$$\lambda = \frac{2 - \left(p_{\text{open}}^{\text{min}} + p_{\text{open}}^{\text{max}}\right)}{e^{-\beta\epsilon}\left(p_{\text{open}}^{\text{min}} + p_{\text{open}}^{\text{max}}\right)}. \tag{S5.31}$$

The limit $1 \ll e^{-\beta\epsilon} \ll \left(\frac{K_C}{K_O}\right)^m$ suggests that we Taylor expand this formula to lowest order about $e^{-\beta\epsilon}\left(\frac{K_O}{K_C}\right)^m \approx 0$ and $e^{-\beta\epsilon} \approx \infty$, which yields

$$\begin{aligned}
[EC_{50}] &\approx K_O \frac{\frac{K_C}{K_O}\left(\left(1-\frac{1}{2+e^{-\beta\epsilon}}\right)^{1/m}\right)}{\frac{K_C}{K_O}\left(\frac{1}{2+e^{-\beta\epsilon}}\right)^{1/m}-1} \\
&\approx K_O \frac{\frac{K_C}{K_O}\left(1-e^{\beta\epsilon/m}\right)}{\frac{K_C}{K_O}e^{\beta\epsilon/m}-1} \\
&\approx K_O \frac{\frac{K_C}{K_O}}{\frac{K_C}{K_O}e^{\beta\epsilon/m}} \\
&= K_O e^{-\beta\epsilon/m}. \qquad \text{(S5.32)}
\end{aligned}$$

Thus, the $[EC_{50}]$ decreases exponentially with ϵ, although this effect is diminished with the number of ligand binding sites m. The precise relationship $[EC_{50}] \propto e^{-\beta\epsilon/2}$ for the nAChR data is shown in Fig. S5.9C.

Finally, we turn to the effective Hill coefficient, whose exact analytic form is given by [11]

$$h = \frac{m[EC_{50}]\left(K_C-K_O\right)\left(p_{\text{open}}^{\text{min}}+p_{\text{open}}^{\text{max}}\right)\left(2-p_{\text{open}}^{\text{min}}-p_{\text{open}}^{\text{max}}\right)}{\left(p_{\text{open}}^{\text{min}}-p_{\text{open}}^{\text{max}}\right)\left([EC_{50}]+K_O\right)\left([EC_{50}]+K_C\right)}, \qquad \text{(S5.33)}$$

where we have used $p_{\text{open}}^{\text{min}}$ and $p_{\text{open}}^{\text{max}}$ from Eqs. 5.3 and 5.4 as well as the $[EC_{50}]$ formula Eq. S5.30. Again, we make a Taylor series of this expression about $e^{-\beta\epsilon}\left(\frac{K_O}{K_C}\right)^m \approx 0$ and $e^{-\beta\epsilon} \approx \infty$ to obtain the lowest order approximation, which is given by

$$\begin{aligned}
h &\approx m\frac{\frac{K_C}{K_O}+1}{\frac{K_C}{K_O}-1} - m\frac{\left(\frac{1}{2+e^{-\beta\epsilon}}\right)^{-1/m}}{\frac{K_C}{K_O}-1} - m\frac{\frac{K_C}{K_O}\left(\frac{1}{2+e^{-\beta\epsilon}}\right)^{1/m}}{\frac{K_C}{K_O}-1} - 2me^{\beta\epsilon}\frac{\frac{K_C}{K_O}\left(\frac{1}{2+e^{-\beta\epsilon}}\right)^{1/m}}{\frac{K_C}{K_O}-1} \\
&\approx m - m\frac{e^{-\beta\epsilon/m}}{\frac{K_C}{K_O}} - m\frac{\frac{K_C}{K_O}e^{\beta\epsilon/m}}{\frac{K_C}{K_O}} - 2me^{\beta\epsilon}\frac{\frac{K_C}{K_O}e^{\beta\epsilon/m}}{\frac{K_C}{K_O}} \\
&\approx m - m\frac{K_O}{K_C}e^{-\beta\epsilon/m} - me^{\beta\epsilon/m}. \qquad \text{(S5.34)}
\end{aligned}$$

Note that in the second step, we used the stronger constraint that $\frac{K_C}{K_O} \gg 1$, although it is still reasonably satisfied for both the nAChR ($\frac{K_C}{K_O} = 6 \times 10^5$) and CNGA2 ($\frac{K_C}{K_O} = 17$) ion channels considered in the text. By keeping the lowest order term, we recoup Eq. 5.12 that all ion channels have the same sharp response, and that this sharpness increases linearly with the number of ligand binding sites. Fig. S5.9D shows that by keeping the first order terms in Eq. S5.34, the shallower responses of the wild type ($\beta\epsilon^{(0)} = -23.7$) and $n = 4$ mutant ($\beta\epsilon^{(4)} = -4.0$) can also be well approximated.

S5.2.2 Mutations Affecting the Ligand-Channel Dissociation Constants

In this section, we discuss how the leakiness, dynamic range, $[EC_{50}]$, and effective Hill coefficient h vary when the channel-ligand dissociation constants K_O and K_C are perturbed, as can be accomplished by mutating the ligand binding domain. Recall that ligand-gated ion channels are typically closed in the absence of ligand ($\epsilon < 0$) and open when bound to ligand ($K_\text{O} < K_\text{C}$).

Fig. S5.10 shows the four ion channel properties using the parameters of the wild type CNGA2 ion channel ($\beta\epsilon = -3.4$ and $m = 4$ ligand binding sites) and letting the ratio $\frac{K_\text{O}}{K_\text{C}}$ of dissociation constants vary. All four graphs demonstrate that the ion channel's traits are nearly insensitive to changes in the dissociation constants provided that K_O does not approach K_C. In the limit $K_\text{O} \to K_\text{C}$, the ligand no longer drives the ion channel to open, causing the dynamic range to shrink to zero. As such, the behavior of the $[EC_{50}]$ and h in this limit should be considered as artifacts from taking this limit (since neither trait is well defined when the dynamic range shrinks to zero). For reference, the wild type CNGA2 channel has $\frac{K_\text{O}}{K_\text{C}} = 0.06$.

Note that the y-axis in Fig. S5.10C plots $\frac{[EC_{50}]}{K_\text{O}}$, so if both K_O and K_C are reduced by a constant factor, then the $[EC_{50}]$ will also be reduced by this same factor. In the limit $\frac{K_\text{O}}{K_\text{C}} \to 0$, $[EC_{50}] = K_\text{O} \left(2 + e^{-\beta\epsilon}\right)^{1/m} - 1 \approx 1.4 K_\text{O}$ for the wild type CNGA2 channel (see Eq. S5.30).

S5.2.3 The Heterooligomeric CNGA2 Channel

We now consider an ion channel with m subunits, each of which could either be a wild type subunit (with dissociation constants K_O and K_C) or a mutated subunit (with dissociation constants K_O^* and K_C^*). Each subunit contains a single ligand binding site. We compute the leakiness, dynamic range, $[EC_{50}]$, and the effective Hill coefficient h of an ion channel composed of n mutated subunits and $m - n$ wild type subunits. In the text, we analyzed the specific case of the CNGA2 ion channel with $m = 4$ subunits.

We begin by taking the limits of $p_\text{open}(c)$, Eq. 5.15, in the absence of ligand, which is given by

$$p_\text{open}^\text{min} = \frac{1}{1 + e^{-\beta\epsilon}}, \tag{S5.35}$$

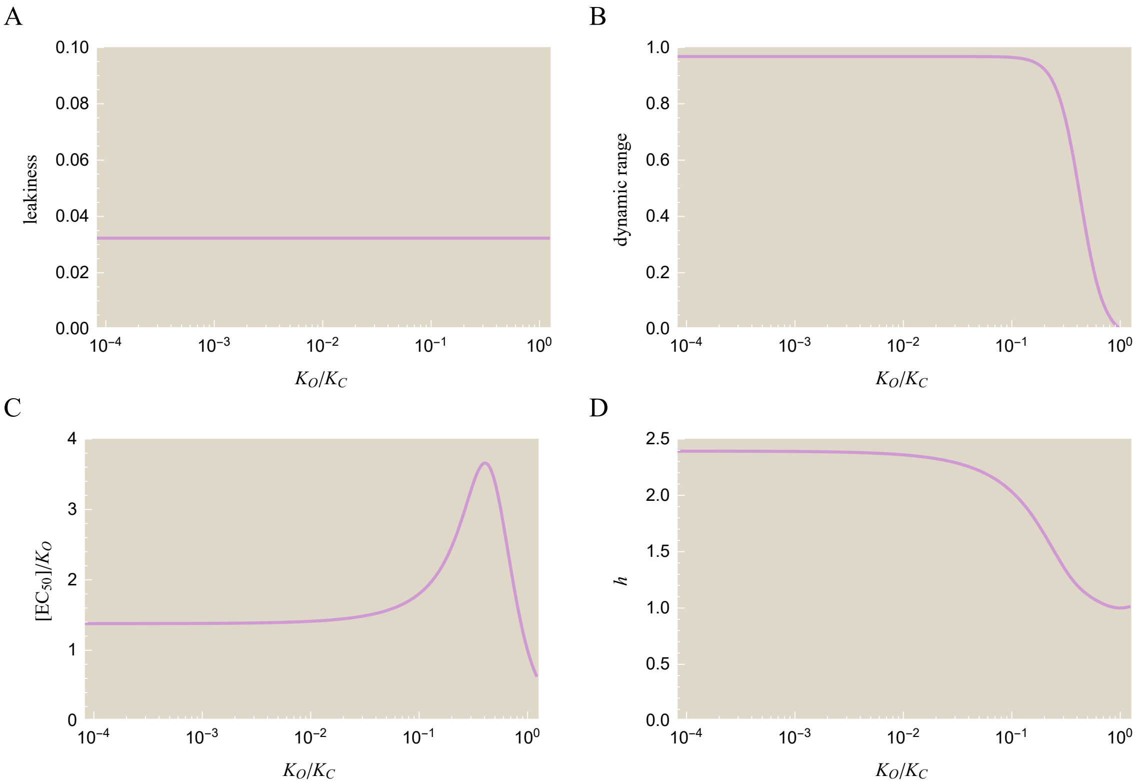


Figure S5.10: CNGA2 properties are robust to changes in the ligand dissociation constants. (A) The leakiness does not depend on either dissociation constant. (B) The dynamic range is near unity for set of dissociation constants where $\frac{K_\mathrm{O}}{K_\mathrm{C}} \leq 0.1$, as was found for both the nAChR and CNGA2 systems. For larger ratios of the dissociation constants, the ligand no longer drives the ion channel to open. (C) When $\frac{K_\mathrm{O}}{K_\mathrm{C}} \leq 0.1$, $[EC_{50}] \approx 1.4 K_\mathrm{O}$ is proportional to K_O but robust to the ratio of dissociation constants. (D) The effective Hill coefficient is also robust to changes in the dissociation constants, with $h \approx 2.4$ when $\frac{K_\mathrm{O}}{K_\mathrm{C}} \leq 0.1$.

and in the presence of saturating levels of ligand, which is given as

$$p_{\mathrm{open}}^{\mathrm{max}} = \frac{1}{1 + e^{-\beta\epsilon} \left(\frac{K_\mathrm{O}}{K_\mathrm{C}}\right)^{m-n} \left(\frac{K_\mathrm{O}^*}{K_\mathrm{C}^*}\right)^{n}}. \tag{S5.36}$$

Throughout this analysis, we assume $K_\mathrm{O} < K_\mathrm{C}$ and $K_\mathrm{O}^* < K_\mathrm{C}^*$ so that ligand binding makes both the wild type and mutant subunits more likely to open.

The two limits of $p_{\mathrm{open}}(c)$ above allow us to directly compute the leakiness and dynamic range of the ion channels. The former is given by

$$\text{leakiness} = \frac{1}{1 + e^{-\beta\epsilon}}, \tag{S5.37}$$

which has an identical form to the leakiness of the MWC model Eq. S5.28. As shown in Fig. S5.11A, the leakiness does not depend explicitly on the number of

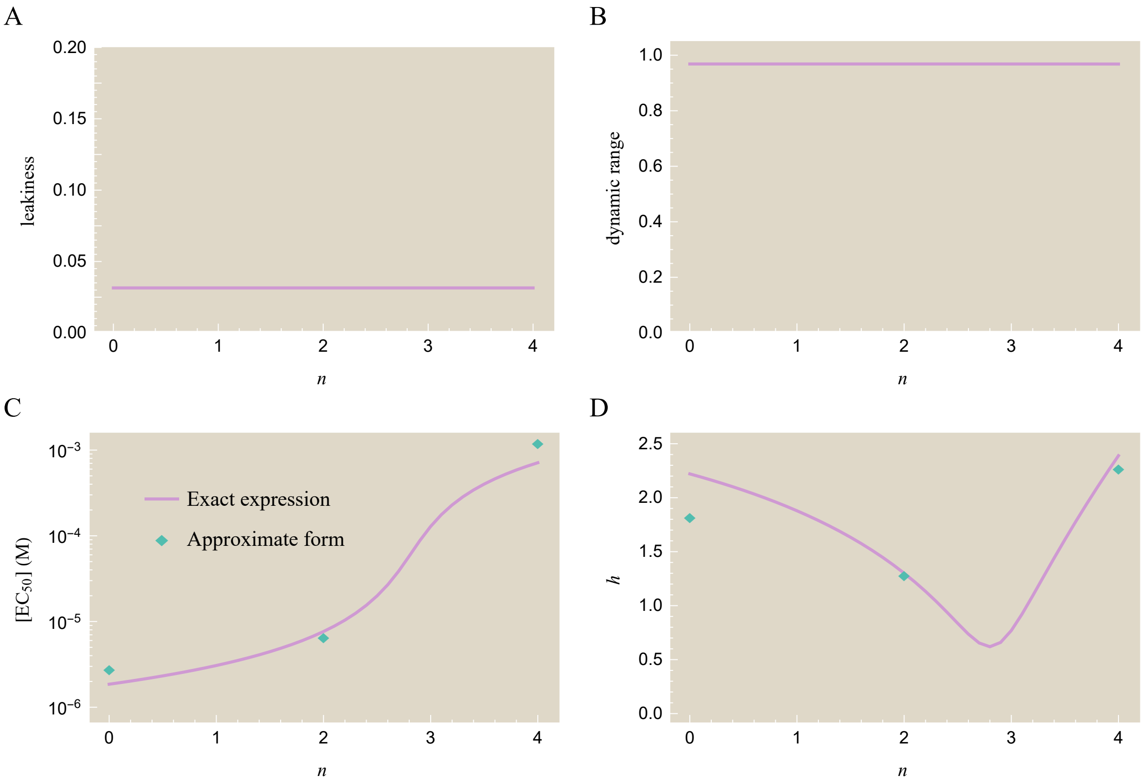


Figure S5.11: Effects of mixing two types of subunits in the CNGA2 ion channel. The CNGA2 ion channel is composed of $m = 4$ subunits, each of which has one ligand binding site. n of these subunits are mutated so as to have weaker ligand binding affinity. (A) The leakiness of the CNGA2 ion channel Eq. S5.37 is uniformly small. (B) All of the mutants have nearly full dynamic range Eq. S5.38 because the open channel dissociations constants (K_O and K_O^*) are significantly larger than the closed channel dissociation constants (K_C and K_C^*). (C) The exact expression (solid, purple) for the $[EC_{50}]$ is shown along with approximations for the $n = 0$, 2, and 4 ion channels (teal diamonds) from Eqs. S5.32 and S5.48. Because the mutated subunits bind poorly to ligand, the $[EC_{50}]$ increases with n. (D) The effective Hill coefficient Eqs. S5.34 and S5.50 can be approximated in the same manner as the $[EC_{50}]$. Although the homooligomeric $n = 0$ and $n = 4$ channels have sharp responses, the effect of combining both types of subunits ($n = 1$, 2, and 3) leads to a flatter response.

mutated subunits n. We next turn to the formula for the dynamic range,

$$\text{dynamic range} = \frac{1}{1 + e^{-\beta\epsilon}\left(\frac{K_O}{K_C}\right)^{m-n}\left(\frac{K_O^*}{K_C^*}\right)^{n}} - \frac{1}{1 + e^{-\beta\epsilon}}. \tag{S5.38}$$

The first term in the dynamic range is approximately 1 because the open state affinities are always smaller than the closed state affinities by at least a factor of ten, and these factors are collectively raised to the m^{th} power. Since these ion channels also exhibit small leakiness, they each have a large dynamic range as shown in Fig. S5.11B.

We next consider approximations for the $[EC_{50}]$ and effective Hill coefficient h. The wild type CNGA2 channel ($n = 0$) will necessarily follow the formulas derived above for the MWC model (Eqs. S5.32 and S5.34). Similarly, the homooligomeric CNGA2 ion channel comprised of all mutated subunits ($n = m$) will have the same formulas as the wild type channel but with $K_{\text{O}} \to K_{\text{O}}^*$ and $K_{\text{C}} \to K_{\text{C}}^*$.

To gain a sense of how the $[EC_{50}]$ and h vary for channels comprised of a mix of wild type and mutant subunits ($1 \le n \le m - 1$), we analyze the $n = m/2$ case (implicitly assuming that m is even) where half the subunits are wild type and the other half are mutated. We begin with the $[EC_{50}]$ formula which by definition is given by

$$\frac{\left[\left(1+\frac{c}{K_{\text{O}}}\right)\left(1+\frac{c}{K_{\text{O}}^*}\right)\right]^{m/2}}{\left[\left(1+\frac{c}{K_{\text{O}}}\right)\left(1+\frac{c}{K_{\text{O}}^*}\right)\right]^{m/2}+e^{-\beta\epsilon}\left[\left(1+\frac{c}{K_{\text{C}}}\right)\left(1+\frac{c}{K_{\text{C}}^*}\right)\right]^{m/2}} = \frac{1}{2}\left(p_{\text{open}}^{\text{min}}+p_{\text{open}}^{\text{max}}\right). \tag{S5.39}$$

Rearranging the terms, we find

$$\lambda\left[\left(1+\frac{c}{K_{\text{O}}}\right)\left(1+\frac{c}{K_{\text{O}}^*}\right)\right]^{m/2} = \left[\left(1+\frac{c}{K_{\text{C}}}\right)\left(1+\frac{c}{K_{\text{C}}^*}\right)\right]^{m/2}, \tag{S5.40}$$

where we have introduced the same (positive) constant λ in Eq. S5.31 from the $[EC_{50}]$ of the standard MWC model,

$$\lambda = \frac{2-\left(p_{\text{open}}^{\text{min}}+p_{\text{open}}^{\text{max}}\right)}{e^{-\beta\epsilon}\left(p_{\text{open}}^{\text{min}}+p_{\text{open}}^{\text{max}}\right)}. \tag{S5.41}$$

Upon raising both sides of Eq. S5.39 to the $\frac{2}{m}$ power, we find the quadratic equation

$$Ac^2 + Bc + C = 0 \tag{S5.42}$$

where

$$A = \frac{\lambda^{2/m}}{K_{\text{O}}K_{\text{O}}^*} - \frac{1}{K_{\text{C}}K_{\text{C}}^*} \tag{S5.43}$$

$$B = \frac{\lambda^{2/m}}{K_{\text{O}}} + \frac{\lambda^{2/m}}{K_{\text{O}}^*} - \frac{1}{K_{\text{C}}} - \frac{1}{K_{\text{C}}^*} \tag{S5.44}$$

$$C = \lambda^{2/m} - 1, \tag{S5.45}$$

which has the exact solution

$$[EC_{50}] = \frac{-B + \sqrt{B^2 - 4AC}}{2A}. \tag{S5.46}$$

To simplify this expression, we note that $|4AC|$ is smaller than B^2 by more than a factor of 10 for the CNGA2 parameter values, so that the square root can be approximated as $\sqrt{B^2 - 4AC} \approx B - \frac{2AC}{B}$, and hence the $[EC_{50}]$ becomes

$$[EC_{50}] \approx -\frac{C}{B} = \frac{1 - \lambda^{2/m}}{\frac{\lambda^{2/m}}{K_O} + \frac{\lambda^{2/m}}{K_O^*} - \frac{1}{K_C} - \frac{1}{K_C^*}}. \tag{S5.47}$$

To further simplify this result, we utilize the relationships $1 \ll e^{-\beta\epsilon} \ll \left(\frac{K_C}{K_O}\right)^{m/2}$ and $1 \ll e^{-\beta\epsilon} \ll \left(\frac{K_C^*}{K_O^*}\right)^{m/2}$ which hold for the CNGA2 parameters. In this limit, $p_{\text{open}}^{\text{min}} \approx 0$, $p_{\text{open}}^{\text{max}} \approx 1$, and $\lambda \approx e^{\beta\epsilon} \ll 1$, so that the formula for the $[EC_{50}]$ becomes

$$\begin{aligned}[EC_{50}] &\approx \frac{1}{\frac{e^{2\beta\epsilon/m}}{K_O} + \frac{e^{2\beta\epsilon/m}}{K_O^*} - \frac{1}{K_C} - \frac{1}{K_C^*}} \\ &\approx \frac{1}{\frac{e^{2\beta\epsilon/m}}{K_O} + \frac{e^{2\beta\epsilon/m}}{K_O^*}} \\ &= e^{-2\beta\epsilon/m} \frac{K_O K_O^*}{K_O + K_O^*}.\end{aligned} \tag{S5.48}$$

Since the mutated CNGA2 subunits have significantly weaker binding affinity ($K_O \ll K_O^*$), $[EC_{50}] \approx e^{-2\beta\epsilon/m} K_O$ where the exponent is twice as large as the homooligomeric case Eq. S5.32. Fig. S5.11C shows how the $[EC_{50}]$ gradually increases as more of mutant subunits are introduced into the ion channel, with the approximation for the $n = \frac{m}{2}$ mutant given by Eq. S5.48 while the $n = 0$ and $n = m$ mutants are described by Eq. S5.32.

We next turn to the effective Hill coefficient h. To greatly simplify the computation, we ignore all of the dissociation constants ($K_C = 20 \times 10^{-6}$ M, $K_O^* = 500 \times 10^{-6}$ M, and $K_C^* = 140 \times 10^{-3}$ M) greater than the $[EC_{50}] \approx 6 \times 10^{-6}$ M, since they all enter $p_{\text{open}}(c)$ as $\left(1 + \frac{c}{K_j}\right)^{m/2}$ with $m = 4$ for CNGA2. Thus, the probability of the channel opening becomes

$$p_{\text{open}}(c) \approx \frac{\left(1 + \frac{c}{K_O}\right)^{m/2}}{\left(1 + \frac{c}{K_O}\right)^{m/2} + e^{-\beta\epsilon}} \tag{S5.49}$$

where only the effect of the smallest dissociation constant ($K_O = 1.2 \times 10^{-6}$ M) is considered. Noting that $K_O \ll K_O^*$, the effective Hill coefficient is given by

$$\begin{aligned} h &\approx \frac{m}{\left(1 + e^{2\beta\epsilon/m}\right)\left(1 + \left(1 + e^{2\beta\epsilon/m}\right)^{m/2}\right)} \\ &\approx \frac{m}{2} - \frac{m}{2}\left(\frac{m}{4} + 1\right) e^{2\beta\epsilon/m}, \end{aligned} \tag{S5.50}$$

where in the second step we used the Taylor expansion about $e^{2\beta\epsilon/m} \ll 1$. Fig. S5.11D shows the effective Hill coefficient together with the approximations for the homooligomeric ($n = 0$ and $n = 4$) ion channel Eq. S5.34 and the $n = 2$ channel given by Eq. S5.50. As discussed in the text, the effective Hill coefficient exhibits a surprising decrease for ion channels comprised of a mix of both types of subunits ($1 \leq n \leq 3$). Qualitatively, this comes about because the subunits of the $n = 0$ wild type channel become sensitive to ligand at concentrations approaching $[EC_{50}]^{(n=0)} \approx e^{-\beta\epsilon/m} K_O$ while the mutant subunits respond at the much larger concentrations $[EC_{50}]^{(n=4)} \approx e^{-\beta\epsilon/m} K_O^*$. Channels containing both subunits consequently have a much flatter response over the range between $e^{-\beta\epsilon/m} K_O$ and $e^{-\beta\epsilon/m} K_O^*$.

S5.3 Data Fitting

In this section, we discuss the fitting procedure used on the nAChR and CNGA2 data sets. All fitting was done using nonlinear regression (NonlinearModelFit in *Mathematica*), and the notebook carrying out this analysis can be found in the supplement of the online publication. A wide array of initial conditions were considered (for example, dissociation constants were sampled in the range $K_D \in [10^{-12}\text{ M}, 10^0\text{ M}]$ and allosteric energies were sampled across $\beta\epsilon \in [-30, 5]$), and the best-fit parameters were chosen from the fit with the largest coefficient of determination R^2. Because all dissociation constants are necessarily positive, we employed the standard trick of fitting the logarithms of dissociation constants, which improves both the fit stability and accuracy.

In S5.3.1, we give more details on the fitting procedure. In S5.3.2, we focus on the related point of the sensitivity of the MWC model parameter values. We compare experimentally measured values from the literature and analyze them in the context of the nAChR and CNGA2 data sets to determine how much flexibility the MWC model has in its ability to capture observed trends. In S5.3.3, we discuss how values such as the $[EC_{50}]$ and effective Hill coefficient can be extracted from experimental measurements.

