MOLECULAR FOUNDATIONS OF ACCESSIBILITY IN GENOTYPE-PHENOTYPE MAPS

by

ANNELIESE JAEHNE MORRISON

A DISSERTATION

Presented to the Department of Chemistry and Biochemistry
and the Division of Graduate Studies of University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

December 2021

DISSERTATION APPROVAL PAGE

Student: Anneliese Jaehne Morrison

Title: Molecular Foundations of Accessibility in Genotype-Phenotype Maps

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy in the Department of Chemistry and Biochemistry by:

James Prell	Chair
Brad Nolen	Core Member
Raghuveer Parthasarathy	Core Member
Patrick Phillips	Institutional Representative
Michael Harms	Advisor

and

Krista Chronister	Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Division of Graduate Studies

Degree awarded December 2021

DISSERTATION ABSTRACT

Anneliese Jaehne Morrison

Doctor of Philosophy

Department of Chemistry and Biochemistry

December 2021

Title: Molecular Foundations of Accessibility in Genotype-Phenotype Maps

How do the biochemical properties of proteins shape evolution? Addressing this question is central to solving issues facing humanity, from rapidly evolving antibacterial and pesticide resistant organisms to achieving predictive protein engineering. Although advancements in sequencing and screening methodologies have made functional characterization of millions of mutations plausible, a predictive understanding of how proteins evolve is still lacking. Understanding how proteins evolve requires detailed knowledge of the map evolution navigates—the genotype-phenotype map—and where major sources of unpredictability, such as epistasis, come from. The genotype-phenotype map is determined by universal physical and biochemical rules. These rules dictate how macromolecules fold into functional forms, how components in regulatory networks interact, and how organism respond to environmental fluctuations. This defines what is—and is not—accessible to evolution. In this dissertation, we address two factors that shape evolutionary accessibility in proteins: 1) how epistasis can arise from the thermodynamic ensemble of macromolecules and 2) how the distribution of protein function in genotype-phenotype maps facilitates the evolution of new functions.

This dissertation includes previously published and unpublished material.

CURRICULUM VITAE

AUTHOR: Anneliese Jaehne Morrison

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR
Florida State University, Tallahassee, FL

DEGREES AWARDED:

Doctor of Philosophy, Chemistry and Biochemistry, 2021, University of Oregon
Bachelor of Science, Chemistry and Biochemistry, 2015, Florida State University

PUBLICATIONS

Morrison A. J. and Harms M. J., Ensemble epistasis is pervasive in the lac repressor (*in preparation*).

Morrison A. J., Wonderlick D. R., and Harms, M. J. (2021). Ensemble epistasis: thermodynamic origins of non-additivity between mutations. *Genetics* 219(1): iyab105.

Wheeler L. C., Anderson J. A., **Morrison A. J.,** Wong C. E., Harms, M. J. (2018) Conservation of specificity in two low specificity proteins. *Biochemistry* 57(5): 684-695.

Sackman A.M., McGee L.W., **Morrison A. J.**, Peirce J., Anisma J., Hamilton H., Sanderbeck S., Newman C., and Rokyta D. R. (2017) Mutation-driven parallel evolution during the first step of adaptation. *Molecular Biology and Evolution* 34(12): 3243-3253.

McGee L. W., Sackman A. M., **Morrison A. J.**, Pierce J., Anisman J., and Rokyta D. R. (2016) Synergistic pleiotropy overrides the cost of complexity in viral adaptation. *Genetics* 202(1): 285-295.

Zhang L., **Morrison A. J.**, and Thibodeau P. H. (2015) Stabilization of serralysin protease from Serratia marcescens by interdomain interactions. *PLoS ONE* 10(9): e0138419.

McGee L. W., Aitchison E. W., Caudle S. B., **Morrison A. J.**, Zheng L., Yang W., and Rokyta D. R. (2014) Payofs, not tradeoffs, in the adaptation of a virus to ostensibly conflicting selective pressures. PLoS genetics 10(10): e1004611.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Mike Harms, for always encouraging and supporting me with kindness, understanding, and patience. I will always be thankful that I was able to grow as a scientist with his guidance. I would like to thank all past and current Harms lab members for all of the advice, feedback, support, and laughter. I would like to extend a special thanks to undergraduate Daria Wonderlick, who I consider to be an immensely talented colleague and whose career I'm excited to watch unfold. I would like to extend thanks to Doug Turnbull, Maggie Weitzmann, and Jeff Bishop of the GC3F for their help with my next-generation sequencing projects and flow cytometry experiments. I would like to thank all my former mentors, Dr. Brian Miller, Dr. Darin Rokyta, Dr. Patrick Thibodeau, Dr. Liang Zhang, Brian Caudle, and Dr. Carl Whittington for inspiring me as a young scientist and supporting me at various points throughout my career. Finally, I would like to give a special thanks to my family, Fabiola Lara, Nicole Paterson, the Paterson family, and Roland and Sasha for the support and adventures throughout graduate school.

I dedicate this work to my mother Karen, my father Jerry, my sisters Rachel, Fabiola, and Nicole, and my brothers Eric, Brandon, and Zach.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

Macromolecules are the fundamental building blocks of biology. DNA encodes the information necessary to build organisms, acting as a blueprint, while proteins and RNA execute the processes required to sustain life. Proteins, in particular, play incredibly diverse roles—from the catalysis of reactions during metabolism to molecular recognition of pathogens during the innate immune response and ion transport during the transmission of nerve impulses in the nervous system.

This rich functional diversity arose from the accumulation of changes to the amino acid sequences of proteins. An amino acid sequence encodes the three-dimensional structure, biophysical and biochemical properties, and function of a protein under a given set of environmental conditions. Changes to an amino acid sequence, called mutations, can often be directly mapped to changes in the biochemical properties of a protein, its function and, in some cases, changes in the morphological, developmental, or physiological properties of an organism. For example, a single mutation in the influenza neuraminidase protein weakens its ability to bind to the antiviral drug oseltamivir, conferring drug resistance[1]. Amino acid changes also underlie many genetic diseases in humans—from sickle cell anemia to cystic fibrosis[2].

Often, more than just a single amino acid change is required to alter function. Despite intense interest across disciplines—from evolutionary biologists to biophysicists and protein engineers—a predictive understanding of how multiple mutations, together, encode structural and functional changes is still lacking. Understanding how mutations map to functional changes is critical to uncovering the molecular mechanisms that underlie macroscopic biological phenotypes, predicting the evolution of new functions such as antibiotic and pesticide resistances, and designing proteins with desirable functions. To achieve such a predictive understanding of biology and evolution, we must have detailed knowledge of how mutations map to molecular level properties and where key sources of unpredictability in this mapping arise.

The rules of biochemistry determine accessibility in the genotype-phenotype map

How do the molecular properties of proteins influence the accessibility of one function from another? Understanding the determinants of accessibility requires knowledge of the mapping between amino acid sequence space and protein function. Sequence space contains the set of all possible amino acid sequences (genotypes). The "rules" set by biochemistry and biophysics determine the distribution of function in sequence space. This distribution strongly determines what genotypes are, and are not, accessible from a given point in sequence space.

A simple word game can be used to illustrate the relationship between the genotype-phenotype map and accessibility. Imagine transforming from **WORD** to **GENE** by sequentially changing single letters from those in **WORD** to those in **GENE**. Meaningful English words are viable, while nonsense words are not. Accessibility of **GENE** from **WORD** requires a connected path of adjacent viable words. In word space,

the distribution of word meaningfulness and connections between them strongly shapes accessible paths from **WORD** to **GENE**[3] (Fig 1.1).

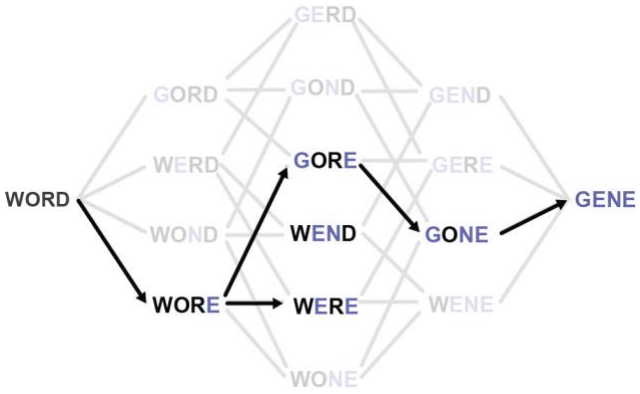


Fig 1.1 John Maynard Smith's word game: an analogy for understanding how the mapping between protein sequence and function shapes the accessibility of new functions. A) An illustration of John Maynard Smith's word game between WORD (black) and GENE (blue)[3]. Transparent intermediate words are non-meaningful, while fully opaque words are meaningful. Connections that lead to/from non-meaningful words are greyed out, while those that connect meaningful words are black.

Analogously to the word space in Fig 1.1, protein sequence space can be imagined as the set of intermediate genotypes between two sequences, and we can map functional properties onto each genotype. Proteins that function at or above some threshold (i.e., for viability or function) are connected to viable genotypes a single mutation away. Sequences with function below the threshold are disconnected. There are also sequences that are functionally viable but not accessible because they are surrounded by inviable sequences (see "WEND" in Fig 1.1). Connected sets of sequences form networks of sequences that are accessible from one another. Because evolutionary trajectories can only traverse connected functional sequences, the distribution of function

is a primary determinant of what is possible and impossible for biology and the protein evolution.

What is the distribution of function in real protein sequence space? This has traditionally been a difficult question for experimentalists to answer because sequence space is inherently vast. In fact, for a very small 100 amino acid protein the number of possible protein sequences in sequence space ($20^{100} \sim 1.3 \times 10^{130}$) is far greater than the number of atoms in the universe[4].

Two main approaches have been taken to study protein sequence space by effectively "shrinking" it to an experimentally tractable size. In one approach, sequence space is sparsely sampled via mutational scanning, directed or experimental evolution experiments[5,6]. These methods yield global information about the mapping between genotype and function, but the resolution is limited by the nature of sparse sampling. Another approach overcomes this limitation by completely characterizing a relatively small volume of sequence space (i.e. all combinations of 9 mutations: $2^9 = 512$ genotypes)[5,7,8]. In this approach, the function of all combinations of a small set of mutations is measured, leading to a high-resolution picture of the mapping between genotype and function. Though higher resolution than the first approach, the second approach is severely limited by the volume of sequence space that can be characterized using current experimental methods. This has led to conclusions about the nature of accessibility in proteins sequence space that may not hold for transitions in function that required navigating much larger volumes of sequence space.

Both approaches have overwhelmingly concluded that intramolecular epistasis tends to constrain the distribution of function in sequence space[9–20]. Epistasis occurs

when the effect of a mutation depends on the presence or absence of other mutations[10,20–22]. Epistasis makes it incredibly difficult to predict the effects of two mutations when introduced in combination. This, in turn, makes it difficult to understand the mapping between genotype and phenotype. Ultimately, this makes protein evolution deeply unpredictable, as past mutations influence the effects of future mutations[11,12,14,15,18–20,22–34].

To build a predictive understanding of protein biochemistry and the evolution of new functions, we must understand 1) the mechanistic origins of epistasis and 2) what the distribution of function looks like in large volumes of protein sequence space and how that influences accessibility.

Epistasis is a ubiquitous feature of proteins that can arise from basic protein biochemistry

Epistasis was first defined in 1909 by Bateson who was describing the dependence of the fitness effects of mutations on the genetic background in which they occurred[22,35]. Later, Fisher defined it as a statistical deviation from the additive combination of two mutations[36]. Since it was originally defined, it has been revealed as a ubiquitous feature of biology, including proteins[10,21,22,24,34]. Epistatic interactions within proteins and other macromolecules have been particularly well-studied for an extensive range of phenotypes—from thermodynamic stability to substrate specificity, allostery, function, and bacterial fitness[12,14,15,18,24,30,37–45]. Such epistasis impairs our ability to understand how sequence changes map to function, imparting unpredictability to our efforts identify the key factors that shape protein evolution and to engineer proteins with desirable functions[10,12,15,23,26,32,46–48].

Though exceedingly common, the specific molecular mechanisms that lead to epistasis are often unclear. However, in recent years an extensive effort has been put forth to uncover how epistasis arises from the biochemical properties of proteins and other macromolecules. Two classes of epistasis have emerged from such studies: nonspecific epistasis and specific epistasis[24,49].

Specific epistasis refers to mutations that impact a small number of other mutations, typically involving direct physical contact mechanisms (i.e., between other residues or ligands) or indirect mechanisms, where a mutation causes a structural change that influences the effect of another distantly located mutation[15,24,42,44,49–51,51–55]. For example, the evolution of glucocorticoid receptor ligand specificity was contingent upon the interaction between a network of two residues—one in direct contact with the new ligand and the second distant, but critical for repositioning the first residue such that it was able to contact the new ligand[44]. Such "specific" epistatic mechanisms are proposed to make evolution historically contingent on a small set of mutations, ultimately constraining evolution[49].

Nonspecific epistasis refers to mutations that additively impact some biophysical property of a protein (i.e., stability or ligand binding affinity) but non-additively impact a biological observable (i.e., expression level or fitness) due to the nonlinear mapping between the two properties[1,24,39,47,49,56–58]. One of the most well-studied mechanisms of such epistasis is stability-mediated epistasis where mutations accumulate that additively affect protein stability but exhibit epistasis at the level of function or fitness once destabilizing mutations accumulate such that the protein unfolds and becomes non-functional[1,12,14,59–63]. Our knowledge of threshold epistasis has led not only to deeper

understanding of specific evolutionary trajectories but also to improving conventional practices in protein engineering, where proteins are pre-stabilized prior to introduction of function-altering mutations[1,59,61,64,65]. Nonspecific epistasis has been proposed to be less constraining than specific epistasis, perhaps even opening new pathways for adaptation towards new functions[24,49].

Increasing our understanding of the molecular sources of epistasis is key to improving prediction efforts and our knowledge of sequence-function relationships. Chapters II and III in this dissertation focus on using theory, computation, and experiment to formally define and test for a particular mechanism of nonspecific epistasis called "*ensemble epistasis*". Such ensemble epistasis arises as a consequence of the nonlinear mapping between an experimental observable and a universal property of macromolecules: the thermodynamic ensemble. Such thermodynamic ensembles are critical to disparate biological processes, such as signaling [66], catalysis and enzyme promiscuity [67–71], and molecular recognition [72,73], and can tune biological output in response to environmental changes [74]. The pervasiveness of epistasis and ensembles in biology suggest that this may be a widespread mechanism of epistasis.

Simple theoretical models indicate that conclusions from small genotype-phenotype maps may not be informative for larger transitions in function

Many of the examples discussed above refer to small, low-dimensional genotype-phenotype maps that largely conclude that the distribution of protein function is heavily constrained by epistasis, as it effectively reduces the number of accessible trajectories through sequence space[5,8,19,75]. Currently, less than 20 studies of combinatorially complete natural evolutionary sequence spaces exist—the largest of which studied all

evolutionary intermediates for a trajectory of 9 mutations (2^9 = 512 genotypes total)[7,8,76]. In a small sequence space, even a small number of non-functional proteins will drastically shrink the number of connection between genotypes and consequently, paths[18,49,77,78]. It is unclear if these properties persist in larger, biologically relevant spaces, where the evolution of new features often occurs.

Theoretical studies have been able to access much greater volumes of sequence space. These studies predict that increasing the size of sequence space increases connectivity between genotypes in two ways: 1) vast sets of genotypes with nearly equal phenotypes span the entire space and 2) indirect paths—paths where substitutions are gained and lost—circumvent inaccessible direct paths[79–88]. Highly connected genotype networks and indirect paths may work together to facilitate evolutionary robustness—the ability to retain function in the face of mutation—as well as innovation, allowing evolution to avoid dead-ends[14,80,83,84,87–93]. Empirical studies of RNA and protein structure evolution have demonstrated evidence for vast genotype networks in sequence space[91,93–97]. Recent advances in high-throughput methodology has led to experimental support of indirect paths facilitating global accessibility, although these studies do not focus on naturally evolved features[86].

These studies suggest that evolution in high-dimensional sequence spaces may be much more relaxed; it may be able to navigate large volumes of connected sets of sequences without encountering evolutionary dead-ends and may be far less constrained by epistasis than currently appreciated. Because there is currently no experimental data for large combinatorially complete sequence spaces, it is unclear if natural protein evolution is highly constrained by epistasis, with very few available evolutionary

trajectories, or if it is more relaxed, with many connected neutral networks of viable sequences and many possible trajectories.

Understanding how the distribution of function and epistasis work together to shape evolution in high-dimensional sequence spaces requires measuring a larger genotype-phenotype map than ever before. Chapter IV of this dissertation will begin to address this gap in knowledge by describing the experimental characterization of a natural evolutionary transition that occurred over 15 substitutions in fluorescence color in coral GFP-like proteins.

Chapter-by-chapter break down

Chapter II lays the formal mathematical and theoretical foundations of a particular mechanism of epistasis called "*ensemble epistasis*". We used a simple analytical model and a virtual deep mutational scan to 1) determine the minimal necessary thermodynamic conditions under which we expect observe ensemble epistasis and 2) determine if it is a plausible mechanism of epistasis in a realistic model of proteins. We found that ensemble epistasis arises when three or more conformations are populated and when mutations have different effects on different conformations. Our virtual deep-mutational scan of the S100A4 protein in ROSETTA tested for the plausibility of satisfying these requirements in a more realistic model of proteins. We found that ensemble epistasis arises in 47% of the mutation pairs examined. We also found that it leads to all types of evolutionarily important classes of epistasis. The pervasive nature of ensemble epistasis in our dataset and the importance of ensembles in biology led us to conclude that ensemble epistasis is likely common in real biological systems. This chapter was published as an article in the journal *Genetics* and was co-authored with Daria Wonderlick and Prof. Michael J Harms.

An important consequence of the work in Chapter II was that it showed that one might test for ensemble epistasis by looking for effector-dependent patterns of epistasis. Chapter III shows the first experimental tests for ensemble epistasis by looking for effector-dependent epistasis in the lac repressor protein. We measured *in vitro* operator binding and *in vivo* gene expression for four mutant cycles of the lac repressor and found that signatures of ensemble epistasis are present in all mutant cycles. We used thermodynamic models to show how specific changes to the underlying ensemble result in specific epistatic patterns. We find that the signal for ensemble epistasis peaks under environmental conditions where many conformations are populated during biologically important functional transitions, here during induction. We conclude that ensemble epistasis is likely an extremely common mechanism of epistasis in real macromolecules that may be identified by measuring epistasis as a function of environmental changes such as effector or ligand concentration. This manuscript is currently in preparation and is co-authored with Prof. Michael J Harms.

Chapter IV describes ongoing work to exhaustively characterize the genotype-phenotype map for a transition in the fluorescence color of GFP-like proteins from *Faviina* corals[53,98]. We use simple theoretical simulations to illustrate that the nature of evolutionary trajectories in small genotype-phenotype maps may be distinct from those in large genotype-phenotype maps. Such considerations led us to the prediction that we may capture this distinction if we measure a 15-site map. We identified this natural evolutionary transition in fluorescence color that occurred over the course of 15 substitutions, resulting in a map that is 96 times larger than any previously characterized genotype-phenotype map[7]. We developed a high-throughput sort-and-sequence protocol

that couples fluorescence activated cell sorting (FACS) with next-generation sequencing to characterize the library.

In our first experimental replicates we were able to characterize the green and red fluorescence intensities of ~7% of the full map (3,676 out of 49,152 genotypes). We find that most of the map is green (~79%) or dead (~20%) and only a tiny fraction is red (~1.5%). We examined the extent of pairwise and third-order epistasis in all possible mutant cycles. We find that epistasis is extensive for both phenotypes and leads to all evolutionarily important classes of epistasis. We conclude that the feature of our subsampled map is consistent with the conclusions of other small genotype-phenotype maps: epistasis makes the landscape rugged, generally constraining accessibility. Future work will look at how the effects of epistasis and the extent of neutral networks shape the full, high-dimensional map. We anticipate that epistasis will have much less of an impact on accessibility as the sheer number of possible pathways and genotypes may generate extensive neutral networks, overwhelming epistatic effects.

CHAPTER II

ENSEMBLE EPISTASIS: THERMODYNAMIC ORIGINS OF NON-ADDITIVITY BETWEEN MUTATIONS

Author Contributions

Anneliese Morrison and Michael Harms conceptualized the study and designed experiments. Michael Harms acquired funding for the study. Anneliese Morrison, Daria Wonderlick, and Michael Harms performed the experiments. Michael Harms administered the project. Anneliese Morrison and Daria Wonderlick analyzed the data. Anneliese Morrison constructed figures. Anneliese Morrison and Michael Harms wrote the manuscript. All authors read and approved the manuscript.

Abstract

Epistasis—when mutations combine non-additively—is a profoundly important aspect of biology. It is often difficult to understand its mechanistic origins. Here we show that epistasis can arise from the thermodynamic ensemble, or the set of interchanging conformations a protein adopts. *Ensemble epistasis* occurs because mutations can have different effects on different conformations of the same protein, leading to non-additive effects on its average, observable properties. Using a simple analytical model, we found that ensemble epistasis arises when two conditions are met: 1) a protein populates at least

three conformations and 2) mutations have differential effects on at least two conformations. To explore the relative magnitude of ensemble epistasis, we performed a virtual deep-mutational scan of the allosteric Ca^{2+} signaling protein S100A4. We found that 47% of mutation pairs exhibited ensemble epistasis with a magnitude on the order of thermal fluctuations. We observed many forms of epistasis: magnitude, sign, and reciprocal sign epistasis. The same mutation pair could even exhibit different forms of epistasis under different environmental conditions. The ubiquity of thermodynamic ensembles in biology and the pervasiveness of ensemble epistasis in our dataset suggests that it may be a common mechanism of epistasis in proteins and other macromolecules.

Introduction

Epistasis—when the effect of a mutation depends on the presence or absence of other mutations—is a common feature of biology. Epistasis can hint at biological mechanism [44,99–103], profoundly shape evolution, and complicate bioengineering that involves simultaneously introducing multiple mutations [32,46,48]. It is therefore important to understand the general mechanisms by which epistasis can arise. Such knowledge will help us better understand biological systems, explain historical evolutionary trajectories, and improve models to predict the combined effects of mutations.

One important class of epistasis is that which occurs between mutations within a single protein. The magnitude of such epistatic interactions, ε, can be quantitatively described as shown in Fig 2.1A; it simply represents the difference in the effect of mutation $a \rightarrow A$ in the ab and aB backgrounds. Sometimes, such epistasis can be understood intuitively. In Fig 2.1B, epistasis arises because the positive charge of mutation $a \rightarrow A$ is adjacent to the negative charge of mutation $b \rightarrow B$. Epistasis occurs as a

result of an electrostatic interaction between charged residues. Sometimes, however, epistasis can be difficult to rationalize. Fig 2.1C shows epistasis between two positions distant in the structure. Where does such epistasis come from? Can it be predicted from an understanding of protein biochemistry?

We and others noted previously that the thermodynamic ensemble of a protein could potentially give rise to non-additive interactions between mutations [104,105]. Proteins exist as ensembles of interchanging conformations, where the probability of seeing an individual conformation is determined by its relative energy. The functional output of a protein is averaged over the functional properties and populations of all individual ensemble conformations [70,106,107]. Mutations can have different effects on each conformation, redistributing their relative probabilities in a nonlinear fashion. The effects of such mutations with respect to an observable would not sum additively, leading to *ensemble epistasis*.

Many important questions about ensemble epistasis remain unanswered. Under what conditions is ensemble epistasis expected to arise? Can it lead to different classes of evolutionarily-relevant epistasis, i.e. magnitude, sign, reciprocal-sign, and high-order? Is it plausible that such epistasis could occur in a real protein, rather than the highly simplified lattice models we used previously? And, finally, are there signals for ensemble epistasis that one might detect experimentally?

To address these questions, we set out to rigorously describe the thermodynamic and mechanistic basis for ensemble epistasis. We identified the minimal set of conditions that are necessary to observe ensemble epistasis: 1) a protein populates three or more conformations and 2) mutations have differential effects on two or more conformations

within the ensemble. We found that this can lead to many types of epistasis, including magnitude, sign, reciprocal sign, and high-order epistasis. From structure-based calculations on the allosteric S100A4 protein, we predict that a large fraction of mutant pairs in real proteins will exhibit ensemble epistasis. We also found that varying the concentration of allosteric effectors could tune epistasis, suggesting one might experimentally detect ensemble epistasis by measuring epistasis at different concentrations of allosteric effectors. We conclude that ensemble epistasis is likely an important determinant of non-additivity between mutations in proteins.

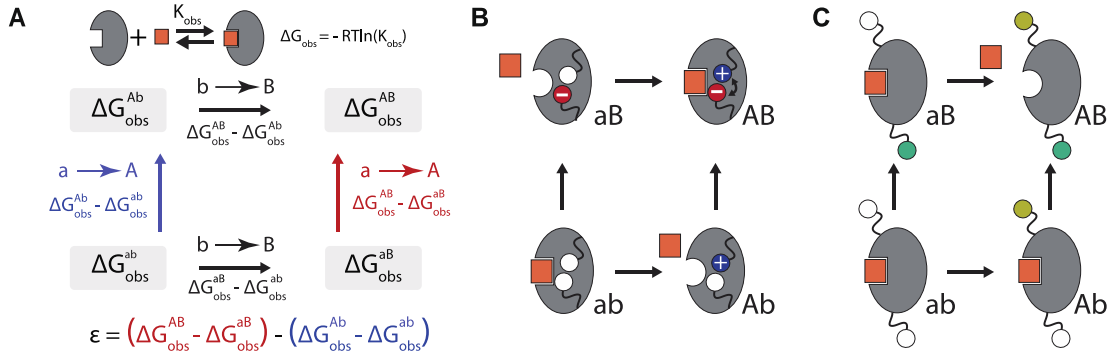


Figure 2.1 Mechanistic and mathematical descriptions of epistasis. A) The mathematical description of epistasis (ε) in ligand binding free energy (ΔG_{obs}) for the mutant cycle between genotypes *ab* and *AB*. ΔG_{obs} measures the strength of the binding interaction between protein (gray) and ligand (orange). We indicate genotypes as superscripts. ε is defined as the difference in the effect of mutation $a \rightarrow A$ in the *aB* background (red text), versus its effect in the *ab* background (blue text). B) Mutant cycle where epistasis is readily understood: the $a \rightarrow A$ and $b \rightarrow B$ mutations introduce charges into the hydrophobic core, destabilizing the protein and disrupting binding of the orange square. Mutations $a \rightarrow A$ and $b \rightarrow B$ lead to a new electrostatic interaction when introduced together (minus and plus signs) restoring stability and binding. C) Mutant cycle with difficult-to-understand epistasis. Mutations at two distant sites (green and yellow spheres) have no effect on binding of the orange square when introduced independently but disrupt binding when introduced together.

Results

Defining the three-conformation ensemble

To understand how the thermodynamic ensemble might lead to epistasis, we first defined a simple quantitative model of a protein exchanging between three conformations *i, j,* and *k*. We defined *i* as the "active" conformation in equilibrium with two "inactive" conformations *j* and *k.* This is a generic model that describes, in broad strokes, a wide variety of functions that depend on conformational change (Fig 2.2A). For example, conformation *i*, but not conformations *j* and *k*, could be capable of catalysis.

We will analyze epistasis in the free energy difference between the active *i* conformation and the inactive conformations, *j* and *k* (ΔG_{obs}). This quantifies how much the active form of the enzyme is favored over the inactive forms. We ΔG_{obs} as follows:

$$\Delta G_{obs} = G_i - \langle G_{j,k} \rangle \tag{4}$$

where G_i is the energy of conformation *i* and $\langle G_{j,k} \rangle$ is the Boltzmann-weighted average of the free energies of conformations *j* and *k* (Fig 2.2B). Importantly, the free energy scale is linear, meaning——in the absence of epistasis——we expect the effects of mutations on ΔG_{obs} to sum.

We will now describe the origin of equation 4. (Some readers may wish to proceed to the next section, *"Mutations can affect multiple conformations in the ensemble"*).

Due to thermal fluctuations, an individual protein molecule will flip between conformations *i, j,* and *k* over time. As a consequence, a population of many protein molecules will exhibit a mixture of conformations. Factors such as the number of

favorable chemical bonds within each conformation determine the frequency of that conformation in the protein population.

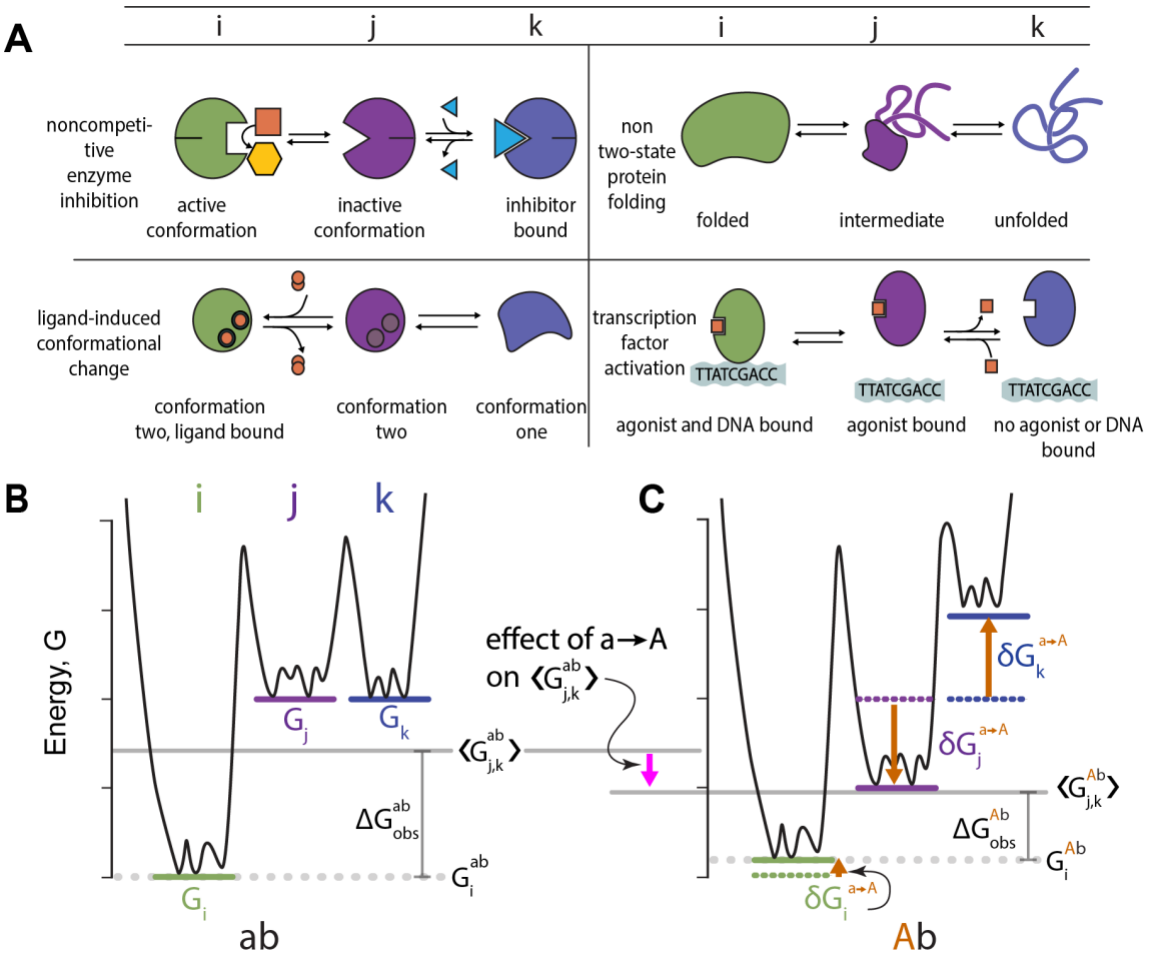


Figure 2.2 Mutations affect multiple ensemble conformations. A) Schematic examples of biological mechanisms in which a protein populates at least three conformations. Columns indicate conformation labels—*i* (green), *j* (purple), or *k* (blue). B) Energy diagram for a hypothetical protein with the *ab* genotype that adopts conformations *i* (green line), *j* (purple line), and *k* (blue line). The solid gray line indicates $\langle G_{j,k} \rangle$ (the average energy of the inactive conformations *j* and *k*) and the dotted gray line indicates G_i^{ab} (the energy of the active conformation *i*). The difference between the solid and dotted gray lines is the observable, ΔG_{obs}^{ab}. C) Hypothetical mutation *a*→*A* changes the energies of conformations *i*, *j*, and *k* and thus ΔG_{obs}. Orange arrows represent the effect of mutation *a*→*A* on individual conformations. For example, $\delta G_j^{a \to A}$ shows the effect on conformation *j*. The mutation has a small effect on *i*, stabilizes *j*, and destabilizes *k*. This leads to a net decrease in $\langle G_{j,k}^{Ab} \rangle$ relative to $\langle G_{j,k}^{ab} \rangle$ (pink arrow), and thus a decrease in ΔG_{obs}^{Ab} relative to ΔG_{obs}^{ab}.

The favorability of each conformation can be quantified by its free energy (*G*). Fig 2.2B shows a free energy landscape for a three-conformation ensemble. The large energy wells correspond to conformations *i*, *j*, and *k*, while the smaller wells correspond to small structural fluctuations within each conformation, such as side-chain rearrangements. Because conformation *i* has a low free energy in this hypothetical example, it will have a much higher frequency in the population than conformations *j* or *k*.

The statistical weight for a given conformation is related to its free energy by the Boltzmann distribution:

$$w_c = e^{-\frac{-G_c}{RT}} \tag{5}$$

where *c* indicates a conformation with free energy G_c, R is the gas constant, and T is the temperature in Kelvin. In the three-conformation ensemble, the frequency of conformation *i* is given by:

$$f_i = \frac{w_i}{w_i + w_j + w_k} = \frac{e^{\frac{-G_i}{RT}}}{e^{\frac{-G_i}{RT}} + e^{\frac{-G_j}{RT}} + e^{\frac{-G_k}{RT}}} \tag{6}$$

Importantly, the frequencies of the conformations are coupled. For example, making conformation *j* more stable (by decreasing G_j)will lower f_i, even if G_i remains the same. This is because individual protein molecules will spend more time in conformation *j* and thus less time, on average, in conformation *i*.

As noted above, we are modeling an ensemble in which conformation *i* is active and conformations *j* and *k* are not. A typical way to quantify activity in such a system is with an equilibrium constant, describing the frequency of *i* relative to *j* and *k*:

$$K_{obs} = \frac{f_i}{f_j + f_k} = \frac{e^{\frac{-G_i}{RT}}}{e^{\frac{-G_j}{RT}} + e^{\frac{-G_k}{RT}}} \tag{7}$$

Equilibrium constants follow a multiplicative scale, meaning that the effects of mutations are expected to multiply rather than add. We will take logarithm of K_{obs} and place the observable on a free-energy scale, where—in the absence of epistasis—mutational effects are expected to add:

$$\Delta G_{obs} = G_i + RTln\left(e^{\frac{-G_j}{RT}} + e^{\frac{-G_k}{RT}}\right) \tag{8}$$

ΔG_{obs} measures the difference in the free energy, at equilibrium, of the active *i* conformation and the inactive *j* and *k* conformations (Fig 2B). We will write the second term as:

$$\langle G_{j,k} \rangle = -RTln\left(e^{\frac{-G_j}{RT}} + e^{\frac{-G_k}{RT}}\right) \tag{9}$$

where the brackets denote the Boltzmann-weighted average. This gives us, finally:

$$\Delta G_{obs} = G_i - \langle G_{j,k} \rangle \,. \tag{10}$$

Mutations can affect multiple conformations in the ensemble

We next considered the effects of mutations. Because each conformation may have different physical interactions, the same mutation may have different effects on different conformations. For the three-conformation ensemble in Fig 2.2B, we thus need terms to describe the effect of the mutation on conformations *i*, *j*, and *k*. To keep track of these effects, we will use the following notation:

- The observable energy for genotype g is ΔG_{obs}^{g} (e.g., ΔG_{obs}^{ab}).
- The energy of conformation c is G_c^g (e.g. G_i^{ab}).
- The energetic effect of mutation $x \rightarrow X$ on conformation c is $\delta G_c^{x \rightarrow X}$ (e.g. $\delta G_j^{a \rightarrow A}$). Unless indicated, mutations are always introduced into the ab genetic background.
- Epistasis within a conformation——meaning the difference in the effect of $a \rightarrow A$ on the energy of conformation c in the ab and aB backgrounds——is $\delta\delta G_c^{ab \rightarrow AB}$.