K_{O} (M)	K_{C} (M)	$\beta\epsilon^{(0)}$	$\beta\epsilon^{(1)}$	$\beta\epsilon^{(2)}$	$\beta\epsilon^{(3)}$	$\beta\epsilon^{(4)}$	R^2
$(0.13 \pm 0.16) \times 10^{-9}$	$(11 \pm 7) \times 10^{-6}$	-23.7 ± 5.9	-19.2 ± 2.5	-14.6 ± 2.5	-8.5 ± 2.3	-4.0	0.995

Table S5.2: Best-fit parameters for the nAChR mutants given the constraint $\boldsymbol{\beta\epsilon^{(4)} = -4.0}$. With this single parameter fixed, the remaining parameters have small uncertainties. R^2 represents the coefficient of determination.

S5.3.1 Fitting Procedure

The fit parameters from Fig. 5.3 are shown in Table S5.2. If the data is fit to the MWC model (Eqs. 5.1 and 5.2) with no constraints, then all of the degenerate parameter sets in Fig. 5.10A would yield equally good fits. For example, any set of degenerate parameters with $K_{\mathrm{O}} \leq 10^{-10}$ M have coefficient of determination $R^2 = 0.995 \pm 0.0002$. In other words, it is impossible to distinguish the actual set of parameter values for nAChR without further information. As highlighted in the main text, one method for lifting this degeneracy is to independently measure one model parameter, which could then be used to fix the remaining parameters. For example, measuring the leakiness of one of the nAChR mutants would fix its corresponding $\beta\epsilon^{(n)}$ parameter, resolving the degeneracy in Fig. 5.10A. The leakiness of the $n = 3$ and $n = 4$ mutants is significantly larger than that of wild type nAChR, and hence should be possible to directly measure experimentally.

Next, we examine how much sloppiness would remain in the system if an experimental measurement fixed one of the $\beta\epsilon$ parameters. To do this, we arbitrarily choose $\beta\epsilon^{(4)} = -4.0$, and we then fit the remaining parameters with this constraint. With the degeneracy now removed from the model, Table S5.2 presents the mean parameter values and the error based on confidence intervals. Note that the remaining MWC parameters are all tightly constrained about their best-fit values, so that there is very little sloppiness left in the system after one parameter value is determined.

A similar fit procedure was used for the CNGA2 data set in Fig. 5.6. Here, we arbitrarily fixed the parameter $K_{\mathrm{C}}^{*} = 140 \times 10^{-3}$ M and fit the remaining parameters, with the results shown in Table S5.3. Again, we find that with one parameter fixed, the remaining parameters are tightly constrained.

S5.3.2 Comparison with Parameters Values from the Literature

In this section, we explore the degree of consistency between multiple independent measurements of the thermodynamic parameters in both the nAChR and CNGA2 systems.

K_O (M)	K_C (M)	K^*_O (M)	K^*_C (M)	$\beta\epsilon$	R^2
$(1.2 \pm 0.1) \times 10^{-6}$	$(20 \pm 3) \times 10^{-6}$	$(500 \pm 100) \times 10^{-6}$	140×10^{-3}	-3.4 ± 0.2	0.997

Table S5.3: Best-fit parameters for the CNGA2 mutants with the constraint $K^*_C = 140 \times 10^{-3}$ M. As in the case of nAChR, we find that once one of the parameters has a fixed value, the degeneracy within the model is lifted and the remaining parameters have small uncertainties.

We begin with the nAChR ion channel, whose allosteric gating parameter $\beta\epsilon$ has been exceedingly difficult to measure, since channel openings in the absence of ligand occur extremely infrequently. Instead of direct measurement, several groups measured the leakiness of nAChR channels with multiple pore mutations. The wild type channel parameter $\beta\epsilon^{(0)} \approx -14.2$ was then extrapolated by assuming that all of these pore mutations only affect the ϵ parameter and have energetically independent effects (i.e. if two mutations change ϵ by $\Delta\epsilon_1$ and $\Delta\epsilon_2$, respectively, then a channel with both mutations would change ϵ by $\Delta\epsilon_1 + \Delta\epsilon_2$) [12, 13]. Subsequently, dose-response curves were used to determine the values of the remaining thermodynamic parameters, namely, $K_O = 25 \times 10^{-9}$ M and $K_C = 150 \times 10^{-6}$ M [5].

We first attempted to use these literature values directly to specify K_O and K_C for the entire class of nAChR mutants. However, using these values the $n = 4$ nAChR mutant cannot be well characterized for any value of $\beta\epsilon^{(4)}$ ($R^2 < 0.5$). Thus, we next examined the sensitivity of the measured $\beta\epsilon^{(0)} = -14.2$ parameter to see how well the full nAChR data set could be fit if this value was slightly altered. Fig. S5.12A demonstrates that if $\epsilon^{(0)}$ is lowered by 4 $k_B T$, the family of nAChR mutants can once again be well characterized ($R^2 > 0.99$) by a single parameter set. In fact, as seen in Fig. S5.12B, even a decrease of 2 $k_B T$ in $\epsilon^{(0)}$ provides a reasonable fit ($R^2 = 0.98$) for the class of nAChR mutants. We note that 2 $k_B T$, roughly the energy of a hydrogen bond, is a very small energy, and this discrepancy may represent a source of error in the assumptions used to determine the $\beta\epsilon^{(0)} = -14.2$ value (e.g. that the effects of multiple channel pore mutations are additive and independent).

We now turn to how the MWC model compares to known literature values for the CNGA2 ion channel. In their paper, Wongsamitkul *et al.* reported single channel measurements for the ratio of the open to closed state for the wild type ($n = 0$) channel, finding

$$\frac{[O]}{[C]} = 1.7 \times 10^{-5} = e^{\beta\epsilon} \tag{S5.51}$$

or equivalently $\beta\epsilon = -11$ [4]. However, other sources have reported values as high

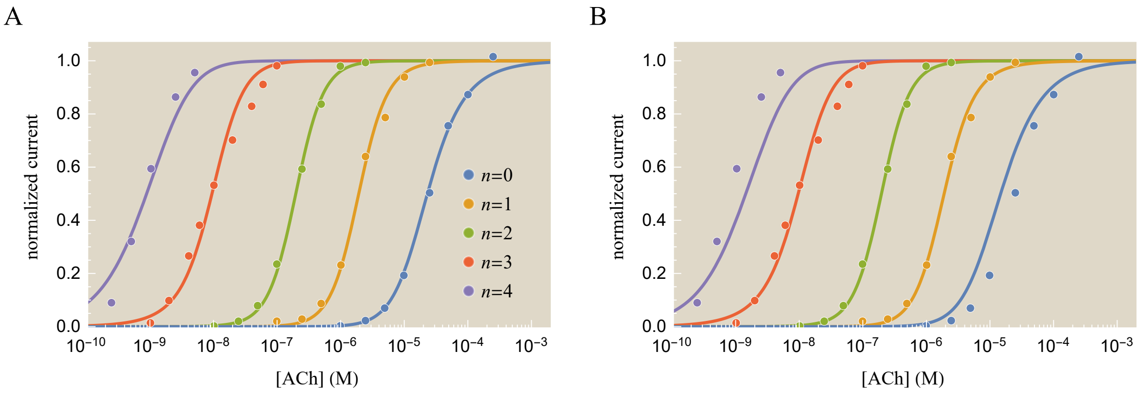


Figure S5.12: nAChR fits can be resolved by slightly perturbing the measured $\boldsymbol{\beta\epsilon}$ value. (A) If the experimentally measured value of $\beta\epsilon^{(0)} = -14.2$ for wild type nAChR is decreased to $\beta\epsilon^{(0)} = -18$, we can characterize all of the nAChR mutants ($R^2 > 0.99$) using a single set of parameters. (B) Even the very modest change to $\beta\epsilon^{(0)} = -16$ enables us to fit most of the data set well ($R^2 = 0.98$).

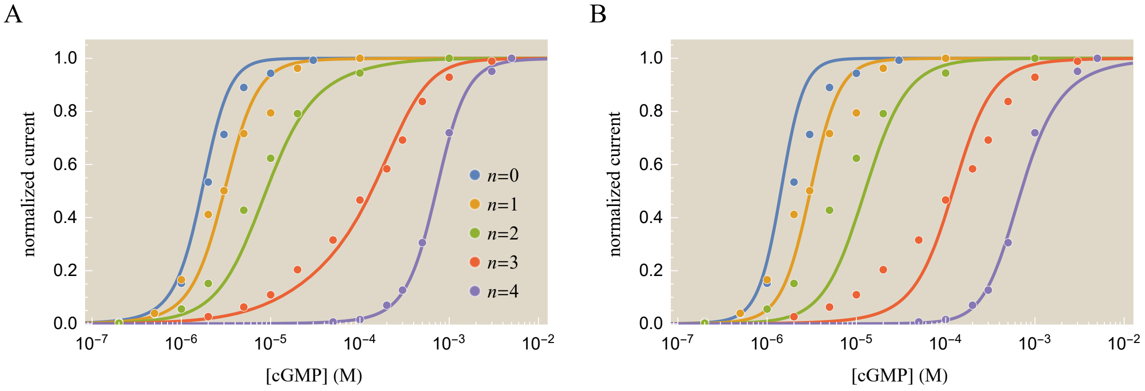


Figure S5.13: CNGA2 fits can also be resolved with slight changes to the measured $\boldsymbol{\beta\epsilon}$ value. (A) Increasing the experimentally measured value of $\beta\epsilon = -11$ to $\beta\epsilon = -5$ permits us to recoup a single set of parameters ($R^2 > 0.99$) for the entire class of mutants. (B) A more modest increase from $\beta\epsilon = -11$ to $\beta\epsilon = -8$ yields a poorer fit ($R^2 = 0.97$).

as $\beta\epsilon = -6$ for this same ion channel [14, 15].

We find that the full spectrum of CNGA2 ion channel mutants can be fit to a single set of thermodynamic parameters (K_O, K_C, K_O^*, K_C^*, and $\beta\epsilon$) when $\beta\epsilon = -6$, as shown in Fig. S5.13A (with $R^2 > 0.99$). Alternatively, using the value $\beta\epsilon = -9$ halfway between the experimental measurements yields markedly worse fits (with $R^2 = 0.97$), as shown in Fig. S5.13B.

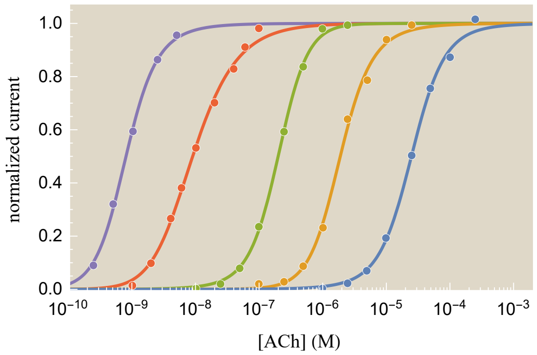


Figure S5.14: Extracting [EC_{50}] and h from the nAChR data. The individual nAChR data sets can be fit to the MWC model in order to interpolate between the data points and extract the best possible [EC_{50}] and h values. Note that each fit is very smooth around the midpoint where normalized current equals ½, which is the key region where both [EC_{50}] and h are computed.

S5.3.3 Fitting the nAChR Mutants with Non-Uniform K_O and K_C

In order to extract the [EC_{50}] and effective Hill coefficient h of an nAChR ion channel from experimental measurements (Fig. 5.3A), the individual data points must be connected in order to precisely infer where the normalized current reaches ½. This interpolation may be done in multiple ways, including connecting the data points with straight line segments or fitting the data to a sigmoidal function. The resulting [EC_{50}] and h values can then be compared to the predictions in Fig. 5.4 which were made while constraining all of the mutants to have the same K_O and K_C dissociation constants (which may have resulted in worse predictions for the characteristics of these mutants).

We chose to interpolate the nAChR data sets by fitting each mutant's data individually to the sigmoidal MWC response Eqs. 5.1 and 5.2. Fig. S5.14 shows how each data set is well fit by an individual MWC response, and that the behavior around the midpoint of each curve when normalized current equals to ½ is well aligned to the data, which gives us confidence that the corresponding [EC_{50}] and h values that are extracted from these curves (see Fig. 5.4C and D) will be precise. Note that the resulting best-fit parameters are not meaningful, as the sole purpose of this plot is to interpolate between the given data points in order to extract the best possible [EC_{50}] and h values.

S5.4 Predicting the behavior of mutants using the MWC model

This section is intended to explore two related questions. First, experiments on nAChR ion channels with single point mutations in different subunits hint at the possibility that each mutation incurs the same energetic cost, as described by the MWC $\epsilon^{(n)}$ parameter (see Eq. 5.20). In S5.4.1, we explore how well this hypothesis accords with the data and the predictive power that it grants the MWC model of nAChR. In S5.4.2, we turn to the CNGA2 ion channels where the opposite hypothesis holds true, namely, that the gating energy ϵ is unaltered by subunit mutations while the remaining MWC parameters are impacted by these mutations. We again examine how a subset of the CNGA2 mutant data captures the behavior of the entire class of mutants.

S5.4.1 nAChR

The wild type nAChR ion channel is characterized by the three MWC parameters K_O, K_C, and $\beta\epsilon^{(0)}$ (see Eqs. 5.1 and 5.2), all three of which can be fit from the wild type data set ($n = 0$ in Fig. 5.3A). We further postulate that the L251S mutations will only change the allosteric energy $\beta\epsilon^{(0)}$, leaving the ligand binding affinities K_O and K_C unchanged.

The nAChR data suggests that each L251S mutation increases the gating equilibrium by $\Delta\epsilon$ per mutation, so that $\beta\epsilon^{(n)} = \beta\epsilon^{(0)} + n\Delta\epsilon$. We aim to find to what extent this hypothesis holds true. Specifically, we note that after the wild type data set fixes K_O, K_C, and $\beta\epsilon^{(0)}$, using one additional data set can fix $\Delta\epsilon$, enabling us to extrapolate the $\beta\epsilon^{(n)}$ values for the remaining mutants. For example, in Fig. 5.9A of the main text, we used $\beta\epsilon^{(0)} = -23.7\ k_\text{B}T$ and $\beta\epsilon^{(4)} = -4.0\ k_\text{B}T$ to determine $\Delta\epsilon = -4.9\ k_\text{B}T$, from which we determined $\beta\epsilon^{(1)} = -18.8\ k_\text{B}T$, $\beta\epsilon^{(2)} = -13.9\ k_\text{B}T$, and $\beta\epsilon^{(3)} = -8.9\ k_\text{B}T$. The resulting predictions characterized the data sets for the $n = 1, 2, 3$ nAChR mutants remarkably well ($R^2 = 0.985$).

Note that this same procedure could work with any two nAChR data sets. For example, we could use the $n = 1$ mutant's data to determine the MWC parameters K_O, K_C, and $\beta\epsilon^{(1)}$, and then use the $n = 2$ data set to determine $\Delta\epsilon$ and extract the remaining $\beta\epsilon^{(n)}$ values. Fig. S5.15 demonstrates the resulting predictions when all ten possible input pairs are used to predict the remaining three mutant dose-response curves. The corresponding best-fit parameters are given in Table S5.4. In each case, the two input curves used to extract the MWC parameters are shown as solid curves, while the three predicted responses are shown as dashed lines.

	K_O (M)	K_C (M)	$\beta\epsilon^{(0)}$	$\beta\epsilon^{(1)}$	$\beta\epsilon^{(2)}$	$\beta\epsilon^{(3)}$	$\beta\epsilon^{(4)}$	R^2
Fig. S5.15A	0.3×10^{-9}	60×10^{-6}	-22.4	-17.8	**-13.2**	**-8.6**	**-4.1**	0.950
Fig. S5.15B	20×10^{-9}	80×10^{-6}	-14.0	**-9.5**	-5.0	**-0.5**	**4.0**	0.868
Fig. S5.15C	20×10^{-9}	80×10^{-6}	-13.7	**-8.6**	**-3.5**	1.6	**6.7**	0.839
Fig. S5.15D	0.1×10^{-9}	80×10^{-6}	-23.8	**-18.8**	**-13.9**	**-8.9**	-4.0	0.985
Fig. S5.15E	10×10^{-9}	10×10^{-6}	**-13.7**	-9.5	-5.4	**-1.2**	**2.9**	0.867
Fig. S5.15F	20×10^{-9}	10×10^{-6}	**-13.9**	-8.8	**-3.7**	1.4	**6.6**	0.839
Fig. S5.15G	0.1×10^{-9}	10×10^{-6}	**-23.8**	-18.8	**-13.9**	**-8.9**	-4.0	0.983
Fig. S5.15H	20×10^{-9}	200×10^{-3}	**-17.0**	**-10.9**	-4.8	1.4	**7.5**	0.729
Fig. S5.15I	0.1×10^{-9}	3×10^{-6}	**-25.0**	**-19.7**	-14.5	**-9.2**	-4.0	0.930
Fig. S5.15J	0.1×10^{-9}	10×10^{-9}	**-20.7**	**-16.6**	**-12.5**	-8.4	-4.3	0.063

Table S5.4: nAChR parameter predictions from two input data sets. Data sets from the two plain text $\beta\epsilon^{(n)}$ columns (shown as solid lines in their corresponding figures) were used to determine the K_O and K_C dissociation constants for the entire class of mutants and to linearly extrapolate the energies (bold text) of the remaining mutants using Eq. 5.20. R^2 indicates the goodness of fit for the three predicted curves (shown as dashed lines in the corresponding figures).

Most of the predictions do an especially good job of predicting the behavior of the intermediary $n = 1$, 2, and 3 mutants, while predictions for the two outer data sets $n = 0$ (wild type) and $n = 4$ are likely to be worse. This follows the general rule that interpolation – predicting values inside the domain of the training set – is more reliable than extrapolation. This suggests that when trying to make predictions for a similar family of mutants, it is most beneficial to acquire data for the extreme cases (i.e. the $n = 0$ and $n = 4$ data sets). In terms of the overall fit performance on the three unknown data sets in each of the ten plots, four of the fits have $R^2 > 0.9$ while four others have $0.9 > R^2 > 0.8$. This fit performance is improved if three or four input data sets are used to predict the remaining dose-response curves.

Interestingly enough, when these predictions fail (most notably in Fig. S5.15J), it occurs because the fitting captures the local details of (and noise in) the input data sets, which throws off the extrapolation to the remaining ion channel mutant. This concept is reminiscent of over-fitting in computer science. Indeed, it suggests that contrary to our intuition, using a more generalized model which has more degrees of freedom and is able to capture the tiny nuances of each individual data set even more precisely would do worse at predicting the global behavior of this class of mutants. In other words, having a coarse-grained model of the system with fewer parameters may provide a better opportunity to correctly predict protein behavior.

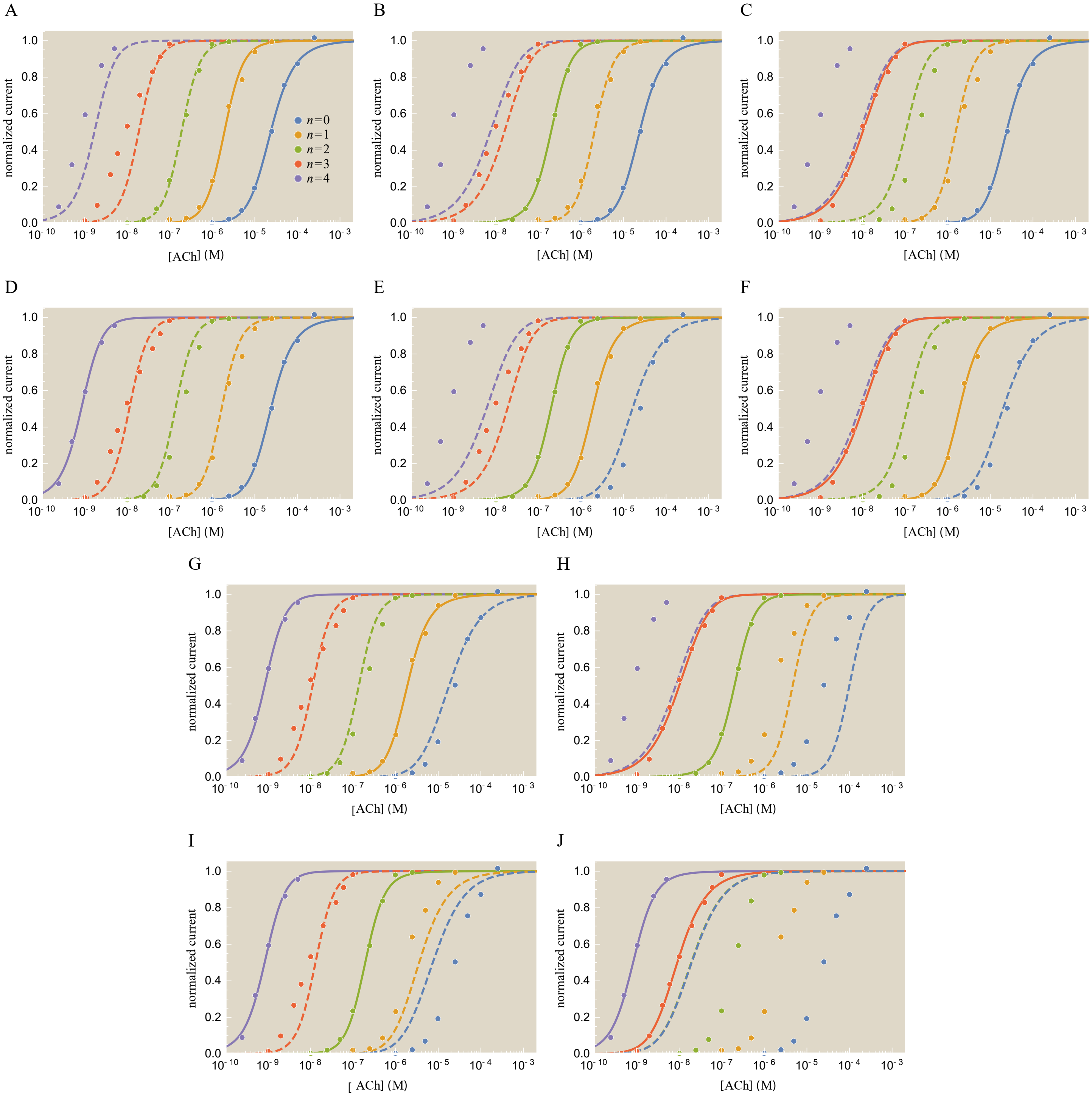


Figure S5.15: Predicting nAChR mutants using different training sets. The MWC parameters for the entire class of nAChR mutants can be fit from two data sets (solid lines). Using these parameters, the dose-response curves of the remaining three mutants can be predicted (dashed lines) without any further fitting. The best-fit parameters are listed in Table S5.4.

S5.4.2 CNGA2

The wild type CNGA2 ion channel has 4 identical subunits with ligand affinity K_{O} in the open state and K_{C} in the closed state. The free energy difference between the closed and open states is given by ϵ. A mutation was introduced in the ligand binding site of any subunit, which results in new dissociation constants K_{O}^* in the open state and K_{C}^* in the closed state, but which will leave the free energy difference ϵ unchanged. Once all of the MWC parameters are known, a CNGA2 mutant with n mutated subunit and $4 - n$ wild type subunits is fully described using Eq. 5.15 with $m = 4$.

One conceptually simple route to resolving the MWC parameters is to first fix the wild type parameters K_{O}, K_{C}, and ϵ using the wild type data set ($n = 0$) and then fix the two mutant dissociation constants K_{O}^* and K_{C}^* from the $n = 4$ data set. From these parameters, the intermediate mutants $n = 1$, 2, and 3 would all follow from Eq. 5.15. Yet, as in the case of nAChR, any two data sets could be used to fix the parameter values. In fact, in this system all five thermodynamic parameters (K_{O}, K_{C}, K_{O}^*, K_{C}^*, and ϵ) could be fit using a *single* data set from one of the $n = 1$, 2, or 3 mutants, since the dose-response curve Eq. 5.15 for such a mutant would contain all five parameters.

Fig. S5.16 shows the predictions (dashed lines) generated by fitting the MWC parameters to all possible input pairs (solid lines), with the best-fit parameters given in Table S5.5. As was found for the nAChR ion channels, the worst predictions resulted from data sets that are very close together (for example, when both input parameters came from $n = 2$, $n = 3$, or $n = 4$), which results in poor extrapolations for the remaining mutant data sets. Surprisingly, the prediction based on the $n = 0$ and $n = 4$ data set, which could be expected to be one of the best fits, was also poor. That said, the majority of the predictions were quite accurate ($R^2 > 0.96$), once again demonstrating the power of the simple statistical mechanical model we have employed.

References

[1] Mirzaev I, Gunawardena J. Laplacian Dynamics on General Graphs. Bull. Math. Biol. 2013;75(11):2118–2149. doi:10.1007/s11538-013-9884-8.

[2] Unwin N. Acetylcholine Receptor Channel Imaged in the Open State. Nature. 1995;373(6509):37–43. doi:10.1038/373037a0.

[3] Labarca C, Nowak MW, Zhang H, Tang L, Deshpande P, Lester HA. Channel

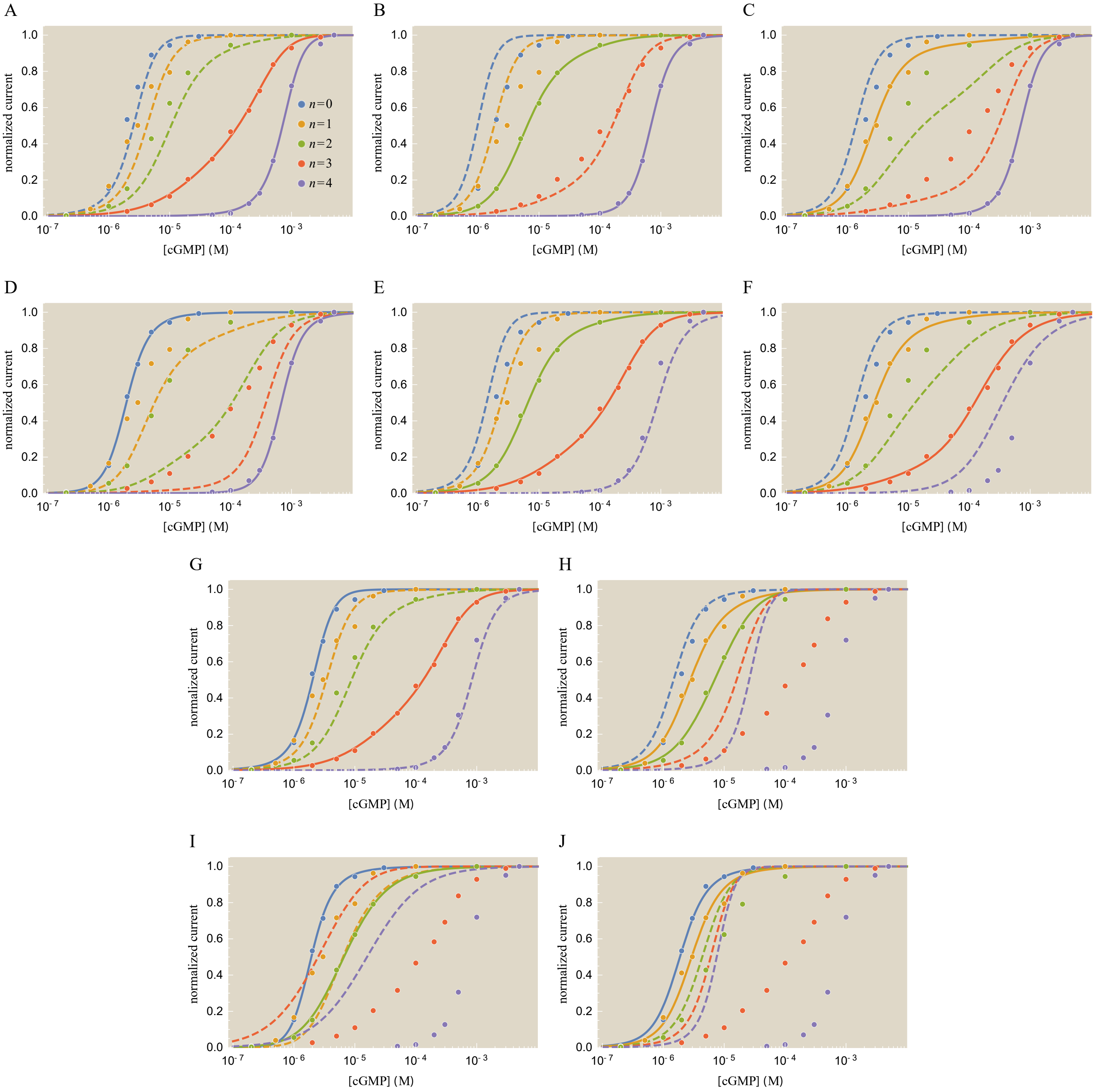


Figure S5.16: Predicting CNGA2 mutants using different training sets. As was found for nAChR, two data sets (solid lines) are sufficient to extract the MWC parameters for the whole class of CNGA2 mutants, which can then be used to extrapolate the responses of the remaining mutants (dashed lines). The best-fit parameters are listed in Table S5.5.

	K_{O} (M)	K_{C} (M)	K_{O}^{*} (M)	K_{C}^{*} (M)	$\beta\epsilon$	R^2
Fig. S5.16A	1.5×10^{-6}	35×10^{-6}	470×10^{-6}	70×10^{-3}	-3.6	0.983
Fig. S5.16B	0.3×10^{-6}	15×10^{-6}	180×10^{-6}	3×10^{-3}	-5.5	0.962
Fig. S5.16C	0.5×10^{-6}	6×10^{-6}	260×10^{-6}	5×10^{-3}	-4.6	0.962
Fig. S5.16D	0.3×10^{-6}	5×10^{-6}	120×10^{-6}	2×10^{-3}	-6.5	0.857
Fig. S5.16E	0.6×10^{-6}	20×10^{-6}	290×10^{-6}	2×10^{-3}	-4.3	0.982
Fig. S5.16F	0.5×10^{-6}	5×10^{-6}	60×10^{-6}	170×10^{-6}	-4.6	0.978
Fig. S5.16G	1×10^{-6}	30×10^{-6}	370×10^{-6}	4×10^{-3}	-3.9	0.990
Fig. S5.16H	0.5×10^{-6}	4×10^{-6}	15×10^{-6}	140×10^{-3}	-4.1	0.883
Fig. S5.16I	0.1×10^{-6}	3×10^{-6}	20×10^{-6}	20×10^{-6}	-10.2	0.640
Fig. S5.16J	0.5×10^{-6}	4×10^{-6}	3×10^{-6}	140×10^{-3}	-4.6	0.713

Table S5.5: CNGA2 parameter predictions from two input data sets. Two data sets (shown as solid lines in the corresponding figures) were used to determine the thermodynamic parameters for the entire class of mutants. R^2 indicates the goodness of fit for the three predicted curves (shown as dashed lines in the corresponding figures).

Gating Governed Symmetrically by Conserved Leucine Residues in the M2 Domain of Nicotinic Receptors. Nature. 1995;376(6540):514–516. doi:10.1038/376514a0.

[4] Wongsamitkul N, Nache V, Eick T, Hummert S, Schulz E, Schmauder R, et al. Quantifying the Cooperative Subunit Action in a Multimeric Membrane Receptor. Sci. Rep. 2016;6(1):20974. doi:10.1038/srep20974.

[5] Auerbach A. Thinking in Cycles: MWC Is a Good Model for Acetylcholine Receptor-Channels. J. Physiol. 2012;590(1):93–98. doi:10.1113/jphysiol.2011.214684.

[6] Gunawardena J. A Linear Framework for Time-Scale Separation in Nonlinear Biochemical Systems. PLoS ONE. 2012;7(5):e36321. doi:10.1371/journal.pone.0036321.

[7] Krashia P, Moroni M, Broadbent S, Hofmann G, Kracun S, Beato M, et al. Human $\alpha3\beta4$ Neuronal Nicotinic Receptors Show Different Stoichiometry If They Are Expressed in Xenopus Oocytes or Mammalian HEK293 Cells. PLoS ONE. 2010;5(10):e13611. doi:10.1371/journal.pone.0013611.

[8] George AA, Lucero LM, Damaj MI, Lukas RJ, Chen X, Whiteaker P. Function of Human $\alpha3\beta4\alpha5$ Nicotinic Acetylcholine Receptors Is Reduced by the $\alpha5$(D398N) Variant. J. Biol. Chem. 2012;287(30):25151–25162. doi:10.1074/jbc.M112.379339.

[9] Filatov GN, White MM. The Role of Conserved Leucines in the M2 Domain of the Acetylcholine Receptor in Channel Gating. Mol. Pharmacol. 1995;48(3):379–84.

[10] Martins BMC, Swain PS. Trade-Offs and Constraints in Allosteric Sensing. PLoS Comput. Biol. 2011;7(11):1–13. doi:10.1371/journal.pcbi.1002261.

[11] Marzen S, Garcia HG, Phillips R. Statistical Mechanics of Monod-Wyman-Changeux (MWC) Models. J. Mol. Biol. 2013;425(9):1433–1460. doi:10.1016/j.jmb.2013.03.013.

[12] Nayak TK, Purohit PG, Auerbach A. The Intrinsic Energy of the Gating Isomerization of a Neuromuscular Acetylcholine Receptor Channel. J. Gen. Physiol. 2012;139(5):349–58. doi:10.1085/jgp.201110752.

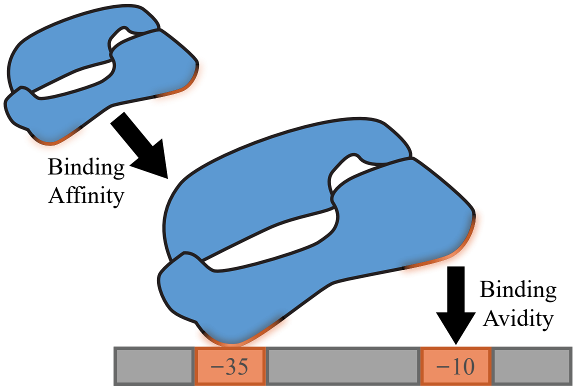
[13] Jackson MB. Kinetics of Unliganded Acetylcholine Receptor Channel Gating. Biophys. J. 1986;49(3):663–72. doi:10.1016/S0006-3495(86)83693-1.

[14] Nache V, Schulz E, Zimmer T, Kusch J, Biskup C, Koopmann R, et al. Activation of Olfactory-Type Cyclic Nucleotide-Gated Channels Is Highly Cooperative. J. Physiol. 2005;569(Pt 1):91–102. doi:10.1113/jphysiol.2005.092304.

[15] Tibbs GR, Goulding EH, Siegelbaum SA. Allosteric Activation and Tuning of Ligand Efficacy in Cyclic-Nucleotide-Gated Channels. Nature. 1997; 386(6625):612–615. doi:10.1038/386612a0.

Chapter 6

HOW THE AVIDITY OF RNA POLYMERASE BINDING TO THE -35/-10 PROMOTER SITES AFFECTS GENE EXPRESSION



Einav T, Phillips R. How the Avidity of Polymerase Binding to the -35/-10 Promoter Sites Affects Gene Expression. Proceedings of the National Academy of Sciences. *In press*

In 2016, Caltech Professor Pamela Bjorkman gave an amazing presentation that started an incredibly fun collaboration between our groups to analyze her synthetic HIV antibodies. Our goal was to design an optimal antibody capable of binding HIV with two arms, a significant improvement over the body's natural antibodies that can typically only bind HIV with only one arm. The quantity that we wanted to maximize, called avidity, quantifies the enhancement achieved by when two binding sites are involved. Inspired by that work, I wondered whether this same principle of avidity could play a role in our lab's area of expertise, namely, transcription. After some discussions with my labmate Suzy Beeler, I learned that the perfect data set had been published that very month. And the rest, as they say, is written below.

6.1 Abstract

Although the key promoter elements necessary to drive transcription in *Escherichia coli* have long been understood, we still cannot predict the behavior of arbitrary novel promoters, hampering our ability to characterize the myriad of sequenced regulatory architectures as well as to design new synthetic circuits. This work builds upon a beautiful recent experiment by Urtecho *et al.* who measured the gene expression of over 10,000 promoters spanning all possible combinations of a small set of regulatory

elements. Using this data, we demonstrate that a central claim in energy matrix models of gene expression – that each promoter element contributes independently and additively to gene expression – contradicts experimental measurements. We propose that a key missing ingredient from such models is the avidity between the -35 and -10 RNA polymerase binding sites and develop what we call a *multivalent model* that incorporates this effect and can successfully characterize the full suite of gene expression data. We explore several applications of this framework, namely, how multivalent binding at the -35 and -10 sites can buffer RNAP kinetics against mutations and how promoters that bind overly tightly to RNA polymerase can inhibit gene expression. The success of our approach suggests that avidity represents a key physical principle governing the interaction of RNA polymerase to its promoter.

6.2 Introduction

Promoters modulate the complex interplay of RNA polymerase (RNAP) and transcription factor binding that ultimately regulates gene expression. While our knowledge of the molecular players that mediate these processes constantly improves, more than half of all promoters in *Escherichia coli* still have no annotated transcription factors in RegulonDB [1] and our ability to design novel promoters that elicit a target level of gene expression remains limited.

As a step towards taming the vastness and complexity of sequence space, the recent development of massively parallel reporter assays has enabled entire libraries of promoter mutants to be simultaneously measured [2–4]. Given this surge in experimental prowess, the time is ripe to reexamine how well our models of gene expression can extrapolate the response of a general promoter.

A common approach to quantifying gene expression, called the *energy matrix model*, assumes that every promoter element contributes additively and independently to the total RNAP (or transcription factor) binding energy [3]. This model treats all base pairs on an equal footing and does not incorporate mechanistic details of RNAP-promoter interactions such as its strong binding primarily at the -35 and -10 binding motifs (shown in Fig. 6.1A). A newer method recently took the opposite viewpoint, designing an RNAP energy matrix that only includes the -35 element, -10 element, and the length of the spacer separating them [5], neglecting the sequence composition of the spacer or the surrounding promoter region.

Although these methods have been successfully used to identify important regulatory elements in unannotated promoters [6] and predict evolutionary trajectories [5], it is clear that there is more to the story. Even in the simple case of the highly-studied *lac* promoter, such energy matrices show systematic deviations from measured levels of gene expressions, indicating that some fundamental component of transcriptional regulation is still missing [7].

We propose that one failure of current models lies in their tacit assumption that every promoter element contributes independently to the RNAP binding energy. By naturally relaxing this assumption to include the important effects of avidity, we can push beyond the traditional energy matrix analysis in several key ways including: (*i*) We can identify which promoter elements contribute independently or cooperatively without recourse to fitting, thereby building an unbiased mechanistic model for systems that bind at multiple sites. (*ii*) Applying this approach to RNAP-promoter binding reveals that the -35 and -10 motifs bind cooperatively, a feature that we attribute to avidity. Moreover, we show that models that instead assume the -35 and -10 elements contribute additively and independently sharply contradict the available data. (*iii*) We show that the remaining promoter elements (the spacer, UP, and background shown in Fig. 6.1A) do contribute independently and additively to the RNAP binding energy and formulate the corresponding model for transcriptional regulation that we call a *multivalent model*. (*iv*) We use this model to explore how the interactions between the -35 and -10 elements can buffer RNAP kinetics against mutations. (*v*) We analyze a surprising feature of the data where overly-tight RNAP-promoter binding can lead to decreased gene expression. (*vi*) We validate our model by analyzing the gene expression of over 10,000 promoters in *E. coli* recently published by Urtecho *et al.* [8] and demonstrate that our framework markedly improves upon the traditional energy matrix analysis.

While this work focuses on RNAP-promoter binding, its implications extend to general regulatory architectures involving multiple tight-binding elements including transcriptional activators that make contact with RNAP (CRP in the *lac* promoter [9]), transcription factors that oligomerize (as recently identified for the *xylE* promoter [6]), and transcription factors that bind to multiple sites on the promoter (DNA looping mediated by the Lac repressor [10]). More generally, this approach of categorizing which binding elements behave independently (without resorting to fitting) can be applied to multivalent interactions in other biological contexts including novel materials, scaffolds, and synthetic switches [11, 12].

6.3 Results

6.3.1 The -35 and -10 Binding Sites give rise to Gene Expression that Defies Characterization as Independent and Additive Components

Decades of research have shed light upon the exquisite biomolecular details involved in bacterial transcriptional regulation via the family of RNAP σ factors [13]. In this work, we restrict our attention to the σ^{70} holoenzyme [8], the most active form under standard *E. coli* growth condition, whose interaction with a promoter includes direct contact with the -35 and -10 motifs (two hexamers centered roughly 10 and 35 bases upstream of the transcription start site), a spacer region separating these two motifs, an UP element just upstream of the -35 motif that anchors the C-terminal domain (αCTD) of RNAP, and the background promoter sequence surrounding these elements.

Urtecho *et al.* constructed a library of promoters composed of every combination of eight -35 motifs, eight -10 motifs, eight spacers, eight backgrounds (BG), and three UP elements (Fig. 6.1A) [8]. Each sequence was integrated at the same locus within the *E. coli* genome and transcription was quantified via DNA barcoding and RNA sequencing. One of the three UP elements considered was the absence of an UP binding motif, and this case will serve as the starting point for our analysis.

The energy matrix approach used by Urtecho *et al.* posits that every base pair of the promoter will contribute additively and independently to the RNAP binding energy [8], which by appropriately grouping base pairs is equivalent to stating that the free energy of RNAP binding will be the sum of its contributions from the background, spacer, -35, and -10 elements (see Appendix S6.1). Hence, the gene expression (GE) is given by the Boltzmann factor

$$\text{GE} \propto e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}}+E_{\text{-35}}+E_{\text{-10}})}. \tag{6.1}$$

Note that all E_js represent free energies (with an energetic and entropic component); to see the explicit dependence on RNAP copy number, refer to Appendix S6.1. Fitting the 32 free energies (one for each background, spacer, -35, and -10 element) and the constant of proportionality in Eq. 6.1 on 25% of the data enables us to predict the expression on the remainder of the $8 \times 8 \times 8 \times 8 = 4{,}096$ promoters (see Methods).

Fig. 6.2A demonstrates that Eq. 6.1 leads to a poor characterization of these promoters ($R^2 = 0.57$, parameter values listed in Appendix S6.2), suggesting that critical features of gene expression are missing from this model. One possible resolution is to assume that the level of gene expression saturates for very strong promoters at r_{max} and for very weak promoters at r_0 (caused by background noise or spurious transcription, see Appendix S6.2), namely,

$$\text{GE} = \frac{r_0 + r_{\text{max}} e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}}+E_{\text{-35}}+E_{\text{-10}})}}{1 + e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}}+E_{\text{-35}}+E_{\text{-10}})}}. \tag{6.2}$$

Since Eq. 6.2 still assumes that each promoter element contributes additively and independently to the total RNAP binding energy, it also makes sharp predictions that markedly disagrees with the data (see Appendix S6.3). Inspired by these inconsistencies, we postulated that certain promoter elements, most likely the -35 and -10 sites, may not contribute synergistically to RNAP binding.

To that end, we consider a model for gene expression shown in Fig. 6.1B where RNAP can separately bind to the -35 and -10 sites. RNAP is assumed to elicit a large level of gene expression r_{max} when fully bound but the smaller level r_0 when unbound or partially bound. Importantly, the Boltzmann weight of the fully bound state contains the free energy E_{int} representing the avidity of RNAP binding to the -35 and -10 sites. Physically, avidity arises because unbound RNAP binding to either the -35 or -10 sites gains energy but loses entropy, while this singly bound RNAP attaching at the other (-10 or -35) site again gains energy but loses much less entropy, as it was tethered in place rather than floating in solution. Hence we expect $e^{-\beta E_{\text{int}}} \gg 1$, and including this avidity term implies that RNAP no longer binds independently to the -35 and -10 sites.

Our coarse-grained model of gene expression neglects the kinetic details of transcription whereby RNAP transitions from the closed to open complex before initiating transcription. Instead, we assume that there is a separation of timescales between the fast process of RNAP binding/unbinding to the promoter and the other processes that constitute transcription. In the quasi-equilibrium framework shown in Fig. 6.1B, gene expression is given by the average occupancy of RNAP in each of its states, namely,

$$\text{GE} = \frac{r_0 + e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})}\left(r_0 e^{-\beta E_{\text{-35}}} + r_0 e^{-\beta E_{\text{-10}}} + r_{\text{max}} e^{-\beta(E_{\text{-35}}+E_{\text{-10}}+E_{\text{int}})}\right)}{1 + e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})}\left(e^{-\beta E_{\text{-35}}} + e^{-\beta E_{\text{-10}}} + e^{-\beta(E_{\text{-35}}+E_{\text{-10}}+E_{\text{int}})}\right)}. \tag{6.3}$$

We call this expression a multivalent model since it reduces to the energy matrix Eq. 6.1 (with constant of proportionality $r_{\text{max}} E^{-\beta E_{\text{int}}}$) in the limit where gene expression is negligible when the RNAP is not bound ($r_0 \approx 0$) and the promoter is sufficiently weak or the RNAP concentration is sufficiently small that polymerase is

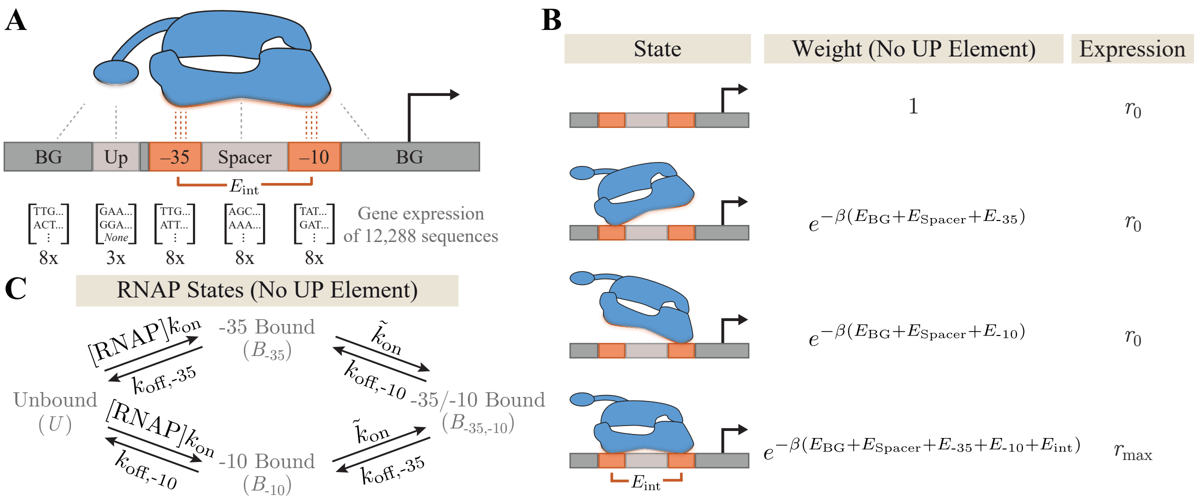

A
BG
Up
–35
Spacer
–10
BG
E_{int}
TTG... ACT... ⋮
GAA... GGA... None
TTG... ATT... ⋮
AGC... AAA... ⋮
TAT... GAT... ⋮
8x
3x
8x
8x
8x
Gene expression of 12,288 sequences
B
State
Weight (No UP Element)
Expression
1
r_0
$e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}}+E_{\text{-35}})}$
r_0
$e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}}+E_{\text{-10}})}$
r_0
$e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}}+E_{\text{-35}}+E_{\text{-10}}+E_{\text{int}})}$
r_{max}
E_{int}
C
RNAP States (No UP Element)
Unbound (U)
-35 Bound ($B_{\text{-35}}$)
-10 Bound ($B_{\text{-10}}$)
-35/-10 Bound ($B_{\text{-35,-10}}$)
$[\text{RNAP}]k_{\text{on}}$
$k_{\text{off,-35}}$
$[\text{RNAP}]k_{\text{on}}$
$k_{\text{off,-10}}$
$\tilde{k}_{\text{on}}$
$k_{\text{off,-10}}$
$\tilde{k}_{\text{on}}$
$k_{\text{off,-35}}$

Figure 6.1: The bivalent nature of RNAP-promoter binding. (A) Gene expression was measured for RNAP promoters comprising any combination of -35, -10, spacer, UP, and background (BG) elements. (B) When no UP element is present, RNAP makes contact with the promoter at the -35 and -10 sites giving rise to gene expression r_0 when unbound or partially bound and r_{max} when fully bound. (C) Having two binding sites alters the dynamics of RNAP binding. k_{on} represents the on-rate from unbound to partially bound RNAP and $\tilde{k}_{\text{on}}$ the analogous rate from partially to fully bound RNAP, while $k_{\text{off},j}$ denotes the unbinding rate from site j.

most often in the unbound state (so that the denominator ≈ 1). The background and spacer are assumed to contribute to RNAP binding in both the partially and fully bound states, an assumption that we rigorously justify in Appendix S6.4.

Fig. 6.2B demonstrates that the multivalent model Eq. 6.3 is better able to capture the system's behavior ($R^2 = 0.91$) while only requiring two more parameters (r_0 and E_{int}) than the energy matrix model Eq. 6.1. The sharp boundaries on the left and right represent the minimum and maximum levels of gene expression, $r_0 = 0.18$ and $r_{\text{max}} = 8.6$, respectively (see Appendix S6.5). The multivalent model predicts that the top 5% of promoters will exhibit expression levels of 7.6 (compared to 8.5 measured experimentally) while the weakest 5% of promoters should express at 0.2 (compared to the experimentally measured 0.1). In addition, this model quickly gains predictive power, as its coefficient of determination only slightly diminishes ($R^2 = 0.86$) if the model is trained on only 10% of the data and used to predict the remaining 90%.