We will now consider the effect of mutation $a \rightarrow A$ on ΔG_{obs}^{ab} (Fig 2.2C). The three terms that describe its effect are $\delta G_i^{a \rightarrow A}$, $\delta G_j^{a \rightarrow A}$, and $\delta G_k^{a \rightarrow A}$. Fig 2.2C shows how a hypothetical mutation $a \rightarrow A$ might change the ensemble: it has a small effect on conformation *i*, stabilizes *j*, and destabilizes *k*. We would describe the effect of the mutation mathematically as:

$$\Delta G_{obs}^{ab} = (G_i^{ab} + \delta G_i^{a \rightarrow A}) - \langle G_{j,k}^{Ab} \rangle \tag{11}$$

where

$$\langle G_{j,k}^{Ab} \rangle = -RT \left(e^{-\frac{(G_j^{ab} + \delta G_j^{a \rightarrow A})}{RT}} + e^{-\frac{(G_k^{ab} + \delta G_k^{a \rightarrow A})}{RT}} \right). \tag{12}$$

The mutation in Fig 2.2C stabilizes $\langle G_{j,k}^{Ab} \rangle$ relative to $\langle G_{j,k}^{ab} \rangle$ because conformation *j* becomes so much more favorable. As a result, the ΔG_{obs}^{Ab} is lower than ΔG_{obs}^{ab} (Fig 2.2C).

The next step is to describe the effect of introducing two mutations simultaneously. To isolate epistasis that arises solely from changes to the thermodynamic

ensemble, we will start by assuming that mutations are additive within each conformation. By this we mean that $G_c^{AB} = G_c^{ab} + \delta G_c^{a\to A} + \delta G_c^{b\to B}$. There are no epistatic contributions of the form $\delta\delta G_c^{ab\to AB}$ reflecting physical interactions within each conformation of the sort seen in Fig 2.1B. This means any epistasis we observe arises solely from the ensemble. We will revisit this simplifying assumption later.

Using this framework, we can describe the combined effects of mutations $a\to A$ and $b\to B$ on ΔG_{obs} as the following:

$$\Delta G_{obs}^{AB} = (G_i^{ab} + \delta G_i^{a\to A} + \delta G_i^{b\to B}) - \langle G_{j,k}^{AB} \rangle \tag{13}$$

where

$$\langle G_{j,k}^{AB} \rangle = -RT\left(e^{-\frac{(G_j^{ab}+\delta G_j^{a\to A}+\delta G_j^{b\to B})}{RT}} + e^{-\frac{(G_k^{ab}+\delta G_k^{a\to A}+\delta G_k^{b\to B})}{RT}}\right). \tag{14}$$

The thermodynamic ensemble can lead to epistasis

To understand the nature of epistasis arising from such a system, we must map the thermodynamic model in Equation 13 to epistasis. Table 1 shows the mapping between each genotype and its thermodynamic description, $\Delta G_{obs}^{genotype}$. We will treat epistasis as the quantitative difference between the effects of mutation a→A in the *ab* and *aB* backgrounds (Fig 2.1A):

$$\varepsilon = (\Delta G_{obs}^{AB} - \Delta G_{obs}^{aB}) - \left(\Delta G_{obs}^{Ab} - \Delta G_{obs}^{ab}\right). \tag{15}$$

We can substitute the thermodynamic equations for each ΔG_{obs} from Table 1 into Equation 15. Upon simplifying this expression (supplementary text, section 1.1 Appendix A), we obtain:

$$\varepsilon = -\left[\left(\langle G_{j,k}^{AB}\rangle - \langle G_{j,k}^{aB}\rangle\right) - \left(\langle G_{j,k}^{Ab}\rangle - \langle G_{j,k}^{ab}\rangle\right)\right]. \tag{16}$$

All terms associated with conformation *i* cancel. We are left with a description of ε that is only in terms of mutational effects on conformations *j* and *k*.

Our expression for ε is determined by the effects of mutations a→A and b→B on conformations *j* and *k*, not their effects on conformation *i*. Perturbations to the relative populations of *j* and *k* necessarily lead to nonlinear changes in ΔG_{obs} because the logarithmic term in $\langle G_{j,k}\rangle$ cannot be simplified further.

Table 1.1 Map between genotype and the thermodynamic description of $\Delta G_{obs}^{genotype}$

Genotype	$\Delta G_{obs}^{genotype}$	$\langle G_{j,k}^{genotype}\rangle$
ab	$G_i^{ab} - \langle G_{j,k}^{ab}\rangle$	$-RTln\left(e^{-\frac{(G_j^{ab})}{RT}} + e^{-\frac{(G_k^{ab})}{RT}}\right)$
Ab	$(G_i^{ab} + \delta G_i^{a\to A}) - \langle G_{j,k}^{Ab}\rangle$	$-RTln\left(e^{-\frac{(G_j^{ab}+\delta G_j^{a\to A})}{RT}} + e^{-\frac{(G_k^{ab}+\delta G_k^{a\to A})}{RT}}\right)$
aB	$(G_i^{ab} + \delta G_i^{b\to B}) - \langle G_{j,k}^{aB}\rangle$	$-RTln\left(e^{-\frac{(G_j^{ab}+\delta G_j^{b\to B})}{RT}} + e^{-\frac{(G_k^{ab}+\delta G_k^{b\to B})}{RT}}\right)$
AB	$(G_i^{ab} + \delta G_i^{a\to A} + \delta G_i^{b\to B}) - \langle G_{j,k}^{AB}\rangle$	$-RTln\left(e^{-\frac{(G_j^{ab}+\delta G_j^{a\to A}+\delta G_j^{b\to B})}{RT}} + e^{-\frac{(G_k^{ab}+\delta G_k^{a\to A}+\delta G_k^{b\to B})}{RT}}\right)$

Conditions necessary for ensemble epistasis

We next used the thermodynamic description of ensemble epistasis derived above (Equation 16) to ask under what conditions ensemble epistasis is expected to arise. In the supplementary text, we show that there are two necessary conditions for ensemble epistasis:

- The protein populates at least three conformations (supplementary text, Appendix A, section 1.2).
- Mutations have differential effects on conformations *j* and *k* (supplementary text, Appendix A, section 1.3).

To understand what these conditions mean in practice, we calculated ensemble epistasis using equation 16 as a function of the difference in the stabilities of conformations *j* and *k* ($G_j^{ab} - G_k^{ab}$) and the difference in the effects of mutations on conformations *j* and *k* ($\delta G_j^{x \to X} - \delta G_k^{x \to X}$) (Fig 2.3A). In panels B-D, we reveal the underlying ensemble that leads to the epistasis observed in Fig 2.3A. The length of the pink arrows illustrates the effect of mutation a→A in each genetic background, *ab* or *aB*. The difference in the length of the pink arrows for the *ab*→*Ab* and *Ab*→*AB* genotypes measures epistasis, ε.

We can see why multiple conformations are required for ensemble epistasis by comparing points B and C on Fig 2.3A. At point B, only conformation j is appreciably populated for all genotypes (pie charts, Fig 2.3B); at point C, conformations j and k have equal starting populations (pie charts, Fig 2.3C). This difference in the starting populations of j and k leads to different epistatic outcomes. At point B, both ab→Ab and aB→AB depend only on the effect of the mutation on conformation j because it is the

only conformation appreciably populated. The lengths of the pink arrows are equal, indicating that there is no epistasis. At point C, the effect of ab→Ab on $\langle G_{j,k} \rangle$ is moderate because the stabilization of conformation j is offset by the entropic cost of depopulating conformation k. This results in epistasis because when a→A is introduced into the aB background, mutation b→B has already depopulated conformation k. As a result, the effect of aB→AB is determined solely by its stabilization of conformation j, and is thus larger than ab→Ab.

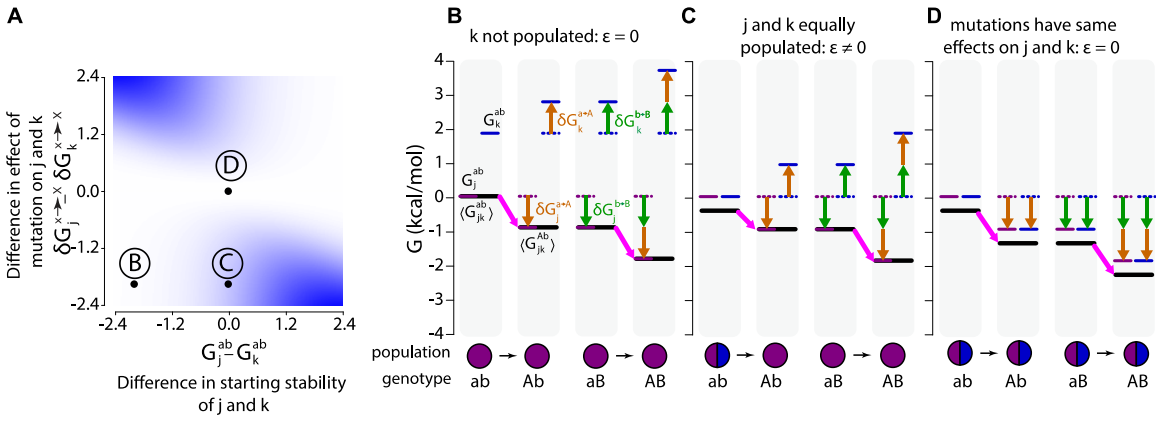


Figure 2.3 Ensemble epistasis arises from redistributed conformational probabilities. A) Epistasis as a function of the difference in the effects of the mutations a→A and b→B on conformations *j* and *k* ($\delta G_j^{x \to X} - \delta G_k^{x \to X}$) in $kcal \cdot mol^{-1}$, y-axis) and the difference in the stability of conformations *j* and *k* for the *ab* genotype ($G_j^{ab} - G_k^{ab}$ in $kcal \cdot mol^{-1}$, x-axis). Color indicates the magnitude of epistasis, ranging from *0* (white) to *1.6* $kcal \cdot mol^{-1}$ (blue). For the whole plot, a→A and b→B had identical effects ($\delta G_j^{a \to A} = \delta G_j^{b \to B}$ and $\delta G_k^{a \to A} = \delta G_k^{b \to B}$). We set $G_j^{ab} = 0\ kcal \cdot mol^{-1}$ and $\delta G_j^{x \to X} = -0.96\ kcal \cdot mol^{-1}$ and then varied G_k^{ab} and $\delta G_k^{x \to X}$ to sample parameter space. All calculations were done at T = 298 K. Panels B-D show the thermodynamic origins for the epistasis at points *B*, *C*, and *D* indicated on panel A. The color scheme is consistent throughout: purple and blue lines are the energies of conformations *j* and *k*, respectively; orange arrows show the effects of mutation a→A; green arrows show the effects of mutation b→B; heavy black lines are the Boltzmann-weighted average energies of *j* and *k*, $\langle G_{j,k} \rangle$; heavy pink arrows are the observed effect of mutation a→A in the genotype indicated below the plot. The difference between the length of the pink arrows in the *ab*→*Ab* and *aB*→*AB* genotypes measures ε. The relative populations of conformations *j* and *k* are shown as a pie chart below the energy diagram.

We can see why differential effects for each mutation are required by comparing points *C* and *D* on Fig 2.3A. At both points, conformations *j* and *k* have equal starting populations (pie charts, Fig 2.3 C-D). At point *C*, the mutations have opposite effects on conformations *j* and *k* (Fig 2.3C); at point *D*, the mutations have identical effects on conformations *j* and *k* (Fig 2.3D). This means that for point *D* the introduction of a→A or b→B shifts the total energy landscape, but does not change the relative proportions of *j* and *k*. As a result, mutation a→A has the same effect regardless of background (compare pink arrows, Fig 2.3D).

Ensembles can lead to magnitude epistasis, sign-epistasis, and reciprocal sign-epistasis

We next asked if the ensemble could lead to different evolutionarily relevant classes of epistasis: magnitude, sign, and reciprocal sign epistasis. In magnitude epistasis, only the magnitude of a mutation's effect changes when another mutation is introduced. In sign epistasis, the same mutation has a positive effect in one background and a negative effect in another. Finally, in reciprocal sign epistasis, both mutations exhibit sign epistasis.

We surveyed the parameter space for the effects of mutations on each conformation while tracking the magnitude and type of epistasis observed (Fig 2.4A). We set the initial energies of conformations *j* and *k* to be equal ($G_j^{ab} = G_k^{ab} = 0$). We then calculated epistasis using equation 16 as a function of the difference in the effects of mutations a→A and b→B on *j* and *k*.

We found four regimes, corresponding to magnitude, sign, reciprocal sign, and no epistasis. To understand the origins of these three regimes, we studied the thermodynamic

ensembles that lead to epistasis at the points indicated *C*, *D*, and *E*. At this slice of parameter space, mutation a→A destabilizes conformation *j* by $0.35\ kcal \cdot mol^{-1}$ and stabilizes conformation *k* by $-0.35\ kcal \cdot mol^{-1}$. The effect of this mutation on the ensemble in the *ab* background is shown in Fig 2.4B: the mutation mildly stabilizes $\langle G_{j,k} \rangle$.

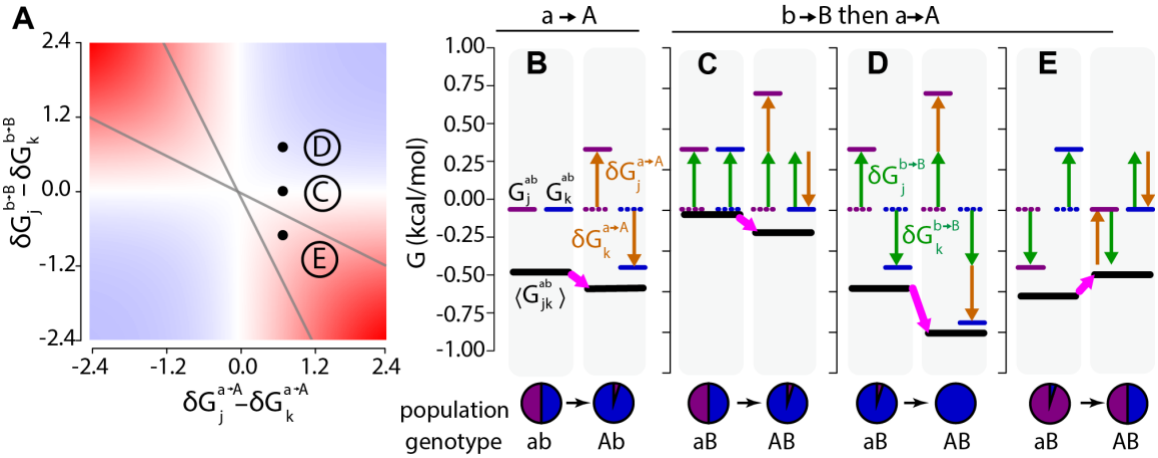


Figure 2.4 Ensemble epistasis arises when mutations have different effects on different conformations. A) Epistasis calculated for a three-conformation ensemble that starts with $G_j^{ab} = G_k^{ab} = 0$. The differences in the effects of mutations a→A and b→B on conformations *j* and *k* are indicated on the x- and y-axes. The magnitude of epistasis is indicated by the color, ranging from *+1.6* (dark red) to *0* (white) to *-1.6* $kcal \cdot mol^{-1}$ (dark blue). Gray lines delineate regions of reciprocal sign (red regions within the lines) and sign epistasis (red regions outside of the lines). All calculations were done at T = 298 K. Panels B-E show the thermodynamic origins of the epistasis indicated by points *C*, *D*, and *E* on panel A. The effect of mutation a→A is constant in all panels; the effect of mutation b→B differs depending on the scenario. The color scheme is consistent with Fig 2.2. B) The effect of a→A in the *ab* background. a→A destabilizes *j* and stabilizes *k*, stabilizing $\langle G_{j,k}^{AB} \rangle$. C) Scenario C: no epistasis. b→B has the same effect on conformations *j* and *k*. D) Scenario D: a→A and b→B act synergistically to destabilize *j* and stabilize *k*. E) Scenario E: a→A and b→B have opposite effects on conformations *j* and *k*.

At point *C*, we see no epistasis (Fig 2.4A). We can see why this occurs in Fig 2.4C. Mutation b→B destabilizes both *j* and *k* by *0.35* $kcal \cdot mol^{-1}$. Because mutation b→B does not have differential effects on each conformation, $\langle G_{j,k} \rangle$ is globally shifted by

$+0.35\ kcal \cdot mol^{-1}$. Introducing a→A and b→B together yields no epistasis because both the *ab* and *aB* genotypes have identical configurations—the observed effect comes only from mutation a→A (compare pink arrows in Fig 2.4B and Fig 2.4C).

At point *D*, we observe magnitude epistasis (Fig 2.4A). We can see why this occurs in Fig 2.4D. Mutations a→A and b→B have synergistic effects on each conformation: *k* is stabilized while *j* is destabilized. We see magnitude epistasis because although the relative population of *j* is reduced, it still has weight in the Boltzmann-weighted average stability (compare pink arrows in Fig 2.4B and 2.4D).

At point *E*, we see reciprocal sign epistasis (Fig 2.4A). We can see why this occurs in Fig 2.4E. a→A and b→B have opposite effects on *j* and *k*: a→A destabilizes *j* and stabilizes *k*, while b→B stabilizes *j* and destabilizes *k*. The effects are equal in magnitude but opposite in sign so their combined effects cancel, yielding $\langle G_{j,k}^{AB} \rangle$ equal to that of the *ab* genotype (compare pink arrows in Fig 2.4B and 2.4E). As a result, mutations a→A and b→B have individually stabilizing effects on $\langle G_{j,k} \rangle$ but are destabilizing when combined.

The magnitude and sign regions of Fig 2.4A show distinct patterns with regard to the sign of epistasis observed: mutations in the magnitude region are more stabilizing (positive epistasis) and those in the sign region are more destabilizing (negative epistasis) than anticipated based on single mutational effects. The magnitude region results in positive epistasis because mutations work synergistically to hyper-stabilize one conformation, while greatly destabilizing the other. This results in one conformation having very little weight in the Boltzmann distribution such that the remaining stabilized conformation determines the observable value. In the sign region, each mutation

preferentially stabilizes a different conformation when introduced alone. However, when introduced together, they have opposing effects within a single conformation. The stabilizing effects of each mutation alone on $\langle G_{j,k} \rangle$ cancel, resulting in a less stable double mutant than anticipated.

Ensemble epistasis may be a common feature in protein mutant cycles

Above we showed mathematically that ensemble epistasis can arise when multiple conformations are populated and mutations have different effects on different conformations. We next wanted to address whether these requirements are met in real systems. Multi-conformation ensembles are common in biology and we expect that the first requirement is often met (Fig 2.2A). However, it is not obvious that the requirement for differential effects of mutations is commonly satisfied. We designed a computational test to ask if it was plausible that both of these conditions are met simultaneously in a protein.

We investigated these questions using the allosteric Ca^{2+} signaling protein, human S100A4. S100A4 adopts a three-conformation ensemble, meeting our first requirement to observe ensemble epistasis (Fig 2.5A) [108–110]. In the absence of Ca^{2+} , it favors the "*apo*" conformation (Fig 2.5A, slate); addition of Ca^{2+} stabilizes the "*ca*" conformation with an exposed hydrophobic peptide-binding surface (Fig 2.5A, purple); finally, addition of *peptide* leads to formation the "*capep*" conformation that has both Ca^{2+} and *peptide* bound (Fig 2.5A, green). These structures can be be assigned indices, as in our analytical model: *capep* (*i*), *ca* (*j*), and *apo* (*k*).

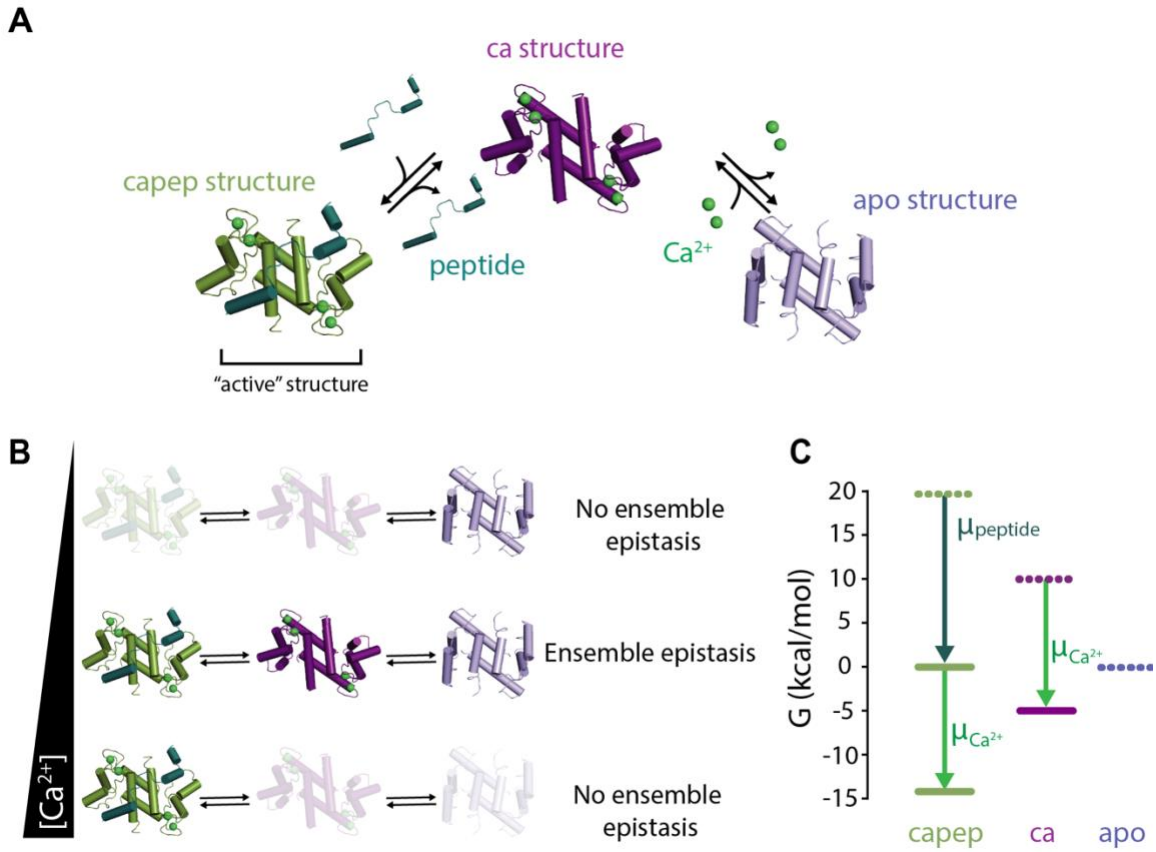


Figure 2.5 Testing for ensemble epistasis in the S100A4 protein. A) Three-conformation ensemble of the S100A4 protein. The apo conformation (*apo*, slate, PDB: 1M31) is in equilibrium with the Ca^{2+} bound (*ca*, purple, PDB: 2Q91) and Ca^{2+}/*peptide* bound (*capep*, green, PDB: 5LPU) conformations when Ca^{2+} (lime green spheres) and *peptide* (dark green) are present. B) The relative populations of the *apo*, *ca*, and *capep* conformations change as Ca^{2+} concentration increases in the presence of saturating peptide. The magnitude of ensemble epistasis observed is Ca^{2+} −dependent, because only some Ca^{2+}concentrations lead to multiple populated conformations. C) Assigned energies ($kcal \cdot mol^{-1}$) of S100A4 conformations. *Apo* is most stable when *peptide*, $\mu_{peptide}$, and Ca^{2+} chemical potentials, $\mu_{Ca^{2+}}$, are zero (dashed lines). *Capep* is stabilized by increasing $\mu_{peptide} = 20\ kcal \cdot mol^{-1}$ (dark green arrow, solid green line). Increasing $\mu_{Ca^{2+}}$ alters the energies of both *ca* and *capep* (lime green arrow, solid lines). All calculations were done at T = 298 K.

We used software for structure-based energy calculations (ROSETTA) to estimate the stability effects of all 3,382 possible single point mutations to the *capep*, *ca*, and *apo* conformations of S100A4. This gives us $\delta G_{capep}^{x \to X}$, $\delta G_{ca}^{x \to X}$, *and* $\delta G_{apo}^{x \to X}$ for every mutation x→X.

We then exploited the allosteric nature of S100A4 to switch between conditions where only single conformations are appreciably populated and where multiple conformations are populated. To model the ensemble, we selected reference concentrations of Ca^{2+} and peptide such that $G^{\circ}_{capep} \gg G^{\circ}_{ca} \gg G^{\circ}_{apo}$ (Fig 2.5C; see methods). We know experimentally that the protein favors the *apo* conformation in the absence of Ca^{2+} and *peptide* [111]. We modeled the signaling behavior of S100A4 by changing the concentrations of Ca^{2+} and *peptide*: $G_{capep} = G^{\circ}_{capep} - 4\mu_{Ca^{2+}} - \mu_{peptide}$ and $G_{ca} = G^{\circ}_{capep} - 4\mu_{Ca^{2+}}$, where $\mu_{Ca^{2+}}$ and $\mu_{peptide}$ are the chemical potentials of Ca^{2+} and *peptide* relative to their reference concentrations (Fig 2.5C). Depending on our choice of $\mu_{Ca^{2+}}$ and $\mu_{peptide}$, we can observe different relative populations of the *capep*, *ca*, and *apo* conformations. For ΔG_{obs}, we used:

$$\Delta G_{obs}^{genotype} = (G_{capep}^{genotype} + RTln\left(e^{-\frac{G_{ca}^{genotype}}{RT}} + e^{-\frac{G_{apo}^{genotype}}{RT}}\right). \quad (17)$$

By analogy to what we derived in Equation 16, epistasis is calculated as:

$$\boldsymbol{\varepsilon} = -\left[\left(\langle \boldsymbol{G}_{\boldsymbol{ca,apo}}^{\boldsymbol{AB}} \rangle - \langle \boldsymbol{G}_{\boldsymbol{ca,apo}}^{\boldsymbol{aB}} \rangle\right) - \left(\langle \boldsymbol{G}_{\boldsymbol{ca,apo}}^{\boldsymbol{Ab}} \rangle - \langle \boldsymbol{G}_{\boldsymbol{ca,apo}}^{\boldsymbol{ab}} \rangle\right)\right] \quad (18)$$

We constructed all 5.6 million pairs of mutations by treating the $\delta G_{capep}^{x \to X}$, $\delta G_{ca}^{x \to X}$, and $\delta G_{apo}^{x \to X}$ ROSETTA values as additive within each conformation, meaning that we calculated the effect of two mutations a→A and b→B in combination on the *apo* conformation, for example, as $G_{apo}^{AB} = G_{apo}^{ab} + \delta G_{apo}^{a \to A} + \delta G_{apo}^{b \to B}$. We made this assumption to isolate epistasis arising solely from changes to the ensemble, as we did in our general thermodynamic model in Equation 13.

Under the assumption of within-conformation additivity, we calculated epistasis in $\langle G_{ca,apo} \rangle$ using Equation 18 as a function of $\mu_{Ca^{2+}}$ at a fixed $\mu_{peptide}$ (see methods for more details). We observed peaks in epistasis at intermediate values of $\mu_{Ca^{2+}}$, where the *capep*, *ca*, and *apo* conformations may all be populated. In contrast, we observed no epistasis at low $\mu_{Ca^{2+}}$ (where only the *apo* conformation is populated) or high $\mu_{Ca^{2+}}$ (where only the *capep* conformation is populated). We observed three basic patterns of $\mu_{Ca^{2+}}$-dependent epistatic magnitude, as exemplified by the three mutant pairs shown in Fig 2.6A: F145R/L109I had no epistasis (left panel) while F145R/F78A had negative epistasis (middle panel) and F145R/M85K had positive epistasis (right panel). Interestingly, the type of epistasis observed—magnitude (dark blue), sign (gold), or reciprocal sign (green)—was also dependent upon $\mu_{Ca^{2+}}$ (Fig 2.6A). This was quite common in our dataset: approximately *61%* of pairs with an epistatic magnitude above $0.6\ kcal \cdot mol^{-1}$ switched epistatic type at least once as $\mu_{Ca^{2+}}$ increased.

We next looked at the magnitude and type of epistasis for all 5.6 million mutation pairs at their peak values over the range of $\mu_{Ca^{2+}}$. We found that 47% of the 5.6 million pairs exhibited epistasis at or above the order of thermal fluctuation, $0.6\ \mathrm{k}cal \cdot mol^{-1}$ (Fig 2.6B). We found that 34% of pairs exhibited magnitude, 12% sign, and 1% reciprocal-sign epistasis at this cutoff. Approximately 11% of pairs exhibited epistasis with a magnitude above $2\ \mathrm{k}cal \cdot mol^{-1}$.

To understand the structural origins of the observed epistasis, we compared the the positions of each mutation from Fig 2.6A in the *apo* (slate, Fig 2.6C) and *ca* (purple, Fig 2.6C) conformations. We first consider F145R. This position is solvent exposed in the *apo* conformation but buried in the *ca* conformation. As a consequence, introducing

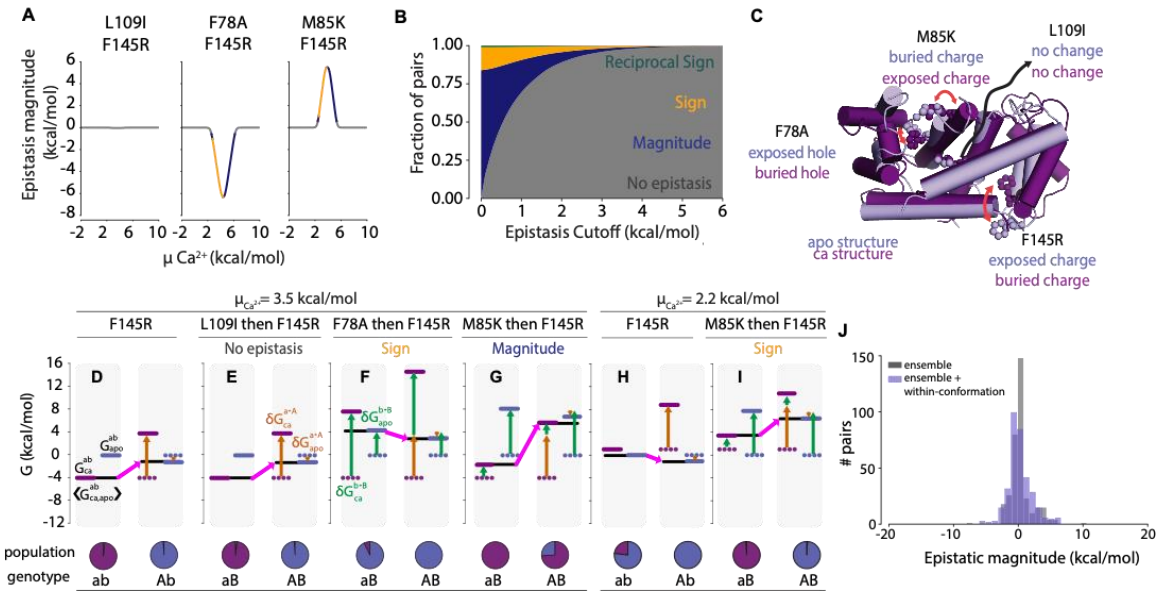


Figure 2.6 The ensemble of S100A4 exhibits ensemble epistasis. A) Epistatic magnitude ($kcal \cdot mol^{-1}$, y-axis) as a function of $\mu_{Ca^{2+}}$ ($kcal \cdot mol^{-1}$, x-axis) for three mutation pairs: L109I/F145R (left panel), F78A/F145R (middle panel), and M85K/F145R (right panel). Color is consistent with epistatic type in panel B. B) Fractional contribution of each epistatic type (y-axis) as a function of epistatic magnitude cutoff ($kcal \cdot mol^{-1}$, x-axis), colored by type: reciprocal sign (green), sign (gold), and magnitude (dark blue). Pairs with epistasis below the cutoff are considered non-epistatic (gray). C) Positions of mutations in the *ca* (purple) and *apo* (slate) conformations. Text indicates their relative environments in each conformation. Red arrows indicate changes in position between the *ca* and *apo* conformations. D-I) Thermodynamic origins of epistasis for three mutation pairs at $\mu_{Ca^{2+}}$ = 3.5 $kcal \cdot mol^{-1}$, (D-G) or $\mu_{Ca^{2+}}$= 2.2 $kcal \cdot mol^{-1}$, (H-I). Ca^{2+} chemical potential is indicated above the panel. Mutation a→A (F145R) is constant; mutation b→B differs in panels E-G and I. The color scheme is consistent throughout: purple and blue lines are the energies of *ca* and *apo*, respectively, while black lines represent $\langle G_{ca,apo}^{genotype} \rangle$; all other colors are consistent with Fig 2.2-2.3. Specific mutations and epistatic classes are indicated at the top of the panel; genotypes and relative populations are below. G) Introduction of mutation F145R (a→A) into the *ab* background at $\mu_{Ca^{2+}}$ = 3.5 $kcal \cdot mol^{-1}$. E) No epistasis scenario: mutations F145R (a→A) and L109I (b→B). F) Sign epistasis scenario: mutations F145R (a→A) and F78A (b→B) G) Magnitude epistasis scenario: mutations F145R (a→A) and M85K (b→B). H) Introduction of mutation F145R (a→A) into the *ab* background at $\mu_{Ca^{2+}}$= 2.2 $kcal \cdot mol^{-1}$. I) Sign epistasis scenario: mutations F145R (a→A) and M85K (b→B). J) Histogram showing the distribution of epistasis between 344 mutant pairs assuming no epistasis between mutations within each conformation (gray) or using calculated epistasis between mutations within each conformation (slate blue).

Arg mildly stabilizes the *apo* conformation, but dramatically destabilizes the *ca* conformation due to burying its charge. Next, L109I is a conservative mutation at a site whose environment is essentially unchanged between the *apo* and *ca* conformations. F78A is solvent exposed in the *apo* conformation but buried in the *ca* conformation. The Phe to Ala mutation is destabilizing to the *ca* conformation due to the loss of hydrophobic contacts. Finally, M85K is buried in the *apo* conformation, but exposed in the *ca* conformation. Mutation to Lys introduces a buried charge, greatly destabilizing it due to the cost of ion desolvation. The differences in the effects of L109I, F78A, and M85K on the *apo* and *ca* conformations cause them to exhibit different types of epistasis when paired with F145R.

F145R exhibits no epistasis when paired with L109I at $\mu_{Ca^{2+}}$ = 3.5 $kcal \cdot mol^{-1}$ (Fig 2.6E). The L109I mutation has a negligible effect on the *apo* and *ca* conformations (genotype *aB*, Fig 2.6E). As a result, F145R has the same effect on $\langle G_{ca,apo} \rangle$ when introduced into both *ab* and L109I (*aB*) backgrounds (compare pink arrows in Fig 2.6D and 2.6E).

Pairing F145R with F78A results in sign epistasis. F78A is destabilizing to both conformations, but much more so to the *ca* conformation (genotype *aB*, Fig 2.6F). Both F78A and F145R preferentially destabilize the *ca* structure, leading to a dramatic decrease in its relative population when introduced together (green arrows, Fig 2.6F). We see sign epistasis because the synergistic destabilization of the *ca* conformation makes $\langle G_{ca,apo}^{AB} \rangle$ only dependent on the stability of the *apo* conformation (compare pink arrows in Fig 2.6D and 2.6F).