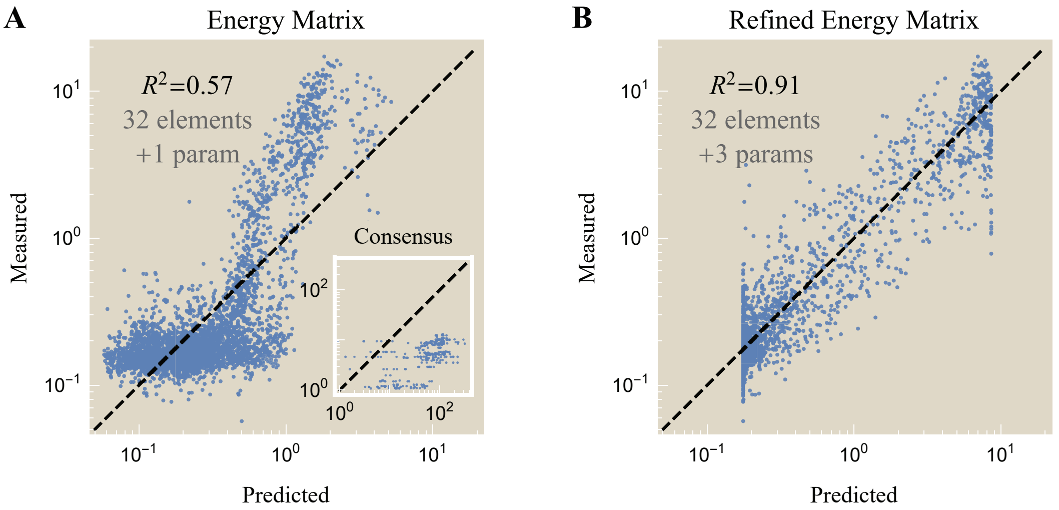


Figure 6.2: Gene expression of promoters with no UP element. Model predictions using (A) an energy matrix (Eq. 6.1) where the -35 and -10 elements independently contribute to RNAP binding and (B) a multivalent model (Eq. 6.3) where the two sites contribute cooperatively. Inset: The epistasis-free nature of the energy matrix model makes sharp predictions about the gene expression of the consensus -35 and -10 sequences that markedly disagree with the data. Parameter values given in Appendix S6.2.

6.3.2 Epistasis-Free Models of Gene Expression Lead to Sharp Predictions that Disagree with the Data

To further validate that the lower coefficient of determination of the energy matrix approach (Eq. 6.1) was not an artifact of the fitting, we can utilize the epistasis-free nature of this model to predict the gene expression of double mutants from that of single mutants. More precisely, denote the gene expression $\text{GE}^{(0,0)}$ of a promoter with the consensus -35 and -10 sequences (and any background or spacer sequence). Let $\text{GE}^{(1,0)}$, $\text{GE}^{(0,1)}$, and $\text{GE}^{(1,1)}$ represent promoters (with this same background and spacer) whose -35/-10 sequences are mutated/consensus, consensus/mutated, and mutated/mutated, respectively, where “mutated” stands for any non-consensus sequence. As derived in Appendix S6.4, the gene expression of these three later sequences can predict the gene expression of the promoter with the consensus -35 and -10 without recourse to fitting, namely,

$$\text{GE}^{(0,0)} = \text{GE}^{(1,1)} \frac{\text{GE}^{(0,1)}}{\text{GE}^{(1,1)}} \frac{\text{GE}^{(1,0)}}{\text{GE}^{(1,1)}}. \tag{6.4}$$

The inset in Fig. 6.2A compares the epistasis-free predictions (x-axis, right-hand side of Eq. 6.4) with the measured gene expression (y-axis, left-hand side of Eq. 6.4). These results demonstrate that the simple energy matrix formulation fails to capture the interaction between the -35 and -10 binding sites. While this calculation cannot readily generalize to the multivalent model since it exhibits epistasis, it is analytically tractable for weak promoters where the multivalent model displays a marked improvement over the energy matrix model (see Appendix S6.3).

6.3.3 RNAP Binding to the UP Element occurs Independently of the Other Promoter Elements

Having seen that the multivalent model (Eq. 6.3) can outperform the traditional energy matrix analysis on promoters with no UP element, we next extend the former model to promoters containing an UP element. Given the importance of the RNAP interactions with the -35 and -10 sites seen above, Fig. 6.3A shows three possible mechanisms for how the UP element could mediate RNAP binding. For example, the C-terminal could bind strongly and independently so that RNAP has three distinct binding sites. Another possibility is that the RNAP αCTD binds if and only if the -35 binding site is bound. A third alternative is that the UP element contributes additively and independently to RNAP binding (analogous to the spacer and background).

To distinguish between these possibilities, we analyze the correlations in gene expression between every pair of promoter elements (UP and -35, spacer and background, etc.) to determine the strength of their interaction. Each model in Fig. 6.3A will have a different signature: The top schematic predicts strong interactions between the -35 and -10, between the UP and -35, and between the UP and -10; the middle schematic would give rise to strong dependence between the -35 and -10 as well as between the UP and -10, while the UP and -35 elements would be perfectly correlated; the bottom schematic suggests that the UP elements will contribute independently of the other promoter elements.

This analysis, which we relegate to Appendix S6.4, demonstrates that the UP element is approximately independent of all other promoter elements ($R^2 \gtrsim 0.6$) as are the background and spacer, indicating that the bottom schematic in Fig. 6.3A characterizes the binding of the UP element. This leads us to the general form of transcriptional regulation by RNAP, shown in Eq. 6.5.

$$\mathrm{GE} = \frac{r_0 + e^{-\beta(E_{\mathrm{BG}}+E_{\mathrm{Spacer}}+E_{\mathrm{UP}})}\left(r_0 e^{-\beta E_{\text{-}35}} + r_0 e^{-\beta E_{\text{-}10}} + r_{\max} e^{-\beta(E_{\text{-}35}+E_{\text{-}10}+E_{\mathrm{int}})}\right)}{1 + e^{-\beta(E_{\mathrm{BG}}+E_{\mathrm{Spacer}}+E_{\mathrm{UP}})}\left(e^{-\beta E_{\text{-}35}} + e^{-\beta E_{\text{-}10}} + e^{-\beta(E_{\text{-}35}+E_{\text{-}10}+E_{\mathrm{int}})}\right)} \tag{6.5}$$

Fig. 6.3B demonstrates how the expression of all promoters containing one of the two UP elements combined with each of the eight background, spacer, -35, and

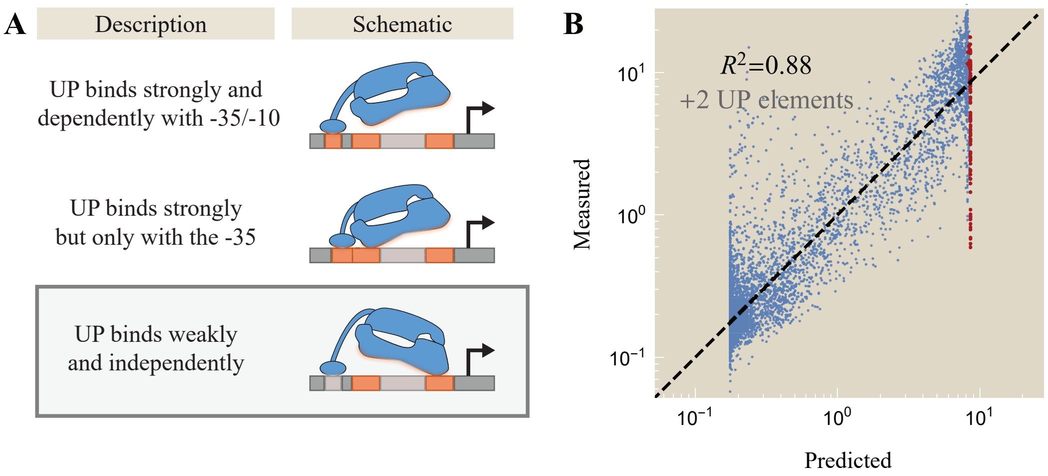


Figure 6.3: The interaction between RNAP and the UP element. (A) Possible mechanisms by which the RNAP C-terminal can bind to the UP element (orange segments represent strong binding comparable to the -35 and -10 motifs; gray segments represent weak binding comparable to the spacer and background). The data supports the bottom schematic (see Appendix S6.4). (B) The corresponding characterization of 8,192 promoters identical to those shown in Fig. 6.2 but with one of two UP binding motifs. Red points represent promoters with a consensus -35 and -10. Data was fit using the same parameters as in Fig. 6.2B and fitting the binding energies of the two UP elements (parameter values in Appendix S6.2).

-10 sequences ($2 \times 8^4 = 8{,}192$ promoters) closely matches the model predictions ($R^2 = 0.88$). We note that the large number of outliers on the left edge of the data may be attributable to noise, since more than half of all promoters have predicted gene expression < 0.2 (see Appendix S6.5). Remarkably, since we used the same free energies and gene expression rates from Fig. 6.2B, characterizing these 8,192 promoters only required two additional parameters (the free energies of the two UP elements). This result emphasizes how understanding each modular component of gene expression can enable us to harness the combinatorial complexity of sequence space.

6.3.4 Sufficiently Strong RNAP-Promoter Binding Energy can Decrease Gene Expression

Although the 12,288 promoters considered above are well characterized by Eq. 6.5 on average, the data demonstrate that the full mechanistic picture is more nuanced. For example, Urtecho *et al.* found that gene expression (averaged over all backgrounds and spacers) generally increases for -35/-10 elements closer to the consensus sequences [8]. In terms of the gene expression models studied above

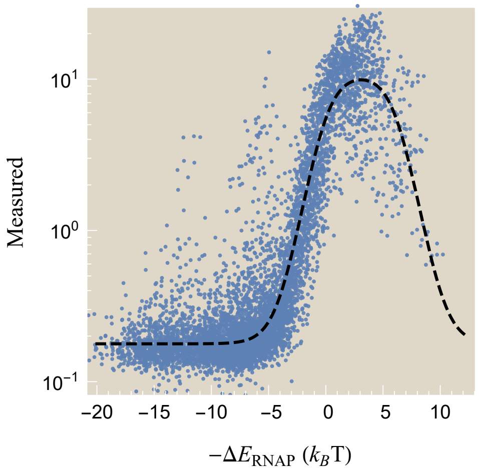


Figure 6.4: Gene expression is reduced when RNAP binds a promoter too tightly. Measured gene expression versus the inferred promoter strength ΔE_{RNAP} relative to the transcription initiation state $\Delta E_{\text{trans}} = -6.2\, k_B T$ (stronger promoters on the right). The dashed line shows the prediction of the multivalent model.

(Eqs. 6.1-6.3), promoters with fewer -35/-10 mutations have more negative free energies $E_{\text{-35}}$ and $E_{\text{-10}}$ leading to larger expression. Yet the strongest promoters with the consensus -35/-10 violated this trend, exhibiting *less* expression than promoters one mutation away. Thus, Urtecho *et al.* postulated that past a certain point, promoters that bind RNAP too tightly may inhibit transcription initiation and lead to decreased gene expression.

The promoters with a consensus -35/-10 are shown as red points in Fig. 6.3B, and indeed these promoters are all predicted to bind tightly to RNAP and hence express at the maximum level $r_{\text{max}} = 8.6$, placing them on the right-edge of the data. Yet depending on their UP, background, and spacer, many of these promoters exhibit significantly less gene expression then expected. Motivated by this trend, we posit that the state of transcription initiation can be characterized by a free energy ΔE_{trans} relative to unbound RNAP that competes with the free energy ΔE_{RNAP} between fully bound and unbound RNAP (see Appendix S6.5), analogous to a non-equilibrium boundary crossing problem with an effective barrier height ΔE_{trans} [14].

Assuming the rate of transcription initiation is proportional to the relative Boltzmann weights of these two states, the level of gene expression r_{max} in Eq. 6.5 will be modified to

$$\frac{r_{\text{max}} + r_0 e^{-\beta(\Delta E_{\text{RNAP}} - \Delta E_{\text{trans}})}}{1 + e^{-\beta(\Delta E_{\text{RNAP}} - \Delta E_{\text{trans}})}}. \tag{6.6}$$

As expected, this expression reduces to r_{max} for promoters that weakly bind RNAP

($e^{-\beta(\Delta E_{\text{RNAP}}-\Delta E_{\text{trans}})} \ll 1$) but decreases for strong promoters until it reaches the background level r_0 when RNAP is glued to the promoter and unable to initiate transcription. Upon reanalyzing the gene expression data with the inferred value $\Delta E_{\text{trans}} = -6.2\, k_B T$ (Appendix S6.5), we find that gene expression diminishes for the strongest RNAP-promoter free energies ΔE_{RNAP} as shown in Fig. 6.4 (stronger promoters to the right). This suggests that for sufficiently strong promoters, the rate limiting step in transcription initiation changes from RNAP dissociation to promoter escape.

6.3.5 The Bivalent Binding of RNAP Buffers its Interaction with DNA against Promoter Mutations

In this final section, we investigate how the avidity between the -35 and -10 sites changes the dynamics of RNAP binding. More specifically, we consider the effective dissociation constant governing RNAP binding when both the -35 and -10 sites are intact and compare it to the case where only one site is capable of binding. To simplify this discussion, we focus exclusively on the case of RNAP binding to the -35 and -10 motifs as shown in the rates diagram Fig. 6.1C, absorbing the effects of the background, spacer, and UP elements into these rates.

At equilibrium, there is no flux between the four RNAP states. We define the effective dissociation constant

$$K_D^{\text{eff}} = \frac{K_{\text{-35}} K_{\text{-10}}}{c_0 + K_{\text{-35}} + K_{\text{-10}}} \tag{6.7}$$

which represents the concentration of RNAP at which there is a 50% likelihood that the promoter is bound (see Appendix S6.6). $K_j = \frac{k_{\text{off},j}}{k_{\text{on}}}$ stands for the dissociation constant of free RNAP binding to the site j and $c_0 = \frac{\tilde{k}_{\text{on}}}{k_{\text{on}}} = [\text{RNAP}]e^{-\beta E_{\text{int}}}$ represents the increased local concentration of singly bound RNAP transitioning to the fully bound state (i.e., E_{int} and c_0 are the embodiments of avidity in the language of statistical mechanics and thermodynamics, respectively). Note that K_D^{eff} is a sigmoidal function of $K_{\text{-10}}$ with height $K_{\text{-35}}$ and midpoint at $K_{\text{-10}} = c_0 + K_{\text{-35}}$.

Fig. 6.5 demonstrates how the effective RNAP dissociation constant K_D^{eff} changes when mutations to the -10 binding motif alter its dissociation constant $K_{\text{-10}}$. When the -35 sequence is weak (dashed lines, $k_{\text{off,-35}} \to \infty$), $K_D^{\text{eff}} \approx K_{\text{-10}}$ signifying that RNAP binding relies solely on the strength of the -10 site. In the opposite limit where RNAP tightly binds to the -35 sequence (solid lines), the cooperativity c_0 and the dissociation constant $K_{\text{-35}}$ shift the curve horizontally and bound the effective

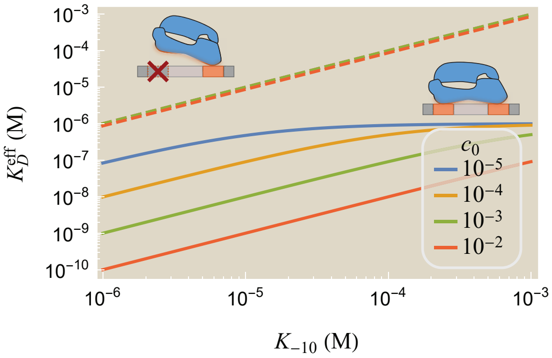


Figure 6.5: The dissociation between RNAP and the promoter. RNAP binding to a promoter with a strong (solid lines, $K_{\text{-}35} = 1\,\mu\text{M}$) or weak (dashed, $K_{\text{-}35} \to \infty$) -35 sequence. c_0 represents the local concentration of singly bound RNAP.

dissociation constant to $K_D^{\text{eff}} \leq K_{\text{-}35}$. This upper bound may buffer promoters against mutations, since achieving a larger effective dissociation constant would require not only wiping out the -35 site but in addition mutating the -10 site. Finally, in the case where the cooperativity c_0 is large, $K_D^{\text{eff}} \approx \frac{K_{\text{-}10}K_{\text{-}35}}{c_0}$ indicating that as soon as one site of the RNAP binds, the other is very likely to also bind, thereby giving rise to the multiplicative dependence on the two K_Ds.

To get a sense for how these numbers translate into physiological RNAP dwell times on the promoter, we note that the lifetime of bound RNAP is given by $\tau = \frac{1}{K_D^{\text{eff}} k_{\text{on}}}$ (see Appendix S6.6). Using $K_D^{\text{eff}} \approx 550\,\text{nM}$ for the *lac* promoter [15] and assuming a diffusion-limited on-rate $10^7 \frac{1}{\text{M}\cdot\text{s}}$ leads to a dwell time of 5 s, comparable to the measured dwell time of RNAP-promoter in the closed complex [16]. It would be fascinating if recently developed methods that visualize real-time single-RNAP binding events probed the dwell time of the promoter constructed by Urtecho *et al.* to see how well the predictions of the multivalent model match experiments [16].

6.4 Discussion

While high-throughput methods have enabled us to measure the gene expression of tens of thousands of promoters, they nevertheless only scratch the surface of the full sequence space. A typical promoter composed of 200 bp has 4^{200} variants (more than the number of atoms in the universe). Nevertheless, by understanding the principles governing transcriptional regulation, we can begin to cut away at this daunting complexity to design better promoters.

In this work, we analyzed a recent experiment by Urtecho *et al.* measuring gene expression of over 10,000 promoters in *E. coli* using the σ^{70} RNAP holoenzyme [8].

These sequences comprised all combinations of a small set of promoter elements, namely, eight -10s, eight -35s, eight spacers, eight backgrounds, and three UPs depicted in Fig. 6.1A, providing an opportunity to deepen our understanding of how these elements interact and to compare different quantitative models of gene expression.

We first analyzed this data using classic energy matrix models which posit that each promoter element contributes independently to the RNAP-promoter binding energy. As emphasized by Urtecho *et al.* and other groups, such energy matrices poorly characterize gene expression (Fig. 6.2A, $R^2 = 0.57$) and offer testable predictions that do not match the data (Appendix S6.3), mandating the need for other approaches [7, 8].

To meet this challenge, we first determined which promoter elements contribute independently to RNAP binding (Appendix S6.4). This process, which was done without recourse to fitting, demonstrated that the -35 and -10 elements bind in a concerted manner that we postulated is caused by avidity. In this context, avidity implies that when RNAP is singly bound to either the -35 or -10 sites, it is much more likely (compared to unbound RNAP) to bind to the other site, similar to the boost in binding seen in bivalent antibodies [17] or multivalent systems [12, 18, 19]. Surprisingly, we found that outside the -35/-10 pair, the other components of the promoter contributed independently to RNAP binding.

Using these findings, we developed a multivalent model of gene expression (Eq. 6.5) that incorporates the avidity of between the -35/-10 sites as well as the independence of the UP/spacer/background interactions. This model was able to characterize the 4,096 promoters with no UP element (Fig. 6.2B, $R^2 = 0.91$) and the 8,192 promoters containing an UP element (Fig. 6.3B, $R^2 = 0.88$). These results surpass those of the traditional energy matrix model (Fig. 6.2A, $R^2 = 0.57$), only requiring two additional parameters that could be experimentally determined (e.g., the interaction energy E_{int} arising from the -35/-10 avidity and the level of gene expression r_0 of a promoter with a scrambled -10 motif, a scrambled -35 motif, or with both motifs scrambled).

These promising findings suggest that determining which components bind independently is crucial to properly characterize multivalent systems. It would be fascinating to extend this study to RNAP with other σ factors [13] as well as to RNAP mutants with no αCTD or that do not bind at the -35 site [20, 21]. Our model would predict that polymerases in this last category with at most one strong binding site should conform to an energy matrix approach.

Quantitative frameworks such as the multivalent model explored here can deepen our understanding of the underlying mechanisms governing a system's behavior. For example, while searching for systematic discrepancies between our model prediction and the gene expression measurements, we found that promoters predicted to have the strongest RNAP affinity did not exhibit the largest levels of gene expression (thus violating a core assumption of nearly all models of gene expression that we know of). This led us to posit a characteristic energy for transcription initiation that reduces the expression of overly strong promoters (Fig. 6.4). In addition, we explored how having separate binding sites at the -35 and -10 elements buffers RNAP kinetics against mutations; for example, no single mutation can completely eliminate gene expression of a strong promoter with the consensus -35 and -10 sequence, since at least one mutation in both the -35 and -10 motifs would be needed (Fig. 6.5).

Finally, we end by zooming out from the particular context of transcription regulation and note that multivalent interactions are prevalent in all fields of biology [22], and our work suggests that differentiating between independent and dependent interactions may be key to not only characterizing overall binding affinities but to also understand the dynamics of a system [23]. Such formulations may be essential when dissecting the much more complicated interactions in eukaryotic transcription where large complexes bind at multiple DNA loci [24, 25] and more broadly in multivalent scaffolds and materials [11, 12].

6.5 Methods

Gene expression was measured as the ratio of RNA to DNA barcodes [8]. We fit both the energy matrix and multivalent models on 75% of the data and characterized the predictive power on the remaining 25%, repeating the procedure 10 times. The coefficient of determination R^2 was calculated for $y_{\text{data}} = \log_{10}(\text{gene expression})$ to prevent the largest gene expression values from dominating the result (see Appendix S6.2).

6.6 Acknowledgements

We thank Guillaume Urtecho and Sriram Kosuri for their marvelous manuscript that inspired this work and for their engaging discussion and collaborative spirit. We also thank Suzy Beeler, Vahe Galstyan, Peng (Brian) He, and Zofii Kaczmarek for many helpful discussions. This work was supported by the Rosen Center at Caltech and the National Institutes of Health through 1R35 GM118043-01 (MIRA).

References

[1] Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muniz-Rascado L, Garcia-Sotelo JS, et al. RegulonDB Version 9.0: High-Level Integration of Gene Regulation, Coexpression, Motif Clustering and Beyond. Nucleic. Acids. Res. 2016;44(D1):D133–D143. doi:10.1093/nar/gkv1156.

[2] Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D, Shendure J. High-Resolution Analysis of DNA Regulatory Elements by Synthetic Saturation Mutagenesis. Nat. Biotechnol. 2009;27(12):1173–1175. doi:10.1038/nbt.1589.

[3] Kinney JB, Murugan A, Callan CG, Cox EC. Using Deep Sequencing to Characterize the Biophysical Mechanism of a Transcriptional Regulatory Sequence. Proc. Natl. Acad. Sci. USA. 2010;107(20):9158–9163. doi:10.1073/Proc.Natl.Acad.Sci.USA.1004290107.

[4] Inoue F, Ahituv N. Decoding Enhancers Using Massively Parallel Reporter Assays. Genomics. 2015;106(3):159–164. doi:10.1016/J.YGENO.2015.06.005.

[5] Yona AH, Alm EJ, Gore J. Random Sequences Rapidly Evolve into De Novo Promoters. Nat. Commun. 2018;9(1):1530. doi:10.1038/s41467-018-04026-w.

[6] Belliveau NM, Barnes SL, Ireland WT, Jones DL, Sweredoski MJ, Moradian A, et al. Systematic Approach for Dissecting the Molecular Mechanisms of Transcriptional Regulation in Bacteria. Proc. Natl. Acad. Sci. USA. 2018; doi:10.1073/Proc.Natl.Acad.Sci.USA.1722055115.

[7] Forcier TL, Ayaz A, Gill MS, Jones D, Phillips R, Kinney JB. Measuring Cis-Regulatory Energetics in Living Cells Using Allelic Manifolds. elife. 2018;7. doi:10.7554/Elife.40618.

[8] Urtecho G, Tripp AD, Insigne KD, Kim H, Kosuri S. Systematic Dissection of Sequence Elements Controlling σ70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in *Escherichia coli*. Biochem. 2019;58(11):1539–1551. doi:10.1021/acs.biochem.7b01069.

[9] Kuhlman T, Zhang Z, Saier MH, Hwa T. Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. Proc. Natl. Acad. Sci. USA. 2007; 104(14):6043–6048. doi:10.1073/Proc.Natl.Acad.Sci.USA.0606717104.

[10] Boedicker JQ, Garcia HG, Johnson S, Phillips R. DNA Sequence-Dependent Mechanics and Protein-Assisted Bending in Repressor-Mediated Loop Formation. Phys. Biol. 2013;10(6):066005. doi:10.1088/1478-3975/10/6/066005.

[11] Varner CT, Rosen T, Martin JT, Kane RS. Recent Advances in Engineering Polyvalent Biological Interactions. Biomacromolecules. 2015;16(1):43–55. doi:10.1021/bm5014469.

[12] Yan GH, Wang K, Shao Z, Luo L, Song ZM, Chen J, et al. Artificial Antibody Created by Conformational Reconstruction of the Complementary-Determining Region on Gold Nanoparticles. Proc. Natl. Acad. Sci. USA. 2018;115(1):E34–E43. doi:10.1073/Proc.Natl.Acad.Sci.USA.1713526115.

[13] Feklístov A, Sharon BD, Darst SA, Gross CA. Bacterial Sigma Factors: A Historical, Structural, and Genomic Perspective. Annu. Rev. Microbiol. 2014; 68(1):357–376. doi:10.1146/annurev-micro-092412-155737.

[14] Roy S, Garges S, Adhya S. Activation and Repression of Transcription by Differential Contact: Two Sides of a Coin. J. Biol. Chem. 1998;273(23):14059–62. doi:10.1074/JBC.273.23.14059.

[15] Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, et al. Transcriptional Regulation by the Numbers: Models. Curr. Opin. Genet. Dev. 2005; 15(2):116–124. doi:10.1016/j.gde.2005.02.007.

[16] Wang F, Redding S, Finkelstein IJ, Gorman J, Reichman DR, Greene EC. The Promoter-Search Mechanism of *Escherichia coli* RNA Polymerase Is Dominated by Three-Dimensional Diffusion. Nat. Struct. Mol. Biol. 2013; doi:10.1038/nsmb.2472.