F145R exhibits magnitude epistasis when paired with M85K. The M85K mutation is greatly destabilizing to the *apo* conformation and slightly destabilizing to the *ca* conformation (green arrows, Fig 2.6G). Combining both mutations causes a decrease in the stability of both conformations and a net destabilization of $\langle G_{ca,apo}^{AB} \rangle$, leading to the observation of magnitude epistasis (pink arrows, Fig 2.6G).

Intriguingly, a slight decrease from $\mu_{Ca^{2+}}$ = 3.5 $\mathrm{k}cal \cdot mol^{-1}$ to $\mu_{Ca^{2+}}$ = 2.2 $\mathrm{k}cal \cdot mol^{-1}$ switches the type of epistasis from magnitude to sign for the F145R/M85K pair (compare Fig 2.6D/G to Fig 2.6H/I). The switch is solely due to the change in the relative energies of the *ca* and *apo* conformations in the *ab* genotype: the *ca* conformation is slightly stabilized relative to the *apo* conformation. The introduction of F145R stabilizes the *apo* conformation, resulting in net stabilization of $\langle G_{ca,apo}^{Ab} \rangle$. M85K destabilizes both conformations, destabilizing $\langle G_{ca,apo}^{aB} \rangle$. When both mutations are combined, $\langle G_{ca,apo}^{AB} \rangle$ is further destabilized, resulting in the observation of sign epistasis (compare pink arrows in Fig 2.6H and Fig 2.6I).

Ensemble epistasis is robust to addition of epistasis from structural contacts

We next wanted to ask how the relative magnitude of epistasis changes when we allow epistasis to arise from both the ensemble and structural contacts. We used ROSETTA to calculate the within-conformation interaction energies of 344 mutant pairs. We then re-calculated the stability of each conformation *c* as:

$$G_c^{AB} = G_c^{ab} + \delta G_c^{a \to A} + \delta G_c^{b \to B} + \delta\delta G_c^{ab \to AB} \tag{19}$$

where $\delta\delta G_c^{ab\to AB}$ is the interaction energy within the conformation calculated by ROSETTA. The values of $\delta\delta G_c^{ab\to AB}$ had a mean and standard deviation of 9.3 ± 9.8 k$cal \cdot mol^{-1}$. We used these new values to calculate ε in $\langle G_{ca,apo} \rangle$. Fig 2.6J shows how the distribution of epistatic magnitude changes when we allow non-additivity to arise from the ensemble alone versus both the ensemble and structural contacts. We found that 24% of the 344 mutation pairs exhibit epistasis on the order of 0.6 k$cal \cdot mol^{-1}$, with an average magnitude of 0.97 k$cal \cdot mol^{-1}$ when we allow epistasis to arise only from the ensemble. When we allowed epistasis to arise from structural contacts in addition to the ensemble, we found that 35% of pairs exhibited epistasis on the order of 0.6 k$cal \cdot mol^{-1}$, with an average magnitude of 1.4 k$cal \cdot mol^{-1}$. The addition of within-conformation contacts widens the distribution relative to the ensemble-only dataset, yielding a modest increase in the average epistatic magnitude. Ensemble epistasis thus seems to be an important source of epistasis, even for proteins that also exhibit epistasis from structural contacts within each conformation.

Discussion

We found that epistasis can arise from a fundamental property of proteins and other macromolecules: the thermodynamic ensemble. Previously we observed ensemble epistasis using lattice models, but the conditions under which it arises and if they are plausibly met in more realistic models of proteins remained unresolved [105]. Here we used a simple—but general—thermodynamic model to study the how the ensemble leads to epistasis. Ensemble epistasis arises because mutations can affect any conformation in the ensemble. Since observables are averaged over the entire ensemble, they cannot be separated into additive components.

Ensemble epistasis should be pervasive in biology

We expect ensemble epistasis in systems where 1) at least three conformations are populated and 2) mutations have differential effects on at least two conformations. The first requirement may be common: multi-conformation ensembles often underlie biological function, from allostery to fold-switching (Fig 2.2A) [107]. The commonality of the second requirement, however, is not as obvious. We tested for the plausibility of meeting the second requirement by modeling the effects of mutations on different conformations of the S100A4 protein. S100A4 is a Ca^{2+} signaling protein that adopts three conformations, meeting the requirement for multiple populated conformations (Fig 2.5A). We identified mutations that had differential effects on both inactive conformations, which satisfied the second requirement. Nearly half of the mutant pairs exhibited epistasis above 0.6 k$cal \cdot mol^{-1}$, suggesting that—at least in principle—ensemble epistasis should be detectable in real proteins (Fig 2.6A).

There is mounting indirect evidence of links between epistasis and thermodynamic ensembles. For example, in TEM-1 β-lactamase, two adaptive mutations were identified that independently increased structural heterogeneity and function. Together the mutations exhibited epistasis, shifting the ensemble into a dominantly non-productive structure [112]. Epistasis also underlies changes in dynamics that caused functional divergence between Src and Abl kinases and the evolution of fold-switching proteins [113,114].

Recently, a thermodynamic model was used to decompose mutational effects on the GB1 protein [39]. A three-structure ensemble model was able to explain much of the epistasis observed in the dataset. The remaining epistasis pointed towards residues that

contribute to functionally important structural dynamics. This approach yielded mechanistic information about the system. Notably, the mathematical framework of the thermodynamic ensemble is not limited to proteins and other macromolecules—it has been used to describe much more complex biological systems like signaling networks and bacterial communities [115–120].

Relationship to threshold epistasis

Ensemble epistasis is related to——but conceptually distinct from——threshold epistasis. Threshold epistasis describes non-additivity arising from the accumulation of destabilizing mutations. Below some threshold stability, the fraction of folded protein molecules drops and any function encoded by the folded structure is lost [64,121–124]. The same mutation could have no effect on a high stability protein but be highly deleterious to a low stability protein. Both ensemble and threshold epistasis arise because the protein can populate more than one conformation; however, at this point, the two mechanisms for epistasis diverge.

To make this concrete, consider the activity of an enzyme. Enzyme activity is proportional to the fraction of enzyme molecules that are in the active form. Mutations that have an additive, linear effect on thermodynamic stability will have a non-additive, nonlinear effect on the fractional population of the active form (equation 8). As such, we can observe epistasis between mutations at the level of enzyme activity simply because we are describing a nonlinear function (activity) with a linear model (equation 16) [34,125]. If we transform the nonlinear fractional population scale (equation 6) onto a linear free energy scale (equation 8), threshold epistasis disappears. One can describe the non-additive, nonlinear effects of mutations on activity as additive, linear effects on stability.

This is not to say threshold epistasis does not matter---phenotype and fitness often depend on nonlinear fractional populations---but rather that it is possible to analyze the data in a way that removes epistasis.

Ensemble epistasis, however, cannot be removed by transforming the data onto a linear scale. We describe the observable (ΔG_{obs}) and the effects of mutations ($\delta G_c^{x \to X}$) on the same linear free energy scale. But because mutations have different effects on different conformations, these linear perturbations are re-weighted in nonlinear fashion, thus leading to irreducible epistasis.

Ensemble epistasis may shape evolution

Though it remains to be seen, we expect that ensemble epistasis plays an important role in shaping protein evolution. We have shown that simple ensembles give rise to magnitude, sign, and reciprocal sign epistasis (Fig 2.4), and that they may give rise to high-order epistasis. Sign and reciprocal sign epistasis are particularly important; they can decrease accessible evolutionary trajectories and are required for the presence of multiple peaks in fitness landscapes [16,26,45,56,126–131]. High-order epistasis can alter accessibility and can facilitate the bypassing of evolutionary dead-ends in genotype-phenotype maps, making evolution deeply unpredictable [23,128,132,133].

Aside from giving rise to evolutionarily-relevant classes of epistasis, we anticipate that ensemble epistasis occurs under physiologically relevant—and thus evolutionarily important—conditions. Ensemble epistasis is maximized when multiple conformations are populated (Fig 2.6A): exactly within the concentration regime where macromolecules act as molecular switches. Further, we found in our S100A4 calculations that we could see changes in the type of epistasis observed as we changed the amount of

allosteric effector, $\mu_{Ca^{2+}}$ (Fig 2.6A). This suggests that ensemble epistasis could play a critical role in shaping the availability of evolutionary trajectories—possibly even in an environment-dependent manner. A small change in the concentration of an effector could open or close new evolutionary trajectories. A similar phenomenon has been observed in allosteric proteins where ligands can act as agonists or antagonists in response to changes in environment, ultimately via changes in the thermodynamic ensemble [134].

Detecting ensemble epistasis

Our work predicts ensemble epistasis is common. How would one detect it experimentally? Effector- or environment-dependent epistasis may be a signal of ensemble epistasis. One straightforward experimental test for ensemble epistasis would be to perturb the thermodynamic ensemble by tuning environmental factors such as effector concentration (Fig 2.5B). For S100A4, we observed distinct effector-dependent patterns of epistasis for mutation pairs, where the amount of epistasis we observed changed with the addition of Ca^{2+} (Fig 2.6A). Ensemble epistasis should be maximized at concentrations where many distinct conformations are populated (i.e. at concentrations where functional transitions occur) and minimized when mutations can impact only a single conformation (i.e. low $\mu_{Ca^{2+}}$). Environmental-dependent epistasis has been noted previously, possibly pointing to an underlying ensemble epistasis [127,135–141].

Additionally, one might test for ensemble epistasis by measuring the temperature dependence of epistasis. If the free energy of each conformation does not change with temperature, the predictions are straightforward. For very low temperatures, only the deepest energy well—corresponding to the most stable conformation—should be populated, preventing ensemble epistasis. At very high temperature, all conformations

will have the same statistical weight, and thus will be equally populated regardless of free energy (Equation 6). But, because of this fact, mutations will not redistribute the populations of the conformations—meaning there will be no ensemble epistasis. For intermediate temperature values, we might expect appreciable temperature-dependent effects on ensemble epistasis. Unfortunately, the free energy of each conformation is not constant with temperature for most proteins [142]. As such, we would expect the effects of ensemble epistasis are convolved with changes in the enthalpy and entropy of each conformation—making temperature-dependent experiments difficult to interpret.

Conclusion

Our results reveal that a universal property of proteins and other macromolecules, the thermodynamic ensemble, can lead to epistasis. While the pervasiveness of ensemble epistasis in biology remains unknown, we anticipate that it is widespread. First, ensemble epistasis is maximized under the physiological conditions where biologically important, ensemble-mediated functions occur. Second, even a simple, three-conformation system can lead to a rich variety of epistasis, suggesting that the necessary conditions for ensemble epistasis are met for many proteins. And, third, structure-based calculations using experimentally solved protein structures revealed the potential for rampant ensemble epistasis. As such, we anticipate that ensemble epistasis plays important roles in shaping protein biology and evolution.

Materials and Methods

For the S100A4 epistasis analysis, we used three published structures for S100A4: the apo structure (PDB 1M31), the Ca^{2+} bound structure (PDB 2Q91), and the structure bound to both Ca^{2+} and a peptide extracted from Annexin A2 (PDB 5LPU). We removed

all non-Ca^{2+} small molecules (including waters) and edited the files to have an identical set of non-hydrogen atoms for the S100A4 chains (trimming any residues before alanine 2 and after phenylalanine 93 in the uniprot sequence, P26447). We arbitrarily selected the first NMR model for the apo structure. Using ROSETTA (Linux build 2018.33.60351), we generated five independent, pre-minimized structures for each of the conformations (*apo, ca,* and *capep*). We then used the "cartesian_ddg" binary to introduce each mutation three times into each of these five pre-minimized structures, yielding 15 calculated ΔG values for each mutation in each of the three conformations [143]. Finally, we averaged the 15 values for each mutation in each conformation. We assumed the units of these ΔG values were in $kcal \cdot mol^{-1}$[144].

For a given genotype, we described the free energy of the calcium-bound form as a function of calcium chemical potential ($\mu_{Ca^{2+}}$) with the expression $G^{\circ}_{ca}(\mu_{Ca^{2+}}) = G^{\circ}_{ca} - 4\mu_{Ca^{2+}}$. G°_{ca} is a constant describing both the relative stability of the "open" form of the protein relative to the "closed" form and the affinity of the open form for Ca^{2+}. We treated the free energy of the apo form as $G^{\circ}_{apo}(\mu_{Ca^{2+}}) = G^{\circ}_{apo} - 4\mu_{Ca^{2+}}$, where G°_{apo} measures the free energy of the apo form. For convenience, we set $G^{\circ ab}_{apo} = 0\ kcal \cdot mol^{-1}$ and $G^{\circ ab}_{ca} = 10\ kcal \cdot mol^{-1}$ for $\mu_{ca} = 0\ kcal \cdot mol^{-1}$. This models the fact that, at some reference $[Ca^{2+}]$, the "closed" form is favored over the "open" form. As $[Ca^{2+}]$, increases, $G_{ca}(\mu_{Ca^{2+}})$ becomes more negative and eventually becomes more favorable than G_{apo}. To verify that this result was not due to the choice of G°_{ca}, we re-ran our analysis for different values of G°_{ca}. We found that changing the value of G°_{ca} has little impact on the magnitude of epistasis we observe. Its main effect is changing the μ_{ca}

value at which the maximum magnitude of epistasis is observed (see Appendix A, Supplementary Fig A1).

We modeled the effects of mutations as changes to G°_{ca} and G°_{apo}. For the *Ab* genotype, for example, we would write:

$$G^{Ab}_{ca}(\mu_{Ca^{2+}}) = G^{\circ}_{ca} - 4\mu_{Ca^{2+}} + \delta G^{a\rightarrow A}_{ca}$$

$$G^{Ab}_{apo}(\mu_{Ca^{2+}}) = G^{\circ}_{apo} + \delta G^{a\rightarrow A}_{apo}$$

$$\langle G^{Ab}_{ca,apo}\rangle(\mu_{Ca^{2+}}) = -RTln\left(e^{-\frac{G^{\circ}_{ca}-4\mu_{Ca^{2+}}+\delta G^{a\rightarrow A}_{ca}}{RT}} + e^{-\frac{G^{\circ}_{apo}+\delta G^{a\rightarrow A}_{apo}}{RT}}\right)$$

where $\delta G^{a\rightarrow A}_{ca}$ and $\delta G^{a\rightarrow A}_{apo}$ are the energetic effects of mutation a→A on the *ca* and *apo* conformations, respectively. See Appendix A, Section 2 of the supplementary text for further information, including a derivation of the model.

Data Availability

Appendix A contains all referenced derivations and proofs in the text. Fig S1 demonstrates that our epistatic analysis of human S100A4 is not sensitive to our assumptions about the affinity of the protein for calcium. All analyses and ROSETTA input files can be downloaded directly from https://github.com/harmslab/ensemble_epistasis.

Bridge to Chapter III

This chapter addressed the formal mathematical foundations and plausibility of epistasis arising from the thermodynamic ensemble. First, a simple analytical model was used to derive the minimal conditions required to observe ensemble epistasis: 1) a protein populates at least three conformations and 2) mutations have differential effects on at

least two conformations. The requirement that multiple conformations are populated is expected to be quite common amongst proteins and other macromolecules. The commonality of the second requirement, however, was not obvious. To assess whether both conditions could simultaneously be plausibly met in real macromolecules, a virtual-deep mutational scan was performed on the allosteric Ca^{2+} signaling protein S100A4 protein. Almost half of the 5.6 million mutation pairs exhibited ensemble epistasis with a magnitude on the order of thermal fluctuations, leading to the conclusion that ensemble epistasis is likely a prominent source of non-additivity in real macromolecules and is likely common in biological systems. An important outcome of this work was that it showed that one might test for ensemble epistasis in a real macromolecule by looking for effector-dependent patterns of epistasis. Chapter III shows the first experimental tests for ensemble epistasis using effector-dependent operator binding in the lac repressor protein. We show that signatures of ensemble epistasis are present in all measured mutant cycles. Using thermodynamic models, we decompose the effects of mutations on the ensemble of the lac repressor protein and find that the peak in the effector-dependent epistasis curve corresponds to environmental conditions where many conformations are populated.

CHAPTER III

ENSEMBLE EPISTASIS IS PERVASIVE IN THE LAC REPRESSOR

Author Contributions

Anneliese Morrison and Michael Harms conceptualized the study and designed experiments. Michael Harms acquired funding for the study. Anneliese Morrison performed the experiments. Michael Harms administered the project. Anneliese Morrison and Michael Harms analyzed the data. Anneliese Morrison constructed figures. Anneliese Morrison wrote the chapter.

Introduction

Non-additivity between mutations—epistasis—is a ubiquitous feature of biology. Epistasis imparts unpredictability to evolution and protein engineering efforts, as the effect of a mutation depends on the presence or absence of other mutations [23,32,46,48,105,132,145]. Its presence across all scales of biological complexity, from individual macromolecules to entire microbial communities, suggests that epistasis is an extraordinarily general feature of biology, yet the underlying molecular mechanisms that cause it are often unclear [21,24,40,122,145–151].

Previously, we argued that intramolecular epistasis can arise from a general feature of macromolecules: the thermodynamic ensemble [105,152]. Thermodynamic ensembles are characterized by the set of interchanging structural conformations a

macromolecule can adopt under a specific environmental condition [153,154]. For example, the ensemble of the bacterial lac repressor protein is shown in Fig 3.1A. The ensemble consists of multiple conformations: a high DNA affinity conformation, a low affinity conformation, a DNA bound conformation, and effector bound conformations. In the absence of an effector, it populates the high affinity conformation and is bound to DNA. When an effector is present, it preferentially binds to and stabilizes the low affinity conformation, such that it releases DNA, facilitating gene transcription. Such thermodynamic ensembles are critical to many biological processes, such as signaling [66], catalysis and enzyme promiscuity [67–71], and molecular recognition [72,73], and can tune biological output in response to environmental changes [74].

Epistasis can arise from the thermodynamic ensemble because mutations can have different effects on each conformation in a multi-conformation ensemble. Such mutations can re-distribute the relative populations of each conformation. Observables measured in bulk are averaged over the properties and relative populations of each ensemble conformation [70,107,153]. Because observables are nonlinearly related to the relative population of all structures in the ensemble, we observe ensemble epistasis [105,152].

Many questions regarding the prevalence of such "*ensemble epistasis*" in real biological systems remain unanswered. Can we detect ensemble epistasis in an explicitly biological context? How common is environment-dependent epistasis in real mutant cycles? Can we map changes in the ensemble to specific patterns of environment-dependent epistasis?

We recently showed theoretically that ensemble epistasis requires two necessary conditions be met: 1) a macromolecule populates at least three conformations and 2)

mutations have different effects on at least two of the conformations [152]. Using structure-based calculations, we showed that it is plausible for both conditions to be met simultaneously in a protein. We might therefore detect ensemble epistasis by looking for environment-dependent epistasis in real proteins.

Here, we measured the magnitude of ensemble epistasis in a simple biological system: the lac repressor. As predicted by our ensemble epistasis model, the magnitude of epistasis observed varied with the amount of effector added in every mutant cycle we investigated, with close agreement between both *in vivo* and *in vitro* results. To further understand the biochemical underpinnings of each epistatic pattern, we modeled changes in the thermodynamic ensemble directly using a Monod-Wymann-Changeux (MWC) model. This allowed us to decompose the effects of mutations on the equilibrium constants that dictate the relative population of each conformation. Our findings illustrate that epistasis arising from the thermodynamic ensemble is a plausible biological mechanism of epistasis in proteins. We anticipate that ensemble epistasis is present in most macromolecules, pointing to general nonlinearity arising from a general underlying ensemble structure and not to specific insightful interactions.

Results

An experimental test of ensemble epistasis using the lac repressor

We selected the lac repressor as a model system to look for evidence of ensemble epistasis (Fig 3.1A-B). The lac repressor has a well-characterized thermodynamic ensemble, a readily measurable biological output, and was previously shown to exhibit effector-dependent epistasis[155–167]. The receptor exists in two main forms, H and L (for

High and Low DNA affinity), which can interact with operator DNA and effector in different combinations. We can experimentally manipulate the relative population conformations by adding the potent effector IPTG, which tunes the ensemble in predictable ways. Based on previous work[168], multiple conformations (H·DNA, H, and L) are populated with no IPTG present (Fig 3.1A) while a single conformation (L·2E) is populated at 1 mM IPTG (Fig 3.1B). Because ensemble epistasis requires multiple conformations, we would predict ensemble epistasis at low, but not high, IPTG concentrations[152].

We selected three well-characterized mutations to investigate ensemble epistasis in this protein: M42I, H74A, and K84L[169–172]. These mutations are physically distant in the structure (Fig 3.1B) and do not disrupt the ability of the repressor to bind operator[169–172]. These mutations have no previously reported epistatic interactions between them. To potentially reveal both pairwise and three-way interactions between mutations, we studied eight lac repressor constructs: wildtype, three single mutants (M42I, H74A, K84L), three double mutants (H74A+K84L, M42I+H74A, M42I+K84L), and the triple mutant (M42I+H74A+K84L).

We first tested for IPTG-dependent epistasis of the lac repressor *in vivo* using a gene reporter assay. We placed YFP under control of the lac operon and then measured the ability of different lac repressor variants to control YFP expression in an IPTG-dependent manner[168]. Fig 3.2A shows the induction curves for all eight genotypes, organized by mutant cycle. Several genotypes were "leaky", allowing YFP transcription at low IPTG concentrations. Genotypes containing the K84L mutation had severely diminished

induction responses. Four out of the eight genotypes (wildtype, H74A, M42I, and M42I + H74A) exhibited strong IPTG-dependent changes in YFP expression.

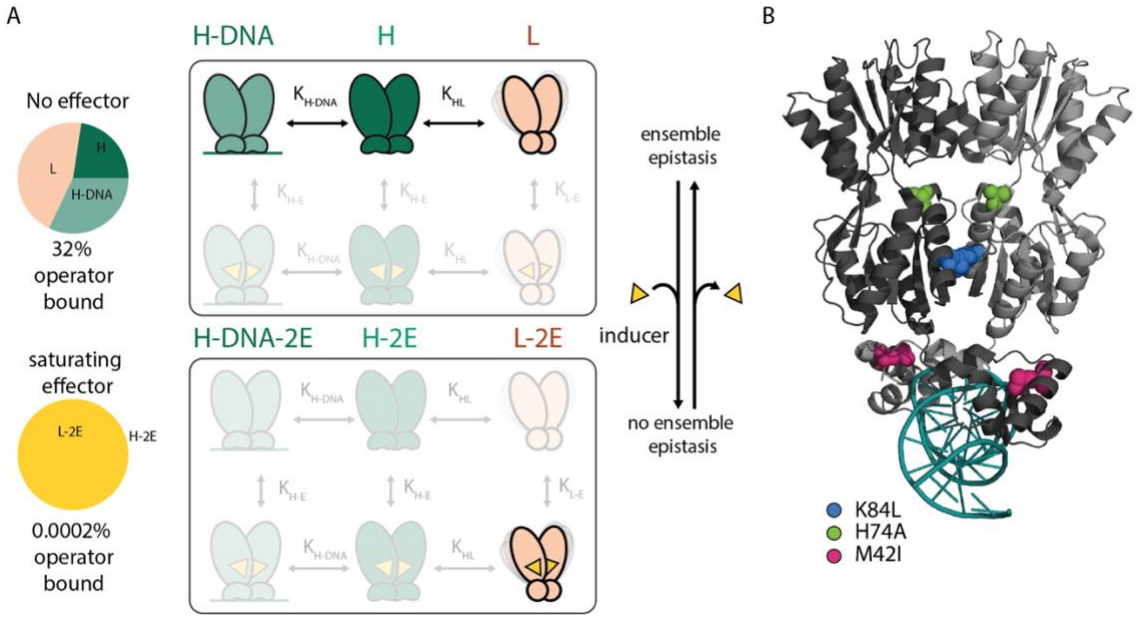


Figure 3.1 Ensembles can lead to epistasis. A) The lac repressor adopts three conformations: high-affinity bound to operator (H-DNA, light teal), high-affinity unbound (H, teal), and low-affinity (L, peach). The relative population of each conformation is shown in the pie chart in the presence (top) or absence (bottom) of inducer (gold triangle). The fraction of operator bound ([H-DNA]/$[DNA]_{tot}$) and the operator binding energy ($\Delta G_{bind}, kcal/mol$) is shown below each pie chart. Equilibrium constants for these calculations were defined using Maximum Likelihood parameter estimates for the wildtype genotype, where $K_{H\text{-}DNA}$: 2.2, $K_{H\text{-}L}$: 0.1, $K_{H\text{-}E}$: 6×10^{-6}, $K_{L\text{-}E}$: 9×10^{-4}, and $K_{L\text{-}DNA}$: 1×10^{-10}. Calculations were done at 120 nM protein, 50 nM operator, and either 0 or 1 mM inducer. B) A hypothetical pair of mutations showing epistasis in ΔG_{bind} because the mutations later equilibrium constants ($K_{H\text{-}DNA}$ and $K_{H\text{-}E}$) and re-distribute the relative populations of each conformation. Calculations done at 120 nM protein, 50 nM operator and 1 μM inducer. Pie charts represent the relative population of each conformation (peach = L, teal = H, light teal = H-DNA + H-2E-DNA, and gold = L-2E). All calculations were done using values for the equilibrium constants defined for panel A. ΔG_{bind} for each genotype (*ab*, *Ab*, *aB*, and *AB*) is shown next to the corresponding pie chart. The effect of mutation A in the *ab* and *aB* backgrounds is shown next to each downward arrow, corresponding to the introduction of mutation A. C) Locations of the mutations used in this study shown in the lac repressor dimer (dark grey and grey) bound to operator (teal); PDB code: 1efa [155]. M42I is shown as dark pink spheres, H74A as green spheres, and K84L as blue spheres.

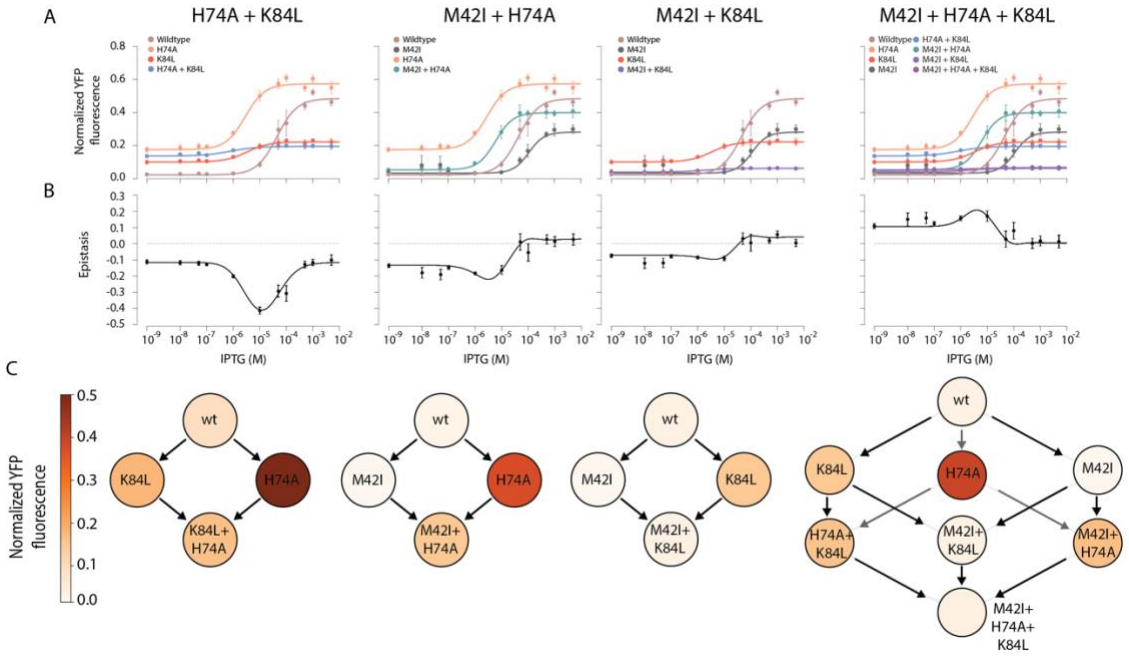


Figure 3.2 Mutant cycles display distinct patterns of effector-dependent epistasis *in vivo*. A) Full effector-response curves for each mutant cycle, with IPTG concentration (in nM) on the x-axis and normalized YFP fluorescence on the y-axis. Each color represents a single genotype. Average data points are represented by filled circles with standard deviation shown as error bars. Lines are Hill fits/linear fits. B) Epistatic magnitude (y-axis) as a function of IPTG (in M, x-axis). Circles represent the magnitude of epistasis averaged over 100 bootstrapped datasets with standard deviation shown as error bars. Lines are the magnitude of epistasis in Hill/linear fits (e.g., epistasis calculated using the smooth lines in panel A). C) Phenotypes for each mutant cycle at the maximum of the epistasis curve. The color bar indicates normalized YFP fluorescence.

We observe effector-dependent epistasis in vivo

We next calculated epistasis as a function of effector concentration for these mutant cycles. We measured all mutational effects relative to the wildtype genotype (M42, H74, K84). For each mutant pair (a→A/b→B), we calculated epistasis as the difference in the effect of mutation a→A in the wildtype (ab) versus mutant (aB) genetic backgrounds. This is given by:

$$\varepsilon_{AB} = (YFP_{AB} - YFP_{aB}) - (YFP_{Ab} - YFP_{ab}) \quad 1.$$

where YFP indicates fluorescence at a given IPTG concentration and the subscripts indicate the genotype (lowercase and uppercase letters indicate wildtype and mutant, respectively). For the three-way mutant, we defined high-order epistasis as the difference between the observed fluorescence of the triple mutant and the predicted fluorescence given the individual and pairwise effects of the three mutations[9]. This is given by:

$$\varepsilon_{ABC} = YFP_{ABC} - (YFP_{ABc} + YFP_{AbC} + YFP_{aBC}) + (YFP_{Abc} + YFP_{aBc} + YFP_{abC}) - YFP_{abc}. \quad 2.$$

Fig 3.2B shows epistasis as a function of IPTG concentration for the mutant cycles. At low IPTG concentrations ($< 10^{-7}$ M), we see a moderate amount of epistasis for all four mutant cycles. At intermediate IPTG concentrations (10^{-7} M < [IPTG] < 10^{-3}), we see a peak in epistasis. The peak coincides with the IPTG concentrations over which the switch-like induction response occurs (Fig 3.2A). At high IPTG concentrations ($> 10^{-3}$ M), we see a decrease in the magnitude of epistasis, which approaches zero in three out of the four mutant cycles.

The peak magnitude of the epistasis in these curves is substantial. The total induction of the wildtype receptor goes from 0.0 to 0.5 between 0 and 1 mM IPTG (Fig 3.2A, left). For the H74A+K84L cycle, the peak epistasis is -0.4 (Fig 3.2B, left)—meaning the magnitude of the epistasis is fully 80% as large as the total dynamic range for the wildtype repressor. This is not limited to the pairwise epistatic curves: the high-order epistasis from the three-way mutant peaks with magnitude near +0.2 (Fig 3.2B, right)—40% of the total induction of the wildtype repressor.

The MWC model gives molecular level insights into the thermodynamic ensemble.

The effector-dependent epistasis observed *in vivo* is consistent with ensemble epistasis. We next wanted to take a more direct approach to dissect the molecular level details underlying each epistatic pattern. Understanding how the ensemble gives rise to each epistatic pattern requires a quantitative model of the map between environmental conditions and the ensemble composition. Recently, Sochor et al derived and tested a Monod, Wyman, and Changeux (MWC) style model of the lac repressor [168]. This model has twelve conformations, whose concentrations are determined by five equilibrium constants: K_{HL} , K_{HE} , K_{LE} , $K_{H\text{-}DNA}$, $K_{L\text{-}DNA}$ (Figure 3.3A) [168].

A requirement of fitting this model to experimental data is precise knowledge of the total concentrations of all species in the linked equilibria—protein, operator, and effector. While the approximate concentrations are known in our *in vivo* assay, we needed a more precise quantitative dataset as input to the MWC model. We therefore expressed and purified all eight lac repressor mutant and measured operator binding *in vitro* using a fluorescence polarization assay. We measured operator binding at ~6-10 protein concentrations and four IPTG concentrations for each variant (Appendix B, Supplementary Fig B3). To estimate model parameters for each genotype, we used a Bayesian Markov-Chain Monte Carlo (MCMC) strategy to sample over model parameters consistent with all observations for each genotype.

Because this results in many non-unique solutions to the system of equations in the model, we took an approach similar to Sochor et al. to reduce the number of independent constants fit by the model[168]. First, we set $K_{L\text{-}DNA}$ to 1×10^{-10}nM^{-1}, in accordance with previous literature and observations that the protein has very weak

affinity for DNA in the L conformation[168]. Then, rather than fitting K_{HL} and $K_{H\text{-}DNA}$ independently, we fit the ratio ($\frac{K_{HL}}{K_{H-DNA}}$)—K_{ratio}—because our estimates for these parameters covaried strongly.

This left the model with three parameters to fit: K_{ratio}, K_{HE}, and K_{LE}. Figure 3.3B shows the raw binding curve and model fit for the wildtype lac repressor at different concentrations of effector. Once such a solution to the system of equations is found, we can use the fit parameters to calculate the concentration of each species in the linked equilibrium shown in Fig 3.3A, as well as the fractional saturation of operator at any concentration of protein, effector, and operator (Fig 3.3C).

Our epistasis model assumes that the effects of mutations are on an additive, linear scale [34]. We therefore calculated epistasis after doing a scale-transformation as applying a linear model to a dataset where mutational effects combine on a nonlinear scale (i.e., multiplicative) may be misleading as it spuriously maps non-additivity arising from a mismatch in scale to specific interactions [34]. We therefore transformed our data onto a linear, energetic scale, ΔG_{bind}:

$$\Delta G_{bind} = -RTln\left(\frac{[protein-operator]}{[protein]_{free}\times[operator]_{free}}\right),$$

where R is the gas constant and T is the temperature. Figure 3D shows how ΔG_{bind} changes for the wildtype genotype as a function of effector.

For all genotypes, K_{ratio} was well-defined and had minimal covariance with other parameters. For many genotypes where there was no induction response K_{HE} and K_{LE} were not well-constrained.