[17] Klein JS, Bjorkman PJ. Few and Far Between: How HIV May Be Evading Antibody Avidity. PLoS Pathog. 2010;6(5):e1000908. doi:10.1371/journal.ppat.1000908.

[18] Banjade S, Rosen MK. Phase Transitions of Multivalent Proteins Can Promote Clustering of Membrane Receptors. elife. 2014;3. doi:10.7554/Elife.04123.

[19] Huang J, Zeng X, Sigal N, Lund PJ, Su LF, Huang H, et al. Detection, Phenotyping, and Quantification of Antigen-Specific T Cells Using a Peptide-MHC Dodecamer. Proc. Natl. Acad. Sci. USA. 2016;113(13):E1890–7. doi:10.1073/Proc.Natl.Acad.Sci.USA.1602488113.

[20] Kumar A, Malloch RA, Fujita N, Smillie DA, Ishihama A, Hayward RS. The Minus 35-Recognition Region of *Escherichia coli* Sigma 70 Is Inessential for Initiation of Transcription at an Extended Minus 10 Promoter. J. Mol. Biol. 1993;232(2):406–418. doi:10.1006/JMBI.1993.1400.

[21] Minakhin L, Severinov K. On the Role of the *Escherichia coli* RNA Polymerase Sigma 70 Region 4.2 and Alpha-Subunit C-Terminal Domains in Promoter Complex Formation on the Extended -10 GalP1 Promoter. J. Biol. Chem. 2003;278(32):29710–8. doi:10.1074/jbc.M304906200.

[22] Gao A, Shrinivas K, Lepeudry P, Suzuki HI, Sharp PA, Chakraborty AK. Evolution of Weak Cooperative Interactions for Biological Specificity. Proc. Natl. Acad. Sci. USA. 2018;201815912. doi:10.1073/Proc.Natl.Acad.Sci. USA.1815912115.

[23] Stone JD, Artyomov MN, Chervin AS, Chakraborty AK, Eisen HN, Kranz DM. Interaction of Streptavidin-Based Peptide-MHC Oligomers (tetramers) with Cell-Surface TCRs. J. Immunol. 2011;187(12):6281–6290. doi:10.4049/jimmunol.1101734.

[24] Goardon N, Lambert JA, Rodriguez P, Nissaire P, Herblot S, Thibault P, et al. ETO2 Coordinates Cellular Proliferation and Differentiation During Erythropoiesis. EMBO J. 2006;25(2):357–66. doi:10.1038/sj.emboj.7600934.

[25] Levine M, Cattoglio C, Tjian R. Looping Back to Leap Forward: Transcription Enters a New Era. Cell. 2014;157(1):13–25. doi:10.1016/J.CELL.2014.02.009.

Chapter S6

SUPPLEMENTARY INFORMATION FOR HOW THE AVIDITY OF RNA POLYMERASE BINDING TO THE -35/-10 PROMOTER SITES AFFECTS GENE EXPRESSION

S6.1 The Energy Matrix Model

S6.1.1 Translating between an Energy Matrix with Base Pair Resolution and Promoter Element Resolution

In this section, we discuss how an energy matrix model with base-pair resolution can be translated into an equivalent model with the resolution of promoter elements. The former model purports that the RNAP-promoter binding energy is composed of independent and linearly additive contributions from each base pair. More precisely, if at position j the base b_j (either A, T, C, or G) contributes a free energy $E_j^{(b_j)}$ to RNAP binding, then the total free energy of binding is given by $\sum_j E_j^{(b_j)}$ as shown in Fig. S6.1.

By breaking this sum up over the positions j demarking the -35 ($-35 \leq j \leq -30$), spacer ($-29 \leq j \leq -13$), -10 ($-12 \leq j \leq -7$), UP ($-59 \leq j \leq -38$; where "no UP" used a random sequence that did not enhance gene expression), and background (all the remaining base pairs between $-120 \leq j < 30$) elements, we achieve an energy matrix model where the free energies E_{BG}, $E_{\text{-35}}$, E_{Spacer}, and $E_{\text{-10}}$ represent the sum of all base pair contributions of the particular sequence considered. For simplicity, the UP element is not explicitly drawn in the figure.

As shown for two sample sequences in Fig. S6.1, modifying the -35 sequence while keeping the rest of the promoter unchanged leads to a different $E_{\text{-35}}$ but keeps E_{BG}, E_{Spacer}, and $E_{\text{-10}}$ unchanged. The expression of the full suite of 12,288 promoters studied in this work can be determined from the free energies of the three UP elements and the eight backgrounds, spacers, -35s, and -10s.

S6.1.2 Characterizing the Dependence of Gene Expression on RNAP Copy Number

In this section, we explicitly write the dependence of RNAP copy number embedded within the free energies of RNAP binding in Eqs. 6.1 and 6.2, thereby making contact with previous models of gene regulation [1]. To that end, we consider P RNAP

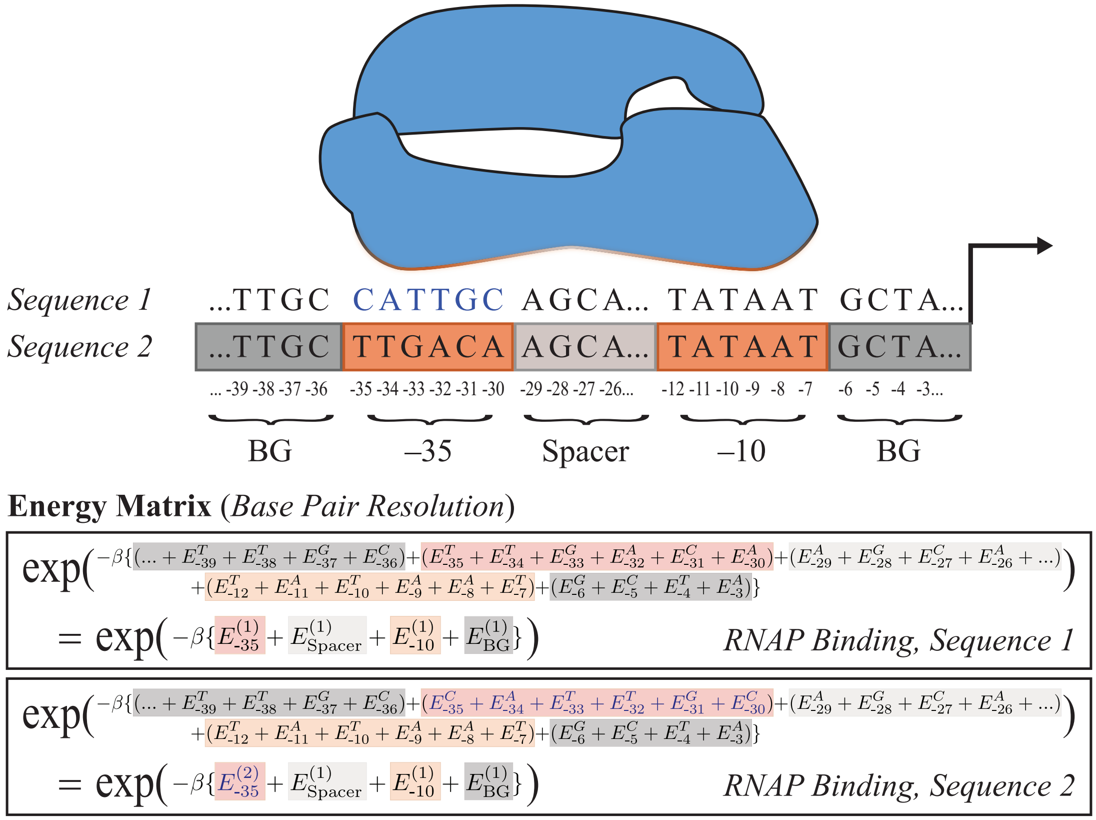

Sequence 1
...TTGC CATTGC AGCA... TATAAT GCTA...
Sequence 2
...TTGC TTGACA AGCA... TATAAT GCTA...
... -39 -38 -37 -36
-35 -34 -33 -32 -31 -30
-29 -28 -27 -26...
-12 -11 -10 -9 -8 -7
-6 -5 -4 -3...
BG
–35
Spacer
–10
BG
Energy Matrix (Base Pair Resolution)
$\exp\left(-\beta\{(\ldots+E^T_{-39}+E^T_{-38}+E^G_{-37}+E^C_{-36})+(E^T_{-35}+E^T_{-34}+E^G_{-33}+E^A_{-32}+E^C_{-31}+E^A_{-30})+(E^A_{-29}+E^G_{-28}+E^C_{-27}+E^A_{-26}+\ldots)+(E^T_{-12}+E^A_{-11}+E^T_{-10}+E^A_{-9}+E^A_{-8}+E^T_{-7})+(E^G_{-6}+E^C_{-5}+E^T_{-4}+E^A_{-3})\}\right)$
$=\exp\left(-\beta\{E^{(1)}_{-35}+E^{(1)}_{\text{Spacer}}+E^{(1)}_{-10}+E^{(1)}_{\text{BG}}\}\right)$
RNAP Binding, Sequence 1
$\exp\left(-\beta\{(\ldots+E^T_{-39}+E^T_{-38}+E^G_{-37}+E^C_{-36})+(E^C_{-35}+E^A_{-34}+E^T_{-33}+E^T_{-32}+E^G_{-31}+E^C_{-30})+(E^A_{-29}+E^G_{-28}+E^C_{-27}+E^A_{-26}+\ldots)+(E^T_{-12}+E^A_{-11}+E^T_{-10}+E^A_{-9}+E^A_{-8}+E^T_{-7})+(E^G_{-6}+E^C_{-5}+E^T_{-4}+E^A_{-3})\}\right)$
$=\exp\left(-\beta\{E^{(2)}_{-35}+E^{(1)}_{\text{Spacer}}+E^{(1)}_{-10}+E^{(1)}_{\text{BG}}\}\right)$
RNAP Binding, Sequence 2

Figure S6.1: An energy matrix model with base pair resolution translates into an energy matrix model with promoter element resolution. Each promoter element contributes to RNAP binding with free energy given by the sum of its free energies from its base pairs. The two sample sequences shown only differ in their -35 sequence (highlighted blue in Sequence 1), resulting in different values of $E^{(1)}_{-35}$ and $E^{(2)}_{-35}$ but the same free energies for the remaining promoter elements.

molecules that are free to bind anywhere along a bacterial genome with N_{NS} non-specific base pairs (i.e., potential RNAP binding sites outside of our promoter of interest). Let $\Delta\epsilon$ be the average energy difference between RNAP bound to the specific promoter versus at any other location along the genome. By definition, the free energy of RNAP binding considered in this work is given by both the entropic and energetic contributions of this binding, namely,

$$e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}}+E_{\text{-35}}+E_{\text{-10}})} \equiv \frac{P}{N_{\text{NS}}} e^{-\beta\Delta\epsilon} = e^{-\beta\left(\Delta\epsilon - k_B T \log(P/N_{\text{NS}})\right)}. \qquad \text{(S6.1)}$$

Because the gene expression for each promoter generated by Urtecho *et al.* was measured under the same experimental condition, the RNAP copy number is consistent across all constructs, and hence the constant $k_BT \log\left(P/N_{\text{NS}}\right)$ can be absorbed into the free energies. If these measurements were repeated under experimental conditions where the RNAP copy number is halved ($P \to \frac{P}{2}$), the total free energy of RNAP binding considered in this work would need to be correspondingly modified from $\left(E_{\text{BG}} + E_{\text{Spacer}} + E_{\text{-35}} + E_{\text{-10}}\right) \to \left(E_{\text{BG}} + E_{\text{Spacer}} + E_{\text{-35}} + E_{\text{-10}} + k_B T \log 2\right)$.

S6.2 Model Fitting and Parameter Values

The energy matrix model (Eq. 6.1) was solved as a least-squares problem that only fit the promoters in Fig. 6.2A with no UP element. The multivalent model (Eq. 6.5) was fit using nonlinear regression on promoter sequences with and without an UP element in order to obtain a single, self-consistent set of parameters capable of capturing the data in Fig. 6.2B and Fig. 3B. The fitting for both models is presented in a supplementary Mathematica notebook available online (doi: 10.22002/D1.1242).

The coefficient of determination R^2 was calculated for $y_{\text{data}} = \log_{10}(\text{gene expression})$ to prevent the largest gene expression values from dominating the result. We trained both the energy matrix and multivalent models on 75% of the data and characterized the predictive power on the remaining 25%, repeating the procedure 10 times. The exact form used was

$$R^2 = 1 - \frac{\sum_{j=1}^{N} \left(y_{\text{data}}^{(j)} - y_{\text{predicted}}^{(j)}\right)^2}{\sum_{j=1}^{N} \left(y_{\text{data}}^{(j)} - \bar{y}_{\text{data}}\right)^2} \tag{S6.2}$$

where $y_{\text{predicted}}$ is the vector of the N measurements of $\log_{10}(\text{gene expression})$ predicted by the model and $\bar{y}_{\text{data}} = \frac{1}{N} \sum_{j=1}^{N} y_{\text{data}}^{(j)}$ is the average of the logarithmic gene expression data. In this form, the R^2 represents the fraction of variance in the measured gene expression data that arises from the variance in the predicted gene expression data. To test the predictive power of each model, we also trained both models on only 10% of the data and used it to predict the gene expression of the remaining 90% of promoters. We found that the coefficient of determination R^2 only slightly decreased from 0.57 → 0.54 for the energy matrix model and from 0.91 → 0.86 for the multivalent model when fitting on this much smaller training set, demonstrating that these models require no more than a thousand promoters to reach their full predictive power.

Table S6.1 shows the parameter values inferred by the energy matrix model (Fig. 6.2A) and multivalent model (Figs. 6.2B and 6.3B). Due to the large number of parameters involved, both models exhibit parameter degeneracy [2] where disparate sets of parameters yield nearly identical results. For example, all of the free energies of the spacer elements can be increased by an arbitrary amount provided that the free energies of all background elements are decreased by this same amount (with similar degeneracy holding between other pairs of promoter elements). To circumvent this degeneracy, one -35, one spacer, one -10, one UP, and one background element (denoted by asterisks in Table S6.1) were fixed to their corresponding value in the energy matrix model, and as such, the parameters below may not represent the binding energies of the promoter elements, but rather only one possible embodiment of these values.

We point out that our model coarse-grains kinetic details of transcription (e.g., transcription initiation, elongation, transcriptional bursting) into the levels of gene expression r_j shown in Fig. 6.1B. Modifying the promoter sequence (i.e., considering different spacers or backgrounds) may well change these rates, although our model assumes that such changes only affect the RNAP-promoter binding affinity. If experiments measure the changes in these kinetic rates, they could either be incorporated into the r_j or into an expanded model that explicitly takes these steps of transcription into account [3].

The small but nonzero r_0 term in our model (Fig. 6.1B) represents the background level of gene expression arising from promoters that lack an RNAP-binding site. Urtecho *et al.* measured 500 negative controls, sequences from the *E. coli* genome that have no promoter or RNA-seq activity, whose expression was nonzero and centered around 0.15 (see Fig. 6.2E of Ref. [4]), comparable to our inferred $r_0 = 0.18$ value. This nonzero expression may arise from instrumental noise or spurious transcription, and we elected to model it using a nonzero r_0 rather than background subtracting it in order to present the data on a log-scale (upon background subtraction, some gene expression measurements would be negative due to experimental noise which would have precluded log-plots). We note that log-fitting is likely to more accurately portray how gene expression proceeds in the cell, since most endogenous promoters exhibit low gene expression while synergistic effects between promoter elements can play a significant role; in contrast, linear fitting would downplay the importance of all but the strongest promoters.

Lastly, we note from Urtecho *et al.* that the UP elements were named because they increased transcription by 136-fold and 326-fold *in vivo* relative to the physiological *rrnb* P1 UP element. Thus, it follows that the free energy of the 326-fold UP should be smaller than that of the 136-fold UP which should be smaller than the free energy of having no UP element, as seen in Table S6.1. Additionally, we point out that all spacer elements are 17 bp long; RNAP binding is highly dependent on this length, and promoters with longer or shorter spacers may influence the -35 and -10 binding free energies. Lastly, the sequence composition of all spacers and backgrounds is given in Ref. [4].

S6.3 Comparing the Energy Matrix and Refined Energy Matrix Models of Gene Expression

S6.3.1 An Epistasis-Free Energy Matrix Model with Saturation does not Capture the Trends in Gene Expression Exhibited by the Data

As shown in Fig. 6.2A, the simplest model where gene expression is proportional to $e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}}+E_{\text{-35}}+E_{\text{-10}})}$ (in the absence of an UP element) fails to characterize the data ($R^2 = 0.57$). In contrast, the multivalent model in Fig. 6.2B quantitatively matches the behavior of the spectrum of promoters ($R^2 = 0.91$). Thus, it behooves us to examine what properties of the latter model are necessary to achieve this concordance with the data.

To that end, we consider an intermediate model where gene expression is given by

$$\text{GE} = \frac{r_0 + \tilde{r}_{\text{max}} e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}}+E_{\text{-35}}+E_{\text{-10}})}}{1 + e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}}+E_{\text{-35}}+E_{\text{-10}})}} \tag{S6.3}$$

where r_0 represents the minimum level of gene expression in the absence of RNAP binding, $\tilde{r}_{\text{max}}$ denotes the amount of gene expression when RNAP is fully bound to the promoter, and the E_j represent the free energy contribution of the promoter element j. Note that this model represents the limit of a very strong interaction energy in Eq. 6.3 ($e^{-\beta E_{\text{int}}} \gg 1$ with $\tilde{r}_{\text{max}} = r_{\text{max}} e^{-\beta E_{\text{int}}}$) where RNAP is either unbound or fully bound to the promoter.

Fig. S6.2A demonstrates that the data is well characterized by Eq. S6.3 ($R^2 = 0.91$). Therefore, one key feature missing from the simplest energy matrix model description Eq. 6.1 was that gene expression will saturate once RNAP binding becomes sufficiently strong (or, mathematically, that the denominator $1 + e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}}+E_{\text{-35}}+E_{\text{-10}})}$ must include the RNAP binding term). Note that the results of this energy matrix model with saturation are nearly identical to the results of the multivalent model in Fig. 6.2B. Indeed, since the inferred interaction energy $E_{\text{int}} = -6.3\, k_B T$ between the -35 and -10 sites is large and negative (see Table S6.1), it is not surprising that the two models produce similar results for the majority of promoters.

Intuitively, the difference between these two models will emerge in their predictions for promoters with weak expression. As we will show below, the energy matrix model with saturation (Eq. S6.3) is epistasis-free: given the gene expression of any initial promoter and two mutants of that promoter, we can predict the expression of the double mutant. If, for example, the initial promoter exhibits weak gene expression and the two mutants exhibit a medium level of gene expression, then the double mutant would be predicted to exhibit a large amount gene expression. As will be explained below, the resulting predictions shown in Fig. S6.2B are highly damning. On the other hand the multivalent model (Eq. 6.3) predicts a more complex relationship between these four promoters, and in the Appendix S6.3.2 we examine an analytically tractable limit to show that this model better recapitulates the gene expression measurements.

We proceed by utilizing the epistasis-free nature of Eq. S6.3. A key feature of the following analysis is that it will not require any model fitting, and hence for the remainder of this Appendix we proceed as if we have no knowledge of the parameter values in Table S6.1. To begin, we approximate the values of $r_0 \approx 0.2$ and $\tilde{r}_{\text{max}} \approx 10$ from the gene expression data (the minimum and maximum y-values in Fig. S6.2A, averaging by eye to account for noise). These two values, together with the gene expression measurements for every construct, will be sufficient to make our epistasis-free predictions without explicitly determining any of the E_j.

As in the main text, denote the gene expression $\text{GE}^{(0,0)}$ of a promoter with the consensus -35 and -10 sequences (and any background or spacer sequence). Let $\text{GE}^{(1,0)}$, $\text{GE}^{(0,1)}$, and $\text{GE}^{(1,1)}$ represent promoters (with this same background and spacer) whose -35/-10 sequences are mutated/consensus, consensus/mutated, and mutated/mutated, respectively. Eq. S6.3 can be inverted to determine

$$e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}}+E_{\text{-35}}+E_{\text{-10}})} = \frac{\text{GE} - r_0}{\tilde{r}_{\text{max}} - \text{GE}} \equiv f(\text{GE}) \qquad \text{(S6.4)}$$

for the double mutant with $\text{GE}^{(1,1)}$ as well as the two singly mutated promoters with $\text{GE}^{(1,0)}$ and $\text{GE}^{(0,1)}$, where we have defined the function f for convenience. Importantly, since the -35 and -10 binding energies additively and independently contribute to the RNAP-promoter free energy, the left-hand side of Eq. S6.4 for the unmutated construct is given by $f(\text{GE}^{(1,1)})\frac{f(\text{GE}^{(0,1)})}{f(\text{GE}^{(1,1)})}\frac{f(\text{GE}^{(1,0)})}{f(\text{GE}^{(1,1)})}$ (exactly analogous to Eq. 6.4 for the simple energy matrix model). Therefore, its gene expression is predicted to be

$$\text{GE}^{(0,0)} = \frac{r_0 + \tilde{r}_{\text{max}} f(\text{GE}^{(1,1)})\frac{f(\text{GE}^{(0,1)})}{f(\text{GE}^{(1,1)})}\frac{f(\text{GE}^{(1,0)})}{f(\text{GE}^{(1,1)})}}{1 + f(\text{GE}^{(1,1)})\frac{f(\text{GE}^{(0,1)})}{f(\text{GE}^{(1,1)})}\frac{f(\text{GE}^{(1,0)})}{f(\text{GE}^{(1,1)})}}. \qquad \text{(S6.5)}$$

Fig. S6.2B shows the results of these epistasis-free predictions. Because Eq. S6.5 applies to *any* pairs of -35 and -10 elements with the same BG and spacer, there

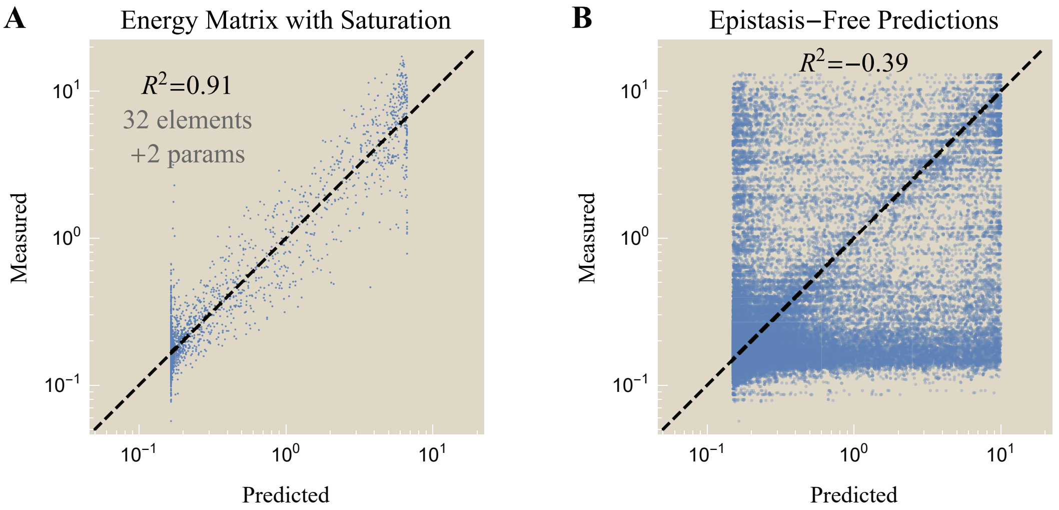


Figure S6.2: Gene expression represented by an energy matrix model with saturation. (A) Characterization of the same promoters as in Fig. 6.2 using the energy matrix model with saturation (Eq. S6.3) with essentially identical fit quality as the multivalent model. (B) Since this model assumes that the RNAP-promoter binding energy is epistasis-free (with the -35 and -10 binding sites contributing additively and independently to the RNAP binding energy), the gene expression of double mutants can be predicted from the expression of single mutants without resorting to fitting (Eqs. S6.4 and S6.5). The large deviations demonstrate that the energy matrix with saturation cannot characterize the gene expression of these constructs.

is a combinatorial explosion of predictions, providing a solid test for this model. As can be seen, aside from the plethora of data points with correctly-predicted low gene expression in the bottom-left corner of the plot, there are large swathes of data points that do not fall on the expected diagonal line, indicating that the epistasis-free prediction in Eq. S6.3 cannot accurately capture the gene expression of the constructs considered here. In the next section, we show that the multivalent model is better equipped to characterize these cases. Notably, these results indicate that although a model may fit the majority of data on average (as in Fig. S6.2A), it may nevertheless make spurious predictions. Such hidden gems may go unnoticed when a pure-fitting mentality is applied to the wealth of data that is becoming increasingly easy to generate.

S6.3.2 A Multivalent Model outperforms the Energy Matrix Model in the Limit of a Weak -35 or Weak -10 RNAP Binding Site

In section S6.3.1, we showed that an energy matrix model with saturation Eq. S6.3 is epistasis-free and hence makes sharp predictions that are inconsistent with the data (Fig. S6.2B). In this section, we consider the multivalent model Eq. 6.3 where binding to the -35 and -10 sites is no longer independent. Because this latter model exhibits epistasis, we will restrict our analysis to the limit of weak promoters with no UP element where we can approximate the multivalent model and compare its results to the energy matrix model with saturation. As before, we proceed without referencing the parameter values in Table S6.1 to emphasize that this analysis can be done without recourse to fitting.