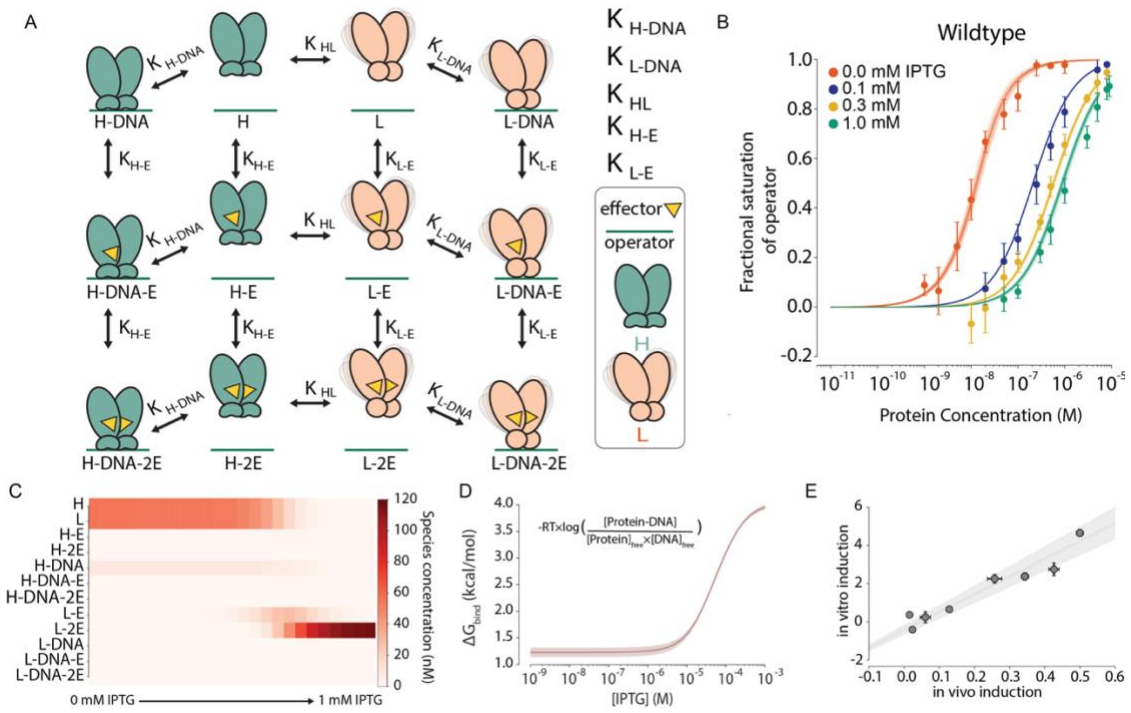


Figure 3.3: Using a thermodynamic model to decompose mutational effects on the lac repressor ensemble. a) Monod-Wymann and Changeux model of the lac repressor ensemble. B) Fractional saturation of operator as a function of protein concentration (in M) for wildtype. Circles represent averaged fractional saturation values from measured data points and error bars represent standard deviation. Clouds of lines represent fractional saturation calculated using 50 sets of sampled parameters from Bayesian MCMC fits. Solid line represents the parameter estimate from fits. The color of each dataset corresponds to the effector concentration it was measured at: 0 mM IPTG (orange), 0.1 mM IPTG (navy blue), 0.3 mM IPTG (gold), and 1.0 mM (teal). C) Heatmap showing the concentration of each species from panel A as a function of IPTG calculated from the model fit parameters at 120 nM total protein and 10 nM total operator and using the fit parameter estimates. D) Operator binding energy (in kcal/mol, y-axis) as a function of IPTG concentration (in M, x-axis) for the wildtype lac repressor. The cloud represents the standard deviation of ΔG_{bind} calculated using 50 sampled sets of fit parameters, as in panel A. The solid line represents the average calculated ΔG_{bind} value over 50 sampled sets of fit parameters. Species concentrations from panel C were used to calculate the logarithm of the apparent association constant, with the equation for ΔG_{bind} shown in the upper left corner (T = 303 K). E) Correlation between measured *in vivo* (x-axis) and *in vitro* (y-axis) induction phenotypes. *In vitro* induction is calculated as the change in operator binding between 0 mM IPTG and 1 mM IPTG. Here ΔG_{bind} was calculated using fit parameters from fits without *in vivo* data added. Error bars on the y-axis represent the standard deviation in $\Delta G_{bind,\ 1\ mM} - \Delta G_{bind,\ 0\ mM}$ over all 50 sampled datasets. Error bars on the x-axis represent the standard deviation in $(YFP_{0\ mM} - YFP_{1\ mM})$ over 100 bootstrapped datasets using *in vivo* data. Line indicates the best fit line using orthogonal distance regression, $R^2 = 0.88$.

To check that our fit parameters were potentially biologically meaningful, we calculated induction for each genotype ($\Delta G_{bind,1\ mM\ IPTG} - \Delta G_{bind,0\ mM\ IPTG}$) and compared it to the measured *in vivo* induction phenotypes (Fig 3.3E). Overall, we find good agreement between the simulated induction dataset and the measured in vivo dataset despite K_{HE} and K_{LE} not being well constrained.

To better resolve K_{HE} and K_{LE}, we used both the *in vivo* and *in vitro* datasets as inputs to the MWC model. Because the operator sequences used in the *in vivo* and *in vitro* experiments differed, we added a multiplication factor that scales K_{ratio} as a fourth fit parameter. Addition of the *in vivo* data helped to constrain K_{HE} and K_{LE} for all eight genotypes (Appendix B, Supplementary Fig B3). Even with the addition of *in vivo* data, K_{HE} and K_{LE} co-vary.

Changes to the ensemble lead to effector-dependent epistasis in ΔG_{bind}

We next wanted to understand how epistasis changes as a function of effector for *in vitro* phenotypes. We first used the MWC model fit parameters for each genotype to calculate the concentration of each species as a function of effector (Appendix B, Supplementary Fig B4). We then determined ΔG_{bind} as a function of effector as we did with the wildtype genotype in Fig 3.3C, show in Fig 3.4A for the M42I + H74A mutant cycle. Finally, we calculated epistasis in ΔG_{bind} as a function of IPTG concentration. Fig 3.4B shows the effector-dependent epistasis curves for the M42I + H74A mutant cycle.

We next wanted to connect the effector-dependent epistasis we observe with underlying changes in the thermodynamic ensemble. We looked at how changes in the relative populations of different conformations—namely unbound, operator bound, effector bound, and both operator and effector bound as shown in Fig 3.4C—map to the

epistatic signal we observed in Fig 3.4B. We investigated the molecular mechanisms giving rise to epistatic patterns by looking at how the effect of a mutation H74A changes in the presence of the M42I mutation for the observable, ΔG_{bind}, and the relative population of each species. We selected three regions of the effector-dependent epistasis curve to investigate: low IPTG (1 nM IPTG), medium IPTG (5 μM IPTG), and saturating IPTG (1 mM IPTG), as shown by points D, E, and F in Fig 3.4B.

In the low effector regime, represented by point C, we see negative epistasis in ΔG_{bind} at a magnitude of ~1 kcal/mol (Fig 3.4B). The relative population of each conformation from the reduced linked equilibria shown in Fig 3.4C is shown in Fig 3.4D for the wildtype, H74A, M42I, and M42I+H74A genotypes. The only conformations that are appreciably populated at 1 nM IPTG are the unbound and operator bound conformations (Fig 3.4D).

H74 is located near the cluster of residues responsible for effector binding and mutation to alanine is known to increase the repressor's affinity for inducer, while decreasing its affinity for operator (Fig 3.1B, green spheres) [169]. We can see these effects in the population plot for H74A and in its destabilizing effect on ΔG_{bind} (Fig 3.4D, bottom panel pink arrow). M42 is located within the DNA binding domain and mutation to isoleucine has been shown to directly stabilize the DNA-binding domain, increasing its affinity for operator (Fig 3.1B, pink spheres) [171]. When the H74A mutation is introduced into the M42I background, we see that it does not have the same effect as it did in the wildtype background (Fig 3.34D bottom panel, compare pink arrows). We see epistasis because the M42I mutation directly stabilizes the DNA-binding domain, resulting in the operator binding energy of the double mutant is identical to the M42I genotype.

At point E, we are near the peak in the magnitude of epistasis for this mutant cycle. The equilibria at the top of Fig 3.4E shows that all conformations are appreciably populated at this effector concentration. The population plot shows that the H74A genotypes ensemble configuration is distinct from the wildtype and M42I genotypes. The H74A mutation has preferentially stabilized the effector bound conformations, resulting in a decrease in the relative population of the operator bound conformations, and a decrease in ΔG_{bind} (Fig 3.5E bottom panel, pink arrows). When H74A is introduced into the M42I background, the population plot shows that the M42I + H74A genotype stabilizes both the effector and effector + operator bound conformations. We see epistasis in ΔG_{bind} because of the joint stabilization of the effector + operator bound conformations (Fig 3.4E bottom panel, compare pink arrows). The epistasis signal we observe is ensemble epistasis—it occurs as a result of the re-weighting of the ensemble by H74A to favor effector bound conformations and M42I to favor operator bound conformations.

At point F, there is no longer a signal for epistasis. When the H74A mutation is introduced into the wildtype background, it has little effect on ΔG_{bind} because the only conformation appreciably populated is the effector bound conformation and the protein is fully induced (Fig 3.4F). The ensemble and phenotype of the double mutant genotype is identical to the M42I genotype, where the direct stabilization of the DNA-binding domain increases the population of the effector + operator bound conformations—stabilizing ΔG_{bind}—relative to the wildtype genotype. When introduced into the M42I background the H74A mutation has no effect on ΔG_{bind} and we observe no epistasis (Fig 3.4F bottom panel, compare pink arrows).

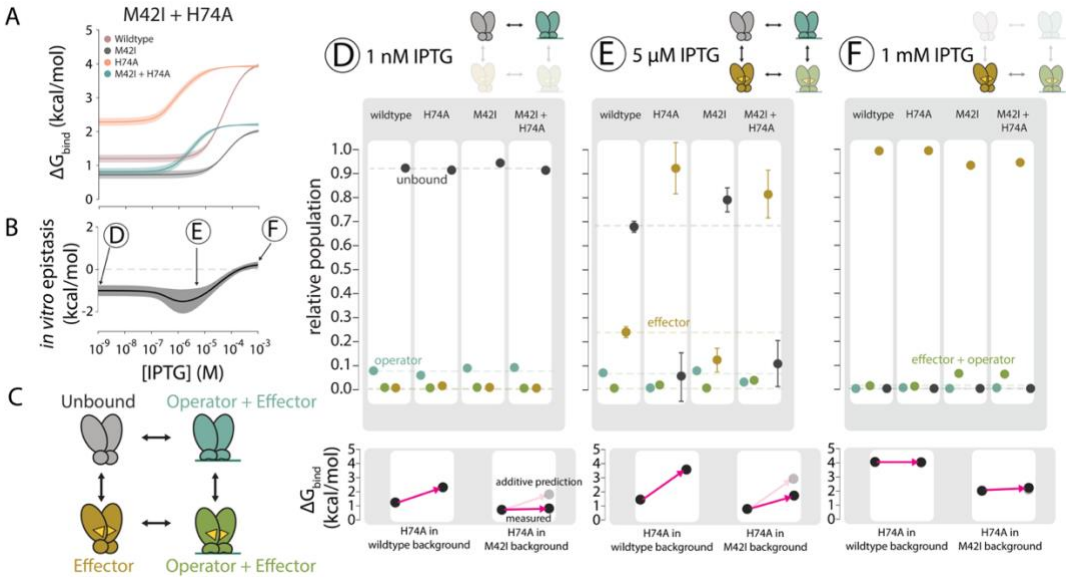


Figure 3.4: Molecular basis of effector-dependent epistasis in the M42I + H74A mutant cycle

A) Operator binding energy (ΔG_{bind} in kcal/mol, y-axis) as a function of IPTG (in M, x-axis) for the M42I + H74A mutant cycle. Each color represents a single genotype. The solid represents the average binding curve over 100 sets of sampled parameters from Bayesian MCMC fits. Shaded areas represent the average $\pm$ standard deviation. All calculations were done with $[protein]_{total}$ = 120 nM and $[operator]_{total}$ = 10 nM. B) Epistasis in ΔG_{bind} (in kcal/mol, y-axis) as a function of IPTG concentration (in M, x-axis). The solid black line represents the average epistasis curve from 100 sets of sampled parameters, as in panel A. Shaded areas represent the average $\pm$ standard deviation. Points D, E, and F are used to refer to diagrams in panels D-F. The grey dashed line indicates where epistasis equals 0 kcal/mol. C) Simplified linked equilibria diagram where the unbound conformations (grey; H and L) are in equilibrium with the operator bound (teal; H-DNA, L-DNA) effector bound (gold; H-E, H-2E, L-E, and L-2E), and operator + effector bound conformations (green; H-DNA-E, H-DNA-2E, L-DNA-E, L-DNA-2E). D) Relative population of each conformation in panel A at 1 nM IPTG. Equilibria diagram at the top right corner shows which conformations are appreciably populated at each effector concentration. Circles represent the average relative population over 100 sampled datasets (as in panel A) and error bars represent the standard deviation. Genotype is indicated above each block. The color scheme for each conformation is the same as in panel C: operator bound shown in teal, unbound in grey, effector bound in gold, and operator + effector bound in green. The relative populations of each conformation for the wildtype genotype are shown as transparent dashed lines for comparison. Black dots in the bottom panel shows the binding energy (ΔG_{bind}, kcal/mol) for each genotype. The bottom left panel shows the H74A mutation introduced into the wildtype background, with its effect on ΔG_{bind} shown by the pink arrow. The bottom right panel shows the H74A mutation in the M42I background. The transparent pink arrow and dot represents the additive prediction, while the opaque represents the measured effect. D) Relative population (top panel) and epistasis in ΔG_{bind} (bottom

panel) diagrams at 5 μM IPTG. E) Relative population (top panel) and epistasis in ΔG_{bind} (bottom panel) diagrams at 5 mM IPTG.

We observe effector-dependent epistasis in all mutant cycles

We measure changes in epistasis as a function of effector in all three remaining mutant cycles. Fig 3.5A shows operator binding energies for each genotype in the H74A + K84L, M42I + K84L, and M42I + H74A + K84L mutant cycles. Fig 3.5B shows the corresponding effector-dependent epistasis curves in ΔG_{bind}. For comparison, the *in vivo* epistasis curves are shown below in Fig3.5C. There is excellent agreement between the epistatic patterns observed *in vivo* and *in vitro*, suggesting that they share similar underlying molecular mechanisms. All three mutant cycles show similar general underlying trends, where at moderate concentrations of IPTG, many conformations are populated, and we see a peak in epistasis. At high concentrations of IPTG, only a single conformation is appreciably populated—namely, effector bound conformations—and we observe lower magnitudes of epistasis.

We find that in all mutant cycles, epistasis peaks as a consequence of the differences in the induction response between different mutants. For example, in the H74A + K84L mutant cycle, H74A strongly induces, having a much higher affinity for effector, whereas K84L has a very diminished induction response. The K84L mutation is known to drastically reduce induction, presumably by stabilization of the region of the dimer interface that is critical for the allosteric transition between the repressed and induced states [173]. If we compare the effect of the H74A mutation introduced into the wildtype versus K84L backgrounds, we see that it has almost no effect on induction in the K84L background. We see a similar trend in the M42I + K84L mutant cycle. When

M42I is introduced into the wildtype background, we observe slightly reduced induction and a stronger affinity for operator. When it is introduced into the K84L background, we observe higher operator affinity, but diminished induction response.

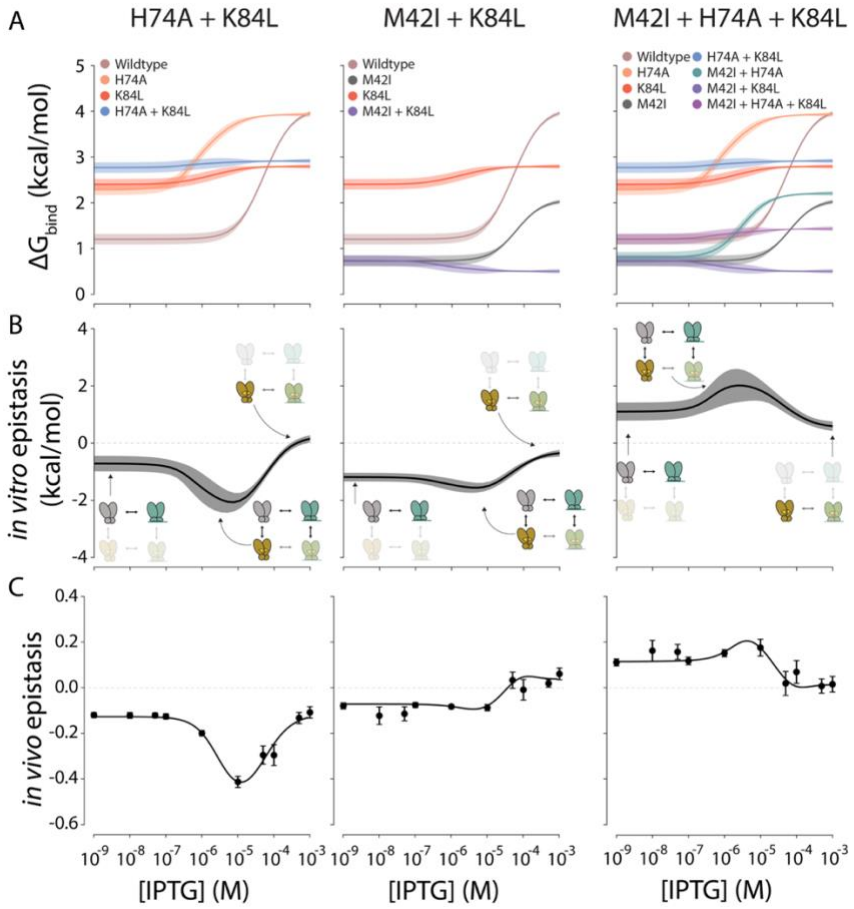


Figure 3.5 Effector dependent epistasis *in vitro*. A) Epistasis in normalized YFP fluorescence from the *in vivo* datasets. As in Fig 3.2C, circles represent the magnitude of epistasis averaged over 50 bootstrapped datasets with standard deviation shown as error bars. Lines are the magnitude of epistasis in Hill (or linear) fits. B) Epistasis in ΔG_{bind} (in kcal/mol, y-axis) as a function of IPTG concentration (in M, x-axis). The solid black line represents the average epistasis curve from 100 sets of sampled parameters from the Bayesian MCMC fits, as shown in Fig 3.3. Shaded areas represent the average $\pm$ standard deviation. Simplified equilibria show approximately what conformations are appreciably populated in each region of the epistasis curve with color scheme and species type identical to Fig 3.4C. C) Operator binding energy (ΔG_{bind} in kcal/mol, y-axis) as a function of IPTG (in M, x-axis) for each mutant cycle. Each color represents a single genotype. The solid represents the average binding curve over 100 sets of sampled parameters. Shaded areas represent the average $\pm$ standard deviation. All calculations were done with $[\text{protein}]_{\text{total}}$ = 120 nM and $[\text{operator}]_{\text{total}}$ = 10 nM.

Discussion

Ensemble epistasis is present both *in vivo* and *in vitro* for all measured mutant cycles of the lac repressor. Previously we showed using simple analytical models that the thermodynamic ensemble could, at least in theory, give rise to epistasis under certain conditions and we predicted that one could detect such ensemble epistasis by looking for environment-dependent epistasis in real macromolecules [152]. However, it was unclear how frequently we would be able to detect ensemble epistasis outside of simple computational models of proteins. Here, we show that we observe signatures of ensemble epistasis in an explicitly biological context by measuring effector-dependent epistasis between in the lac repressor protein. Thermodynamic modeling of the lac repressor indicates that when mutations impact different protein conformations, the relative populations of each conformation can change, resulting in epistasis at the level of our observable.

Ensemble epistasis is likely common in real mutant cycles

Thermodynamic ensembles are an extraordinarily general feature of biology. Such ensembles are often critical for biological function, from the response of transcription factors and cell receptors to environmental changes to post-translational modifications (such as phosphorylation) during signaling cascades [67–69,74]. Our previous theoretical and computational work suggested that 1) epistasis can arise from ensembles and 2) such ensemble epistasis should be pervasive in biological systems where thermodynamic ensembles underlie an observable. It was still unclear, however, if the magnitude of ensemble epistasis would be detectable in a real biological system and how frequently pairs of mutations would exhibit it.

Here, we experimentally tested for effector-dependent epistasis——a key signature of ensemble epistasis——in four mutant cycles of the lac repressor protein. Our observable, defined as operator binding, is dependent upon the relative population of the high-DNA affinity (H) conformation. The relative population of H can change modulated by changes in the population of other ensemble conformations (Fig 3.1A). We chose three structurally distant mutations that were known to have different effects on different conformations: M42I stabilizes operator bound conformations, H74A stabilizes effector bound conformations, and K84L stabilizes effector and operator bound conformations, drastically reducing induction (Fig 3.1C, Fig 3.4A) [169,171–174]. Despite no *a priori* knowledge that mutant cycles containing these mutations would exhibit ensemble epistasis, we observed effector-dependent epistasis in all measured mutant cycles both *in vivo* and *in vitro*. Our *in vitro* dataset revealed that the measured epistatic interactions were quite large, on the order of ~1-2 kcal/mol (Fig 3.4A). Because environment-dependent epistasis was ubiquitous in measured mutant cycles, we anticipate that ensemble epistasis will be pervasive in mutant cycles where thermodynamic ensembles underlie the observable.

There is a rapidly growing body of literature that shows that epistasis is frequently environment dependent in biological systems **[175]**. For example, a recent study of the evolutionary transition between an ancestral dihydrocoumarin hydrolase and a derived methyl-parathion hydrolase showed that epistatic interactions between historical mutations changed both magnitude and sign depending on the presence of different divalent metal ions **[176]**. Environment-dependent epistatic interactions such as this have been observed in diverse macromolecular systems—from catalytic RNA's **[177,178]**, drug

resistance enzymes [137,179,180], yeast tRNAs [181], cis-regulatory elements [182–184], and transcription factors [167,185]. Environment-dependent epistasis has been observed extensively in more complex biological systems, for instance in experimental evolution experiments [136,186,187]. While the definitive link to ensemble mechanisms in these studies is unclear, the presence of environment-dependence in macromolecular systems may point to an underlying ensemble mechanism.

Ensemble epistasis is maximized where important functional transitions occur

We found both *in vivo* and *in vitro* that the magnitude of ensemble epistasis is maximized under the physiological conditions where macromolecules transition between distinct, biologically important functions (Fig 3.3 and Fig 3.4). Here, this corresponds to IPTG concentrations where many conformations are populated because the system is transitioning from primarily the unbound and operator bound conformations to effector bound conformations (Fig 3.3 and Fig 3.5). Similar to previous work in the lac repressor and our work in the S100A4 protein, we observe switching in epistatic type, primarily between magnitude and sign epistasis, as we approach saturating effector concentrations [152,167](Fig 3.5).

Epistasis peaking in transitional regions, where conformational diversity is also maximized, could be profoundly important for biological function, and consequently, evolution [24,124,175,188,189]. Conformational diversity has been linked to many important topics in evolutionary biology: phenotypic plasticity [190,191], increased rates of evolution [192], and increased robustness and evolvability [189,193]. A recent study compared the trajectories of *E. coli rpoB* mutants under fluctuating antibiotic concentrations to those

constant antibiotic concentrations. In addition to pervasive environment-dependent epistasis, they found that there was more diversity in the solutions to survival in the fluctuating environments, indicating that there were more paths accessible, and that specific mutational trajectories were contingent upon intermediate environments, meaning that those trajectories are only accessible if an intermediate environment was encountered [179]. Ensemble epistasis may therefore be a pervasive type of epistasis that couples mutational effects to environmental changes, thus strongly shaping evolution.

The facilitation of evolvability by the thermodynamic ensemble may mirror what has been found in more complex biological systems. For example, intrinsically disordered regions of the Potato virus Y (PVY) Viral genome-linked (VPg) protein were found to be critical to evolve to overcome host resistance, whereas VPg proteins engineered to be less disordered were unable to restore host infection [194]. While the importance of ensemble epistasis in evolution remains to be shown, one might test for it in a similar way, by comparing the adaptive capacity of a protein with and without an ensemble.

Other possible origins of epistasis

In this study we were able to understand the molecular origins of epistasis by pairing careful biochemical experiments with thermodynamic modeling (Fig 3.3-3.5). There are, of course, many other forms of intramolecular epistasis, e.g., direct electrostatic contacts between amino acids and threshold epistasis[24]. Here, we focused on a specific mechanism, ensemble epistasis, by choosing mutations that were structurally distant (Fig 3.1B).

Other mutant cycles of the lac repressor and different systems may exhibit multiple types of epistasis simultaneously. Further work is needed to distinguish different mechanisms of epistasis that underlie specific patterns, which will require detailed structural and biochemical studies in simple macromolecular systems. If mutational effects on the equilibrium constants that dictate the relative populations of each conformation are known and fully resolved, we expect that the effects on equilibrium constants will be additive in the absence of epistasis arising from physical interactions. Thus, one might distinguish ensemble and contact mechanisms of epistasis looking for leftover non-additivity at the level of equilibrium constants.

We might improve phenotype prediction by accounting for protein biochemistry

Moving forward, we might increase the accuracy of phenotype predictions by accounting for the biochemical and biophysical properties of macromolecules. In the instance of ensemble epistasis, it has historically been difficult, if not impossible, to experimentally measure the properties of the conformational ensembles of a single protein, let alone the tens-thousands that might be required to uncover the relationship between mutations and a specific ensembles energetic configuration. One avenue might be to use a combination of molecular dynamics simulations and biophysical measurements [107,195]. This too, however, can become intractable for larger, more complex proteins.

Recently, an approach was taken that applied a thermodynamic model to ligand binding in the bacterial GB1 protein [39]. The authors were able to describe a large portion of the observed epistasis by applying a three-conformation model to their dataset, accounting for the stability differences between folded and unbound, folded and bound,

and unfolded structures upon mutation. This approach explained much of the nonlinearity in their dataset and yielded previously unknown information about the number of conformations describing the GB1 protein ensemble.

Key questions remain about ensemble epistasis in biology. How common is it in other biological systems? How high order do such interactions go? Do we need to account for it to gain a deeper, more predictive understanding of biology? To begin answering these questions one could scale up our approach by pairing thermodynamic modeling with quantitative high-throughput experimental measurements of effector-dependent phenotypes to characterize many mutant cycles at once. Recent advances in high-throughput protein characterization methods make these question experimentally tractable, as it is now possible to characterize hundreds to tens of thousands of proteins at once [196–199].

Conclusions

Our work shows that effector-dependent epistasis—a key signature of ensemble epistasis—is pervasive in mutant cycles of the lac repressor protein in an explicitly biological context. The role of ensemble epistasis in shaping biology and evolution remains to be seen, but we anticipate that it can profoundly influence how macromolecules and organisms respond to environmental changes. Our results show that its magnitude is large enough to be detected *in vivo* and that all combinations of mutations investigated gave rise to effector-dependent epistasis, though we had no a priori knowledge that these pairs would give rise to ensemble epistasis. We also found that ensemble epistasis peaks in regions that are critical for important biological functions upon environmental perturbations, such as presence of an effector. We expect that

ensemble epistasis is as ubiquitous and fundamental of a feature in shaping biology as thermodynamic ensembles.

Materials and methods

Molecular biology and plasmid construction

For all *in vitro* measurements, we used the tetrameric wildtype lac repressor gene in the plasmid phg165c which obtained from Dr. Liskin Swint-Kruse's lab (addgene plasmid #90058) [170]. The naturally occurring T109 isoform was used as the wildtype genetic background for all measurements. We used the Quikchange lightning mutagenesis kit to introduce the following four mutations in the lac repressor protein in all possible combinations: M42I, H74A, and K84L. All eight lac repressor mutants were cloned into the pBAD vector with a C-terminal His-tag using a sequence and ligation independent cloning (SLIC) protocol. His-tagged lac repressor pBAD constructs were transformed into ccdB Survival T1R cells for expression.

For all *in vivo* measurements, we cloned the wildtype lac repressor sequence into the PAM5087 vector (addgene plasmid #85123) downstream of the constitutive lacIq promoter using SLIC cloning [200]. The PAM5087 vector contains the YFP gene downstream of the O1 lac operator sequence and a trc promoter. We generated a wildtype repressor construct c-terminally tagged with the mCherry protein. A 24 amino-acid rigid linker sequence (GSLAEAAAKEAAAKEAAAKAAAAS) was cloned in frame between the two genes to prevent protein-protein interactions [201]. The mCherry-tagged wildtype lac repressor gene was then cloned into the PAM5087 vector. All eight genotypes were cloned upstream of the mCherry gene. Finally, we created a DEL construct in the PAM5087 background by introducing an N-terminal stop codon in the wildtype lac

repressor protein. All PAM5087 constructs were transformed into BLIM cells for all *in vivo* experiments.

Protein expression and purification

For protein expression, we inoculated 1.5 L 2xYT (24 g/L Tryptone, 15 g/L Yeast Extract, 7.5 g/L NaCl) supplemented with 50 μg/mL kanamycin with a 1:100 dilution of overnight cultures grew at 37℃, 250 RPM until an OD_{600} between 0.4-0.6 was reached. Protein expression was induced by addition of 0.2% (v/v) L-arabinose and 0.2% glycerol. Cultures were grown for three hours at 37℃, 250 RPM. Cells were harvested by centrifugation at 3000 RPM, 4℃. Cell pellets were resuspended in HisA buffer (46.6 mM Na_2HPO_4, 3.4 mM NaH_2PO_4, 500 mM NaCl, 20 mM imidazole, 2.5% glycerol, pH 8.0) supplemented with 17.5 U lysozyme/g pellet and 17.5 U DNaseI/g pellet (ThermoFisher, 90082 and 900823). Cells were lysed using sonication with a 30% duty cycle and 55% amplitude in an ice slurry. Lysate was loaded onto two connected 5 mL HisTrap HP columns (Sigma, GE17-5248-02). Protein was purified using an ion exchange gradient protocol by washing with 30 mL HisA buffer and eluting using a 70 mL gradient to 100% HisB (46.6 mM Na_2HPO_4, 3.4 mM NaH_2PO_4, 500 mM NaCl, 500 mM imidazole, 2.5% glycerol, pH 8.0) on an AKTA Prime FPLC at 4℃. Fractions containing eluted protein were pooled and dialyzed into Binding buffer (42.24 mM Na_2HPO_4, 7.75 mM NaH_2PO_4, 150 mM NaCl, 1 mM TCEP, 10% glycerol, pH 7.6) overnight at 4℃. Purified protein stocks were confirmed to be >90% pure by SDSPAGE. Protein stocks were flash frozen as pellets in liquid nitrogen and stored at -80℃.

***In vivo* expression assays**

All 16 lac repressor mutants downstream of the lacIq promoter in the PAM5087 vector, which contains the YFP gene downstream of the O1 operator. Each construct was transformed into BLIM cells (Addgene bacterial strain #35609). BLIM cells are derived from BL26 Blue cells (Novagen) with both the lac operon and F' episome, which contains the lac repressor gene, removed [202]. LB media supplemented with 50 μg/mL kanamycin were inoculated from PAM5087 construct glycerol stocks and grown at 37℃, 250 RPM overnight. Overnight cultures were used to inoculate 10 mL cultures in pre-growth M9 media (47.8 mM Na_2HPO_4, 22 mM KH_2PO_4, 8.6 mM NaCl, 18.7 mM NH_4Cl, 10 mM $NaHCO_3$, 0.0025% (w/v) thiamine-HCl, 0.2% (w/v) Casamino Acids, 0.01X Basal Eagle Medium, 0.04% glycerol, 2 mM $MgSO_4$, 1 mM $CaCl_2$) supplemented with 50 μg/mL kanamycin. Pre-growth cultures were grown for 5-6 hours at 37℃, 250 RPM. Black-walled clear-bottom 96-well plates (Corning Catalog #: 9018) were filled with 180 μl of overnight M9 media (47.8 mM Na_2HPO_4, 22 mM KH_2PO_4, 8.6 mM NaCl, 18.7 mM NH_4Cl, 10 mM $NaHCO_3$, 0.0025% (w/v) thiamine-HCl, 0.2% (w/v) Casamino Acids, 0.01X Basal Eagle Medium, 0.8% glycerol, 2 mM $MgSO_4$, 1 mM $CaCl_2$) supplemented with 50 μg/mL kanamycin. We added varying amounts of IPTG to overnight M9 media to achieve the following final concentrations: 0 nM, 50 nM, 10 uM, 50 uM, 500 uM, 1 mM, 5 mM, 10 mM, 50 mM, and 500 mM. Wells were inoculated with 20 μl of each pre-growth culture. Clear lids with condensation rings (Millipore Sigma Catalog #: CLS3931-50EA) were sealed onto plates with parafilm to minimize evaporation. The plate reader (Synergy H1) was pre-heated to 37℃ and YFP fluorescence (485 nm/530 nm, gain = 100, measurements/data point = 10), mCherry fluorescence (575 nm/610 nm,

measurements/data point = 10), and OD_{600} (measurements/data point = 8) measurements were taken for 16 hours at intervals of 20 minutes (read height = 7 mm, continuous double orbital shaking at frequency = 282 cpm).

In vivo data processing

After blank subtraction we fit a 2nd-order polynomial to each biological replicate so that we could average normalized data at specific OD_{600} values across replicates. We normalized each replicate using the polynomial fit from a positive control cell line that contains the PAM5087 plasmid but lacks the lac repressor, representing the maximum level of YFP fluorescence possible under each experimental condition. All downstream *in vivo* analyses were conducted at an $OD_{600} \sim 0.53$, near mid-log phase. We then fit each dataset to a Hill model or linear model, using an AIC test for model selection. The Hill model was defined as follows:

$$Normalized\ YFP = YFP_{max} \times \left(\frac{x^n}{K_d + x^n}\right) + YFP_{min}$$

where YFP_{max} corresponds to the maximum YFP fluorescence measured, x is the IPTG concentration, K_d is the apparent dissociation constant, n is the Hill coefficient, and YFP_{min} is the minimum YFP fluorescence measured. For the AIC test, we calculated AIC score for the Hill model and the linear model as follows:

$$AIC = 2k - d \times \log\left(\frac{RSS}{d}\right)$$

where k is the number of model parameters ($k_{Hill} = 3$, $k_{linear} = 2$), d is the number of observations, and RSS is the sum of squared residuals, or the discrepancy between the observed data and the values predicted by the model. The AIC scores for each model were compared and we selected the model with the lower AIC score.

Estimating lac repressor concentrations *in vivo*

To estimate the concentration of the lac repressor in the *in vivo* experiments, we took an approach similar to Sochor et al [168]. We cloned the lac repressor-mCherry construct into pBAD with a c-terminal His-tag and purified the protein as above with the untagged constructs. We constructed a BSA calibration curve using SDSPAGE and gel densitometry. We then used the linear relationship between band density and protein concentration to estimate the concentration of the purified lac repressor-mCherry stock. A calibration curve between protein concentration and mCherry fluorescence was constructed by serially diluting the protein concentration stock (on the same day as the protein concentration vs band density calibration curve) and measuring mCherry fluorescence (excitation: 575 nm, mission: 610 nm) in 96-well black-walled clear-bottom plates (Corning Catalog #: 9018). The linear relationship between concentration and emission intensity at 610 nm was used to convert mCherry signal in all *in vivo* experiments to protein concentration per well.

We converted the concentration of lac repressor per well to an intracellular protein concentration following an approach similar to Sochor et al [168]. Briefly, we measured the OD_{600} for serial dilutions of BLIM cells in 96-well black-walled clear-bottom plates (Corning Catalog #: 9018). We then plated each aliquot on LB agar supplemented with KAN. We fit a line to the data to determine the relationship between OD_{600} on the plate reader and number of cells. We used this calibration curve to calculate the number of cells per well in all *in vivo* experiments.

We estimated the intracellular volume of an *E. coli* cell to be $1x10^{-15}$ L ([168,203]Sochor, Kubitschek & Friske, 1986). The intracellular volume per well was found

by multiplying the number of cells in each well by the intracellular volume. The intracellular protein concentration was then calculated by dividing the concentration of Lac-mCherry per well by the intracellular volume per well. This represents the total concentration of intracellular protein able to bind the O1 operator sequence.

Fluorescence polarization measurements

Thawed protein pellets were buffer exchanged into fresh 1X Binding Buffer three times at 15℃, 5000 RPM (42.24 mM Na_2HPO_4, 7.75 mM NaH_2PO_4, 150 mM NaCl, 1 mM TCEP, 10% glycerol, pH 7.6) using 3 kDa Microsep centrifugal filters (Pall Corporation, #MCP004C41). Scatter-corrected protein concentrations were calculated using A_{280}, A_{320}, and A_{340} measurements. Lac repressor titrations were prepared in triplicate technical replicates for each biological replicate in black 96-well plates (ThermoScientific catalog #265301). IPTG in 1X Binding Buffer was added to each well to achieve one of the following final concentrations: 0 mM, 0.1 mM, 0.3 mM, or 1 mM. An 18-mer oligo tagged at the 5' end with fluorescein, FI-Oid (5'-[FI]ATTGTGAGCGCTCACAAT-3') was ordered HPLC purified from Eurofins MWG and resuspended to 100 μM in 1X TE (pH 8.0) [204,205]. Resuspended FI-Oid was annealed at a final concentration of 10 μM dsDNA using a thermocycler as follows: 1) 1 cycle at 95℃, 5 minutes 2) 70 cycles at 95℃ (-1℃/cycle), 1 minute/cycle. Annealed, double-stranded FI-Oid was then diluted to 100 nM in 1X Binding Buffer and added to each well to achieve a final concentration of 10 nM. Plates were kept covered to protect the fluorescent probe from light and degassed in a temperature-controlled centrifuge pre-equilibrated to 29℃ at 300 rcf for 15 minutes. The plate reader (Spectramax i3 equipped with a fluorescence polarization detection cartridge) was pre-equilibrated to 29℃ for all

experiments. After degassing, samples were shaken in the temperature-equilibrated plate reader (medium orbital) for 15 minutes to ensure the binding reaction reached equilibrium. Fluorescence polarization measurements were taken at an excitation wavelength of 485 nm and an emission wavelength of 535 nm, 6 flashes per read, and 500 ms integration time. Each plate contained four blank wells containing only buffer plus 10 nM FI-Oid and either 0, 0.1, 0.3, or 1 mM IPTG. We calculated a ΔmP value, corresponding to the change in polarization upon the addition of protein, by subtracting the blank from each polarization measurement. We fit the ΔmP values with a single-site binding model to obtain fractional saturation of operator as a function of protein concentration under each IPTG condition.

Lac repressor ensemble modeling

To extract information about the conformations populated by the lac repressor, we used a previously validated thermodynamic model of the lac respressor ensemble[168]. This model describes the lac repressor with two conformations—high (H) and low (L) DNA affinity—each of which can form a complex with effector (E) or DNA (D). Because the lac repressor behaves as a dimer, the repressor dimer can populate states with no effector bound, one effector bound, or two effectors bound. A single dimer binds to a single DNA molecule, meaning each lac repressor is bound to DNA bound or free states. The model assumes the receptor remains a dimer under all conditions. Overall, this model has 16 possible species and five equilibrium constants.

We implemented this model as a function that returns the concentrations of all relevant species given the values of the thermodynamic parameters and total species concentrations of effector, DNA, and lac repressor. This function encodes the

thermodynamic relationships derived by Sochor et al. and enforces mass-balance relationships (e.g. the concentrations of the free and bound DNA species must sum to the total DNA concentration)[168]. Internally, this function guesses values for free effector and then iterates to self-consistency between the thermodynamic and mass-balance relationships. The software implementing this method—written in a combination of the C and Python programming languages— is available online at https://github.com/harmslab/lacmwc.

To estimate the values of the thermodynamic parameters consistent with our binding and induction data, we used a Bayesian Markov-Chain Monte Carlo (MCMC) strategy to sample over parameter combinations. We analyzed all experimental conditions simultaneously for each genotype, globally estimating the thermodynamic parameters for each genotype. We used the following likelihood function:

$$\mathcal{L} = -\frac{1}{2}\sum_{i}\left[\frac{\left(\theta(\vec{p},\vec{c})_{i,calc} - \theta_{i,obs}\right)^2}{\sigma_i^2} + ln(2\pi\sigma_i^2)\right] \quad 3.$$

where $\theta_{i,obs}$ and σ_i are the mean and standard deviation of the measured fractional saturation at condition *i*; $\theta_{i,calc}$ is the calculated fractional saturation under these conditions given a vector of parameters and total concentrations. We constrained all equilibrium constants to greater than zero; otherwise, we used uninformed priors. We generated 4000 MCMC samples from 10 MCMC walkers for each genotype, discarding the first 40 as burn in. We checked for convergence by comparing results from multiple independent runs. We propagated our Bayesian MCMC samples forward through all downstream population and epistasis analyses, leading to the "clouds" of fit lines in all estimates. We used the emcee 3.1.0[206] and the "likelihood" python libraries (https://github.com/harmslab/likelihood) for

these calculations. We implemented our model in Python 3.9 extended with numpy 1.21.1[207], pandas 1.3.2[208],and scipy 1.6.2[209].

Far-UV CD Spectroscopy

Far-UV circular dichroism spectra (200–250 nm) were collected on a Model J-815 CD spectrometer (Jasco) with a 1 mm quartz cell (Starna Cells, Inc.) at 25°C. We prepared samples for the untagged and mCherry-tagged wildtype lac repressor constructs at a concentration of 5 μM in 1X Binding Buffer. We collected three scans and averaged to reduce noise. A buffer blank was subtracted from each spectrum and the raw ellipticity was converted to mean molar ellipticity using protein concentrations and the residue length per construct.

Bridge to Chapter IV

In Chapters II and III, we showed how one can use a combination of theory, computation, and experiment to understand how the molecular properties of proteins shape biology and evolution. Chapter III specifically tested hypotheses generated from theoretical and computational explorations of the relationship between the thermodynamics of macromolecular ensembles and epistasis. We observed signatures of ensemble epistasis—environment-dependent epistasis—both *in vivo* and *in vitro* for the lac repressor protein. We found that all mutant cycles measured showed evidence of ensemble epistasis and used thermodynamic models to understand the origins of each epistatic signal. An important outcome of this chapter is that ensemble epistasis is of a magnitude that is 1) large enough to be measured in an explicitly biological context, meaning that it likely shapes biology and 2) it appears to be very common in mutant

cycles, suggesting that it may be pervasive in biology. The work in Chapter IV we again combine theory and experiment to understand how biochemistry and epistasis more generally shape the evolution of new protein functions in GFP-like proteins from *Faviina* corals. In Chapter IV, simple theoretical models led us to hypothesize that the properties of evolution in large genotype-phenotype maps may be fundamentally different than those in the currently studied small genotype-phenotype maps. This informed our decision to study a natural evolutionary transition in function that occurred over 15 substitutions, resulting in a map that is 96X bigger than the largest currently available dataset. From our current dataset where we have sparsely sampled the full genotype-phenotype map, we find that 1) green fluorescence is a much more common phenotype than red fluorescence, 2) the effects of mutations are highly background-dependent, and 3) epistasis in 2-site maps is pervasive.

CHAPTER IV

MEASURING THE GENOTYPE-PHENOTYPE MAP FOR AN EVOLUTIONARY TRANSITION IN CORAL FLUORESCENCE COLOR

Author Contributions

Anneliese Morrison and Michael Harms conceptualized the study and designed experiments. Michael Harms acquired funding for the study. Anneliese Morrison performed the experiments. Michael Harms administered the project. Anneliese Morrison and Michael Harms analyzed the data. Anneliese Morrison constructed figures. Anneliese Morrison wrote the chapter.

Introduction

A complete, predictive understanding of how proteins acquire new functions is a central goal of molecular evolution and protein engineering. Achieving this requires detailed knowledge of the map evolution navigates—the genotype-phenotype map—and where key sources of unpredictability—such as epistasis, or when the effect of a mutation depends on the presence or absence of other mutations—come from. The connectivity of the genotype-phenotype map is determined by universal physical and biochemical rules. Ultimately, this defines what is—and is not—accessible to evolution by shaping evolutionary trajectories through viable, functional genotypes while avoiding non-functional genotypes.

We can use a simple "word game" to illustrate how the distribution of function shapes evolution in protein sequence space, where we transform one word into another by changing a single letter at a time[210] (Fig 1A). English words are analogous to protein sequences, and single letter changes are analogous to mutations. For the word space in Fig 1, we can see that "evolving" from WORD to GENE is heavily constrained by how meaningful (or functional) intermediate words are. Because there are so few meaningful words (only 7 out of 2^4=16), only one out of 24 (4!) possible trajectories are accessible.

Though protein sequence space is vast, experimental studies of combinatorially complete binary genotype-phenotype maps have been technically limited to small volumes of sequence space, i.e., on the order of three to nine substitutions, or 2^3=8 genotypes [211] to 2^9=512 genotypes[7]. Such studies have broadly led to the conclusions about protein sequence space that suggest 1) it is heavily constrained by epistasis and 2) there are very few accessible evolutionary trajectories[7,17,23,26,27,41,44,55,56,103,122,128,131,132,136,145,146,211–221].

Many naturally evolved traits occurred over the span of many mutations, requiring an evolutionary search of vast regions of the genotype-phenotype map. Theoretical and computational work suggests that larger genotype-phenotype maps may have important consequences for protein evolution: vast "neutral" networks of connected genotypes with nearly identical phenotypes may arise due to the scaling relationship between the number of genotypes (2^L) and the number of forward trajectories (L!), potentially rendering many evolutionary trajectories accessible[222–227] (Fig 4.1B-C). This means that studies of small genotype-phenotype maps may be insufficient to understand protein evolution. Our simple theoretical simulations shown in Fig 4.1C tell us that we

may need to measure much larger genotype-phenotype maps (i.e., L>12) to understand more general features of natural protein evolution.

The current lack of data for high-dimensional genotype-phenotype maps leaves many open questions. How do biochemical constraints shape the distribution of function in large genotype-phenotype maps? Do features like neutral networks and indirect paths increase the accessibility of novel phenotypes? How does the size and connectivity of neutral networks change as a function of map size? How does high-order epistasis impact accessibility as we increase map size? Does the magnitude of epistasis decrease with increasing map size, i.e., do 14^{th}-order epistatic interactions matter?

In this chapter, we set out address this gap by exhaustively characterizing a combinatorially complete 15-site genotype-phenotype map for a natural evolutionary transition in fluorescence color in GFP-like proteins from *Faviina* corals[98,228]. To accomplish this, we developed a quantitative sort-and-sequence (sort-seq) protocol to measure the green and red fluorescence intensity of tens of thousands of pooled variants by combining fluorescence activated cell sorting with next generation sequencing. In our first biological replicates, we characterized ~7% (3,676 out of 49,152 genotypes) of the full genotype-phenotype map. We find that a large fraction of the measured genotype-phenotype map exhibits green fluorescence (~79%), while only ~1.5% exhibit red fluorescence. The remaining ~20% of observed genotypes exhibit completely broken chromophores. As we progress through the evolutionary trajectory by the introduction of derived amino acids, we see a trend towards decreased green fluorescence and increased red fluorescence. There are extensive background-dependent effects of single mutations in both phenotypes, for example, more mutations are beneficial for redness when

introduced into the Q62H background vs when introduced into any genetic background. We find extensive pairwise and third-order epistasis between mutations. Finally, we observe all types of evolutionarily important classes of epistasis, indicating that the genotype-phenotype map may be fairly rugged.

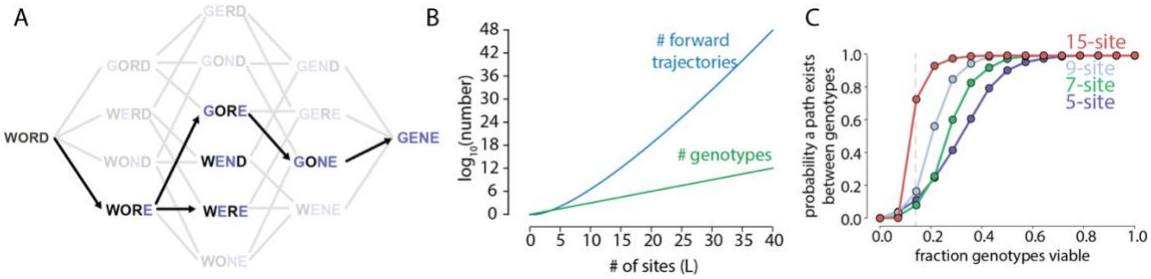


Fig 4.1 Protein evolution in large genotype-phenotype maps may have different properties than small maps. A) An illustration of John Maynard Smith's word game between WORD (black) and GENE (blue). Transparent intermediate words are non-meaningful, while fully opaque words are meaningful. Connections that lead to/from non-meaningful words are greyed out, while those that connect meaningful words are black. B) An illustration of the scaling-relationship between the number of sites (L, x-axis) in a genotype-phenotype map and the number of genotypes (green) and forward trajectories (blue) (log-scale, y-axis). C) A simple theoretical result showing the relationship between the probability of traversing a genotype-phenotype map (y-axis) and the fraction of viable genotypes (x-axis) for different map sizes of 2^L, where L=5 (slate blue), 7 (green), 9 (light blue), 15 (red).

Materials and Methods

Library design and construction

The ancestral GFP-like protein gene was cloned into the pQE-30 plasmid (Qiagen, Appendix C Supplementary Section 1 and Supplementary Fig C1)[229]. It was cloned C-terminal to the EBFP2 gene and expressed as a fusion protein. A rigid alpha-helical linker sequence was (GSLAEAAAKEAAAKEAAAKAAAAS) inserted between the two fluorescent proteins to reduce Förster resonance energy transfer (FRET) between

them[201,229] (Appendix C, Supplementary Fig C2). Following closely to Sarkysian et al, we designed our construct so that we could link each genotype with a unique 25 base pair "molecular barcode" at the C-terminal end of the protein coding sequence[229]. We added several restriction sites throughout the construct (NheI at the N-terminus and EagI/NotI C-terminal to the molecular barcode) and took advantage of naturally occurring restriction sites (FspI, SfoI, BseYI, BsaAI, PvuII, and BbsI) to do barcode association experiments, which are described in more detail in the next section (see Appendix C, Supplementary Fig C1-C2 for construct design). We cloned the following constructs into the pQE-30 vector as non-fusion proteins (i.e., lacking the EBFP2-rigid linker sequence) to serve as fluorescence minus one (FMO) controls in all FACS experiments: pQE-30-EBFP2, pQE-30-AncestralGFP, and pQE-30-AncestralDerived.

For library construction, the ancestral GFP-like protein sequence was codon-optimized for expression in *E. Coli*. We ordered a combinatorially complete library with all 15 mutations introduced into the ancestral GFP-like protein background from Genewiz using combinatorial trimer synthesis technology (Fig 4.2A, Appendix C Supplementary Table C1)[53]. Trimer synthesis allowed us to control the frequency of amino acid at each of the 15 sites, for example, at binary sites the ancestral and derived site frequencies should be around 50%, while at the triple site each amino acid should be around 33.3%. Molecular barcodes were added during gene synthesis by addition of 25 degenerate nucleotides C-terminal to the GFP-like library (Appendix C, Supplementary Fig C2). After gene synthesis, the entire library was cloned into the pQE-30-EBFP2 vector, downstream of the rigid linker sequence (Appendix C, Supplementary Fig C1). The library was transformed into XL10-Gold ultracompetent cells and stored at -80°C as

individual 200 μl glycerol stocks for sort-seq experiments. The final library diversity was estimated to be around 1,000,000 clones, meaning that there are approximately 1,000,000 unique barcodes in our library that map to the known 49,152 library sequences.

Barcode association

Because we were technically unable to sequence entire length of the GFP-like protein using current next generation sequencing technology, we used paired-end next generation sequencing platforms (Illumina Miseq paired-end 300 and Illumina NovaSeq paired-end S1 300 cycle and S4 300 cycle) to construct a mapping between each library variant and its corresponding unique 25 bp molecular barcodes. The library was digested into five fragments by incubating library DNA per digest at 37°C overnight with the following restriction enzymes: FspI-HF/EagI-HF (producing a 724 bp fragment), SfoI-HF/EagI-HF (620 bp), BseYI-HF/EagI-HF (498 bp), BbsI-HF/EagI-HF (352 bp), BsaAI-HF/EagI-HF (320 bp), and PvuII/BbsI/EagI (209 bp). This produced five fragments of the library physically linked to molecular barcodes (see Appendix C, Supplementary Fig C2). Fragments were run on a 1% agarose gel at ~80V until resolved and purified using a gel extraction and purification kit (GeneJet Catalog #: K0691). Purified fragments were prepared for Illumina sequencing using a PCR-free workflow via blunt-end ligation of Illumina TruSeq-like adapters to each fragment (KAPA HyperPrep Kit, Roche). A PCR-free workflow is critical for preparation of these samples, as we saw evidence of extensive PCR crossover when any amplification steps were included in our workflow because the GFP-like protein library is low diversity. We modified a single step in the protocol provided by the kit: we extended the blunt-end ligation to be overnight at 4°C

instead of 20 minutes at 20°C. We found that extending this step improved the overall yield of adapter-ligated fragments. Post adapter ligation, we purified each reaction using a 1.2X bead-cleanup (Omega Bio-Tek MagBind TotalPure NGS beads, SKU #: M1378-01). Some samples required an additional bead clean up or a BluePippin size selection step to remove residual adapter dimer.

We prepared custom Python sequencing analysis scripts that mapped each GFP-like protein fragment to its unique molecular barcode and then aligned fragments with identical barcodes to reconstruct full-length variant sequences.

1. All paired-end sequences underwent the following quality control steps:
    a. Any bases with a phred score below 5 were replaced with 'N''s to indicate that they are ambiguous.
    b. Both the forward/reverse reads must match (within 1 base pair) the expected 5'/3' sequence.
    c. Reverse reads must be long enough to "gap" the sequence correctly (i.e., is Y217 present or not).
    d. The spacer region between the GFP-like gene and the barcode must be identical in length and within 3 base pairs of the expected sequence.
    e. The barcode must have a length of 50 base pairs and all must be unambiguous (i.e., must have phred score > 5).
    f. Sequences meeting the above criteria were gapped to account for the presence/absence of Y217 and aligned to the reference (ancestral GFP) sequence.

g. The identities of aligned sequences were then extracted for all 15 postions in the library by comparing the aligned sequence to the ancestral GFP reference sequence to identify if the sites were ambiguous, ancestral, or derived.
h. If the forward and reverse read identities differed at more than 10% of the sites, both sequences were thrown out.

2. Amino acid sequences passing QC were mapped to their corresponding 25 base pair barcode
3. All barcodes were clustered with Bartender using a hamming distance of 3^{230}. All sequences corresponding to the same barcode "centroid" were averaged, where a centroid is the consensus barcode sequence for all clustered barcodes. Positions in each sequence were counted as ambiguous (represented by '?') if the frequency of a given position in the consensus sequence was below 90%.

Out of the 373,556,262 sequences that passed quality control, we observed 9,060,526 unique barcodes prior to clustering and 2,203,019 unique centroids after clustering. Barcode centroids mapped to 161,365 ambiguous sequences and 25,349 unambiguous sequences.

Library expression and photoactivation

To begin expression, we used glycerol stocks of each construct to inoculate a 10 mL culture of LB supplemented with 35 μg/mL CAM and 100 μg/mL AMP. Starter cultures were incubated at 37°C, 250 RPM overnight. In the morning, starter cultures were used to inoculate a culture for protein expression. Newly inoculated cultures were grown at 37°C, 250 RPM in 50 mL LB supplemented with 35 μg/mL CAM and 100

μg/mL AMP until an OD_{600} of ~0.4-0.8 was reached. Expression was then induced by the addition of 1 mM IPTG. Cultures were incubated at 37d°C, 250 RPM for 4 hours. Cells were then pelleted at 5000 RPM, 4°C for 10 minutes. The supernatant was discarded, and the pellet was resuspended in 1X PBS supplemented with 1% BSA that was pre-chilled to 4°C. The library (and all compensation controls—for example, pQE30-DerivedGFP) was then photoactivated for 10 minutes at room temperature while stirring under UV flashlight (385-395 nm) supplied with new batteries for each experiment (uvBEAST Black Light UV Flashlight, Amazon Catalog #: B078Y6G469). Prior to each sort-seq experiment, we checked that the extent of protein expression and photoactivation was comparable to past experiments by measuring the emission spectra of each construct (for blue fluorescence: ex = 383 nm, em = 400-500 nm; for green fluorescence: ex = 490 nm, em = 500-590 nm; and for red fluorescence: ex = 550, em = 560-620 nm). All spectra were measured on a SpectraMax i3 plate reader in 96-well black-walled, clear-bottomed plates (ThermoScientific item #: 265301).

For two of the biological replicates of the sort-seq experiment, we used a plating method to pre-culture the library prior to expression. We streaked out the library from glycerol stocks onto 15 mm LB agar/CAM/AMP plates and incubated them overnight at 37°C. In the morning, we resuspend colonies in ~20 mL LB/CAM/AMP and used the resuspension to inoculate 50 mL LB/CAM/AMP for expression. All subsequent steps were identical to the liquid overnight culture method. We found that the only difference between preparing the library on solid media vs liquid agar was that we observed a small "high blue" population in our FACS data. Upon sorting and sequencing ~10 such cells, we found that these either had extensive rearrangements in the plasmid (i.e., EBFP2

inserted in the middle of the GFP sequence) or did not prime, indicating that these suffered from rearrangements as well (see Appendix C, Supplementary Fig C4).

Fluorescence activated cell sorting

All FACS experiments took place on the same day as protein expression. Each experiment required the expression of the GFP-like protein library as well as several FMO compensation controls: pQE-30 (empty) as a negative control, pQE-30-EBFP2, pQE-30-AncestralGFP, and pQE-30-DerivedGFP (photoactivated). After photoactivation of the library and the pQE-30-DerivedGFP construct, all constructs were washed two times as follows: pellet at 5000 RPM, 4°C for 5 minutes, discard supernatant, and resuspend in ~20 mL pre-chilled 1X PBS supplemented with 1% BSA. The OD_{600} of each sample was measured on a SpectraMax i3 plate reader in 96-well black-walled, clear-bottomed plates (ThermoScientific item #: 265301). These samples were then diluted to an OD_{600} of ~0.1 in 10 mL pre-chilled 1X PBS supplemented with 1% BSA. These samples were placed on ice until FACS experiments were performed. All FACS experiments were started within an hour of preparing samples.

To characterize the fluorescence of each variant in the GFP-like protein library, we performed three biological replicates of the FACS experiments, where we sorted cells based on green and red fluorescence intensity. All FACS experiments were done on a SONY SH800 cytometer. We used the 405 nm, 488 nm, and 561 nm lasers with the FL1 (450/50, blue channel), FL2 (510/20, green channel), and FL3 (585/30, red channel) detectors and 100 μm sorting chips. The sample and sorting chambers were set to 5°C and sample agitation was turned on. We performed compensation using all FMO

controls, i.e., an autofluorescence negative control (pQE-30 (empty)), an EBFP2 control (pQE-30-EBFP2), a green GFP-like protein control (pQE-30-AncestralGFP), and a red GFP-like protein control (pQE30-DerivedGFP, photoactivated). After compensation, we measured ~5000 events for rainbow calibration beads, which were prepared by addition of ~2-3 drops of beads to 1-2 mL 1X PBS/1%BSA (Spherotech Catalog #: RCP-30-5A). We included this control to measure variation in signal (from the cytometer hardware rather than biological sample variation) for experiments run on different days. We found there was not a significant change in signal intensity for each sort-seq experiment (Appendix C, Supplementary Fig C3). We then proceeded to two-way cell sorting. First, we used a linear gate to select the cell population that exhibited homogeneous EBFP2 fluorescence as a control to ensure that variation in green or red fluorescence intensity was due to changes in the GFP-like protein, not differences in protein expression. This "normal blue intensity" cell population represented >95% of the entire cell population. Cells in the "normal blue intensity" population were then sorted into 5-6 separate populations based on their intensity in the green channel and 4-5 populations based on their intensity in the red channel. We used FACS data for ten randomly selected clones from the library to determine the approximate width of the fluorescence distributions for our system. We used this to inform how many bins we used in each channel. We sorted between 500,000 to 1,000,000 events per gate. All two-way sorting was done into 5 mL Eppendorf tubes containing 1 mL pre-chilled 1X PBS/1% BSA. All gates were selected evenly across the logarithmic scale. We additionally sorted cells with low green fluorescence and low red fluorescence (i.e., with intensity below our lowest intensity gates) to identify genotypes with broken chromophores. All Eppendorf tubes containing

sorted populations were gently vortexed placed on ice, and immediately centrifuged at 7000-75000 rcf for 10 minutes at 4°C. All supernatant was carefully removed and the pellet (which was visible with 500,000+ sorted events) was resuspended with ~50 μl sterile double deionized H_2O. Samples were stored at -20°C until barcode extraction.

Barcode extraction and sequencing

After each sort-seq experiment, we prepared each sorted population for next generation sequencing (NovaSeq paired-end S1 300 cycle (biological replicates 1 and 2) and paired-end S4 300 cycle (biological replicate 3) as follows:

1. Sorted cells were thawed and boiled.
2. Barcodes were PCR amplified in triplicate using Phusion polymerase (NEB catalog #: M0530S), yielding a 105 bp fragment containing the molecular barcode.
3. Post amplification, samples were pooled and purified using a bead clean-up (Omega Bio-Tek MagBind TotalPure NGS beads, SKU #: M1378-01).
4. TruSeq-like adapters were ligated to barcodes using a PCR-free workflow (KAPA hyperprep kit). After ligation, we purified samples using a 1.2X bead clean-up. Some samples required additional clean-up steps to remove residual adapter dimer.

In our first sort-seq sequencing run (NovaSeq S1 300 cycle, biological replicates 1 and 2), we found that our PCR-free workflow led to extensive index hopping. For the third biological replicate of our sort-seq experiments, we added an additional PCR step using the KAPA Hyperprep reagents and protocol. PCR reactions were purified using a

1X bead cleanup. Adapter ligated samples were submitted to the University of Oregon Genomics and Cell Characterization Facility for sequencing. See Appendix C, Supplementary Section 2.

Barcode sequencing analysis pipeline

Raw de-multiplexed next-generation sequencing data was first subjected to the following quality control steps:

1. Both reads had to have a mean Phred score above 19.
2. The 25 bp barcode was extracted from both reads using the known barcode flanking sequence. If an extracted barcode from one of the reads had a length <25 bp, neither read was used.
3. Barcodes from both reads had to be either identical or only have a single mismatch.

We then extracted the raw number of counts per barcode, per sorted population. Barcodes were linked with known library genotypes using the barcode-genotype mapping data. Briefly, all barcodes from both sort-seq biological replicates and all barcode-genotype mapping runs were clustered using the Bartender clustering package[230]. Two barcodes were placed in the same cluster if they differed at a maximum of three sites. This clustering step resulted in a list of barcodes per genotype, allowing us to calculate the total number of counts per genotype in each sorted population. Raw counts in each sorted population were considered below the limit of detection if they were below a frequency of 1×10^{-5} (see Fig 4.3H for details). All other counts were then converted to

frequencies by dividing counts by the total number of observations in each sorted population.

Extracting phenotypes from genotype frequencies across sorted populations

We used a simple mean method to estimate the mean and standard deviation of the green and red fluorescence intensities for each genotype, as in Peterman et al[231]. Essentially, we took the weighted average of all sorted cells. We assume that for a hypothetical cell in a gate *j*, the cell fluorescence, f_j, is as follows:

$$f_j = b\sqrt{L_j \times U_j}$$

where L_j and U_j represent the upper and lower fluorescence boundaries for bin *j* and *b* represents a location within the width of bin *j*. Peterman et al found that setting b equal to

$$b = \frac{w}{\left(e^{\frac{w}{2}} - e^{-\frac{w}{2}}\right)}$$

minimized bias[231], where *w* equals bin width. We calculated the mean fluorescence, *v*, from bin *j* to bin *m* as follows:

$$v = \frac{1}{N}\sum_{j=1}^{m} r_j \times f_j$$

where *N* represents the total frequency of a genotype over all bins, r_j is the frequency of a genotype in bin_j. We calculated the standard deviation as follows:

$$sd = \sqrt{\frac{1}{N}\sum_{j=1}^{m} r_j\left(f_j - v\right)^2}.$$

Results

Measuring the genotype-phenotype map for a natural evolutionary transition in coral GFP-like protein fluorescence color

Our goal is to understand how increasing the size of the genotype-phenotype map impacts the accessibility of novel functions. This requires characterizing a large volume of sequence space for a functional transition. Characterization of a large genotype-phenotype map required identifying a system with 1) sufficient phylogenetic information for an ancestral sequence reconstruction (ASR) and 2) a phenotypic transition that took place over >12 substitutions and 3) a phenotype that is easily quantitatively measured in a high-throughput manner.

We focused on a map size of L > 12 because we anticipate that this will allow us to measure differences in accessibility between large and small maps that would not be apparent in smaller maps due to the scaling relationship between map size and the number of possible trajectories. Fig 4.1C illustrates how this relationship might impact the probability that an accessible trajectory exists between a derived and ancestral genotype as a function of map size, where phenotypes are randomly assigned to each genotype as either “dead” or “alive”. If we compare the 5-site and 15-site spaces, we find that even if relatively few genotypes are viable, for example ~14% (Fig 4.1C, dashed grey line), there is an 11% chance that an accessible path exists across the 5-site space, while there is an ~71% chance that an accessible path exists in the 15-site space. This behavior arises because as we increase map dimensionality, the number of possible

forward trajectories grows rapidly from 5! = 120 forward trajectories to 15!=1.3×10^{12} (Fig 4.1B).

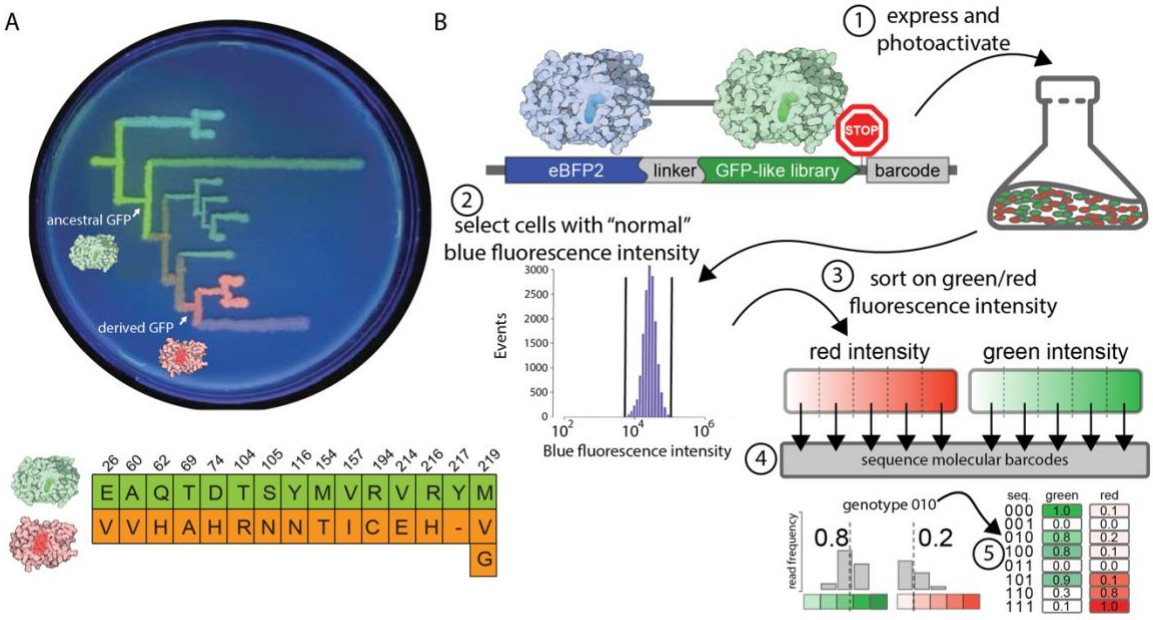


Fig 4.2 Constructing and measuring the genotype-phenotype map for a transition in fluorescence color. A) Figure adapted from Ugalde et al.[98] The fifteen substitutions that converted the green ancestral to derived red GFP-like protein are shown below. B) Sort-seq experimental scheme: 1) the GFP-like library is expressed in bacteria as a fusion protein containing a molecular barcode, 2) we control for expression variability and identify broken chromophores by gating on the homogeneous EBFP2-expressing population, 3) we sort the "normal" EBFP2 population into bins of varying green/red fluorescence intensity, 4) we sequence the molecular barcodes from each pool and quantify green/red fluorescence using the distribution of genotype frequencies across bins, and 5) we use quantitative phenotypes to construct a complete genotype-phenotype map.

The family of GFP-like proteins from *Faviina* corals is one such protein meeting our criteria. Previously, the Matz group used ASR to study an evolutionary transition between an ancestral green GFP-like protein to a derived photoactivatable red GFP-like protein[98,228] (Fig 4.1B). Out of the 37 possible amino acid substitutions between the ancestral and derived proteins, they found that 12 substitutions were sufficient to give rise to a photoactivatable red fluorescent protein from the green ancestor[228] (Fig 4.1C).

Similar to other protein genotype-phenotype maps, this evolutionary transition exhibits extensive epistasis. For example, Q62H mutation is required for the chemistry of photoactivation and chromophore maturation for the red phenotype. However, the mutation alone in the ancestral background is insufficient to result in red fluorescence, indicating that this position is highly epistatic[98,228,232]. Additionally, one critical amino acid substitution (M219G) could not occur in a single mutation in the nucleotide sequence, indicating that other intermediate amino acids, or indirect trajectories, may have been required for this evolutionary transition.

We selected 15 substitutions that were found to be the most highly associated with the observation of red fluorescence in a previous study to construct our map[228] (Fig 4.2A and Appendix C, Supplementary Table C1). At site 219 we included a third possible amino acid, valine, in addition to the ancestral (methionine) and derived (glycine) amino acids. Sites with more than two amino acids allow for mutational reversions during adaptation, or indirect trajectories[133,222,223]. Such indirect trajectories have shown to increase accessibility[133,222,223]. Inclusion of this site in our map will allow us to evaluate the relative importance of indirect paths in natural evolutionary transition as a function of map size. This resulting map contains 14 single-step mutations and one two-step mutation, or ($2^{14} \times 3 =$) 49,152 possible genotypes between the ancestral and the derived GFP-like protein sequences.

We developed a sort-seq method to quantitatively characterize the genotype-phenotype map for all evolutionary intermediates between the ancestral green and derived red GFP-like proteins (Fig 4.2B). Sort-seq methods are extremely powerful in

that they yield high-throughput, quantitative phenotype measurements by combining fluorescence activated cell sorting (FACS) and next-generation sequencing[229,231]. Briefly, we first express our library in bacteria. We photoactivate the library under broad spectrum UV light to induce maturation of the red chromophore. On the FACS instrument, we select the population of cells with uniform EBFP2 fluorescence as 1) a control for variable protein expression and 2) to identify genotypes with "broken" chromophores. We then sort the population into bins of different fluorescence intensity in the red and green channels. Post-sorting, we isolate and sequence barcodes from each population of known green or red fluorescence intensities. Because genotypes are sorted multiple times, each genotype has a corresponding histogram of read counts spread over one or more bins with known intensity. This histogram of read frequency per bin is used to calculate quantitative estimates of the green and red fluorescence intensities for each genotype in the library (Fig 4.2B).

Proof-of-principle sort-seq experiment

First, we wanted to confirm that our sort-seq strategy shown in Fig 4.2B met the following criteria: 1) it yields phenotype estimates that match known values and 2) the post-sorting sample preparation did not distort input genotype frequencies. We also wanted to know what our limit of detection was for observing variants that are present at low frequency in the library.

To demonstrate our sort-seq strategy works, we randomly sampled and sequenced five clones from the library and individually characterized their intensities in the green channel on the cytometer. We then prepared three mixtures of bacteria expressing each

clone. Prior to our sort-seq experiment, we took a sample of each mixture for sequencing as a control to ensure that the barcode frequencies derived from our sequencing data match the known input frequencies. During our sorting experiment, we selected the population of EBFP2-expressing cells and divided that population into four subpopulations based on their fluorescence intensities in the green channel. Fig 4.3A-B shows an example of our binning strategy in our trial sort-seq experiment for one of the biological replicates. After sorting based on green fluorescence intensity, we extract barcodes and submitted them for sequencing.

Fig 4.3B-D shows the correlation between the known and measured input frequencies for each genotype in the pre-sort samples. We found that we can accurately calculate the frequencies of genotypes in the input pool from sequencing data, meaning that our sample preparation protocol does not distort the frequencies of genotypes in our samples.

Fig 4.3F shows the barcode frequencies for each clone per bin for mixture 1 (bins shown in Fig 4.3B). Fig 4.3G shows how well the mean phenotypes we inferred matches the known values for each clone. We find that for all genotypes the inferred green fluorescence intensities are within the standard deviation of known values, indicating that our sort-seq experiments yield informative estimates of fluorescence intensities.