We define $\text{GE}^{(1,1)}$, $\text{GE}^{(1,0)}$, $\text{GE}^{(0,1)}$, and $\text{GE}^{(0,0)}$ as in section S6.3.1, but we will restrict our attention to promoters where the original sequence exhibits low gene expression ($\text{GE}^{(1,1)} \lesssim 0.25$) and the two mutants exhibit medium gene expression ($0.25 \lesssim \text{GE}^{(1,0)}, \text{GE}^{(0,1)} \lesssim 1.0$). For such cases, we expect that the predicted gene expression $\text{GE}^{(0,0)}$ of the double mutant will be larger in the multivalent model (Eq. 6.3) than the energy matrix model with saturation (Eq. S6.3) due to the avidity between the -35 and -10 sites. In other words, the multivalent model acknowledges that the -35 and -10 sites bolster each other and consequently predicts larger gene expression when both sites exhibit even a moderate capability of binding.

As discussed in section S6.3.1, $\text{GE}^{(0,0)}$ is exactly given by Eq. S6.5 in the energy matrix model with saturation. Applying that result to the present case of weak promoters ($\text{GE}^{(1,1)} \lesssim 0.25$) with medium-strength singly mutants ($0.25 \lesssim \text{GE}^{(1,0)}, \text{GE}^{(0,1)} \lesssim 1.0$), Fig. S6.3A shows that this model generally underestimates the gene expression of these promoters. This serves as a promising indicator that the avidity of RNAP binding is missing from such an approach.

We next turn to the more complex multivalent model. Because RNAP exhibits epistasis within this framework, the relationship between gene expression is more complex and hence we only roughly approximate $\text{GE}^{(0,0)}$. To that end, it behooves us to generalize the levels of gene expression in Fig. 1(B) so that RNAP bound only at the -35 site leads to a gene expression level of $r_{\text{-35}}$ while RNAP bound only at the -10 site elicits $r_{\text{-10}}$ gene expression (satisfying $r_0 < r_{\text{-35}}, r_{\text{-10}} \leq r_{\text{max}}$), leading to

$$\text{GE} = \frac{r_0 + e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})}\left(r_{\text{-35}}e^{-\beta E_{\text{-35}}} + r_{\text{-10}}e^{-\beta E_{\text{-10}}} + r_{\text{max}}e^{-\beta(E_{\text{-35}}+E_{\text{-10}}+E_{\text{int}})}\right)}{1 + e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})}\left(e^{-\beta E_{\text{-35}}} + e^{-\beta E_{\text{-10}}} + e^{-\beta(E_{\text{-35}}+E_{\text{-10}}+E_{\text{int}})}\right)}. \tag{S6.6}$$

In the main text, we assumed that $r_{\text{-35}} = r_{\text{-10}} = r_0$ for simplicity (and because relaxing this assumption does not qualitatively change any of our results). Here, we will keep these more general rates, as it will aid in the following analysis.

Using Eq. S6.6, we can approximate gene expression for our four promoters of interest,

$$\mathrm{GE}^{(1,1)} \approx r_0 \tag{S6.7}$$

$$\mathrm{GE}^{(0,1)} \approx \frac{r_0 + r_{\text{-}35} e^{-\beta(E_{\mathrm{BG}}+E_{\mathrm{Spacer}})} e^{-\beta E_{\text{-}35}}}{1 + e^{-\beta(E_{\mathrm{BG}}+E_{\mathrm{Spacer}})} e^{-\beta E_{\text{-}35}}} \tag{S6.8}$$

$$\mathrm{GE}^{(1,0)} \approx \frac{r_0 + r_{\text{-}10} e^{-\beta(E_{\mathrm{BG}}+E_{\mathrm{Spacer}})} e^{-\beta E_{\text{-}10}}}{1 + e^{-\beta(E_{\mathrm{BG}}+E_{\mathrm{Spacer}})} e^{-\beta E_{\text{-}10}}} \tag{S6.9}$$

$$\mathrm{GE}^{(0,0)} \approx \frac{r_0 + r_{\max} e^{-\beta(E_{\mathrm{BG}}+E_{\mathrm{Spacer}})} e^{-\beta(E_{\text{-}35}+E_{\text{-}10}+E_{\mathrm{int}})}}{1 + e^{-\beta(E_{\mathrm{BG}}+E_{\mathrm{Spacer}})} e^{-\beta(E_{\text{-}35}+E_{\text{-}10}+E_{\mathrm{int}})}}. \tag{S6.10}$$

In Eq. S6.7, we used the fact that the promoter is very weak ($\mathrm{GE}^{(1,1)} \lesssim 0.25$) to infer that RNAP is unable to bind at either the -35 or -10 sites ($e^{-\beta E_{\text{-}35}}, e^{-\beta E_{\text{-}10}} \ll 1$). Since replacing the -35 site slightly improves gene expression ($0.25 \lesssim \mathrm{GE}^{(0,1)} \lesssim 1.0$), we only keep the -35 binding term in Eq. S6.8 but continue to neglect the -10 terms (assuming that binding to the -10 is sufficiently unfavored that it overwhelms the avidity term, $e^{-\beta(E_{\text{-}10}+E_{\mathrm{int}})} \ll 1$). Analogous statements hold for $\mathrm{GE}^{(1,0)}$ and the -10 site in Eq. S6.9. Lastly, when both the -35 and -10 sites are replaced (Eq. S6.10), the fully bound RNAP state will dominate over the two partially bound states due to avidity.

For every set of four promoters satisfying our criteria, we can use Eq. S6.7 to infer r_0, Eq. S6.8 to solve for $e^{-\beta(E_{\mathrm{BG}}+E_{\mathrm{Spacer}})} e^{-\beta E_{\text{-}35}}$ (in terms of $r_{\text{-}35}$), and Eq. S6.9 to solve for $e^{-\beta(E_{\mathrm{BG}}+E_{\mathrm{Spacer}})} e^{-\beta E_{\text{-}10}}$ (in terms of $r_{\text{-}10}$). In addition, we can directly estimate $r_{\max} \approx 10$ directly from the maximum gene expression of all promoters. Combining these statements, we can rewrite Eq. S6.10 as

$$\mathrm{GE}^{(0,0)} \approx \frac{r_0 + r_{\max} A e^{-\beta(E_{\mathrm{BG}}+E_{\mathrm{Spacer}})} e^{-\beta E_{\text{-}35}} e^{-\beta(E_{\mathrm{BG}}+E_{\mathrm{Spacer}})} e^{-\beta E_{\text{-}10}}}{1 + A e^{-\beta(E_{\mathrm{BG}}+E_{\mathrm{Spacer}})} e^{-\beta E_{\text{-}35}} e^{-\beta(E_{\mathrm{BG}}+E_{\mathrm{Spacer}})} e^{-\beta E_{\text{-}10}}} \tag{S6.11}$$

with the unknown quantity $A = e^{\beta(E_{\mathrm{BG}}+E_{\mathrm{Spacer}}-E_{\mathrm{int}})}$. Therefore, the three unknown constants $r_{\text{-}35}$, $r_{\text{-}10}$, and A would permit us to predict $\mathrm{GE}^{(0,0)}$ using single and double mutant data within the multivalent model. To facilitate this, we coarsely approximate that the partially bound RNAP states give rise to intermediate expression levels $r_{\text{-}35} = r_{\text{-}10} \approx \sqrt{r_0 r_{\max}} \approx 1$ and that the average energy of a background and spacer sequence is negligible compared to the favorable interaction energy which is on the order of several $k_B T$ leading to $A \approx e^{-\beta E_{\mathrm{int}}} \approx 100$. Fig. S6.3B demonstrates that the multivalent model predicts larger gene expression often closer to $r_{\max} \approx 10$. Although this approximate result for the multivalent model exhibits scatter about the predicted diagonal line, it nevertheless show a marked improvement over the energy matrix model, supporting the notion that avidity is a key concept when predicting the gene expression of mutations that greatly weaken or greatly strengthen the -35 and -10 sites.

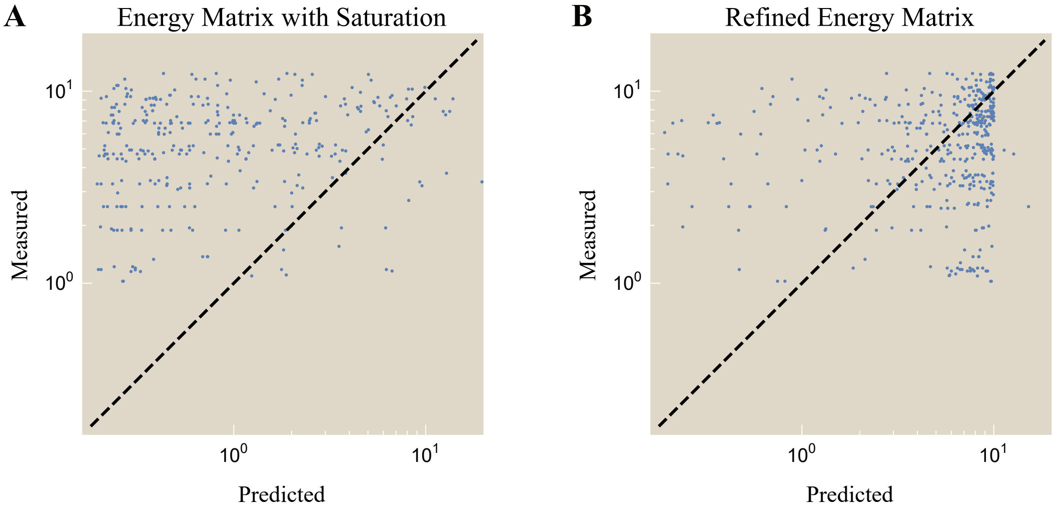


Figure S6.3: Relating gene expression measurements with minimal fitting. Using gene expression measurements for a weak promoter and two single mutants with higher gene expression, we can predict the expression of the double mutant and compare it to data. (A) The epistasis-free energy matrix model with saturation (Eq. S6.3) underestimates the gene expression, suggesting that the avidity between the -35 and -10 sites is missing from this analysis. (B) The multivalent model Eq. 6.3 predicts higher gene expression levels that better characterize the data.

S6.4 Interactions Between the Different Promoter Elements

In this section, we extend the analysis shown in the Fig. 6.2A inset to determine the strength of interactions between every pair of promoter elements as shown in Fig. S6.5A. As an example, Fig. S6.4 considers the combinations of a promoter with two possible -35 motifs ($-35^{(1)}$ or $-35^{(2)}$) and two possible spacers ($\text{Spacer}^{(1)}$ or $\text{Spacer}^{(2)}$) with the same UP, -10, and background sequences.

Suppose that the -35 and spacer elements contribute independently to gene expression (GE) so that we can write $\text{GE} = f_1(E_{-35})f_2(E_{\text{Spacer}})$ as the product of two functions f_1 and f_2 (in the standard energy matrix model, $f_1(E) = f_2(E) = e^{-\beta E}$). This independence implies that the system has no epistasis, namely,

$$\text{GE}^{(0,0)} = \text{GE}^{(1,1)} \frac{\text{GE}^{(0,1)}}{\text{GE}^{(1,1)}} \frac{\text{GE}^{(1,0)}}{\text{GE}^{(1,1)}}. \tag{S6.12}$$

Thus, for all possible pairs of -35 and spacer elements, we can compare the predicted gene expression given by Eq. S6.12 with the experimental measurements to discern whether these two segments of the promoter contribute independently to gene expression. In the following analysis, we will also restrict ourselves to promoters where GE $> 10^{-0.5}$ for all four mutants to ensure that the measurements are within the dynamic range of the experiment (so that we can be certain we are analyzing gene expression measurements and not noise).

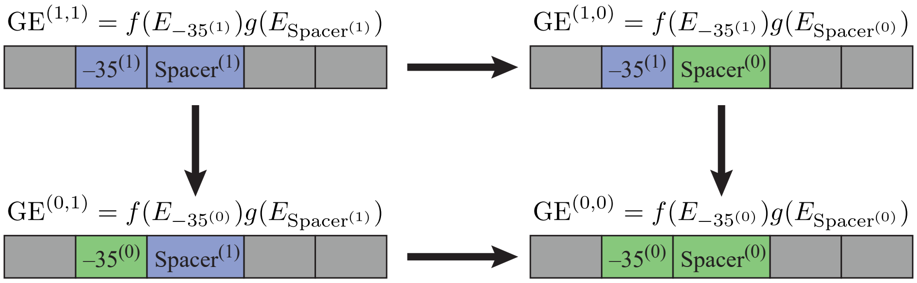
S6.4.1 Characterizing Promoters with no UP Element

We first carry out this analysis on the 4,096 promoters with no UP elements as shown in Fig. S6.5. In each plot, we compare the epistasis-free predicted GE (x-axis) with the measured value (y-axis). If two promoter elements independently contribute to gene expression, their data should fall onto the straight line $y = x$. We can quantify all deviations from such lines using the coefficient of determination R^2, with smaller R^2 values signifying that the promoter elements are not multiplicatively independent.

This analysis shows that while the -35 and -10 elements interact in a fashion discordant with an energy matrix formulation (leading to a negative R^2), the remaining promoter elements interact approximately independently of each other and can be approximated using an energy matrix model. This rigorously justified our sole consideration of the -35 and -10 binding sites in Fig. 6.1B, allowing us to avoid, for example, enumerating states where the RNAP is solely bound to the spacer or the background. Instead, the promoter is well approximated by treating the -35 and -10 motifs as cooperative binding sites while the spacer and background contribute independently to RNAP binding (as per Eq. 6.3).

Lastly, we note that the multivalent model (Eq. 6.3) does not strictly exhibit the multiplicative independence between the -35 and spacer elements (or any of the other weakly interacting promoter elements) that would lead to an $R^2 = 1$ expectation, but as we now show it closely approximates multiplicative independence. First, note that using the parameter values from Table S6.1, the denominator in the multivalent model is approximately 1 because $e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})}\left(e^{-\beta E_{\text{-35}}}+e^{-\beta E_{\text{-10}}}+e^{-\beta(E_{\text{-35}}+E_{\text{-10}}+E_{\text{int}})}\right)$ is $\lesssim 1$ for approximately 90% of the promoters. Additionally, the numerator in the model may be dominated by either of its terms: For weak promoters where $r_0 \gg r_0 e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})}\left(e^{-\beta E_{\text{-35}}}+e^{-\beta E_{\text{-10}}}+\frac{r_{\text{max}}}{r_0}e^{-\beta(E_{\text{-35}}+E_{\text{-10}}+E_{\text{int}})}\right)$ and GE $\approx r_0$. In the opposite case of a strong promoter, $r_0 \ll r_0 e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})}\left(e^{-\beta E_{\text{-35}}}+e^{-\beta E_{\text{-10}}}+\frac{r_{\text{max}}}{r_0}e^{-\beta(E_{\text{-35}}+E_{\text{-10}}+E_{\text{int}})}\right)$ and we can approximate Eq. 6.5 as

$$\text{GE} \approx r_0 e^{-\beta(E_{\text{BG}}+E_{\text{Spacer}})}\left(e^{-\beta E_{\text{-35}}}+e^{-\beta E_{\text{-10}}}+\frac{r_{\text{max}}}{r_0}e^{-\beta(E_{\text{-35}}+E_{\text{-10}}+E_{\text{int}})}\right), \qquad \text{(S6.13)}$$


$\mathrm{GE}^{(1,1)} = f(E_{-35^{(1)}})g(E_{\mathrm{Spacer}^{(1)}})$
–35$^{(1)}$
Spacer$^{(1)}$
$\mathrm{GE}^{(1,0)} = f(E_{-35^{(1)}})g(E_{\mathrm{Spacer}^{(0)}})$
–35$^{(1)}$
Spacer$^{(0)}$
$\mathrm{GE}^{(0,1)} = f(E_{-35^{(0)}})g(E_{\mathrm{Spacer}^{(1)}})$
–35$^{(0)}$
Spacer$^{(1)}$
$\mathrm{GE}^{(0,0)} = f(E_{-35^{(0)}})g(E_{\mathrm{Spacer}^{(0)}})$
–35$^{(0)}$
Spacer$^{(0)}$

$$\boxed{\text{Epistasis-Free Predicted }\mathrm{GE}^{(0,0)} = \mathrm{GE}^{(1,1)}\frac{\mathrm{GE}^{(0,1)}}{\mathrm{GE}^{(1,1)}}\frac{\mathrm{GE}^{(1,0)}}{\mathrm{GE}^{(1,1)}}}$$

Figure S6.4: Quantifying the interactions between promoter elements. If the -35 and spacer promoter elements independently contribute to gene expression, then an epistasis-free prediction of gene expression for the double mutant (bottom right) can be predicted using the gene expression of the other three promoters.

which exhibits the multiplicative independence implied by Eq. S6.12 between the weakly interacting promoter elements. Practically speaking, this means that in creating Fig. S6.5, we only considered data points where gene expression was above the background level that we inferred to be $10^{-0.5}$ based on the gene expression measurements in Fig. 6.2. In summation, Eq. 6.5 exhibits approximate independence between the weakly interacting promoter elements which can be identified as the plots for which $R^2 > 0$ in Fig. S6.5.

S6.4.2 Characterizing Promoters with an UP Element

Here, we extend the analysis in the previous section to a promoter that includes an UP element. As before, we seek to understand whether the UP, -35, spacer, -10, and background elements act independently of each other or whether they interact with avidity to facilitate RNAP binding.

Fig. S6.6 carries out this analysis using all 12,288 sequences from Urtecho *et al.* for every pair of promoter elements [4]. As in the previous section, we find that the -35 and -10 sites do not interact independently (as shown by a negative R^2). We acknowledge that several additional pairs of elements (i.e., -10/BG, -35/BG, and -10/Spacer) exhibit low R^2 values which may arise because: (*i*) Our model only approximately obeys multiplicative independence as discussed in Appendix S6.4.1 (so that $R^2 \approx 1$ even in the absence of experimental noise) or (*ii*) there may be additional interactions between promoter elements that we neglect, such as the importance of the discriminator [5] or weak RNAP binding sites in the background

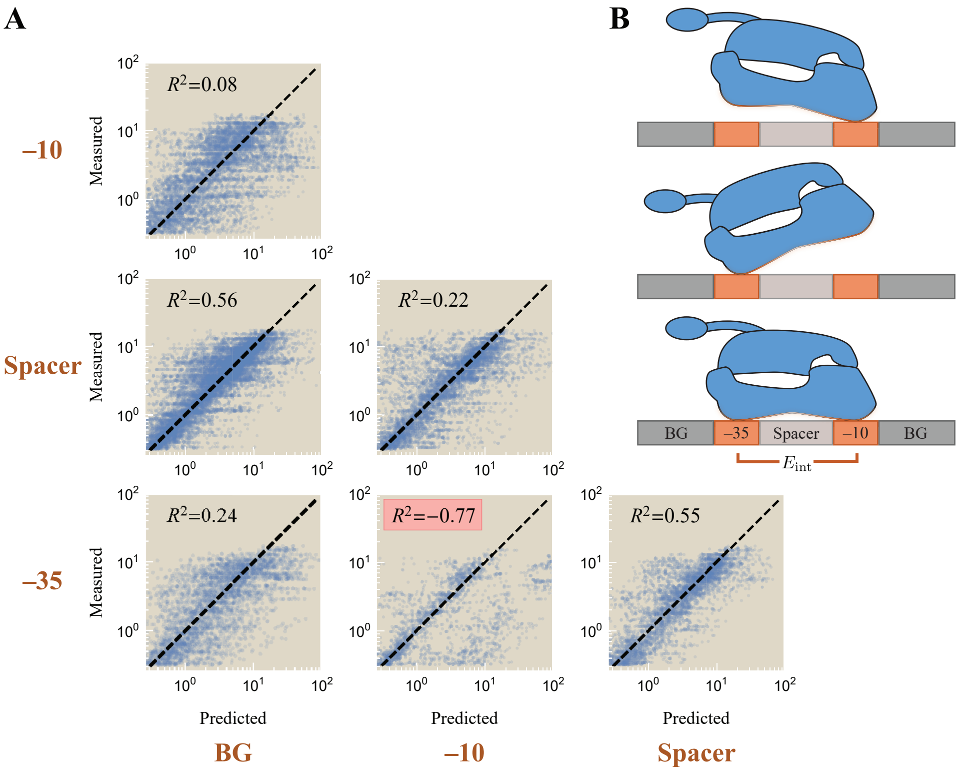


Figure S6.5: Interactions between the promoter elements with no UP binding site. (A) For every pair of elements (brown labels on the left and bottom), the measured gene expression (y-axis) is compared to the epistasis-free prediction (x-axis, Eq. S6.12) assuming that the two promoter elements are independent. Deviations between the predictions and measurements indicate that the two promoter elements interact. Data is plotted with low opacity to better show the general trend across the promoters. (B) The resulting schematic of a promoter with no UP element is that RNAP can bind to either the -35 or -10 sites independently with an avidity interaction when both are bound; the spacer and background (BG) contribute independently to the RNAP binding energy provided RNAP is bound to either the -35 or -10 element.

sequences [6]. We proceed by only considering interactions sufficiently strong to induce a negative R^2 value, namely, the avidity between the -35 and -10 motifs, with our eyes wide open to the possibility that more complex models could attempt to capture the full suite of higher-order interactions.

We end this section by analyzing which of the three schematics shown in Fig. 6.3A best characterizes the binding of the UP element. We note that the UP element appears to be particularly independent ($0.6 \lesssim R^2$) compared all other pairings of elements ($0.1 \lesssim R^2 \lesssim 0.6$), suggesting that the RNAP C-terminal binds weakly provided that either the -35 or -10 motifs are bound (Fig. S6.6B). This supports the bottom schematic in Fig. 6.3 and gives rise to the form of gene expression Eq. 6.5 used in the main text.

To complete this argument, we further note that the middle schematic in Fig. 6.3A would imply that the UP element only binds when the -35 element is bound, which would result in gene expression of the form

$$\mathrm{GE} = r_0 \frac{1 + e^{-\beta(E_{\mathrm{BG}}+E_{\mathrm{Spacer}})}\left(e^{-\beta(E_{\text{-35}}+E_{\mathrm{UP}})} + e^{-\beta E_{\text{-10}}} + \frac{r_{\mathrm{max}}}{r_0} e^{-\beta(E_{\text{-35}}+E_{\mathrm{UP}}+E_{\text{-10}}+E_{\mathrm{int}})}\right)}{1 + e^{-\beta(E_{\mathrm{BG}}+E_{\mathrm{Spacer}})}\left(e^{-\beta(E_{\text{-35}}+E_{\mathrm{UP}})} + e^{-\beta E_{\text{-10}}} + e^{-\beta(E_{\text{-35}}+E_{\mathrm{UP}}+E_{\text{-10}}+E_{\mathrm{int}})}\right)}. \tag{S6.14}$$

In this case, we would expect a low R^2 between the -35 and -10 elements as well as between the UP and -10 elements, but we would have an $R^2 \approx 1$ between UP and -35 since binding of one forces the binding of the other in this model. Given the larger-than-expected $R^2 = 0.56$ value between the UP and -10 elements and the smaller-than-expected $R^2 = 0.62$ value between the UP and -35 elements, this model is unlikely to be correct.

Finally, the top schematic in Fig. 6.3A implies a low R^2 value between the -35 and -10, the UP and -35, and the UP and -10 elements. In this case, all three elements bind strongly and in a highly dependent manner, so that eight RNAP states would need to be considered (with avidity terms between every pair of elements). Because the R^2 values between the UP/-35 and UP/-10 are larger than expected, this model does not appear to properly characterize RNAP binding, leading us to favor the bottom schematic in Fig. 6.3A.

S6.5 RNAP Binding Too Tightly Decreases Gene Expression

All of the gene expression models examined in this work assert that gene expression monotonically increases with the RNAP-promoter binding affinity. In contrast, Urtecho *et al.* found that this monotonic relationship did not hold for the strongest promoters. In other words, gene expression increased as the -35 and -10 motifs approached their consensus sequences (which bind the tightest to RNAP) *except* that promoters with both a consensus -35 and consensus -10 sequence exhibited lower gene expression than the corresponding sequences with one mutation in either motif [4]. This suggests that past a certain point, increasing the RNAP-promoter binding energy causes RNA polymerase to bind top tightly, thereby inhibiting gene expression [7].