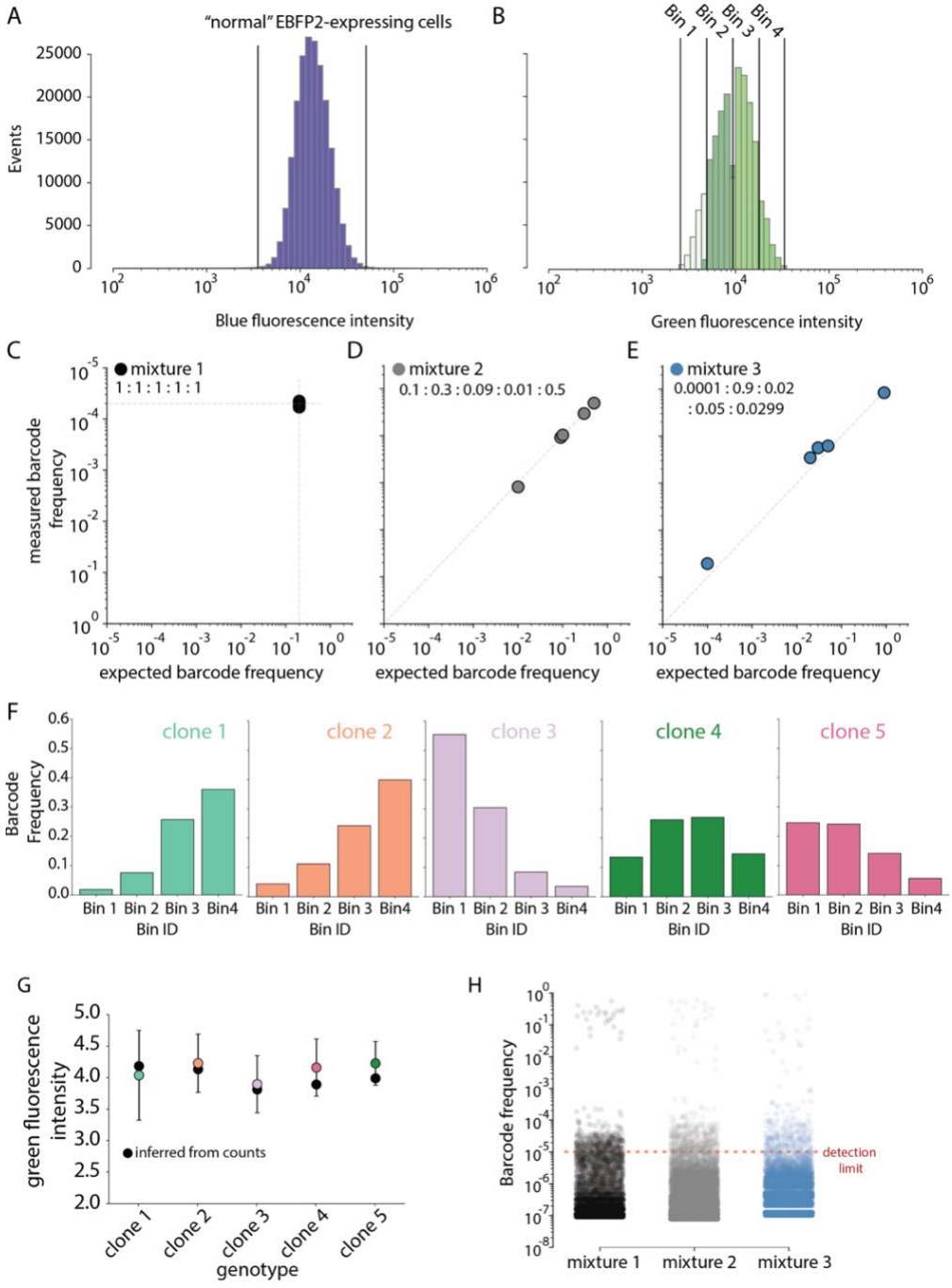
We found that ~0.003-0.006% of reads from each pool were background noise, representing either sequencing errors or contamination from other lanes. We found that the bulk of such sequences were present at a frequency below $\sim 1 \times 10^{-5}$ (Fig 4.3H). We

used this as our limit of detection for our sort-seq datasets. All data in Fig 4.3 are for a single sort-seq trial.

Subsampling the genotype-phenotype map

Next, we performed two full biological replicates of our sort-seq protocol to sparsely characterize the GFP-like protein library in the red and green channels. These small-scale replicates were done to characterize general features of the map and to determine the sequencing depth necessary for full-scale library characterization. The binning strategies for both biological replicates are shown in Appendix C, Supplementary Fig C4-C6. The number of genotypes observed was strongly limited by 1) the occurrence of extensive index hopping during our sequencing run and 2) insufficient depth in our barcode mapping experiments to unambiguously assign each barcode a fully reconstructed genotype.

Fig 4.3 Using sort-seq to infer green fluorescence intensity of mixtures of known genotypes. (next page) A) Example of our gating strategy for identifying the "normal" EBFP2 population from one sort-seq trial experiment. B) Example of our gating strategy in the green channel for one trial. C) Plots showing the expected genotype frequency in one of three pre-sort mixtures (x-axis) and the inferred genotype frequency using next-generation sequencing data (y-axis) for one sort-seq trial. Mixture 1 contains each genotype at equal frequency (1:1:1:1:1), mixture 2 contains each genotype at a relative frequency of 0.1:0.3:0.09:0.01:0.5, and mixture 3 contains each genotype at a relative frequency of 0.0001:0.9:0.02:0.05:0.0299. F) Genotype frequency calculated from sequencing data (y-axis) for each of the five genotypes in mixture 1 as a function of bin ID (x-axis). G) Inferred green fluorescence intensity from data in Panel F (y-axis) for each of the five genotypes. Black datapoints are the average inferred intensities from each of the three mixtures. Colored datapoints are the averaged intensities calculated from FACS data for each genotype individually, where data was averaged from two separate FACS experiments. Error bars represent the standard deviation of the phenotypes inferred directly from FACS data. H) Jitter plot of the barcode frequency (y-axis) from sequencing data for each of the three mixtures (x-axis, mixture 1 shown in black, mixture 2 shown in grey, and mixture 3 in blue). The detection limit used in quality control steps for the sort-seq experiments is shown by the red dashed line (1×10^{-5}).


A
"normal" EBFP2-expressing cells
Events
Blue fluorescence intensity
B
Bin 1
Bin 2
Bin 3
Bin 4
Green fluorescence intensity
C
mixture 1
1 : 1 : 1 : 1 : 1
measured barcode frequency
expected barcode frequency
D
mixture 2
0.1 : 0.3 : 0.09 : 0.01 : 0.5
expected barcode frequency
E
mixture 3
0.0001 : 0.9 : 0.02
: 0.05 : 0.0299
expected barcode frequency
F
clone 1
clone 2
clone 3
clone 4
clone 5
Barcode Frequency
Bin 1 Bin 2 Bin 3 Bin4
Bin ID
G
green fluorescence intensity
inferred from counts
clone 1
clone 2
clone 3
clone 4
clone 5
genotype
H
Barcode frequency
detection limit
mixture 1
mixture 2
mixture 3

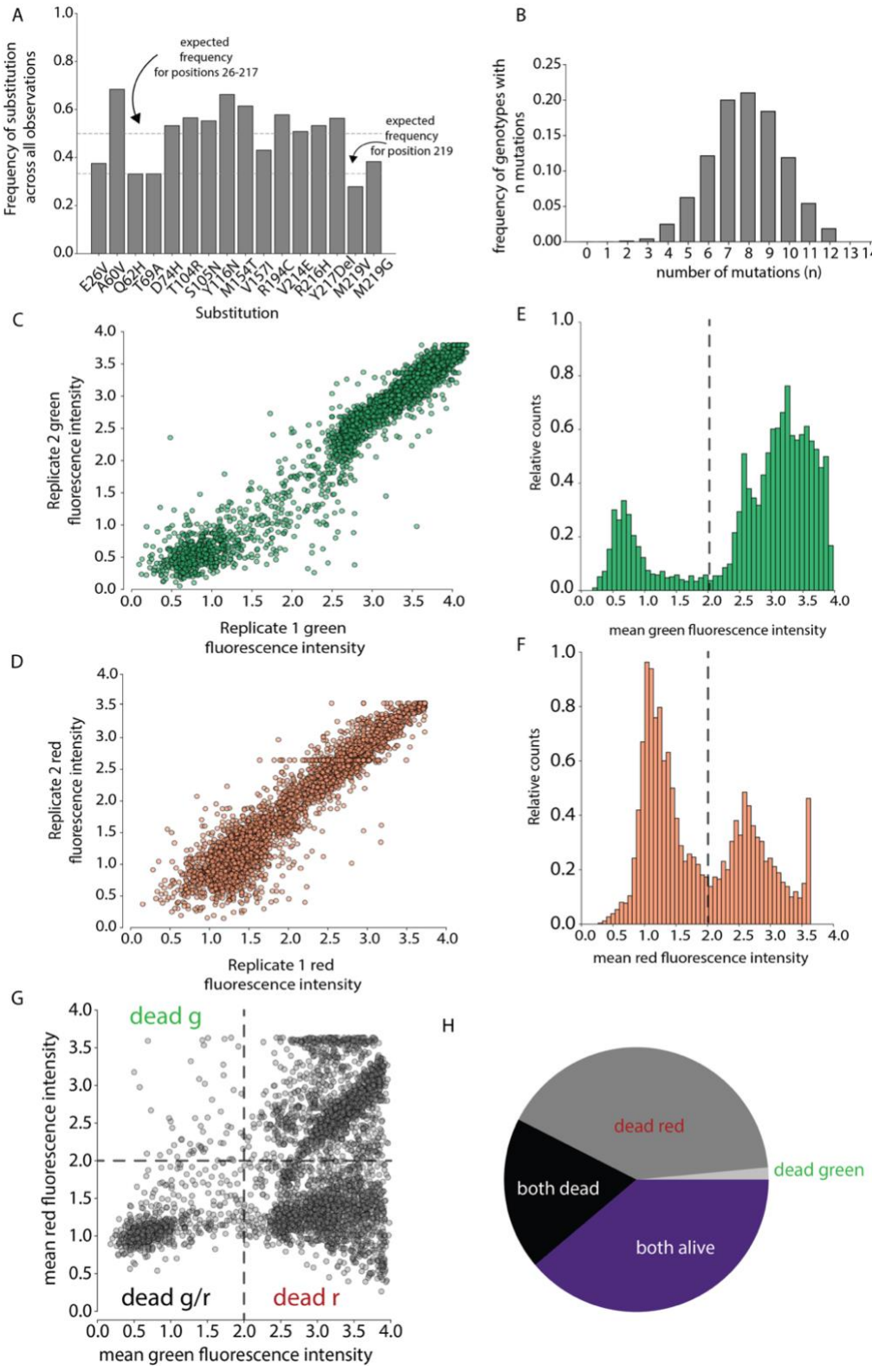
Given our sequencing limitations, we proceeded to extracting barcodes from our sequencing data and used our barcode mapping to determine which genotypes were observed in each dataset. We see ~17,500 unambiguous genotypes and ~33,600 ambiguous genotypes with at least one read in both replicates. After quality control, we find that 3,676 unambiguous genotypes and 175 ambiguous genotypes have quantitative phenotype values in the green and red channels for at least one replicate. Fig 4.4A shows that the frequency of each mutation in the unambiguous dataset deviates from our expectation of ~0.5 for binary positions (substitutions 26-217) and ~0.333 for the triple position (substitution 219), suggesting that some mutations are overrepresented in our library. We also found that there are few sequences containing <3 and >12 total mutations (Fig 4.4B). We did not observe either the ancestral (wildtype) or the derived GFP-like proteins in either of our datasets.

Figs 4.4C-D show the correlation between the green and red phenotypes in replicates 1 and 2, respectively. We find that there is a strong correlation between the phenotypes inferred from sort-seq data in both channels. In the green channel, we see two distinct populations corresponding to “broken” and bright fluorescent proteins (below a fluorescence intensity of~2, Fig 4.4E). In the red channel, we distinct populations corresponding to broken and bright fluorescent proteins, but the “broken” chromophore population seems to be well-resolved from the bright chromophore population (Fig 4.4F).

Fig 4.4G shows the correlation between mean green and red fluorescence intensities for genotypes observed in both channels. There is a distinct population that exhibits correlated green and red fluorescence (top right quadrant). We will analyze this

subpopulation in greater depth in the next section, but we suspect that this is bleed through from the green channel into the red channel that was not captured by compensation. We find that ~40% of the genotypes observed above 2.0 in both channels, while ~42% are below 2.0 in one channel and ~18 % are below 2.0 in both channels (Fig 4.4H). The observation of high red fluorescence without green fluorescence is quite rare, constituting ~1.5% of the population. For all subsequent analyses, we will treat chromophores in the red and green channels that have a fluorescence intensity below 2.0 to be "dead", or broken, chromophores.

Fig 4.4 Characterizing a subset of the GFP-like protein library. (next page) A) Frequency of each substitution in all genotypes observed (y-axis) with substitutions labelled on the x-axis. The grey dashed lines indicate the expected frequency of each mutation for the two-mutation (50%) and three-mutation (33.3%) sites in the absence of overrepresentation of specific variants in the library. B) Frequency of genotypes with *n* mutations (y-axis) versus the number of substitutions *n* in a given genotype (x-axis). C) Correlation between green fluorescence intensities inferred from replicates 1 (x-axis) and 2 (y-axis). D) Correlation between red fluorescence intensities inferred from replicates 1 (x-axis) and 2 (y-axis). E) Distribution of mean green fluorescence intensity (averaged over both replicates) with relative counts on the y-axis and green fluorescence intensity on the x-axis. The dashed grey line shows the cutoff between "bright" and "broken" chromophores. F) Distribution of mean red fluorescence intensity (averaged over both replicates) with relative counts on the y-axis and red fluorescence intensity on the x-axis. The dashed grey line shows the cutoff between "bright" and "broken" chromophores. G) Mean green fluorescence intensity (x-axis) versus mean red fluorescence intensity (y-axis). The plot is broken up into quadrants using cutoffs in panels F and G to identify bright green and red chromophores (upper left quadrant), bright green/broken red (lower left quadrant, labelled "dead r"), broken green/bright red (upper right quadrant, labelled "dead g"), and broken green/broken red (lower right quadrant, labelled "dead"). H) Pie chart showing the proportion of genotypes falling in each quadrant, colored as follows: black ("dead" quadrant), dark grey ("dead r" quadrant), light grey ("dead g" quadrant), and purple (upper left quadrant).


A
Frequency of substitution across all observations
expected frequency for positions 26-217
expected frequency for position 219
E26V
A60V
Q62H
T69A
D74H
T104R
S105N
Y116N
M154T
V157I
R194C
V214E
R216H
Y217Del
M219V
M219G
Substitution
B
frequency of genotypes with n mutations
number of mutations (n)
C
Replicate 2 green fluorescence intensity
Replicate 1 green fluorescence intensity
D
Replicate 2 red fluorescence intensity
Replicate 1 red fluorescence intensity
E
Relative counts
mean green fluorescence intensity
F
Relative counts
mean red fluorescence intensity
G
dead g
dead g/r
dead r
mean red fluorescence intensity
mean green fluorescence intensity
H
dead red
dead green
both dead
both alive

Effects of mutations on green and red fluorescence intensities

Next, we wanted to examine how green and red fluorescence intensities change as a function of mutation. We began by looking at the effect of the Q62H mutation. The Q62H mutation is located within the chromophore and has been shown to be critical for the evolution of photoactivatable red fluorescence in coral GFP-like proteins[98,228,232,233]. Upon exposure to UV-light the polypeptide backbone is cleaved between the amide nitrogen and alpha carbon of H62 and a double bond is formed between the alpha and beta carbons of H62[233–235] (Fig 4.5A). As a result, the pi-electron conjugation of the neighboring tyrosine residue extends to H62 and the fluorescence spectrum of the protein is red-shifted[233–235].

Due to the critical nature of the Q62H mutation for the observation of red fluorescence, we expect to see lower red fluorescence intensities in genotypes without the Q62H mutation than in genotypes with the Q62H mutation. Fig 4.5B-C shows the correlation between green and red fluorescence in the +Q62H and -Q62H backgrounds. Surprisingly, we see a distinct high red/high green fluorescence population in the background lacking the critical Q62H mutation (Fig 4B, red circle). Given the lack of a similar population in the +Q62H population and the nature of the correlation and the critical nature of the Q62H mutation for the underlying chemistry of the red chromophore, we suspected that the source of this high green/high red -Q62H population was bleed through from the green channel into the red channel. We often see bleed through of this type, as 1) there is overlap in the green and red fluorescence emission spectra and 2) the green chromophore is much brighter than the red chromophore. This

makes it difficult to fully decompose the two signals given the filter set on our instrument. For the remainder of the analysis, we replaced the red fluorescence intensities for this subpopulation with a value corresponding to the broken red chromophore population (i.e., we took the mean of the population below an intensity of 2.0 in Fig 4.4F). However, more characterization is needed to confirm that this subpopulation is arising from bleed through from the green to red channel.

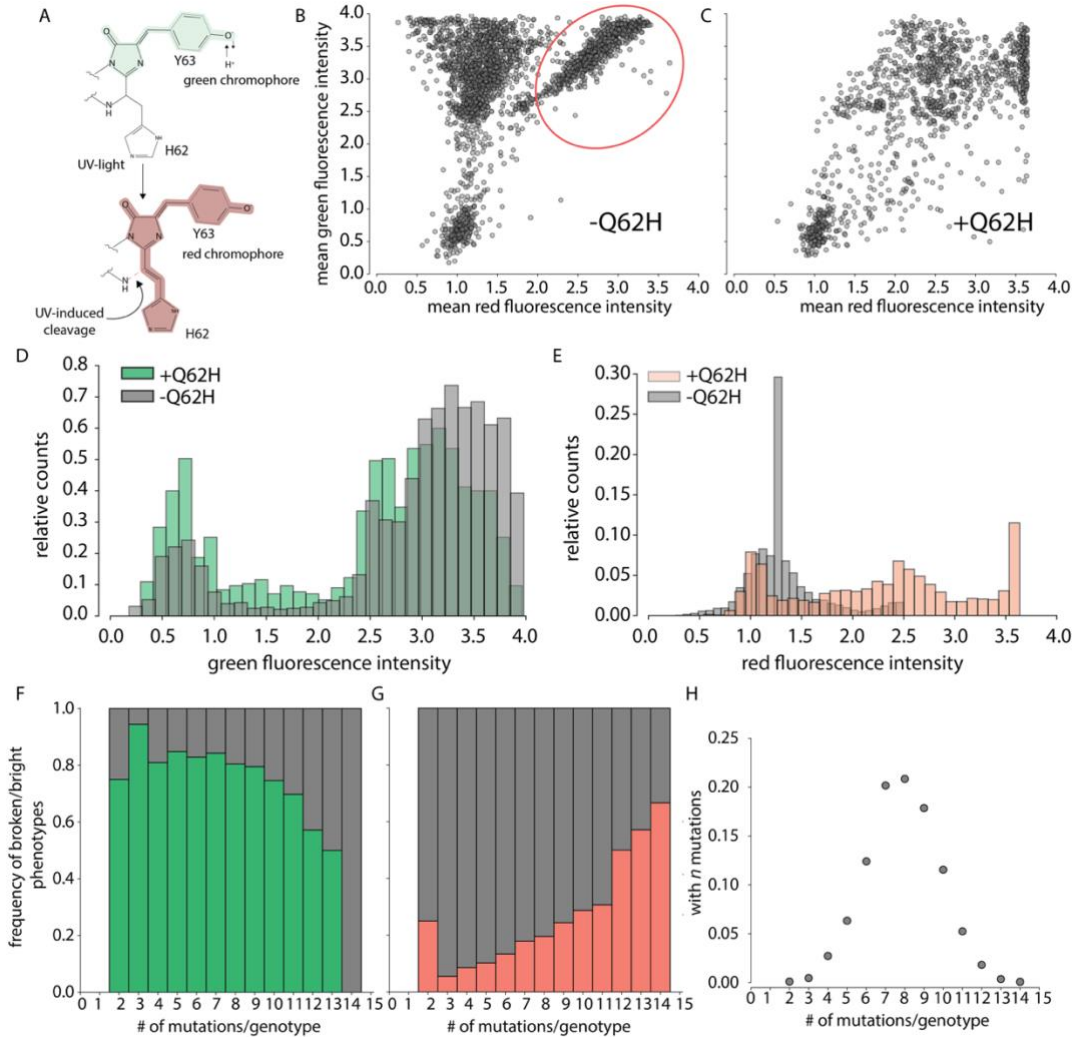
After applying our compensation correction, we find that ~1.5% are red (but not green), ~60.6% are green (but not red), ~18.6% are completely broken, and ~19.3% are above 2.0 in both channels. The observation of bright red fluorescence is relatively rare in our dataset, while the observation of greenness is very common (in total, ~79% of the observations).

Fig 4.5D-E shows the distribution of fluorescence intensities in the red and green channels for the +/-Q62H backgrounds after filtering to correct for bleed through. The -Q62H background (grey) is broadly beneficial for green fluorescence. The +Q62H background (green and red) reduces green fluorescence intensity and increases red fluorescence relative to the -Q62H background. We found a single genotype lacking the Q62H mutation (E26V/A60V/D74H/T104R/S104N/Y116N/M154T/V157I/R194C/Y217Del/M219G) that still exhibited high red fluorescence intensity after applying our filtering cutoffs. Upon examination of the frequency of observations per bin for this genotype, it's unclear if this is a real signal or if there is under sampling in the green channel, leading to

spuriously low inferred green fluorescence intensity (Appendix C Supplementary Fig C7).

We next looked at the frequency of broken versus bright chromophores as a function of the number of mutations present (Fig 4.5F-G). We find that as we increase the number of mutations, the frequency of genotypes with bright red fluorescence increases from <10% to ~60% while the frequency of genotypes with broken green chromophores increases from ~20% to ~50%. Fig 4.5H shows the relative number of genotypes observed with *n* mutations. Next, we examined how individual mutations impacted green and red fluorescence intensities in all backgrounds. We first isolated all the single-mutation maps available in our dataset. We found that there were 4,227 possible single-mutation maps.

Fig 4.5 Effects of mutations on green and red fluorescence intensities. (next page) A) UV-induced photoactivation mechanism for the Q62H containing red fluorescence GFP-like proteins (adapted from Lukyanov et al[234]). B) Mean green fluorescence intensity (y-axis) versus mean red fluorescence intensity (x-axis) for all genotypes lacking the critical Q62H chromophore mutation. The red circle shows genotypes that are both bright green and bright red even in the absence of the critical mutation, indicating that this population is arising from bleed through from the green to red channel. C) Mean green fluorescence intensity (y-axis) versus mean red fluorescence intensity (x-axis) for all genotypes containing the critical Q62H mutation. D) Distribution of green fluorescence intensities for the +Q62H (green) and -Q62H (grey) backgrounds with relative counts on the y-axis and intensity on the x-axis. E) Distribution of red fluorescence intensities for the +Q62H (salmon) and -Q62H (grey) backgrounds with relative counts on the y-axis and intensity on the x-axis. This is post-filtering for the population circled in red in panel B. F) Frequency of broken (grey) or bright (green) chromophores (y-axis) versus the number of mutations per genotype (x-axis). G) Frequency of broken (grey) or bright (salmon) chromophores (y-axis) versus the number of mutations per genotype (x-axis). H) Frequency of genotypes containing *n* mutations (y-axis) vs number of mutations (x-axis).


A
Y63
green chromophore
H62
UV-light
Y63
red chromophore
UV-induced cleavage
H62
B
mean green fluorescence intensity
mean red fluorescence intensity
-Q62H
C
mean red fluorescence intensity
+Q62H
D
+Q62H
-Q62H
relative counts
green fluorescence intensity
E
+Q62H
-Q62H
relative counts
red fluorescence intensity
F
frequency of broken/bright phenotypes
# of mutations/genotype
G
# of mutations/genotype
H
with n mutations
# of mutations/genotype

We calculated the effect of introducing a mutation at position x in each background it was observed in. Mutational effects were classified into three bins: neutral (effect is between zero $\pm$ the mean standard deviation of green/red intensities), beneficial (positive magnitude), or deleterious (negative magnitude). Fig 4.6A shows the fraction of backgrounds where the introduction of a given mutation was neutral, beneficial, or deleterious in for the green and red phenotypes. Supplementary Fig C8 shows the percentage of backgrounds each mutation was observed in (Appendix C).

As anticipated, we see that the Q62H mutation is the most beneficial for the observation of the red phenotype (Fig 4.6A). When averaged over all backgrounds, most of the other mutations are neutral or slightly more beneficial than deleterious. We find that a group of mutations (A60V, Q62H, S105N, Y116N, R194C, R216H, and Y217Del) are primarily deleterious in for green fluorescence. We find that the intermediate substitution at position 219 (M219V) is beneficial for green fluorescence on average, while the derived substitution (M219G) is mostly neutral. The Y116N mutation appears to be almost entirely deleterious for green fluorescence.

We were intrigued by the overall neutral nature of mutational effects on red fluorescence shown in Fig 4.6A. Previous work showed that there are extensive background-dependent effects in this set of mutations, especially with respect to Q62H[228].

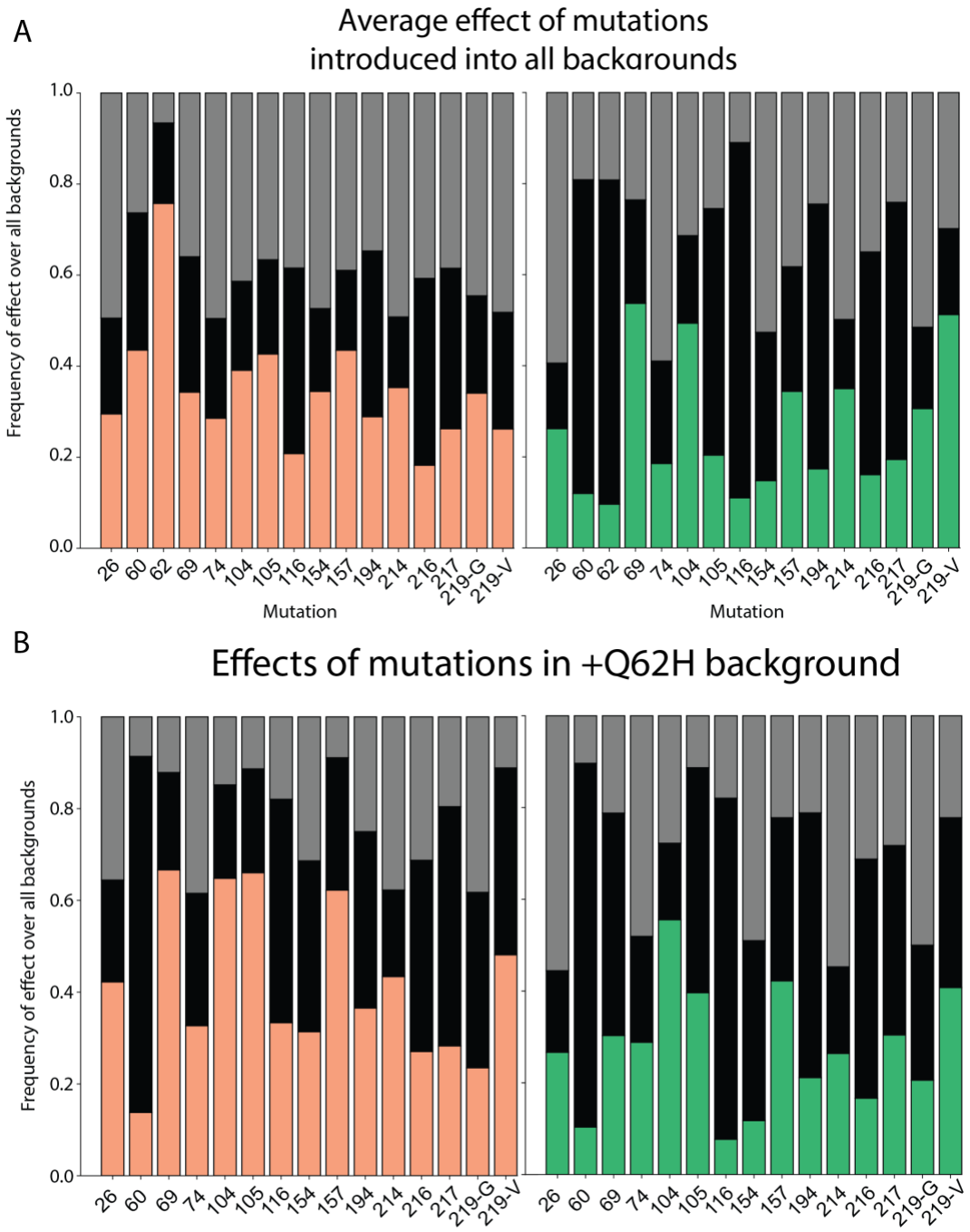


Fig 4.6 Effects of mutations are background dependent. A) Frequency of mutational effects that are neutral (grey), deleterious (black) or beneficial (shown in salmon for red fluorescence and in green for green fluorescence). Mutational effects are averaged over all backgrounds when a mutation, *x*, is introduced (x-axis). B) Frequency of mutational effects that are neutral (grey), deleterious (black), or beneficial (shown in salmon for red fluorescence and in green for green fluorescence) in the +Q62H background when a mutation, *x*, is introduced (x-axis).

We began investigating background-dependent effects by investigating the effect of introducing mutations into the +Q62H background (Fig 4.6B). We find that in the Q62H background, several mutations (namely, T69A, T104R, S105N, and V157I) become much more beneficial for red fluorescence than they appeared to be when averaged over their introduction into all backgrounds. Overall, mutational effects on green fluorescence are relatively invariant between Fig 4.6A and 4.6B.

Finally, we looked at the effects of each mutation on green and red fluorescence in different genetic backgrounds, as we did in Fig 4.7. We find that all derived mutations increase red fluorescence on average when the Q62H mutation is introduced (see Appendix C, Supplementary Figs C9-C22 for effects on red fluorescence per background). Y116N is, without exception, deleterious for green fluorescence when introduced in all backgrounds (see Appendix C, Supplementary Figs C9-C22 for effects on green fluorescence per background). Y116N is mostly neutral with respect to red fluorescence when introduced into all backgrounds. We find that the intermediate amino acid substitution at position 219 (M219V) is almost always more beneficial for green fluorescence than the derived substitution (M219G), while they are almost approximately identical in their effects on red fluorescence.

Epistatic interactions between mutations

Given the extensive background-dependence of single mutational effects shown in the previous section, as well as by others[98,228], we wanted to understand the nature of epistatic interactions between mutations in the library. We started by constructing all possible combinatorially complete two mutation (L=2, $2^2 = 4$ genotypes), three mutation

(L=3, 2^3 = 8 genotypes), and four mutation (L=4, 2^4 = 16 genotypes) binary mutant cycles using our dataset of 3,676 genotypes with quantitative phenotypes for green and red fluorescence. This resulted in 1,159 L=2 mutant cycles, 113 L=3 mutant cycles, and a single L=4 mutant cycle. We repeated this to calculate all maps containing the 3-site position, which resulted in 52 L=1 mutant cycles and 14 L=2 mutant cycles.

We began by calculating epistasis in the red and green phenotypes in all L=2 binary mutant cycles. We calculated mutational effects relative to the defined "wildtype" genotype, which is the genotype with the fewest derived substitutions. We defined pairwise epistasis as the difference between the phenotype of the double mutant and the additive prediction from each single mutational effect. We defined third-order epistasis as the difference between the triple mutant and the additive prediction from single mutational effects plus all pairwise interactions[132]. For comparisons between the two phenotypes, we converted epistasis values to z-scores, Z_{epi}, as follows:

$$Z_{epi} = \frac{(x_{epi} - \mu_{epi})}{\sigma_{epi}},$$

where x_{epi} is the green/red epistatic magnitude for a specific set of mutations, μ_{epi} is the mean epistatic magnitude for green/red fluorescence, and σ_{epi} is the standard deviation in epistatic magnitude for green/red fluorescence.

Fig 4.7A shows the distribution of Z_{epi} for all pairwise (salmon) and three-way (dark red) interactions for red fluorescence, while Fig 4.7B shows pairwise (light green) and three-way (dark green) interactions for green fluorescence. The distributions of magnitude for pairwise versus third-order interactions are overall quite similar for both

phenotypes, both in range and overall magnitude. The red fluorescence phenotype seems have a similar overall distribution to the green fluorescence phenotype but seems to have more pairs that give rise to strong epistatic interactions, while the green distribution seems to contain more non-epistatic pairs. The distributions of the pairwise epistatic Z-scores for each mutation are very similar to the distribution for all mutations (Fig 4.7C-D). We see a particularly large spread in Z_{epi} scores for specific mutations in for the red phenotype, especially A60V and Q62H (Fig 4.7D). Appendix C, Supplementary Figs C23-C24 show the distribution of Z_{epi} scores for green and red fluorescence for each genotype.

We first focused on pairwise epistasis. We found that the mean pairwise epistatic Z-score for green fluorescence was $3.9 \times 10^{-17} \pm 1.0$, with a maximum of 3.9 and a minimum of -3.4. We found that the mean pairwise epistatic magnitude for red fluorescence was $-2.5 \times 10^{-17} \pm 0.9$, with a maximum of 5.3 and a minimum of -6.9.

The most strongly positive epistatic pair for green fluorescence was T104R + Y116N while the most strongly negative was M154T + R194C. Both pairs exhibit reciprocal sign epistasis. For the positively epistatic pairs, each mutation abolishes green fluorescence individually, but they restore green fluorescence when introduced together. For the negatively epistatic pair, each mutation restores fluorescence from a non- fluorescent wildtype genotype, yet they destroy fluorescence when introduced together. Y116N is proposed to enlarge a solvent-bearing cavity near the chromophore, which we know from Fig 4.6A is broadly deleterious for green fluorescence, while T104R is proposed to stabilize the chromophore backbone[232].

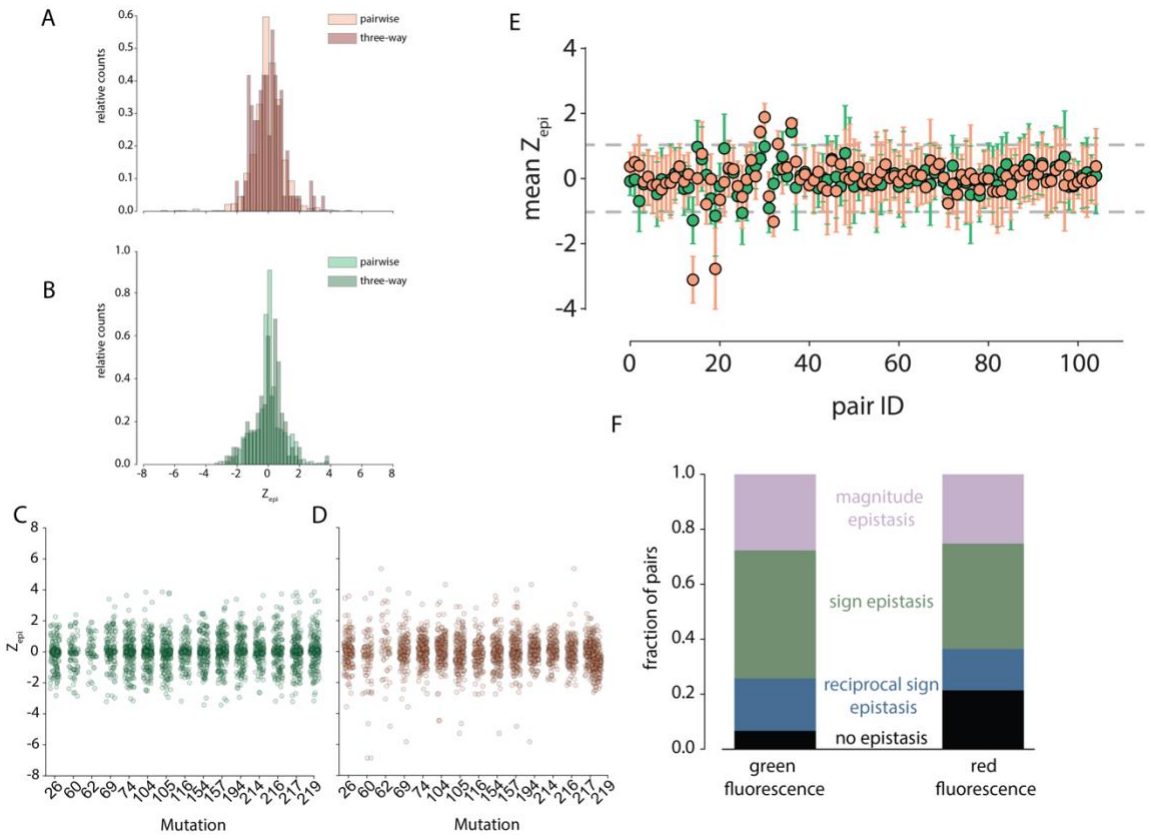


Fig 4.7 Epistasis is common in the GFP-like protein library. A) Distribution of Z_{epi} in red fluorescence for pairwise (light salmon) and three-way (dark red) interactions. B) Distribution of the Z_{epi} in green fluorescence for pairwise (light green) and three-way (dark green) interactions. C-D) Pairwise Z_{epi} (y-axis) as a function of mutation (x-axis) for C) green and D) red fluorescence. E) Pairwise Z_{epi} (y-axis) for each unique pair of mutations (x-axis) averaged over all backgrounds. Green circles represent the average epistasis Z-score in green fluorescence, while red circles represent the average epistasis Z-score in red fluorescence. Green/red error bars represent the standard deviation. F) Proportion of pairs (y-axis) giving rise to magnitude (lilac), sign (green), reciprocal sign (blue), and no epistasis (black) for the green and red phenotypes (x-axis). Pairs with epistatic magnitudes above -1 or below 1 were considered non-epistatic.

These two mutations may jointly increase green fluorescence when introduced together by compensating for each individually destabilizing effect. M154T is proposed to stabilize the overall fold of the protein while R194C is known to increase the

flexibility of the chromophore region[232]. It is not clear from structural considerations why these would produce such a strong, negative epistatic interaction.

The most strongly positive epistatic pair for red fluorescence was Q62H + Y217Del, while the most strongly negative was Q62H + A60V. The identities of the maximally epistatic red pairs are unsurprising. The Y217Del mutation has been shown to be critical for promoting red fluorescence via promoting flexibility in the chromophore residues[232]. The A60V mutation is located directly adjacent to the chromophore and its effect is known to be highly dependent on genetic background[228]. In fact, Fig 4.6B shows that it is often deleterious in the Q62H background.

Next, we wanted to see if there were pairs of mutations that were highly epistatic when averaged over all genetic backgrounds. Fig 4.7E shows the Z-score values of all 105 unique pairs of mutations in our dataset, averaged over all backgrounds for green and red fluorescence. Most pairs are, on average, very close to zero or are non-epistatic when averaged over all backgrounds. We do see several genotypes that seem to be highly epistatic, especially for epistasis in red (below an epistatic magnitude of -1 and above a magnitude of +1). The Q62H mutation ubiquitously present amongst all positively interacting pairs for both red and green fluorescence (red: Q62H/T104R, Q62H/S105N, Q62H/V157I, Q62H/R216H and green: Q62H/R216H). For negatively interacting pairs, we frequently observe the A60V mutation for both phenotypes (red: A60V/Q62H, A60V/Y116N, Q62H/M154T and green: A60V/Q62H, A60V/Y116N, AND A60V/Y217Del). A60V appears to be a highly negatively epistatic substitution across all

backgrounds for both green and red fluorescence. Q62H seems to be more epistatic in the context of red fluorescence.

Finally, we looked at the frequency of pairs that lead to evolutionarily important classes of epistasis: magnitude, sign, and reciprocal sign. In magnitude epistasis, only the size of an effect changes when two mutations are combined. In sign epistasis, the sign of one of the mutations reverses (i.e., if beneficial alone, it is deleterious in combination with another mutation). In reciprocal sign epistasis, both mutations change sign. For both phenotypes, we find that the overall proportion of mutations giving rise to sign, reciprocal, and magnitude epistasis are very similar, with the exception that the red phenotype tends to be less epistatic (Fig 4.7F). Intriguingly, both sign and reciprocal sign epistasis, which are particularly important types of epistasis, are commonly observed amongst mutation pairs.

Conclusions and future directions

Here, we combined theory and high-throughput experiments to design and characterize a genotype-phenotype map for a natural evolutionary transition that is 96 times larger than the current largest map[7]. Our current dataset is a sparse subsampling of the full map, which has allowed us to assess the global characteristics of the map while estimating the read depth needed the full experiment. We inferred green and red fluorescence intensities from sort-seq data for ~7% of the library. Most of the library appears to be green (~79%). The bulk of the remaining library is completely non-fluorescent (~19.5%), while only ~1.5% is red. We find that mutational effects are highly background dependent for both phenotypes, especially with respect to the presence of the

Q62H mutation in the red phenotype. For low-dimensional submaps (i.e., L=2) we find that epistasis is pervasive and tends to have stronger negative effects on the red phenotype. Together, our data suggests that the subsampled genotype-phenotype map may be rugged and constrained by epistatic interactions, much like other small genotype-phenotype maps. More work must be done with the full dataset, however, to compare the relative effects of epistasis and neutral networks on trajectories as a function of map size.

Mutational effects are background dependent

Previous work has shown that the substitutions in this evolutionary trajectory are highly context-dependent, at least in a subset of possible genetic backgrounds[53,98,232]. Thus, we looked at effect of introducing each mutation into all possible backgrounds in the sparsely sampled map. We found that specific mutations were strongly beneficial or deleterious over all backgrounds, for instance Q62H was beneficial for red fluorescence (>75% of genotypes) and Y116N was universally deleterious for green fluorescence (~80% of genotypes) (Fig 4.6A).

The Y116N mutation exhibited the most consistently negative impact on green fluorescence, regardless of background, out of any mutation (Fig 4.6A). Y116N is particularly important for the red fluorescent phenotype[53,232]. It has been shown to increase the volume of a cavity near the chromophore that's important for solvation and increasing flexibility near the chromophore[232]. It will be interesting to see how the Y116N mutation shapes accessibility and if it is an important determinant of evolutionary trajectories in the full map.

The strong beneficial association of the Q62H mutation with red fluorescence is unsurprising, as it is a fundamental requirement for the chemistry of the red chromophore (Fig 4.5A)[53,98,232,234]. We find that there is asymmetry in how Q62H impacts red fluorescence in the presence of other mutations. For example, Fig 4.6B shows that the introduction of the A60V mutation into Q62H-containing backgrounds is most often deleterious for red fluorescence. However, if we look at the effect of introducing Q62H into the A60V background, we see that it is beneficial for red fluorescence (Appendix C, Supplementary Fig C10). The A60V mutation has been shown to stabilize and rigidify the core of the protein, exhibiting a negative impact on red fluorescence earlier in the evolutionary trajectory but becoming beneficial later on[53,232]. We anticipate that the Q62H substitution is the strongest determinant of accessibility from green to red in the full map, though that is still to be shown.

Epistasis is prevalent in small submaps

The strong context-dependent effects of mutations discussed in the previous section (and shown in Fig 4.6) can be extremely important for shaping evolutionary trajectories, as they indicate that there must be a specific order in which mutations can accumulate[5,11,19,20,23,24]. Such epistatic interactions strongly determine the local topology of the genotype-phenotype map[9,20,23,24]. We analyzed all possible two- and three-mutation submaps in our dataset to look at the extent of epistasis in small submaps. We found that such low-dimensional submaps are highly epistatic. This epistasis falls into classes of all types, namely, sign, reciprocal sign, and magnitude epistasis. Sign and reciprocal sign are particularly extreme forms of epistasis that have been proposed to primarily decrease

accessible evolutionary trajectories[9,16,20]. In fact, reciprocal sign epistasis is a topological requirement for multiple peaks in fitness landscapes[20].

Our epistasis analysis revealed specific positions that were frequently epistatic: A60V and Q62H. At least one of these substitutions was present in all strongly epistatic pairs of mutations, with the strongest epistatic pair being A60V and Q62H. A60V was almost exclusively associated with negative epistatic interactions, which is consistent with our data shown in Fig 4.6B. The biophysical effect of A60V is to optimize packing of an alpha helix N-terminal to the chromophore[232]. Flexibility in the chromophore region is has been shown to be a key biophysical factor in the evolution of the red chromophore––the excess stabilization introduced by A60V may likely reduces this flexibility in the absence of other key substitutions[53,232].

Unsurprisingly, Q62H was consistently associated with positive epistatic interactions in the red phenotype. The strongest positive interactions occurred with the T104R, S105N, and V157I substitutions. The positive association of Q62H with these substitutions is unsurprising: T104R and S105N have been shown to work as a network of stabilizing residues near the chromophore, intriguingly, in conjunction with A60V. In fact, A60V becomes more beneficial for red fluorescence in the presence of either of these substitutions[232] (Appendix C, Supplementary Figs C13-C14). V157I has been proposed to prevent chromophore isomerization reactions that result in non-productive red chromophores[232]

Future directions

Our current work shows that epistasis is a core feature of low-dimensional submaps. Future work involves analyzing the full genotype-phenotype map for this evolutionary transition, which is currently awaiting sequencing at the GC3F facilities. The major driving force behind this project is to understand if the conclusions made about evolution in low-dimensional maps are informative for high-dimensional maps, i.e., is epistasis the primary determining factor of accessibility? Are there biochemical features that lead to extensive neutral networks in sequence space as we increase dimensionality?

We plan to do two primary analyses to address these central questions. First, we plan to calculate trajectories through the map as a function of map size. Our simple theoretical simulations in Fig 4.1C indicate that we may see differences in the probability of traversing the genotype-phenotype map as we increase the dimensionality of the map. Our simple model is highly epistatic, as we assigned phenotypes completely at random to each simulated genotype. In the real map, we expect to see correlations between the phenotypes of neighboring genotypes[84], i.e., if a network of genotypes has the Q62H mutation they are likely to be red. We expect this to facilitate the probability of traversing the map by creating extensive neutral networks that grow quickly as a consequence of the relationship between the number of possible genotypes (2^L) and dimensionality (L)[80,83,84,86,223]. We're also interested in the role of the three-site mutation (M219V/G) in creating indirect paths, connecting neutral networks that may otherwise be disconnected if we only considered binary mutations[84,86,222,223]. It will be interesting to see if the importance of indirect paths scales with dimensionality for a natural evolutionary transition in function.

Second, we will look at the effect of epistasis on evolutionary trajectories. We will do this in two ways. We will look at the impact of re-calculating evolutionary trajectories through maps where epistasis has been artificially removed from the dataset and ask if specific epistatic interactions are important for the evolution of red fluorescence. We also will look at this as a function of map size. Our current dataset suggests that epistasis makes the map between green and red fluorescence extremely rugged. Such landscapes can be multi-peaked, causing evolutionary trajectories to end at genotypes that are local, but not global, maxima, and can result in severely decreased accessibility throughout the entire map[9,19,20]. However, there are potentially two major differences between low- and high-dimensional maps with respect to epistasis: 1) high-dimensional spaces are expected to exhibit extensive sets of connected genotypes, or neutral networks, which may offset the decrease in accessibility arising from specific epistatic interactions and 2) high-dimensional spaces can contain high-order epistasis, or interactions between three or more mutations[10,15,23,84,86]. It is not clear if such high-order interactions matter at all in high-dimensional maps, or if they create long-term memory in protein evolution.

Bridge to Chapter V

Chapter IV highlights our ongoing efforts to characterize the genotype-phenotype map for an evolutionary transition in coral fluorescence color. We used simulations to show that—even for a very simple toy system—we expect large genotype-phenotype maps to exhibit distinct properties from small genotype-phenotype maps. We used these theoretical expectations to design a library for a natural evolutionary transition in

fluorescence color that occurred over the course of 15 substitutions, resulting in a library that is 96 times larger than the current largest genotype-phenotype map. We developed high-throughput tools to measure the genotype-phenotype map for this transition.

We show in a sparsely sampled subset of the full genotype-phenotype map that mutational effects are highly background dependent. We also show that epistasis is rampant in low-dimensional 2-site submaps. Such extensive epistasis suggests that the genotype-phenotype map is quite rugged, which is consistent with observations made in studies of small genotype-phenotype maps. We conclude Chapter IV by detailing future work that will be done to test our hypotheses regarding the differences in the contributions of neutral networks and epistasis to accessibility in small versus large genotype-phenotype maps. Chapter V wraps up this dissertation by highlighting how powerful pairing theory with computational and experimental work can be in addressing central questions in evolutionary biology and biochemistry. We also provide insights into future work for studying the evolutionary importance of ensemble epistasis and how the properties of genotype-phenotype maps scale with size.

CHAPTER V

CONCLUSIONS AND FUTURE DIRECTIONS

This dissertation illustrates the power of combining theory and experiment to address central questions in biology how protein sequence maps to function. It demonstrates that simple theoretical and computational models can lead to testable hypotheses and guide the design of careful biochemical experiments to answer questions that have been traditionally difficult to answer in evolutionary biology and biochemistry. We used this approach to 1) reveal a biochemical mechanism of epistasis in proteins and 2) design an experiment to understand how protein biochemistry and epistasis work together to shape the long-term evolution of new functions. This work will help us develop better models of protein evolution, aiding protein engineers and preventing the evolution of pathogens and resistances, and add to our fundamental knowledge of how proteins and biology work.

In Chapter II, we lay the theoretical groundwork and illustrate that the thermodynamic conditions under which we expect ensemble epistasis to arise is likely to be quite common in biology. Our work showed that ensemble epistasis is maximized where many conformations are populated. We found that we could tune environmental parameters to toggle between conditions where many and few conformations were populated, leading us to find a key experimental signature of ensemble epistasis: environment-dependent epistasis. We then took these theoretical predictions into reality by testing for effector-dependent epistasis in the lac repressor protein. We found that such

effector-dependent patterns of epistasis were pervasive in the lac repressor protein both *in vivo* and *in vitro*. From our theoretical, computational, and experimental studies we conclude that ensemble epistasis is likely prevalent source of epistasis in biology.

Moving forward, we could improve our understanding of epistasis by building models that can distinguish between distinct sources of epistasis, for example, distinguishing direct contact epistasis from ensemble epistasis. To accomplish this, one might scale up the approach in Chapter III by combining thermodynamic modeling, molecular dynamics simulations, and high-throughput experiments to decompose the effects of mutations on biophysical protein properties.

Chapter IV zooms out from trying to understand specific types of epistasis to ask if general conclusions made from studies of short-term evolutionary transitions (i.e., the evolution of antibiotic resistance) scale to longer-term evolutionary transitions in function. We used simple theoretical simulations to show that evolution in small genotype-phenotype maps may be under fundamentally different constraints than large genotype-phenotype maps. For example, as the number of substitutions (L) in an evolutionary transition increase, the number of genotypes (2L) and possible trajectories (L!) grows extremely rapidly. As a consequence of this scaling relationship, we predict that vast neutral networks might arise. Such networks might overwhelm the effects of epistasis, which has traditionally been viewed as a key constraining factor in protein evolution in small genotype-phenotype maps.

We used these simple theoretical predictions to inform where we could expect to observe such features of high-dimensional maps. We constructed and designed experimental methods to exhaustively characterize the genotype-phenotype map for a

natural evolutionary transition in fluorescence color in GFP-like proteins from corals. This transition occurred over the course of 15 substitutions, resulting in a map that is 96 times larger than any previously characterized binary map. Our current dataset is a low-dimensional sparse subsampling of the full map. We find that its features are consistent with other low-dimensional maps in that it is extensively epistatic. In future work we will compare the effects of epistasis and map size on evolutionary trajectories to address our current gap in knowledge about evolution in high-dimensional genotype-phenotype maps.

APPENDIX A

SUPPLEMENTARY MATERIAL FOR CHAPTER II

This section includes the supplementary material and supplementary figures referenced in chapter II. Other supplementary files corresponding to analyses and ROSETTA input files, can be found on Github at: https://github.com/harmslab/ensemble_epistasis.

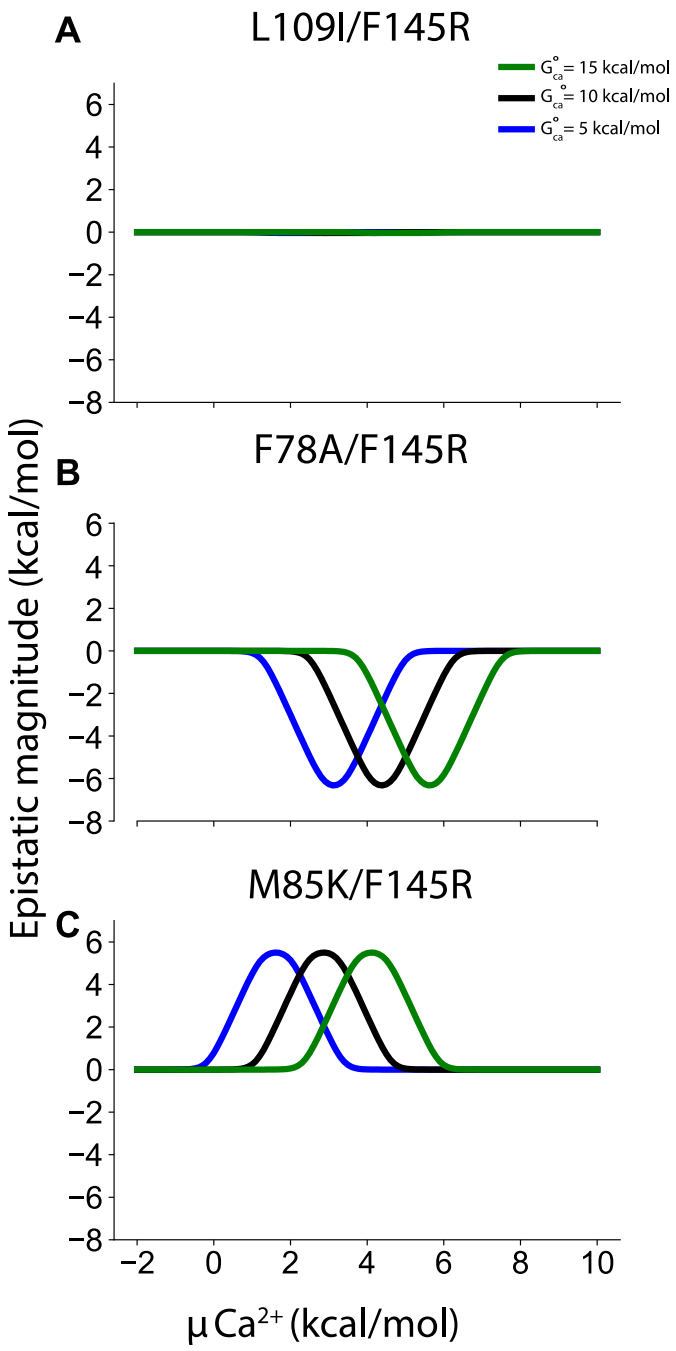


Fig A1: Changing the value of G°_{ca} changes the $\mu_{Ca^{2+}}$ value at which ensemble epistasis is maximized. Each panel shows the magnitude of epistasis on the y-axis and $\mu_{Ca^{2+}}$ (in $kcal \cdot mol^{-1}$)on the x-axis. Each colored curve represents ensemble epistasis observed with a different G°_{ca} value: 5 $kcal \cdot mol^{-1}$) (blue), 10 $kcal \cdot mol^{-1}$) (black), and 15 $kcal \cdot mol^{-1}$) (green). A) $\mu_{Ca^{2+}}$ −dependent epistasis for the L109I/F145R mutation pair, B) the F78A/F145R mutation pair, and C) M85K/F145R mutation pair.

Supplementary Section 1 Necessary conditions for ensemble epistasis

1.1 Ensemble epistasis appears between two mutations in a three-conformation ensemble

We define epistasis between mutations a→A and b→B in ΔG_{obs} as the difference in the effect of a→A in the *ab* and *aB* backgrounds (Fig 2.1A):

$$\varepsilon = -\left(\Delta G_{obs}^{AB} - \Delta G_{obs}^{aB}\right) - \left(\Delta G_{obs}^{Ab} - \Delta G_{obs}^{ab}\right) \quad (1)$$

$\Delta G_{obs}^{genotype}$ is given by Equation 10:

$$\Delta G_{obs}^{genotype} = G_i^{genotype} - \langle G_{j,k}^{genotype} \rangle$$

where

$$\langle G_{j,k}^{genotype} \rangle = -RTln\left(e^{-\frac{G_j^{genotype}}{RT}} + e^{-\frac{G_k^{genotype}}{RT}}\right).$$

We model mutations as having additive effects within each conformation *i*, *j*, or *k*.

ΔG_{obs} for each genotype is shown below (reproducing Table 1 in the main text):

Table A1 Map between genotype and the thermodynamic description of $\Delta G_{obs}^{genotype}$

Genotype	$\Delta G_{obs}^{genotype}$	$\langle G_{j,k}^{genotype} \rangle$
ab	$G_i^{ab} - \langle G_{j,k}^{ab} \rangle$	$-RTln\left(e^{-\frac{(G_j^{ab})}{RT}} + e^{-\frac{(G_k^{ab})}{RT}}\right)$
Ab	$(G_i^{ab} + \delta G_i^{a\to A}) - \langle G_{j,k}^{Ab} \rangle$	$-RTln\left(e^{-\frac{(G_j^{ab}+\delta G_j^{a\to A})}{RT}} + e^{-\frac{(G_k^{ab}+\delta G_k^{a\to A})}{RT}}\right)$
aB	$(G_i^{ab} + \delta G_i^{b\to B}) - \langle G_{j,k}^{aB} \rangle$	$-RTln\left(e^{-\frac{(G_j^{ab}+\delta G_j^{b\to B})}{RT}} + e^{-\frac{(G_k^{ab}+\delta G_k^{b\to B})}{RT}}\right)$
AB	$(G_i^{ab} + \delta G_i^{a\to A} + \delta G_i^{b\to B}) - \langle G_{j,k}^{AB} \rangle$	$-RTln\left(e^{-\frac{(G_j^{ab}+\delta G_j^{a\to A}+\delta G_j^{b\to B})}{RT}} + e^{-\frac{(G_k^{ab}+\delta G_k^{a\to A}+\delta G_k^{b\to B})}{RT}}\right)$

If we substitute the relevant expressions for $\Delta G_{obs}^{genotype}$ into our expression for ε, we get:

$$\varepsilon = \left(\left[(G_i^{ab} + \delta G_i^{a\to A} + \delta G_i^{b\to B}) - \langle G_{j,k}^{AB} \rangle\right] - \left[(G_i^{ab} + \delta G_i^{b\to B}) - \langle G_{j,k}^{aB} \rangle\right]\right)$$

$$- \left(\left[(G_i^{ab} + \delta G_i^{a\to A}) - \langle G_{j,k}^{Ab} \rangle\right] - \left[G_i^{ab} - \langle G_{j,k}^{ab} \rangle\right]\right)$$

$$\varepsilon = G_i^{ab} + \delta G_i^{a\to A} + \delta G_i^{b\to B} - \langle G_{j,k}^{AB}\rangle - G_i^{ab} - \delta G_i^{b\to B} + \langle G_{j,k}^{aB}\rangle - G_i^{ab} - \delta G_i^{a\to A}$$

$$+ \langle G_{j,k}^{Ab}\rangle + G_i^{ab} - \langle G_{j,k}^{ab}\rangle$$

$$\varepsilon = -\langle G_{j,k}^{AB}\rangle + \langle G_{j,k}^{aB}\rangle + \langle G_{j,k}^{Ab}\rangle - \langle G_{j,k}^{ab}\rangle$$

$$\varepsilon = -\left[\left(\langle G_{j,k}^{AB}\rangle - \langle G_{j,k}^{aB}\rangle\right) - \left(\langle G_{j,k}^{Ab}\rangle - \langle G_{j,k}^{ab}\rangle\right)\right]$$

This cannot be simplified further, implying that ε may be non-zero.

1.2 To see ensemble epistasis, it is necessary to have three or more conformations

We can next consider the two-conformation case, where *k* is not populated. In this case:

$$\Delta G_{obs}^{genotype} = G_i^{genotype} - \langle G_j^{genotype}\rangle$$

simplifies to:

$$\langle G_j^{genotype}\rangle = -RTln\left(e^{-\frac{G_j^{genotype}}{RT}}\right) = G_j^{genotype}.$$

As in Section 1.1, we can write a table showing ΔG_{obs} for each genotype:

Table A2 Map between genotype and the thermodynamic description of $\Delta G_{obs}^{genotype}$ for a two-conformation ensemble

Genotype	$\Delta G_{obs}^{genotype}$
ab	$G_i^{ab} - G_j^{ab}$
Ab	$\left(G_i^{ab} + \delta G_i^{a\to A}\right) - \left(G_j^{ab} + \delta G_j^{a\to A}\right)$
aB	$\left(G_i^{ab} + \delta G_i^{b\to B}\right) - \left(G_j^{ab} + \delta G_j^{b\to B}\right)$
AB	$\left(G_i^{ab} + \delta G_i^{a\to A} + \delta G_i^{b\to B}\right) - \left(G_j^{ab} + \delta G_j^{a\to A} + \delta G_j^{b\to B}\right)$

If we substitute the relevant expressions for $\Delta G_{obs}^{genotype}$ into our expression for ε we get

$$\varepsilon = \left(\left[\left(G_i^{ab} + \delta G_i^{a\to A} + \delta G_i^{b\to B}\right) - \left(G_j^{ab} + \delta G_j^{a\to A} + \delta G_j^{b\to B}\right)\right]\right.$$
$$\left. - \left[\left(G_i^{ab} + \delta G_i^{b\to B}\right) - \left(G_j^{ab} + \delta G_j^{b\to B}\right)\right]\right)$$
$$- \left(\left[\left(G_i^{ab} + \delta G_i^{a\to A}\right) - \left(G_j^{ab} + \delta G_j^{a\to A}\right)\right] - \left[G_i^{ab} - G_j^{aA}\right]\right)$$
$$\varepsilon = G_i^{ab} + \delta G_i^{a\to A} + \delta G_i^{b\to B} - G_j^{ab} - \delta G_j^{a\to A} - \delta G_j^{b\to B} - G_i^{ab} - \delta G_i^{b\to B} + G_j^{ab}$$
$$+ \delta G_j^{b\to B} - G_i^{ab} - \delta G_i^{a\to A} + G_j^{ab} + \delta G_j^{a\to A} + G_i^{ab} - G_j^{aA}$$
$$\varepsilon = 0$$

All terms cancel, demonstrating it is necessary to have at least three conformations to observe ensemble epistasis.

1.3 To see ensemble epistasis, it is necessary for mutations a→A and b→B to have different effects on conformations *j* and *k*.

To test the necessity of mutations having differential effects on conformations *j* and *k*, we set $\delta G_j^{b\to B}$=$\delta G_k^{b\to B} = \delta G_{j,k}^{b\to B}$. This means mutation b→B has the same effect on conformations *j* and *k*. In contrast, we left $\delta G_j^{a\to A} \neq \delta G_k^{a\to A}$, meaning a→A has different effects on conformations *j* and *k*. Because b→B has identical effects and a→A has differential effects, this analysis tests whether it is necessary for both mutations to have differential effects to observe ensemble epistasis.

Consider the expression for $\langle G_{j,k}^{aB} \rangle$:

$$\langle G_{j,k}^{aB} \rangle = -RTln\left(e^{-\frac{\left(G_j^{ab}+\delta G_{jk}^{b\to B}\right)}{RT}} + e^{-\frac{\left(G_k^{ab}+\delta G_{jk}^{b\to B}\right)}{RT}}\right)$$

Because $\delta G_{jk}^{b\to B}$ is shared among terms, we can factor it out:

$$\langle G_{j,k}^{aB} \rangle = -RTln\left(e^{-\frac{\left(G_j^{ab}\right)}{RT}} e^{-\frac{\left(\delta G_{jk}^{b\to B}\right)}{RT}} + e^{-\frac{\left(G_k^{ab}\right)}{RT}} e^{-\frac{\left(\delta G_{jk}^{b\to B}\right)}{RT}}\right)$$

$$\langle G_{j,k}^{aB} \rangle = -RTln\left(e^{-\frac{\left(\delta G_{jk}^{b\to B}\right)}{RT}} \left(e^{-\frac{\left(G_j^{ab}\right)}{RT}} + e^{-\frac{\left(G_k^{ab}\right)}{RT}} \right) \right)$$

$$\langle G_{j,k}^{aB} \rangle = -RTln\left(e^{-\frac{\left(\delta G_{jk}^{b\to B}\right)}{RT}} \right) - RTln\left(e^{-\frac{\left(G_j^{ab}\right)}{RT}} + e^{-\frac{\left(G_k^{ab}\right)}{RT}} \right)$$

$$\langle G_{j,k}^{aB} \rangle = \delta G_{jk}^{b\to B} - RTln\left(e^{-\frac{\left(G_j^{ab}\right)}{RT}} + e^{-\frac{\left(G_k^{ab}\right)}{RT}} \right)$$

$$\langle G_{j,k}^{aB} \rangle = \delta G_{jk}^{b\to B} + \langle G_{j,k}^{ab} \rangle.$$

Using the same reasoning, we can factor $\delta G_{jk}^{b\to B}$ out of the expression for $\langle G_{j,k}^{AB} \rangle$:

$$\langle G_{j,k}^{AB} \rangle = \delta G_{jk}^{b\to B} + \langle G_{j,k}^{Ab} \rangle.$$

We can then substitute these simplified expressions for $\langle G_{j,k}^{aB} \rangle$ and $\langle G_{j,k}^{AB} \rangle$ into the expression for ε:

$$\varepsilon = \left[\left(\langle G_{j,k}^{aB} \rangle - \langle G_{j,k}^{AB} \rangle \right) - \left(\langle G_{j,k}^{ab} \rangle - \langle G_{j,k}^{Ab} \rangle \right) \right]$$

$$\varepsilon = \left(\left[\delta G_{jk}^{b\to B} + \langle G_{j,k}^{ab} \rangle \right] - \left[G_{jk}^{b\to B} + \langle G_{j,k}^{Ab} \rangle \right] \right) - \left(\langle G_{j,k}^{ab} \rangle - \langle G_{j,k}^{Ab} \rangle \right)$$

$$\varepsilon = \delta G_{jk}^{b\to B} + \langle G_{j,k}^{ab} \rangle - G_{jk}^{b\to B} - \langle G_{j,k}^{Ab} \rangle - \langle G_{j,k}^{ab} \rangle + \langle G_{j,k}^{Ab} \rangle$$

$$\varepsilon = 0.$$

All terms cancel, demonstrating that it is necessary for both a→A and b→B to have differential effects on conformations *j* and *k* to observe ensemble epistasis.

2 Modeling the calcium-dependence of ensemble populations for S100A4.

2.1 Deriving the model

S100A4 populates both a closed conformation (*M*) and an open conformation (M^*), differentiated by exposure of a hydrophobic cleft by rotation of two helices. In the absence of Ca^{2+}, *M* is favored over M^*. Ca^{2+} binds cooperatively to four sites in the M^*

conformation [111]. The $M^* \cdot (Ca^{2+})$ and *M* species correspond to the "*ca*" and "*apo*" species from the main text. Finally, peptide binds preferentially to the M^* conformation. To model the system, we make the following assumptions:

- *M* is strongly favored over M^* in the absence of Ca^{2+}.
- Ca^{2+} binds cooperatively at four equivalent sites on M^*.
- Ca^{2+} binds cooperatively at four equivalent sites on M^* than *M*, allowing us to neglect the $M \cdot (Ca^{2+})$.
- $Peptide$ binds much more tightly to M^* than *M*, allowing us to neglect any $M \cdot peptide$ conformations.

With these assumptions, we can describe the system with the following scheme and equilibrium constants:

$$M + 4Ca^{2+} \leftrightarrow M^* + 4Ca^{2+} \leftrightarrow M^* \cdot (Ca^{2+})_4$$

$$K_* = \frac{[M^*]}{[M]}$$

$$K_C = \frac{\left[M^* \cdot (Ca^{2+})_4\right]}{[M^*][Ca^{2+}]^4}.$$

The stability of $M^* \cdot (Ca^{2+})_4$ relative to the other protein conformations is given by:

$$G = -RTln\left(\frac{\left[M^* \cdot (Ca^{2+})_4\right]}{[M]+[M^*]}\right).$$

Substitute the equilibrium constants and simplify:

$$G = -RTln\left(\frac{[M^*]K_C[Ca^{2+}]^4}{[M]+[M^*]}\right),$$

$$G = -RTln\left(\frac{K_*[M^*]K_C[Ca^{2+}]^4}{[M]+K_*[M^*]}\right),$$

$$G = -RTln\left(\frac{K_*K_C[Ca^{2+}]^4}{1+K_*}\right).$$

Assume that $K_* \ll 1$, meaning that *M* is highly favored over M^* in the absence of Ca^{2+}:

$$G \approx -RTln\left(\frac{K_* K_C [Ca^{2+}]^4}{1}\right) = -RTln(K_* K_C [Ca^{2+}]^4)$$

$$G = -RTln(K_*) - RTln(K_C) - RTln([Ca^{2+}]^4)$$

$$G = -RTln(K_*) - RTln(K_C) - 4RTln\big([Ca^{2+}]\ \ \big).$$

Setting $\mu_{Ca^{2+}} = RTln([Ca^{2+}])$:

$$G = G_* + G_C - 4\mu_{Ca^{2+}}.$$

where G_* is the stability of M^* relative to *M* in the absence of Ca^{2+}. G_C describes the affinity of the M^* conformation for Ca^{2+}. The terms G_* and G_C, together, describe the intrinsic stability of the active, metal-bound "*ca*" complex at a reference $[Ca^{2+}]$. We therefore define a new constant:

$$G^\circ_{ca} \equiv G_* + G_C$$

The final expression for $G_{ca}(\mu_{Ca^{2+}})$ is:

$$G_{ca}(\mu_{Ca^{2+}}) = G^\circ_{ca} - 4\mu_{Ca^{2+}}.$$

The microscopic free energy of the *apo* (*M*) conformation does not depend on the concentration of Ca^{2+}; therefore, G_{apo} is a constant:

$$G_{apo}(\mu_{Ca^{2+}}) = G^\circ_{apo}.$$

2.2 Setting arbitrary offset

We do not know G°_{ca} or G°_{apo}. We do know, however, that at a low calcium concentration $G^\circ_{apo}(\mu_{Ca^{2+}}) \ll G^\circ_{ca}(\mu_{Ca^{2+}})$ (meaning, the *M* form is favored over M^* at low calcium). We also know that $G^\circ_{ca}(\mu_{Ca^{2+}})$ will increase linearly relative to

$G^{\circ}_{apo}(\mu_{Ca^{2+}})$ as a function of $\mu_{Ca^{2+}}$. If we do not care about the absolute value of $[Ca^{2+}]$ at which the system transitions between favoring $*apo* and *pep*, we can choose arbitrary values for G°_{ca} and G°_{apo} and then still calculate how epistasis should change as a function of $\mu_{Ca^{2+}}$for the protein. For convenience, we set $G^{\circ}_{apo} = 0$ and $G^{\circ}_{ca} = 10$ at $\mu_{Ca^{2+}} = 0$. We tested the sensitivity of our results to our choice of G°_{apo} (Fig S1).