In this Appendix, we explore this phenomenon and develop a model to account for it. More specifically, our model will posit that the state of transcription initiation can

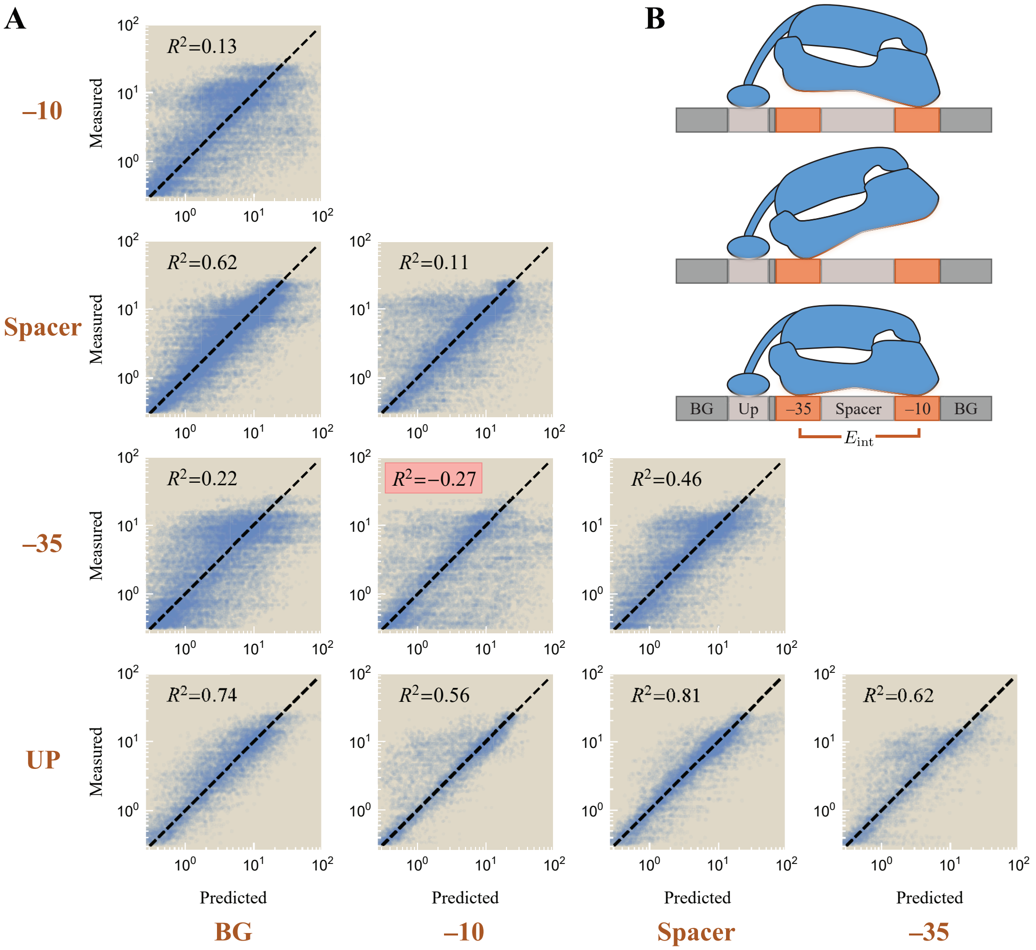


Figure S6.6: Interactions between the promoter elements with an UP binding site. (A) For every pair of elements (brown labels on the left and bottom), the measured gene expression (y-axis) is compared to the epistasis-free prediction (x-axis, Eq. S6.12) assuming that the two promoter elements are independent. (B) The resulting schematic of gene expression where RNAP can bind to either the -35 or -10 sites independently with an avidity interaction when both are bound; the UP, spacer, and background (BG) contribute independently to the RNAP binding energy provided RNAP is bound to either the -35 or -10 element.

be characterized by a free energy so that the probability of initiating transcription versus remaining bound on the promoter is given by the Boltzmann weight of the two states.

To start our analysis, Fig. S6.7A shows the predicted versus measured gene expression of the multivalent model (Eq. 6.5) for promoters with an UP element, with all data plotted with low opacity except the promoters with the lowest or highest levels of predicted gene expression. The sharp left edge of the data is set by the background level of gene expression $r_0 = 0.18$ in the absence of RNAP, while the right edge represents the maximal expression $r_{\text{max}} = 8.6$ of very strong promoters. Note that if the scatter in data on the left edge is attributable to noise, then the outliers on the right edge cannot simply arise from noise, since there are 10x fewer data points and hence we expect 10x fewer outliers (although there are roughly the same number of outliers 2σ, 3σ, and 4σ away from the predicted values on both edges of the plot). This suggests that there is a mechanistic explanation for why the promoters that our model predicts should bind very tightly to RNAP exhibit low gene expression.

We next analyzed whether any promoters increased expression when their -35 or -10 sites were replaced by the consensus sequences, but exhibited decreased gene expression when both the -35 and -10 sites became the consensus sequences. Out of the 12,000 constructs, 850 exhibited this pattern of expression. One possible explanation is that although strong binding helps recruit RNAP to the promoter, overly strong binding could inhibit transcription initiation and decrease gene expression. We note, however, that this is a coarse-grained effective model that neglects molecular details of the transition from the closed to open complex, transcription initiation, and other critical steps of RNAP functioning [8]. Nevertheless, in the multivalent model, the effect of overly strong RNAP-promoter binding must be to decrease the single parameter r_{max} (which represents the level of gene expression when RNAP is fully bound to the promoter), since no other parameters should depend upon the total RNAP-promoter binding strength.

To proceed, we assume that fully bound RNAP with free energy

$$\Delta E_{\text{RNAP}} = E_{\text{BG}} + E_{\text{Spacer}} + E_{\text{UP}} + E_{\text{-35}} + E_{\text{-10}} + E_{\text{int}} \tag{S6.15}$$

relative to the unbound state can initiate transcription by moving into a transcription initiation state with free energy ΔE_{trans} relative to the unbound state as shown schematically in Fig. S6.7B. Intuitively, bound RNAP will always immediately transcribe when $\Delta E_{\text{trans}} - \Delta E_{\text{RNAP}}$ is large and negative, but when the affinity between the RNAP and promoter becomes sufficiently strong (the case depicted in Fig. S6.7B), RNAP will prefer to stay bound to the promoter and not transcribe immediately. We posit that the rate of entering the transcribing state [8], and hence the rate of gene expression r_{max} in Fig. 6.1B, should be modified to

$$\tilde{r}_{\text{max}} \equiv \frac{r_{\text{max}} + r_0 e^{-\beta(\Delta E_{\text{RNAP}} - \Delta E_{\text{trans}})}}{1 + e^{-\beta(\Delta E_{\text{RNAP}} - \Delta E_{\text{trans}})}}, \tag{S6.16}$$

similar to recently proposed scrunching models of transcription initiation [9]. For promoters whose RNAP binding is far weaker than transcription initiation

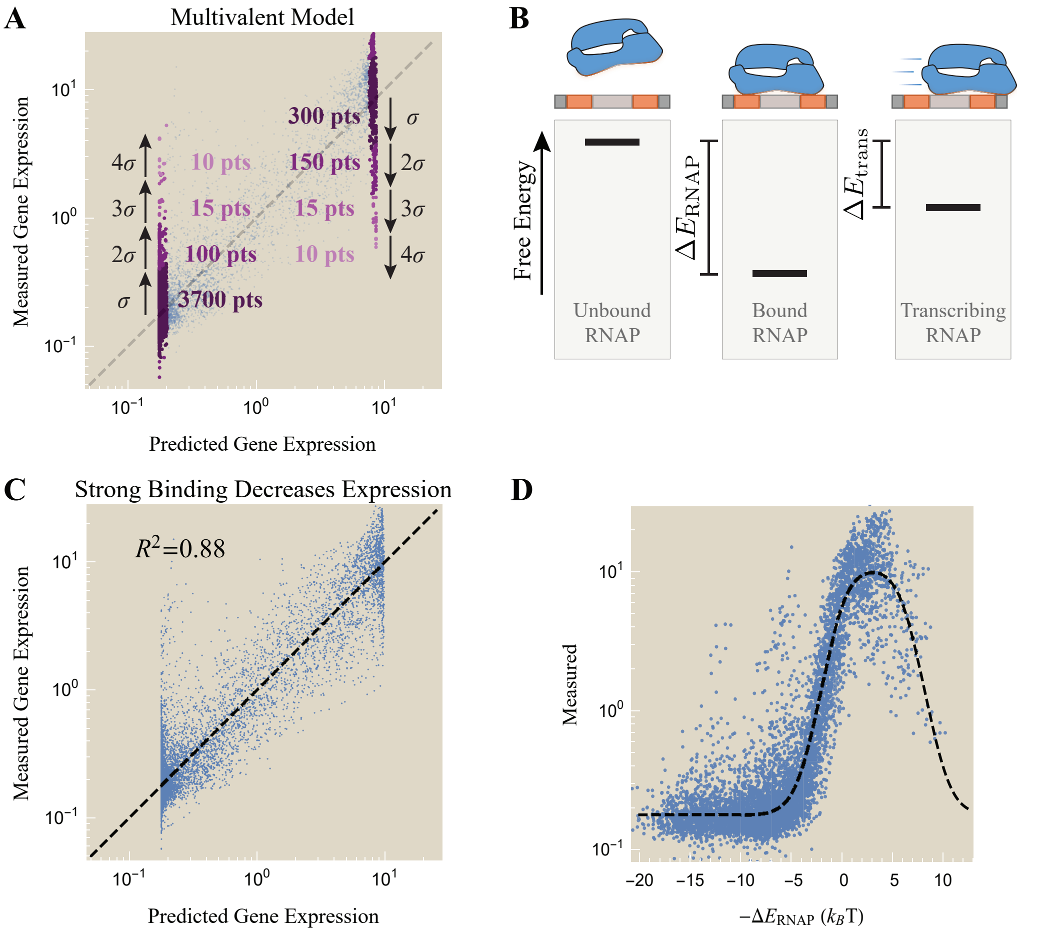


Figure S6.7: Gene expression is reduced for promoters that bind RNAP too tightly. (A) In the multivalent model (Eq. 6.5), although there are 10x fewer points on the right edge of the plot than the left edge, there are the same number of outliers, suggesting a biophysical mechanism for the reduction in gene expression of the strongest promoters. (B) The average level of transcription modeled as a two state system where the bound RNAP state (with free energy ΔE_{RNAP} relative to the unbound state) can enter a transcription initiation state with free energy ΔE_{trans}. (C) Gene expression characterized using the modified maximum level of gene expression using Eq. S6.16 with $\Delta E_{\text{trans}} = -6.2\, k_B T$. (D) Measured gene expression versus the promoter strength ΔE_{RNAP} (stronger promoters on the right because of the minus sign). The dashed line shows the prediction of the multivalent model modified using Eq. S6.16.

($e^{-\beta(\Delta E_{\text{RNAP}}-\Delta E_{\text{trans}})} \ll 1$), this rate reduces to the constant value r_{max}. Increasing the RNAP-promoter affinity decreases ΔE_{RNAP} which leads to a decrease in the level of gene expression. In the limit of an infinitely strong promoter ($\Delta E_{\text{RNAP}} \to -\infty$), RNAP is glued in place and unable to transcribe, thereby reducing the level of gene expression to the background level r_0.

Fig. S6.7C shows the gene expression data refit to the multivalent model with the maximal level of gene expression given by Eq. S6.16 (using $\Delta E_{\text{trans}} = -6.2\, k_B T$ inferred by nonlinear regression). We note that using this model eliminates the sharp right edge of the data (red ellipse in Panel A), signifying that the promoters with extremely tight RNAP binding have shifted left, moving closer to the level of gene expression predicted by the model. Fig. S6.7D compares the predicted $-\Delta E_{\text{trans}}$ for each promoter (using the best fit parameter in Table S6.1) against the measured level of gene expression. To facilitate a comparison with the multivalent model, we overlay this data with the approximate predicted level of gene expression

$$\text{GE} \approx \frac{r_0 + \tilde{r}_{\text{max}} e^{-\beta \Delta E_{\text{RNAP}}}}{1 + e^{-\beta \Delta E_{\text{RNAP}}}}, \tag{S6.17}$$

where we have ignored the two partially bound RNAP states and used the maximum level of gene expression in Eq. S6.16. Although only a small number of promoters exhibits sufficiently strong binding that diminishes their gene expression, the data exhibits a clear downwards trend in this limit.

S6.6 Dynamics of RNAP with Avidity

S6.6.1 Probability of the RNAP States at Equilibrium

In this section, we derive the probabilities of the four RNAP states shown in Fig. 6.1C in equilibrium. RNAP may be unbound (concentration U), singly bound at the -35 site ($B_{\text{-35}}$), singly bound at the -10 site ($B_{\text{-10}}$), or bound to both sites ($B_{\text{-35,-10}}$). These concentrations must obey detailed balance,

$$B_{\text{-35}} = \frac{[\text{RNAP}] k_{\text{on}}}{k_{\text{off,-35}}} U \tag{S6.18}$$

$$B_{\text{-10}} = \frac{[\text{RNAP}] k_{\text{on}}}{k_{\text{off,-10}}} U \tag{S6.19}$$

$$B_{\text{-35,-10}} = \frac{\tilde{k}_{\text{on}}}{k_{\text{off,-10}}} B_{\text{-35}}, \tag{S6.20}$$

as well as the normalization condition

$$[\text{RNAP}] = U + B_{\text{-35}} + B_{\text{-10}} + B_{\text{-35,-10}}. \tag{S6.21}$$

In writing Eqs. S6.18 and S6.19, we have assumed a sufficiently large reservoir of RNAP so that binding to the promoter of interest does not appreciably decrease the concentration of free RNAP (a reasonable assumption in *E. coli* where there are ≈ 2000 RNAP molecules [10]).

Eqs. S6.18-S6.21 can be solved to obtain the concentration of each RNAP state, namely,

$$U = \frac{K_{\text{-35}}K_{\text{-10}}}{K_{\text{-35}}K_{\text{-10}} + K_{\text{-35}}[\text{RNAP}] + K_{\text{-10}}[\text{RNAP}] + c_0[\text{RNAP}]}[\text{RNAP}] \quad (\text{S6.22})$$

$$B_{\text{-35}} = \frac{K_{\text{-35}}[\text{RNAP}]}{K_{\text{-35}}K_{\text{-10}} + K_{\text{-35}}[\text{RNAP}] + K_{\text{-10}}[\text{RNAP}] + c_0[\text{RNAP}]}[\text{RNAP}] \quad (\text{S6.23})$$

$$B_{\text{-10}} = \frac{K_{\text{-10}}[\text{RNAP}]}{K_{\text{-35}}K_{\text{-10}} + K_{\text{-35}}[\text{RNAP}] + K_{\text{-10}}[\text{RNAP}] + c_0[\text{RNAP}]}[\text{RNAP}] \quad (\text{S6.24})$$

$$B_{\text{-35,-10}} = \frac{c_0[\text{RNAP}]}{K_{\text{-35}}K_{\text{-10}} + K_{\text{-35}}[\text{RNAP}] + K_{\text{-10}}[\text{RNAP}] + c_0[\text{RNAP}]}[\text{RNAP}], \quad (\text{S6.25})$$

where we have defined the dissociation constants $K_j = \frac{k_{\text{off},j}}{k_{\text{on}}}$ of free RNAP binding to the site j as well as the effective concentration $c_0 = \frac{\tilde{k}_{\text{on}}}{k_{\text{on}}}$ of singly bound RNAP binding to the remaining promoter site. If we further define the effective dissociation constant

$$K_D^{\text{eff}} = \frac{K_{\text{-35}}K_{\text{-10}}}{c_0 + K_{\text{-35}} + K_{\text{-10}}}, \quad (\text{S6.26})$$

we can rewrite the probability of the unbound state as

$$U = \frac{K_D^{\text{eff}}}{K_D^{\text{eff}} + [\text{RNAP}]}[\text{RNAP}]. \quad (\text{S6.27})$$

From this equation, we see that the promoter is bound 50% of the time ($U = \frac{[\text{RNAP}]}{2}$) when $[\text{RNAP}] = K_D^{\text{eff}}$, as stated in the main text.

S6.6.2 Dynamics of RNAP Unbinding with Avidity

Here, we rederive the results from the previous section by analyzing the dynamics of RNAP binding rather than its equilibrium configuration. This calculation highlights the intimate connection between the effective dissociation constant in Eq. S6.26 and the kinetics of RNAP binding.

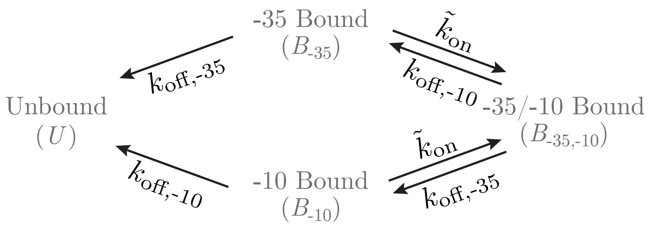


Figure S6.8: Dynamics of RNAP unbinding from the -35 and -10 sites. The avidity between the -35 and -10 sites will prolong the time before RNAP unbinds from the promoter.

To that end, we first compute the probability that a bound RNAP will remain bound after a time t. Since we are only interested in the unbinding process, we consider the rates diagram in Fig. S6.8 where the on-rates from the unbound state have been removed. Following Ref. [11], we assume that the three bound states – RNAP bound to only the -35 site (concentration $B_{\text{-35}}$), only the -10 site ($B_{\text{-10}}$), or to both sites ($B_{\text{-35,-10}}$) – quickly equilibrate and compute the effective off-rate from these bound states to the RNAP unbound state (U). If the three bound states are in equilibrium, then there is no flux between any two states, namely,

$$\tilde{k}_{\text{on}} B_{\text{-35}} = k_{\text{off,-10}} B_{\text{-35,-10}} \tag{S6.28}$$

and

$$\tilde{k}_{\text{on}} B_{\text{-10}} = k_{\text{off,-35}} B_{\text{-35,-10}}. \tag{S6.29}$$

The total concentration of bound RNAP is given by

$$[\text{RNAP}]_{\text{bound}} = B_{\text{-35}} + B_{\text{-10}} + B_{\text{-35,-10}} = B_{\text{-35,-10}} \left(1 + \frac{k_{\text{off,-35}}}{\tilde{k}_{\text{on}}} + \frac{k_{\text{off,-10}}}{\tilde{k}_{\text{on}}} \right). \tag{S6.30}$$

The loss of bound RNAP is caused by unbinding from the two singly bound forms, leading to the effective off-rate

$$\begin{aligned} \frac{d}{dt}[\text{RNAP}]_{\text{bound}} &\equiv -k_{\text{off}}^{\text{eff}} [\text{RNAP}]_{\text{bound}} && \text{(S6.31)} \\ &= -k_{\text{off,-35}} B_{\text{-35}} - k_{\text{off,-10}} B_{\text{-10}} && \text{(S6.32)} \\ &= -\frac{2 k_{\text{off,-35}} k_{\text{off,-10}}}{\tilde{k}_{\text{on}} + k_{\text{off,-35}} + k_{\text{off,-10}}} [\text{RNAP}]_{\text{bound}}. && \text{(S6.33)} \end{aligned}$$

Hence, the dynamics of RNAP unbinding are characterized by

$$[\text{RNAP}]_{\text{bound},t} = [\text{RNAP}]_{\text{bound},0} e^{-k_{\text{off}}^{\text{eff}} t} \tag{S6.34}$$

where the likelihood of remaining bound decreases exponentially according to the timescale $\tau = \frac{1}{k_{\text{off}}^{\text{eff}}}$.

Lastly, to connect this result to the calculations in the previous section, we return to the full model in Fig. 6.1C where unbound RNAP can associate onto the promoter. As in simple monovalent ligand-receptor systems, the effective dissociation constant Eq. S6.26 is related to the off-rate from the bound to unbound states ($k_{\text{off}}^{\text{eff}}$) divided by the on-rate from the unbound to bound states ($2k_{\text{on}}$), namely,

$$K_D^{\text{eff}} = \frac{k_{\text{off}}^{\text{eff}}}{2 k_{\text{on}}}. \tag{S6.35}$$

References

[1] Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, et al. Transcriptional Regulation by the Numbers: Models. Curr. Opin. Genet. Dev. 2005; 15(2):116–124. doi:10.1016/j.gde.2005.02.007.

[2] Wang F, Redding S, Finkelstein IJ, Gorman J, Reichman DR, Greene EC. The Promoter-Search Mechanism of *Escherichia coli* RNA Polymerase Is Dominated by Three-Dimensional Diffusion. Nat. Struct. Mol. Biol. 2013; doi:10.1038/nsmb.2472.

[3] Scholes C, DePace AH, Sánchez Á, Naef F, Bentley D, DePace A, et al. Combinatorial Gene Regulation Through Kinetic Control of the Transcription Cycle. Cell Syst. 2017;4(1):97–108.e9. doi:10.1016/j.cels.2016.11.012.

[4] Urtecho G, Tripp AD, Insigne KD, Kim H, Kosuri S. Systematic Dissection of Sequence Elements Controlling σ70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in *Escherichia coli*. Biochem. 2019;58(11):1539–1551. doi:10.1021/acs.biochem.7b01069.

[5] Feklístov A, Sharon BD, Darst SA, Gross CA. Bacterial Sigma Factors: A Historical, Structural, and Genomic Perspective. Annu. Rev. Microbiol. 2014; 68(1):357–376. doi:10.1146/annurev-micro-092412-155737.

[6] Yona AH, Alm EJ, Gore J. Random Sequences Rapidly Evolve into De Novo Promoters. Nat. Commun. 2018;9(1):1530. doi:10.1038/s41467-018-04026-w.

[7] Ellinger T, Behnke D, Bujard H, Gralla JD. Stalling of *Escherichia coli* RNA Polymerase in the +6 to +12 Region in Vivo is Associated with Tight Binding to Consensus Promoter Elements. J. Mol. Biol. 1994;239(4):455–465. doi:10.1006/JMBI.1994.1388.

[8] Ruff EF, Record MT, Artsimovitch I. Initial Events in Bacterial Transcription Initiation. Biomolecules. 2015;5(2):1035–1062. doi:10.3390/biom5021035.

[9] Henderson KL, Felth LC, Molzahn CM, Shkel I, Wang S, Chhabra M, et al. Mechanism of Transcription Initiation and Promoter Escape by E. coli RNA Polymerase. Proc. Natl. Acad. Sci. USA. 2017;114(15):E3032–E3040. doi:10.1073/pnas.1618675114.

[10] Klumpp S, Hwa T. Growth-Rate-dependent Partitioning of RNA Polymerases in Bacteria. Proc. Natl. Acad. Sci. USA. 2008;105(51):20245–20250. doi:10.1073/Proc.Natl.Acad.Sci.USA.0804953105.

[11] Stone JD, Artyomov MN, Chervin AS, Chakraborty AK, Eisen HN, Kranz DM. Interaction of Streptavidin-Based Peptide-MHC Oligomers (tetramers) with Cell-Surface TCRs. J. Immunol. 2011;187(12):6281–6290. doi:10.4049/jimmunol.1101734.

Description	Parameter	Energy Matrix Model	Multivalent Model
Level of gene expression	r_{max}	0.42	8.6
	r_0	—	0.18
Interaction energy	E_{int}	—	−6.3
-35 motif ($E_{\text{-35}}$)	TTGACA	−1.3	−1.3*
	TTTACA	−0.2	3.3
	ATTACA	0.6	8.3
	TTTACC	0.3	5.7
	TTAAGA	0.6	8.8
	TTGCAA	−0.4	2.5
	CTCAGA	0.7	9.5
	CTTAGA	0.6	9.5
-10 motif ($E_{\text{-10}}$)	TATAAT	−0.9	−0.9*
	AATAAT	−0.1	3.6
	GATAAT	−0.1	3.2
	TATAAA	0.0	4.4
	GATAAC	0.6	9.8
	TATGTT	0.1	4.6
	GTTAAA	0.6	>10
	GTTGTA	0.6	>10
Spacer (E_{Spacer})	P1-6	0.0	0.0*
	lac	0.4	3.9
	ECK125136938	0.0	1.0
	ECK125137104	0.1	1.5
	ECK125137108	0.1	0.7
	ECK125137405	0.1	1.2
	ECK125137640	0.4	3.8
	ECK125137726	−0.1	−0.8
Background (E_{BG})	bg463205:463355	0.0	0.0*
	bg977040:977190	0.1	2.3
	bg991964:992114	0.3	4.2
	bg1163421:1163571	0.1	1.5
	bg3514590:3514740	0.1	1.9
	bg4323949:4324099	0.1	2.5
	bg4427287:4427437	0.2	3.1
	bg4471352:4471502	−0.1	1.6
UP (E_{UP})	No UP	0.0	0.0*
	136-fold UP	—	−2.6
	326-fold UP	—	−3.2

Table S6.1: Parameter values for the models of transcriptional regulation considered in this work. The levels of gene expression (r_0 and r_{max}) are in the same arbitrary units as the experimental measurements while the energies are all in k_BT units (energies that are more negative indicate tighter binding). The original nomenclature from the experimental work is used for each promoter element. Parameter denoted by an asterisk (*) represent values that were fixed to their corresponding value in the energy matrix model to prevent parameter degeneracy.

Chapter 7

ADDENDUM

This short chapter presents additional insights that extend the calculations in the previous chapters and will serve as the foundation for future work.

7.1 Cranking up Complexity: Analyzing the Multi-Step Process of Metabolic Regulation

I met Denis Titov in January 2018, when he had just finished a postdoc in Harvard and was moving to Berkeley as a newly minted professor. Denis's area of expertise was metabolism, the process by which we convert food into energy. Metabolism is ripe with applications to many facets of human health from diet (how calorie restriction extends lifespan) to exercise (why physical activity decreases the risk of heart disease, stroke, and cancer) to aging (how metabolism influences our susceptibility to age-related diseases). Denis had devised an ingenious experimental method that alters the balance of cellular energy carriers (NADH/NAD$^+$, NADPH/NADP$^+$, ATP/ADP) that are ordinarily tightly regulated in cells. The question that Denis proposed to Rob and me was whether we could model the glycolytic pathway (the first step in metabolism shown in Fig. 7.1A) to characterize his system both experimentally and theoretically.

There are many reasons why modeling glycolysis is especially tractable. First, glycolysis is "the *E. coli* of metabolic pathways" – it is one of the simplest and best-studied systems where every component has been extensively characterized. Second, the majority of steps in glycolysis have no regulation and their associated enzyme follows Michaelis-Menten kinetics (which is straightforward to model). However, the key regulatory nodes in the pathway are the three allosteric enzymes hexokinase (HK), phosphofructokinase (PFK), and pyruvate kinase (PK) shown in pink boxes as Fig. 7.1A, and these enzymes have a host of allosteric inhibitors (brown) and activators (green) that regulate their activity. Denis proposed that if we can carefully model the behavior of these key nodes, we can pull together the accumulated knowledge from the other simpler steps of the pathway and create a robust model of glycolysis.

Since my expertise is in characterizing allosteric systems, and my first project looked at allosteric enzymes (Chapter 1), it quickly became apparent that we had a solid collaboration with two complementary skill sets that together make this project tractable. In this way, I began an incredibly fun and fascinating journey with Denis that still continues today.

While this project is ongoing, Fig. 7.1 demonstrates the style of modeling we have carried out thus far. Denis begins by reading the literature to determine all of the allosteric effectors that interact with each key enzyme, enabling us to make a cartoon representation as shown for phosphofructokinase in Panel B. I then mathematize this cartoon using a statistical mechanics framework to quantify how the activity should vary as the concentration of each effector is changed. Denis then mines the last fifty years of papers (an incredible feat!) to create a "gold standard" data set featuring the most pristine, careful experiments that collectively span the range of physiological conditions under which phosphofructokinase operates. A subset of this massive data set is shown in Panels C-F, summarizing 15 different experimental conditions from 5 different papers. Finally, Denis and I test how well the model can reproduce the data, discuss possible modifications, set off to find further data sets to validate our hypotheses, and iterate until we achieve the characterizations shown by the curves in Panels C-F.

These results are noteworthy because reproducibility in biology is fraught with difficulties. Experiments are tough to reproduce across labs (and surprisingly often within the same lab), so there is no reason to think that any model should be able to characterize this diverse set of data that spans multiple years, comes from different labs, and carries out measurements under a range of conditions. Direct conflicts between data sets and systematic discrepancies must be resolved using rigorous, unbiased methods. Our success in doing this for the key nodes in glycolysis has revealed that we can push our models of allosteric systems to these complex cases riddled with multiple substrates and allosteric effectors. We are now in the process of combining all of the steps to see how well they can characterize the full glycolytic pathway. Following that, we will make contact with Denis's experiments that alter the ratio of NADH/NAD$^+$, NADPH/NADP$^+$, and ATP/ADP to predict the outcome of his experiments before he carries them out!

7.2 The Statistical Mechanics of Bivalent Binding

In statistical mechanical models of ligand-receptor binding, the probability that a molecule with concentration c will be bound to a receptor is given by $c/(c + K_D)$

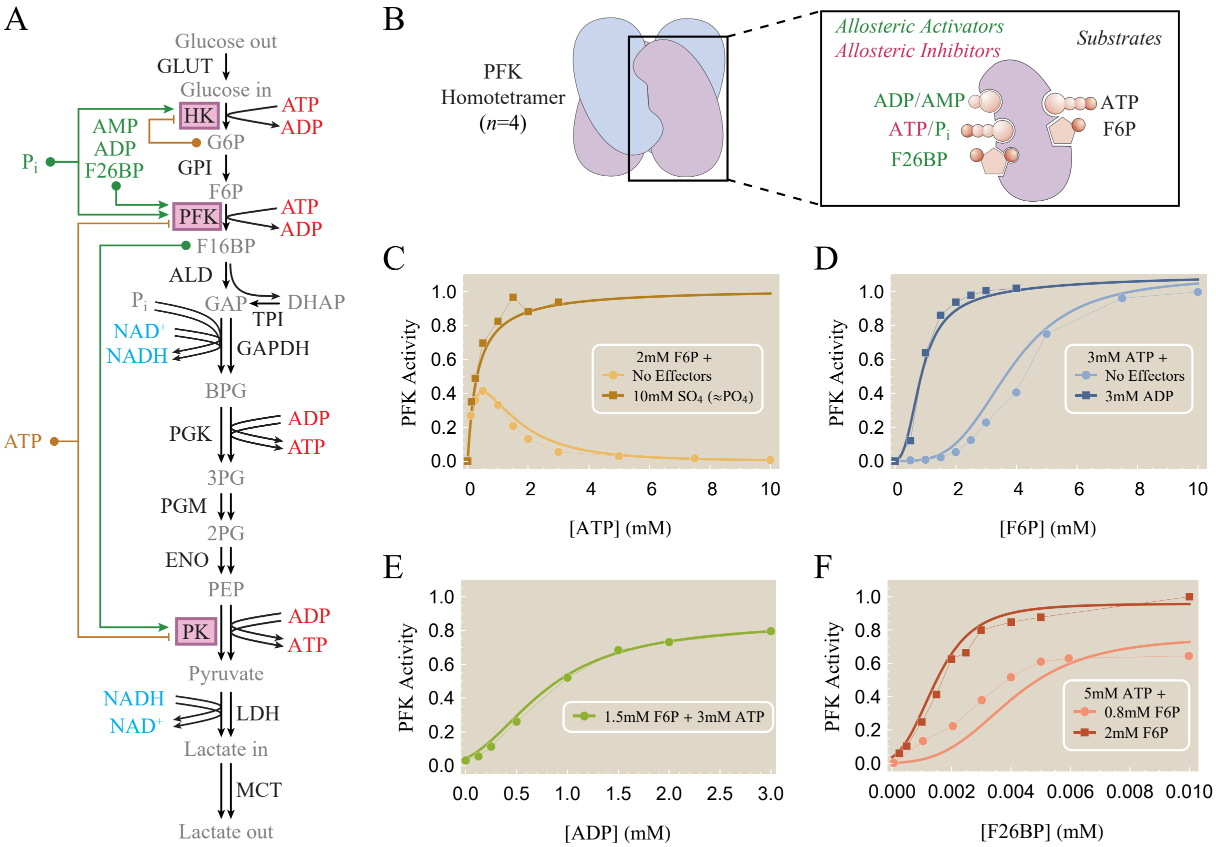


Figure 7.1: Modeling the glycolytic pathway. (A) The steps of mammalian glycolysis together with the enzymes (black), substrates (gray), allosteric regulators (green/brown), and energy carriers (red/teal). (B) Phosphofructokinase is one of the key regulators of glycolysis. It is composed of four identical subunits, each containing binding sites for the substrates ATP and F6P as well as three sets of allosteric sites for the regulators shown in the inset. (C-F) A statistical mechanical model of phosphofructokinase is able to characterize its diverse behavior across a diverse set of physiologically relevant conditions.

where K_D is the dissociation constant (the off-rate divided by the on-rate) of the molecule. When two identical molecules are tethered together, it is conventional wisdom that the dissociation constant will be halved ($K_D \to \frac{K_D}{2}$), since the on-rate has doubled while the off-rate has stayed the same (or equivalently, since either molecule could bind the receptor). In 2018, my amazing labmate Vahe Galstyan showed me that conventional wisdom was not necessarily right.

Before launching into the math, it is worth mentioning that this type of calculation often arises in the context of multivalent binding (the subject of Chapter 6). For example, our bodies harbor multivalent cells ranging from the T cells that enable antibody production, natural killer cells that destroy invaders, and inflamed epithelial cells that recruit other bodily defenses to sites of infection. Different classes of antibodies can have 2, 4, or even 10 arms, so that even if each individual receptor binds weakly, the cell as a whole binds very tightly. Viruses such as influenza harness multivalency by packing their surface with hundreds of binding sites to better stick to their target cells. HIV take the opposite approach and decreases its number of binding sites to fifteen, making it far less infective than influenza but also undercutting our body's immune response by preventing our antibodies from multivalently binding to its surface. The following calculation comparing the dissociation constant of a single ligand versus two tethered ligands applies in each of these contexts.

To proceed, consider a heterodimeric molecule composed of a green and blue ligand tethered together by a flexible linker. The two ligands have the same size, shape, and binding energy to the receptor (in other words, these ligands are identical, but we color them differently to help keep track of particle identity during the calculation). As shown in Fig. 7.2, define the volume of the solution to be V_{sol} and discretize this volume into small boxes with volume V_{box} whose size is comparable to that of either ligand. Thus, rather than thinking of a continuous picture, we will suppose that the green ligand must reside in one of these boxes, which will enable us to compute its entropy.

Let V_u represent the volume accessible to the blue ligand when the green molecule is unbound but fixed in one of the boxes (Fig. 7.2A). Let V_b represent the volume accessible to the blue ligand when the green molecule is bound to the receptor (Fig. 7.2B; this is equivalent to the volume accessible to the green ligand when the blue molecule is bound). We denote the binding energy in the bound state relative to the unbound state as E_{bound}.

In the unbound case, the number of states of the system equals $\Omega_{\text{unbound}} = \frac{V_{\text{sol}}}{V_{\text{box}}}\frac{V_{\text{u}}}{V_{\text{box}}}$ where the two terms account for translational and rotational entropy, respectively. In the case where either the green or the blue ligand is bound, the multiplicity of the states is given by $\Omega_{\text{bound}} = \frac{2V_{\text{b}}}{V_{\text{box}}}$, where the factor of two accounts for either ligand (green or blue) being equally capable of binding. The relative probability of the bound state is given by the ratio of its Boltzmann weight ($\Omega e^{-\beta E}$) to that of the unbound state, namely,

$$\frac{\Omega_{\text{bound}} e^{-\beta E_{\text{bound}}}}{\Omega_{\text{unbound}}} = \frac{\frac{2V_{\text{b}}}{V_{\text{box}}} e^{-\beta E_{\text{bound}}}}{\frac{V_{\text{sol}}}{V_{\text{box}}}\frac{V_{\text{u}}}{V_{\text{box}}}} = 2\frac{V_{\text{box}}}{V_{\text{sol}}}\frac{V_{\text{b}}}{V_{\text{u}}} e^{-\beta E_{\text{bound}}}. \tag{7.1}$$

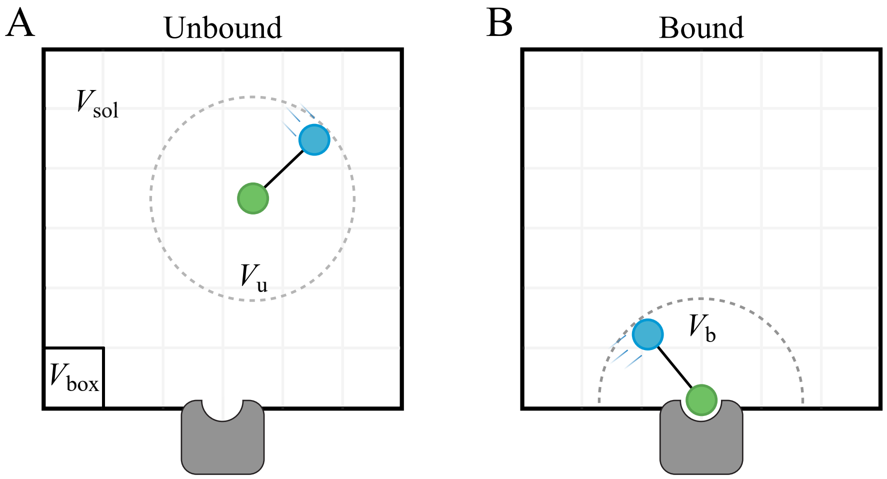


Figure 7.2: The thermodynamics of a dimeric molecule binding to a receptor. A heterodimer (the green and blue balls linked together by a flexible linker) in solution with volume $V_{\rm sol}$ that is discretized into small boxes with volume $V_{\rm box}$ (comparable to the size of the green ball). (A) If the green ball is fixed in place in one of the boxes, the blue ball can reside in a volume $V_{\rm u}$ where the subscript signifies the unbound state. (B) When the green ball is bound to the receptor, the blue ball can access a volume $V_{\rm b}$ where the subscript denotes the bound state.

We next compare this situation with a monovalent ligand (i.e. a green molecule that is not tethered to a blue molecule). Here, the unbound state has multiplicity $\Omega_{\text{unbound}}^{\text{mono}} = \frac{V_{\rm sol}}{V_{\rm box}}$ while the bound state has multiplicity $\Omega_{\text{bound}}^{\text{mono}} = 1$ and relative energy E_{bound}. Therefore, the ratio of Boltzmann weights is given by

$$\frac{\Omega_{\text{bound}}^{\text{mono}} e^{-\beta E_{\text{bound}}}}{\Omega_{\text{unbound}}^{\text{mono}}} = \frac{e^{-\beta E_{\text{bound}}}}{\frac{V_{\rm sol}}{V_{\rm box}}} = \frac{V_{\rm box}}{V_{\rm sol}} e^{-\beta E_{\text{bound}}}. \tag{7.2}$$

The dissociation constant K_D for the bivalent ligand and the analogous K_D^{mono} for a monovalent ligand are proportional to Eqs. 7.1 and 7.2, respectively, with the same constant of proportionality. Therefore $\frac{K_D}{K_D^{\text{mono}}} = \frac{2V_{\rm b}}{V_{\rm box}}$, demonstrating that the dissociation constant changes from $K_D^{\text{mono}} \to \frac{2V_{\rm b}}{V_{\rm box}} K_D^{\text{mono}}$ when two ligands are tethered together. In the case where the binding site is large and planar, $V_{\rm b} = \frac{V_{\rm u}}{2}$ and the dissociation constant is unchanged by this tethering. In contrast, the dissociation constant doubles (as per conventional wisdom) when the blue ligand is free to rotate completely around a bound green ligand, suggesting that the epitope is, for example, a very small and thin stalk. Pamela Bjorkman suggested that an antibody binding to a virus likely approximates this latter case more than the former, suggesting that in the context of antibody-virus binding, the shift of $K_D^{\text{mono}} \to 2K_D^{\text{mono}}$ often seen in models is correct. However, in the many cases where a molecule binds to an embedded pocket (as in some instances of B cell-T cell binding or antibody-pathogen interactions), care should be taken when writing down such equations, and the underlying assumption inherent in this equation should be acknowledged.

7.3 Generalized Data Collapse

In Chapter 2 Eq. 2.11, we presented a framework to collapse fold-change data from simple repression onto the functional form

$$\text{fold-change} = \frac{1}{1 + e^{-\beta F(c)}}, \tag{7.3}$$

where $F(c)$ represents the free energy difference between the (active) repressor bound state and the empty promoter. In this section, we investigate how this data collapse could be extended to other regulatory architectures.

More precisely, the functional form in Eq. 7.3 is well suited to model the fold-change of repressed systems, which naturally lie between 0 (for a highly repressed system) and 1 (for a minimally repressed system). However, other architectures such as simple activation cannot be collapsed using Eq. 7.3 since no value of $F(c)$ could account for fold-change values greater than 1. In this section, we present a general framework to collapse fold-change data from any transcriptional regulation architecture. In the present analysis, we focus on collapsing the probability that RNAP is bound the promoter (p_{bound}) as a function of the effective free energy of the system and then translate this result into fold-change measurements.

We begin by rewriting Eq. 2.1 for p_{bound} as

$$p_{\text{bound}} = \frac{p}{1 + r_A + r_I + p}, \tag{7.4}$$

where we have defined

$$p = \frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_P}, \tag{7.5}$$

$$r_A = \frac{R_A}{N_{NS}} e^{-\beta \Delta\varepsilon_{RA}}, \tag{7.6}$$

and

$$r_I = \frac{R_I}{N_{NS}} e^{-\beta \Delta\varepsilon_{RI}} \tag{7.7}$$

to be the free energy differences between the unbound promoter and the RNAP bound, active-repressor bound, and inactive-repressor bound states shown in Fig. 2.2, respectively. We can rewrite p_{bound} in the form

$$\begin{aligned} p_{\text{bound}} &= \frac{1}{1+\frac{1+r_A+r_I}{p}} \\ &\equiv \frac{1}{1+e^{-\beta F_{\text{bound}}}} \end{aligned} \tag{7.8}$$

where in the last step we defined the free energy of p_{bound} for simple repression,

$$e^{-\beta F_{\text{bound}}} = \frac{1+r_A+r_I}{p}. \tag{7.9}$$

F_{bound} represents the effective free energy difference if all of the RNAP-bound states were grouped together into a single state and all of the RNAP-unbound states were grouped together into another state. Because $0 \leq p_{\text{bound}} \leq 1$, such an F_{bound} will always exist.

Since the parameter p, which includes both the RNAP copy number P and binding energy $\Delta\varepsilon_P$, is often poorly characterized, experimental measurements focus on fold-change. Recall that fold-change is given by the ratio of p_{bound} in the presence and absence of repressor (see Eq. 2.2). In the limit of no repressor, the free energy $F_{\text{bound}}^{(R=0)}$ satisfies

$$e^{-\beta F_{\text{bound}}^{(R=0)}} = \frac{1}{p}. \tag{7.10}$$

Therefore, the fold-change equation Eq. 2.3 can be rewritten as

$$\text{fold-change} = \frac{1}{1+r_A+r_I} = e^{\beta\left(F_{\text{bound}}-F_{\text{bound}}^{(R=0)}\right)}. \tag{7.11}$$

We recover Eq. 2.3 exactly by employing the approximation $r_I \ll 1 + r_A$ from Chapter 2, which assumes that the inactive repressor binds weakly to DNA.

Fig. 7.3 shows the data from Fig. 2.7B collapsed onto this new functional form. Note that since Eq. 7.11 has the form of an exponential function, the data collapses onto a straight line when fold-change is plotted on a logarithmic axis. By definition, $F_{\text{bound}} - F_{\text{bound}}^{(R=0)} \leq 0$ for simple repression, since the fold-change can only be decreased by the addition of repressors. Activation would obey $F_{\text{bound}} - F_{\text{bound}}^{(R=0)} \geq 0$, providing an elegant way to immediately categorize the mechanism of transcriptional regulation within this framework.

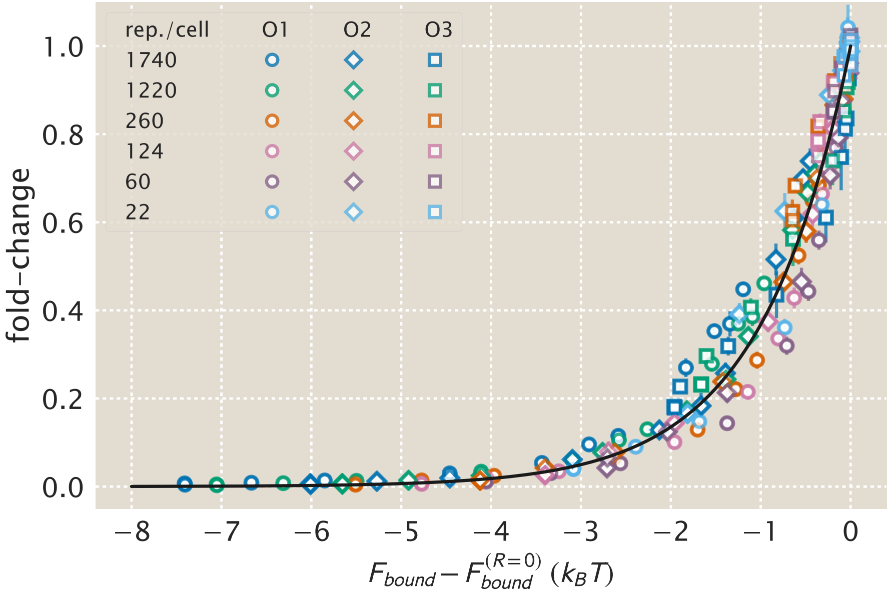


Figure 7.3: A General data collapse procedure. The difference between the effective free energy F_{bound} in the presence and absence of repressor (Eqs. 7.9 and 7.10) is plotted versus fold-change. The data collapses onto the form of the exponential in Eq. 7.11, and this procedure can be readily extended to arbitrary transcriptional regulation architectures.

Chapter 8

CONCLUSION

The proceeding chapters chronicled my journey to model a diverse set of allosteric systems ranging from enzymes to transcription factors to ion channels. This final chapter provides an opportunity to reflect on these accomplishments and to consider how future work may build upon these results to carry the field forward.

8.1 The Power of Quantitative Modeling

In their first week of classes, Caltech freshmen use Newton's laws to compute the parabolic trajectory of a ball thrown into the air. Five weeks later, they utilize these same laws to predict whether the energy and angular momentum of a rolling ball enables it to roll up a step. The following term, they learn that when a ball is thrown at relativistic velocities, Newtonian mechanics breaks down, suggesting that a new mechanism governs objects traveling near the speed of light. The unity of this approach stems from the fact that it is driven by a fundamental theory, rather than by examining each individual phenomenon as a separate conceptual challenge.

Yet in the present age of high-throughput experiments, machine learning, and big data, it is worth pausing to reexamine what it means to truly master a system. For example, is it enough to create a neural network that can fit data? While such progress, fueled by immense creativity and great effort, certainly deserves recognition, in my mind it does not suggest mastery of the material on par with our understanding of Newton's laws.

One of the themes of my work has been to attempt to hold biology to the same lofty standards as physics. The models developed in the preceding chapters emphasized the fundamental principles governing a system and the intuition that arises from tying their behavior to experimentally-tunable parameters. Much as in Caltech's physics classes, such quantitative frameworks enable us to generalize our results and predict how other systems would behave. And when we encounter data that runs counter to our predictions, these models reveal that a novel mechanism is at play that merits further exploration.

This work emphasized that in many instances, modeling can add a layer of understanding that is complementary to what additional experiments could provide.

Notable highlights include: (1) We explained how seemingly counter-intuitive non-monotonic activity curves can arise in the context of enzymes and showed that allostery was sufficient to drive this phenomenon (Figs. 1.12 and 1.14). (2) We demonstrated that biology can be highly predictive by using the induction profile of a transcription factor under one set of conditions to predict its behavior in seventeen other physiologically relevant conditions (Figs. 2.4 and 2.5). (3) We showed that the seemingly complex effects of mutations on input-output responses can be decoupled by mapping the mutations onto a subset of physical parameters governing the system (Figs. 3.5 and S4.1). (4) We made sense of curious patterns in ion channel mutants and confirmed our predictions of how these mutations affected the survival of the host cells via channel leakiness (Figs. 5.6 and S5.5). (5) We pushed our models to their utmost limits and showed that they are able to predict the gene expression of over 10 000 promoters (Figs. 6.2 and 6.3).

Each of these cases exemplifies how biophysical models provide a unique perspective that can uncover novel connections and make sharp predictions about a system's behavior. These projects embody my love of statistical mechanics and my belief that simple models can explain the behavior of our complex world. Most importantly, they weave a narrative of unity by suggesting that disparate biological phenomena can be collapsed down to their underlying mechanisms, which both deepens our understanding of these systems and helps guide future experiments.

8.2 The Future of Quantitative Modeling

In this final section, I turn my gaze from my personal accomplishments and speculate upon how the style of thinking embodied in this work may advance the field of biology in the coming decades.

Being able to predict the effects of general mutations would revolutionize many fields of biology from the study of genetics and evolution to the branches of microbiology and biochemistry. One of my personal interests lies in synthetic biology, where many groups aim to construct novel proteins (e.g. transcription factors that are inducible by a new effector molecule) or create biosensors (e.g. proteins that specifically attack cancer cells). Such efforts often begin by combining components of known molecules, but the resulting proteins may have small signal-to-noise (high leakiness or low saturation) or may only function at physiologically inaccessible concentrations (have an $[EC_{50}]$ that is too large or too small) and thus must be altered by mutations to enhance their functionality.

This problem is being tackled experimentally, with recent high-throughput methods able to measure tens of thousands of mutations *in vivo*, and this number will surely continue to rise in the coming years. Many computational groups are also refining models of mutations, and programs such as Rosetta and Evolutionary Coupling have seen great success in specific contexts.

One feature that I think is missing from such efforts is a theoretical approach that coarse-grains the myriad details in both the experimental and computational methods and searches for general principles that govern how mutations behave. For example, operator mutant data in Chapter 3 suggests that making any mutation in residues 17 or 18 of LacI will shift its input-output response along a one-parameter family of curves (see Fig. 3.4). Such insights could be missed by both of the above approaches by focusing exclusively on the response of each mutant.

An ideal theoretical model would predict what mutations could alter a given input-output response into a desired response. While such a framework is not yet feasible, the projects above hint that it is possible to predict how certain combinations of mutations interact. Another major area of research would be to harness comprehensive mutagenesis studies that examine every point mutation in a protein and map exactly which of the physical parameters are altered by these mutations. Lastly, once point mutation have been examined, it is imperative to consider how multiple mutations interact. A first step in this direction could be to analyze how mutations in the same region of a protein (i.e. two or more mutations in the LacI DNA binding site) affect protein function. Resolving these questions would greatly bolster our ability to navigate the vast space of protein mutations, because once we characterize n individual mutations, we could predict how all 2^n mutants (with or without each mutation) would behave, providing an exponential explosion in our understanding.

In all likelihood, new methods that combine theory, experiment, and computation will lead to revolutions not just in predicting the effects of mutations, but in all areas of biology. There is no doubt that this is the century of the life sciences, as the air sizzles with the excitement sparked by recent discoveries. As I look back on all that I have learned these past six years, I remain awed by nature's beauty, complexity, and ingenuity, and by the immense privilege of being a scientist and getting to tackle some of greatest mysteries of life.