2.3 Modeling mutant cycles

ab genotype

$$G^{ab}_{ca}(\mu_{Ca^{2+}}) = G^{\circ}_{ca} - 4\mu_{Ca^{2+}}$$

$$G^{ab}_{apo} = G^{\circ}_{apo}$$

$$\langle G^{ab}_{ca,apo}\rangle(\mu_{Ca^{2+}}) = -RTln\left(e^{-\frac{(G^{\circ}_{ca}-4\mu_{Ca^{2+}})}{RT}} + e^{-\frac{(G^{\circ}_{apo})}{RT}}\right)$$

Ab genotype:

$$G^{Ab}_{ca}(\mu_{Ca^{2+}}) = G^{\circ}_{ca} - 4\mu_{Ca^{2+}} + \delta G^{a\to A}_{ca}$$

$$G^{Ab}_{apo} = G^{\circ}_{apo} + \delta G^{a\to A}_{apo}$$

$$\langle G^{Ab}_{ca,apo}\rangle(\mu_{Ca^{2+}}) = -RTln\left(e^{-\frac{(G^{\circ}_{ca}-4\mu_{Ca^{2+}}+\delta G^{a\to A}_{ca})}{RT}} + e^{-\frac{(G^{\circ}_{apo}+\delta G^{a\to A}_{apo})}{RT}}\right)$$

aB genotype:

$$G^{aB}_{ca}(\mu_{Ca^{2+}}) = G^{\circ}_{ca} - 4\mu_{Ca^{2+}} + \delta G^{b\to B}_{ca}$$

$$G^{aB}_{apo} = G^{\circ}_{apo} + \delta G^{b\to B}_{apo}$$

$$\langle G_{ca,apo}^{aB} \rangle(\mu_{Ca^{2+}}) = -RTln\left(e^{-\frac{\left(G_{ca}^{\circ} - 4\mu_{Ca^{2+}} + \delta G_{ca}^{b\to B}\right)}{RT}} + e^{-\frac{\left(G_{apo}^{\circ} + \delta G_{apo}^{b\to B}\right)}{RT}}\right)$$

AB genotype:

$$G_{ca}^{AB}(\mu_{Ca^{2+}}) = G_{ca}^{\circ} - 4\mu_{Ca^{2+}} + \delta G_{ca}^{a\to A} + \delta G_{ca}^{b\to B}$$

$$G_{apo}^{AB} = G_{apo}^{\circ} + \delta G_{apo}^{a\to A} + \delta G_{apo}^{b\to B}$$

$$\langle G_{ca,apo}^{AB} \rangle(\mu_{Ca^{2+}}) = -RTln\left(e^{-\frac{\left(G_{ca}^{\circ} - 4\mu_{Ca^{2+}} + \delta G_{ca}^{a\to A} + \delta G_{ca}^{b\to B}\right)}{RT}} + e^{-\frac{\left(G_{apo}^{\circ} + \delta G_{apo}^{a\to A} + \delta G_{apo}^{b\to B}\right)}{RT}}\right)$$

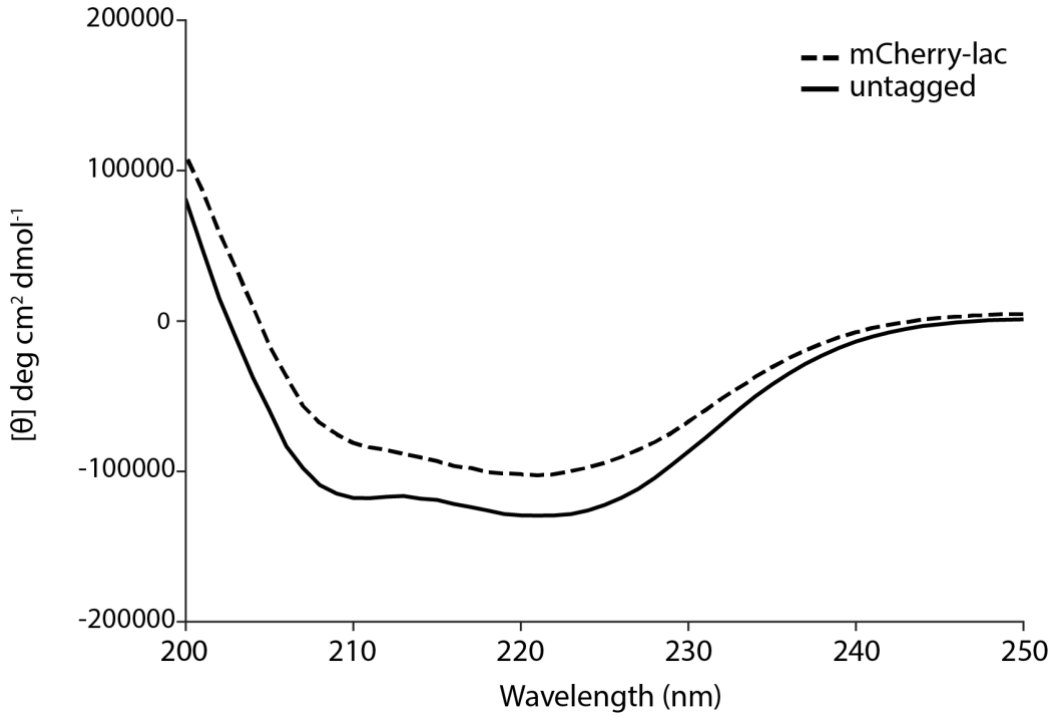
Final expression for $\mu_{Ca^{2+}}$-dependence of ε:

$$\varepsilon(\mu_{Ca^{2+}}) = -\left[\left(\langle G_{ca,apo}^{AB} \rangle - \langle G_{ca,apo}^{aB} \rangle\right) - \left(\langle G_{ca,apo}^{Ab} \rangle - \langle G_{ca,apo}^{ab} \rangle\right)\right].$$

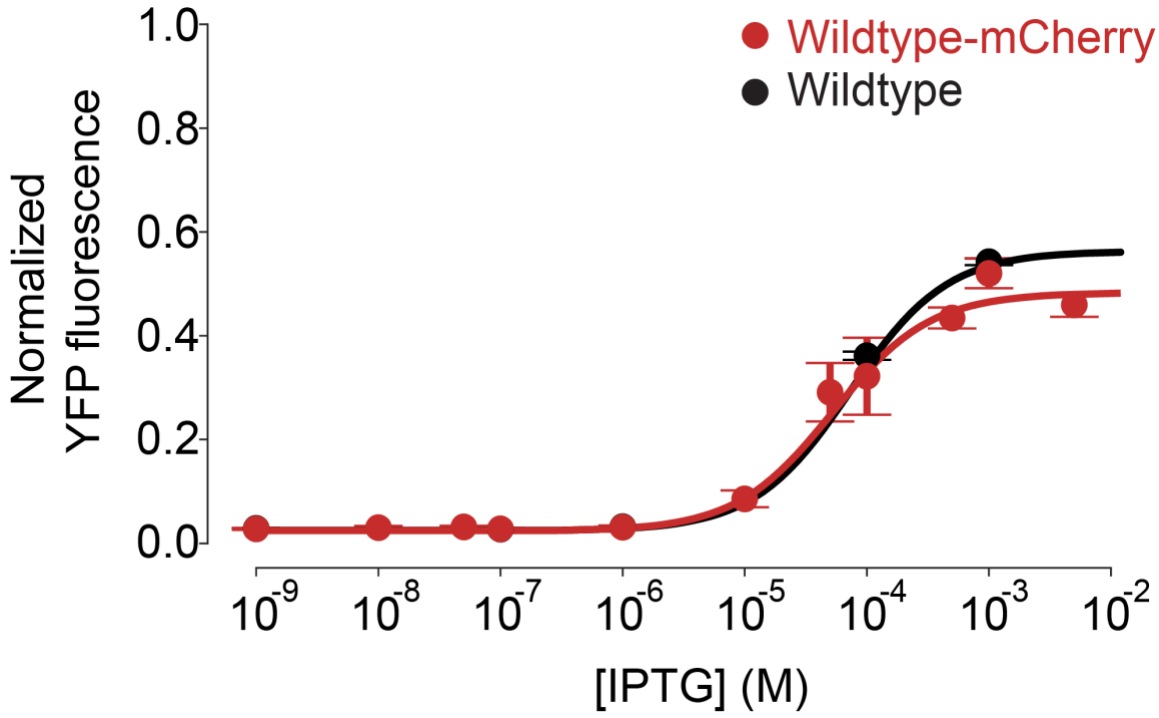
APPENDIX B

SUPPLEMENTARY MATERIAL FOR CHAPTER III

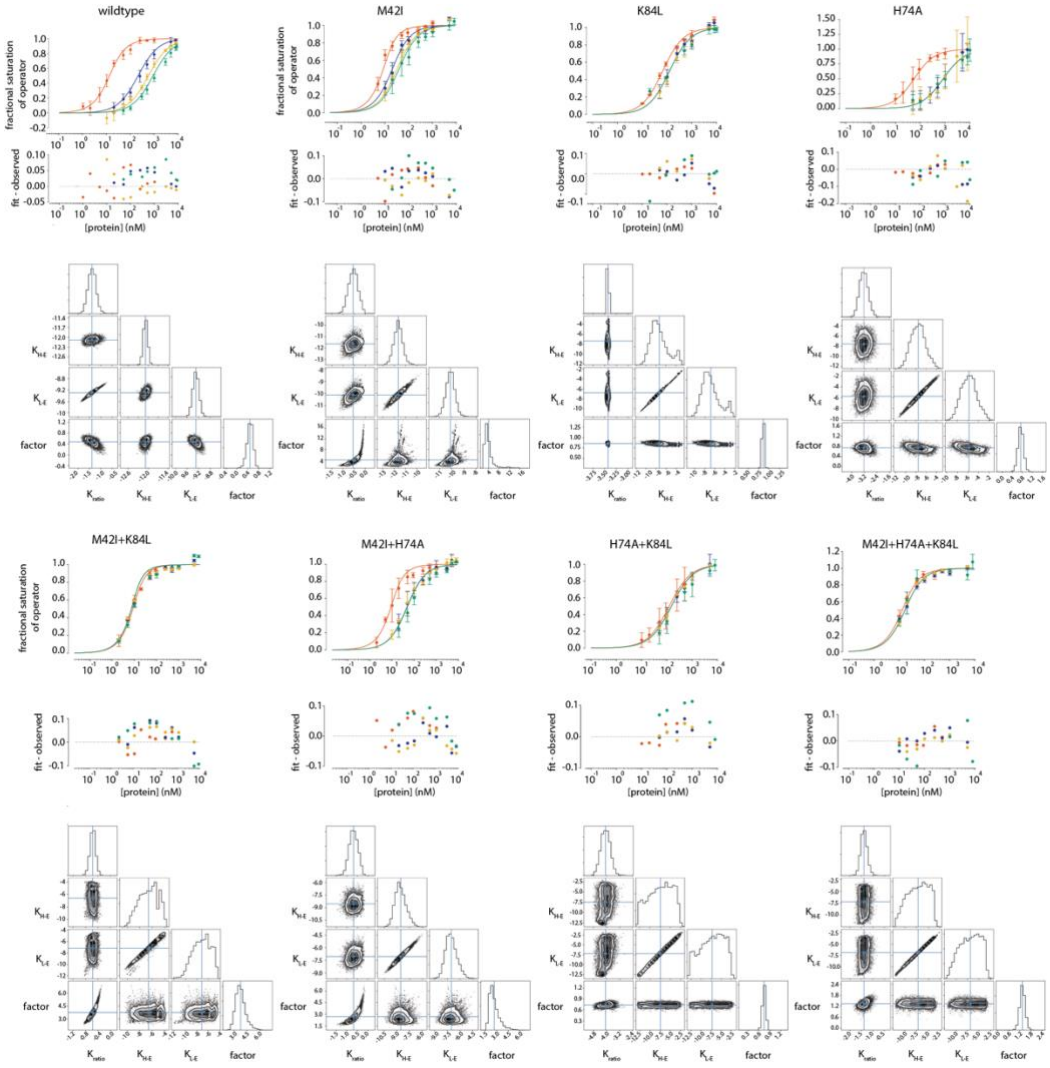
This section includes the supplementary material and supplementary figures referenced in chapter III.



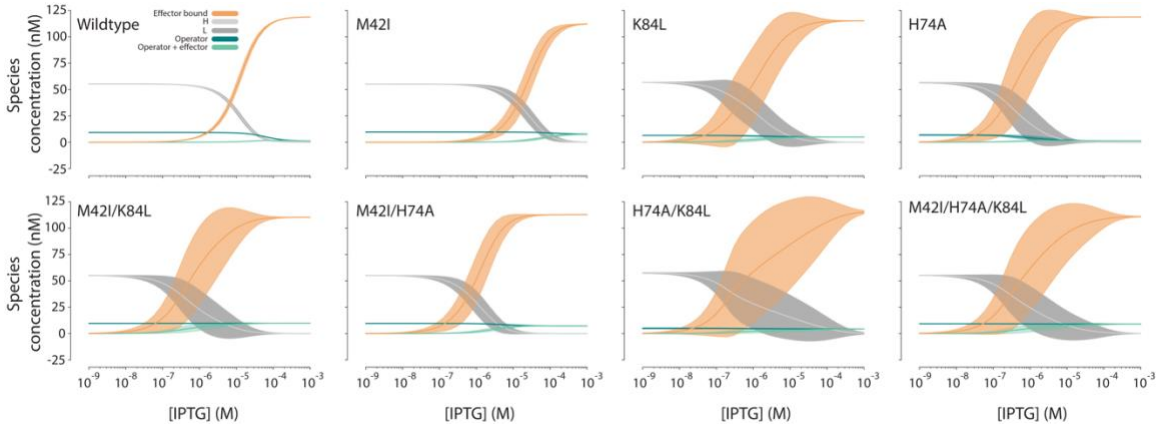
Supplementary Fig B1 Far-UV CD spectra for the mCherry-tagged and untagged wildtype lac repressor. CD spectra of the wildtype lac repressor protein (solid line) and the mCherry-tagged lac repressor (dashed).



Supplementary Fig B2 Induction curves of the mCherry-tagged and untagged wildtype lac repressor The *in vivo* induction curves with normalized YFP fluorescence on the y-axis and IPTG concentration (M) on the x-axis. The wildtype lac repressor is shown in black and the mCherry-tagged lac repressor is shown in dark red.



Supplementary Fig B3 MWC model fits and corner plots for all eight genotypes. The top plots of each row show fractional saturation of operator on the x-axis and protein concentration in nM on the y-axis. Circles are averages of measured datapoints, error bars represent the standard deviation, and solid lines represent the fit curve using the Bayesian MCMC fit parameter estimates. Each color represents measurements made at a different IPTG concentration: 0 mM (red), 0.1 mM (dark blue), 0.3 mM (gold), and 1.0 mM (teal). Residual plots are shown below, with the residual on the y-axis and protein concentration in nM on the x-axis. Corner plots for Bayesian MCMC fits are shown below residual plots for all four fit parameters: K_{ratio}, $K_{H\text{-}E}$, $K_{L\text{-}E}$, and factor.



Supplementary Fig B4 Simulated species concentration as a function of IPTG concentration. Species concentration (y-axis, nM) as a function of IPTG concentration (x-axis, M) for each lac repressor genotype studied. Each group of conformations is shown as follows: unbound conformations (light grey and dark grey; H and L), operator bound (dark teal; H-DNA, L-DNA) effector bound (peach; H-E, H-2E, L-E, and L-2E), and operator + effector bound conformations (light teal; H-DNA-E, H-DNA-2E, L-DNA-E, L-DNA-2E). The solid line represents the average species concentration from 100 sets of sampled parameters from Bayesian MCMC fits. Shaded areas represent the average $\pm$ standard deviation. All calculations were done with $[\text{protein}]_{\text{total}}$ = 120 nM, $[\text{operator}]_{\text{total}}$ = 10 nM.

APPENDIX C

SUPPLEMENTARY MATERIAL FOR CHAPTER IV

This section includes the supplementary material and supplementary figures referenced in chapter IV.

Supplementary Section 1: Construct sequences and design

The codon corrected nucleotide sequences of the ancestral GFP-like protein and the derived GFP-like protein sequence used to construct the library are found below.

>Ancestral GFP-like protein

ATGAGCGTGATCAAAAGCGACATGAAAATCAAACTGCGCATGGAGGGCACC
GTTAACGGCCACAAGTTCGTGATCGAAGGTGAAGGCGAAGGTAAACCGTACG
AAGGCACCCAGACCATGAACCTGAAAGTGAAAGAAGGCGCCCCGCTGCCGT
TTGCCTATGATATCCTGACCACCGCTTTCCAGTACGGTAACCGCGTTTTCACT
AAATACCCTAAGGACATCCCGGACTACTTTAAGCAGAGCTTCCCGGAAGGCT
ACAGCTGGGAACGCAGTATGACCTTTGAGGATGGCGGCATTTGCACCGCCAC
TTCTGATATCACCTTGGAAGGCGACTGCTTCTTCTACGAAATCCGCTTTGACG
GCGTGAACTTTCCGCCGAATGGTCCGGTGATGCAGAAGAAGACACTGAAGTG
GGAGCCGAGCACCGAAAAAATGTACGTGCGCGATGGCGTTCTGATGGGTGAT
GTTAACATGGCACTGCTGCTGGAAGGTGGTGGTCACTACCGCTGCGACTTTA
AGACCACCTACAAAGCAAAGAAAGGCGTGCAGCTGCCGGATTATCATTTTGT
GGACCACCGTATCGAGATCCTGAGCCATGATAAAGACTACAATAATGTTAAA
CTGTATGAACATGCCGTTGCCCGTTACAGCATGCTGCCTAGCCTGGCCAAAG
CCGCCTAATAA

>Derived GFP-like protein

ATGAGCGTGATCAAAAGCGACATGAAAATCAAACTGCGCATGGAGGGCACC

GTTAACGGCCACAAGTTCGTGATCGTTGGTGAAGGCGAAGGTAAACCGTACG

AAGGCACCCAGACCATGAACCTGAAAGTGAAAGAAGGCGCCCCGCTGCCGT

TTGCCTATGATATCCTGACCACCGTTTTCCATTACGGTAACCGCGTTTTCGCTA

AATACCCTAAGCATATCCCGGACTACTTTAAGCAGAGCTTCCCGGAAGGCTA

CAGCTGGGAACGCAGTATGACCTTTGAGGATGGCGGCATTTGCACCGCCCGT

AACGATATCACCTTGGAAGGCGACTGCTTCTTCAACGAAATCCGCTTTGACG

GCGTGAACTTTCCGCCGAATGGTCCGGTGATGCAGAAGAAGACACTGAAGTG

GGAGCCGAGCACCGAAAAAATGTACGTGCGCGATGGCGTTCTGACTGGTGAT

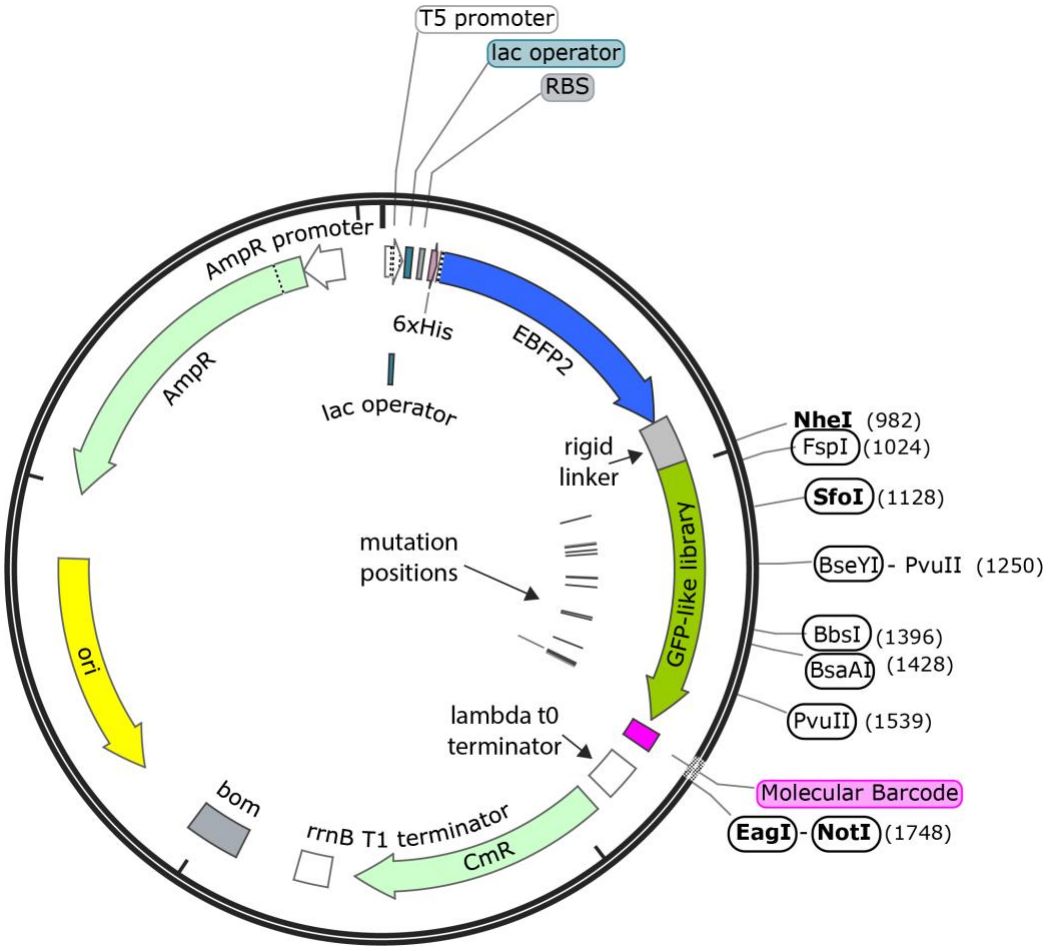
ATCAACATGGCACTGCTGCTGGAAGGTGGTGGTCACTACCGCTGCGACTTTA

AGACCACCTACAAAGCAAAGAAAGGCGTGCAGCTGCCGGATTATCATTTTGT

GGACCACTGCATCGAGATCCTGAGCCATGATAAAGACTACAATAATGTTAAA

CTGTATGAACATGCCGAAGCCCATAGCGTTCTGCCTAGCCTGGCCAAAGCCG

CCTAATAA



Supplementary Fig C1 Full plasmid map for the GFP-like protein library. Relevant construct components are highlighted. All circled restriction enzymes are those used in genotype to barcode association study. Bolded restriction enzymes make single cuts in the construct. Mutation positions relative to the GFP-like library sequence are shown as lines.

In the sequences below, please note that the ‘NNNNNNNNNN…’ portion corresponds to the region where the molecular barcode is located (shown in magenta in Supplementary Fig 1 and 2).

>pQE30-EBFFP2-AncGFP

CTCGAGAAATCATAAAAAATTTATTTGCTTTGTGAGCGGATAACAATTATAAT
AGATTCAATTGTGAGCGGATAACAATTTCACACAGAATTCATTAAAGAGGAG
AAATTAACTATGAGAGGATCGCATCACCATCACCATCACGGATCCATGGTGA
GCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGA
CGGCGACGTAAACGGCCACAAGTTCAGCGTGAGGGGCGAGGGCGAGGGCGA
TGCCACCAACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTG
CCCGTGCCCTGGCCCACCCTCGTGACCACCCTGAGCCACGGCGTGCAGTGCTT
CGCCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATG
CCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCACCT
ACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCA
TCGAGCTGAAGGGCGTCGACTTCAAGGAGGACGGCAACATCCTGGGGCACA
AGCTGGAGTACAACTTCAACAGCCACAACATCTATATCATGGCCGTCAAGCA
GAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACGTGGAGGACGG
CAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGC
CCCGTGCTGCTGCCCGACAGCCACTACCTGAGCACCCAGTCCGTGCTGAGCA
AAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCCGCACCGC
CGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGGGATCCGCATGCGAG
CTCGGTACCGGCAGCCTGGCAGAAGCAGCCGCCAAAGAAGCTGCCGCAAAA

GAAGCCGCCGCAAAGGCCGCCGCAGCCAGTATCGTGCATAATAGCCTGGCTA

GCATGAGCGTGATCAAAAGCGACATGAAAATCAAACTGCGCATGGAGGGCA

CCGTTAACGGCCACAAGTTCGTGATCGAAGGTGAAGGCGAAGGTAAACCGTA

CGAAGGCACCCAGACCATGAACCTGAAAGTGAAAGAAGGCGCCCCGCTGCC

GTTTGCCTATGATATCCTGACCACCGCTTTCCAGTACGGTAACCGCGTTTTCA

CTAAATACCCTAAGGACATCCCGGACTACTTTAAGCAGAGCTTCCCGGAAGG

CTACAGCTGGGAACGCAGTATGACCTTTGAGGATGGCGGCATTTGCACCGCC

ACTTCTGATATCACCTTGGAAGGCGACTGCTTCTTCTACGAAATCCGCTTTGA

CGGCGTGAACTTTCCGCCGAATGGTCCGGTGATGCAGAAGAAGACACTGAAG

TGGGAGCCGAGCACCGAAAAAATGTACGTGCGCGATGGCGTTCTGATGGGTG

ATGTTAACATGGCACTGCTGCTGGAAGGTGGTGGTCACTACCGCTGCGACTTT

AAGACCACCTACAAAGCAAAGAAAGGCGTGCAGCTGCCGGATTATCATTTTG

TGGACCACCGTATCGAGATCCTGAGCCATGATAAAGACTACAATAATGTTAA

ACTGTATGAACATGCCGTTGCCCGTTACAGCATGCTGCCTAGCCTGGCCAAA

GCCGCCTAATAATCGTAGCCCGGGCATCCTAGGANNNNNNNNNNNNNNNNN

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGCGGCCGCGGTACCCCG

GGTCGACCTGCAGCCAAGCTTAATTAGCTGAGCTTGGACTCCTGTTGATAGAT

CCAGTAATGACCTCAGAACTCCATCTGGATTTGTTCAGAACGCTCGGTTGCCG

CCGGGCGTTTTTTATTGGTGAGAATCCAAGCTTGCTTGGCGAGATTTTCAGGA

GCTAAGGAAGCTAAAATGGAGAAAAAAATCACTGGATATACCACCGTTGATA

TATCCCAATGGCATCGTAAAGAACATTTTGAGGCATTTCAGTCAGTTGCTCAA

TGTACCTATAACCAGACCGTTCAGCTGGATATTACGGCCTTTTTAAAGACCGT

AAAGAAAAATAAGCACAAGTTTTATCCGGCCTTTATTCACATTCTTGCCCGCC

TGATGAATGCTCATCCGGAATTTCGTATGGCAATGAAAGACGGTGAGCTGGT

GATATGGGATAGTGTTCACCCTTGTTACACCGTTTTCCATGAGCAAACTGAAA

CGTTTTCATCGCTCTGGAGTGAATACCACGACGATTTCCGGCAGTTTCTACAC

ATATATTCGCAAGATGTGGCGTGTTACGGTGAAAACCTGGCCTATTTCCCTAA

AGGGTTTATTGAGAATATGTTTTTCGTCTCAGCCAATCCCTGGGTGAGTTTCA

CCAGTTTTGATTTAAACGTGGCCAATATGGACAACTTCTTCGCCCCCGTTTTC

ACCATGGGCAAATATTATACGCAAGGCGACAAGGTGCTGATGCCGCTGGCGA

TTCAGGTTCATCATGCCGTTTGTGATGGCTTCCATGTCGGCAGAATGCTTAAT

GAATTACAACAGTACTGCGATGAGTGGCAGGGCGGGGCGTAATTTTTTTAAG

GCAGTTATTGGTGCCCTTAAACGCCTGGGGTAATGACTCTCTAGCTTGAGGCA

TCAAATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTT

GTTTGTCGGTGAACGCTCTCCTGAGTAGGACAAATCCGCCCTCTTGAGCTGCC

TCGCGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGA

GACGGTCACAGCTTGTCTGTAAGCGGATGCCGGGAGCAGACAAGCCCGTCAG

GGCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCGCAGCCATGACCCAGTCAC

GTAGCGATAGCGGAGTGTATACTGGCTTAACTATGCGGCATCAGAGCAGATT

GTACTGAGAGTGCACCATATGCGGTGTGAAATACCGCACAGATGCGTAAGGA

GAAAATACCGCATCAGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGC

TCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATAC

GGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAG

GCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCA

TAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGG

TGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCT

CCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCC

TTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCT

CAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCC

GTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCC

GGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGC

AGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACT

ACGGCTACACTAGAAGGACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGT

TACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCT

GGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAG

GATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAAC

GAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCA

CCTAGATCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATAT

GAGTAAACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTC

AGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTCGTGTAGA

TAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACC

GCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCC

GGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGT

CTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTG

CGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGTCACGCTCGTCGTTTGG

TATGGCTTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATCCC

CCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGA

AGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTC

TCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAA

CCAAGTCATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGC

GTCAATACGGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATC

ATTGGAAAACGTTCTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTGA

GATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTT

ACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAA

AAAAGGGAATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTT

TCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATAT

TTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCG

AAAAGTGCCACCTGACGTCTAAGAAACCATTATTATCATGACATTAACCTAT

AAAAATAGGCGTATCACGAGGCCCTTTCGTCTTCAC

>pQE30-EBFP2-Anc15

CTCGAGAAATCATAAAAAATTTATTTGCTTTGTGAGCGGATAACAATTATAAT

AGATTCAATTGTGAGCGGATAACAATTTCACACAGAATTCATTAAAGAGGAG

AAATTAACTATGAGAGGATCGCATCACCATCACCATCACGGATCCATGGTGA

GCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGA

CGGCGACGTAAACGGCCACAAGTTCAGCGTGAGGGGCGAGGGCGAGGGCGA

TGCCACCAACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTG

CCCGTGCCCTGGCCCACCCTCGTGACCACCCTGAGCCACGGCGTGCAGTGCTT

CGCCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATG

CCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCACCT

ACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCA

TCGAGCTGAAGGGCGTCGACTTCAAGGAGGACGGCAACATCCTGGGGCACA

AGCTGGAGTACAACTTCAACAGCCACAACATCTATATCATGGCCGTCAAGCA

GAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACGTGGAGGACGG

CAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGC

CCCGTGCTGCTGCCCGACAGCCACTACCTGAGCACCCAGTCCGTGCTGAGCA

AAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCCGCACCGC

CGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGGGATCCGCATGCGAG

CTCGGTACCGGCAGCCTGGCAGAAGCAGCCGCCAAAGAAGCTGCCGCAAAA

GAAGCCGCCGCAAAGGCCGCCGCAGCCAGTATCGTGCATAATAGCCTGGCTA

GCATGAGCGTGATCAAAAGCGACATGAAAATCAAACTGCGCATGGAGGGCA

CCGTTAACGGCCACAAGTTCGTGATCGTTGGTGAAGGCGAAGGTAAACCGTA

CGAAGGCACCCAGACCATGAACCTGAAAGTGAAAGAAGGCGCCCCGCTGCC

GTTTGCCTATGATATCCTGACCACCGTTTTCCATTACGGTAACCGCGTTTTCGC

TAAATACCCTAAGCATATCCCGGACTACTTTAAGCAGAGCTTCCCGGAAGGC

TACAGCTGGGAACGCAGTATGACCTTTGAGGATGGCGGCATTTGCACCGCCC

GTAACGATATCACCTTGGAAGGCGACTGCTTCTTCAACGAAATCCGCTTTGAC

GGCGTGAACTTTCCGCCGAATGGTCCGGTGATGCAGAAGAAGACACTGAAGT

GGGAGCCGAGCACCGAAAAAATGTACGTGCGCGATGGCGTTCTGACTGGTGA

TATCAACATGGCACTGCTGCTGGAAGGTGGTGGTCACTACCGCTGCGACTTTA

AGACCACCTACAAAGCAAAGAAAGGCGTGCAGCTGCCGGATTATCATTTTGT

GGACCACTGCATCGAGATCCTGAGCCATGATAAAGACTACAATAATGTTAAA

CTGTATGAACATGCCGAAGCCCATAGCGTTCTGCCTAGCCTGGCCAAAGCCG

CCTAATAATCGTAGCCCGGGCATCCTAGGANNNNNNNNNNNNNNNNNNNN

NNNNNNNNNNNNNNNNNNNNNNNNNNNNGCGGCCGCGGTACCCCGGGTC

GACCTGCAGCCAAGCTTAATTAGCTGAGCTTGGACTCCTGTTGATAGATCCAG

TAATGACCTCAGAACTCCATCTGGATTTGTTCAGAACGCTCGGTTGCCGCCGG

GCGTTTTTTATTGGTGAGAATCCAAGCTTGCTTGGCGAGATTTTCAGGAGCTA

AGGAAGCTAAAATGGAGAAAAAAATCACTGGATATACCACCGTTGATATATC

CCAATGGCATCGTAAAGAACATTTTGAGGCATTTCAGTCAGTTGCTCAATGTA

CCTATAACCAGACCGTTCAGCTGGATATTACGGCCTTTTTAAAGACCGTAAAG

AAAAATAAGCACAAGTTTTATCCGGCCTTTATTCACATTCTTGCCCGCCTGAT

GAATGCTCATCCGGAATTTCGTATGGCAATGAAAGACGGTGAGCTGGTGATA

TGGGATAGTGTTCACCCTTGTTACACCGTTTTCCATGAGCAAACTGAAACGTT

TTCATCGCTCTGGAGTGAATACCACGACGATTTCCGGCAGTTTCTACACATAT

ATTCGCAAGATGTGGCGTGTTACGGTGAAAACCTGGCCTATTTCCCTAAAGG

GTTTATTGAGAATATGTTTTTCGTCTCAGCCAATCCCTGGGTGAGTTTCACCA

GTTTTGATTTAAACGTGGCCAATATGGACAACTTCTTCGCCCCCGTTTTCACC

ATGGGCAAATATTATACGCAAGGCGACAAGGTGCTGATGCCGCTGGCGATTC

AGGTTCATCATGCCGTTTGTGATGGCTTCCATGTCGGCAGAATGCTTAATGAA

TTACAACAGTACTGCGATGAGTGGCAGGGCGGGGCGTAATTTTTTTAAGGCA

GTTATTGGTGCCCTTAAACGCCTGGGGTAATGACTCTCTAGCTTGAGGCATCA

AATAAAACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTT

TGTCGGTGAACGCTCTCCTGAGTAGGACAAATCCGCCCTCTTGAGCTGCCTCG

CGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGAGAC

GGTCACAGCTTGTCTGTAAGCGGATGCCGGGAGCAGACAAGCCCGTCAGGGC

GCGTCAGCGGGTGTTGGCGGGTGTCGGGGCGCAGCCATGACCCAGTCACGTA

GCGATAGCGGAGTGTATACTGGCTTAACTATGCGGCATCAGAGCAGATTGTA

CTGAGAGTGCACCATATGCGGTGTGAAATACCGCACAGATGCGTAAGGAGAA

AATACCGCATCAGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCG

GTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGT

TATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCC

AGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAG

GCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGG

CGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCC

TCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTC

TCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGT

TCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCA

GCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAA

GACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGC

GAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGC

TACACTAGAAGGACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTT

CGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGC

GGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAGGATCTC

AAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAA

CTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGA

TCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAA

ACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGAT

CTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTCGTGTAGATAACTAC

GATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGAC

CCACGCTCACCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGG

CCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAAT

TGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACG

TTGTTGCCATTGCTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCT

TCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTT

GTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGAAGTAAG

TTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTAC

TGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGT

CATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATA

CGGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAA

AACGTTCTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTGAGATCCAGT

TCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCAC

CAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAGGG

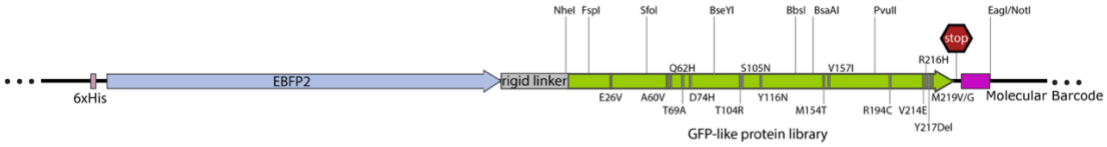
AATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAATATT

ATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGT

ATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGC

CACCTGACGTCTAAGAAACCATTATTATCATGACATTAACCTATAAAAATAG

GCGTATCACGAGGCCCTTTCGTCTTCAC

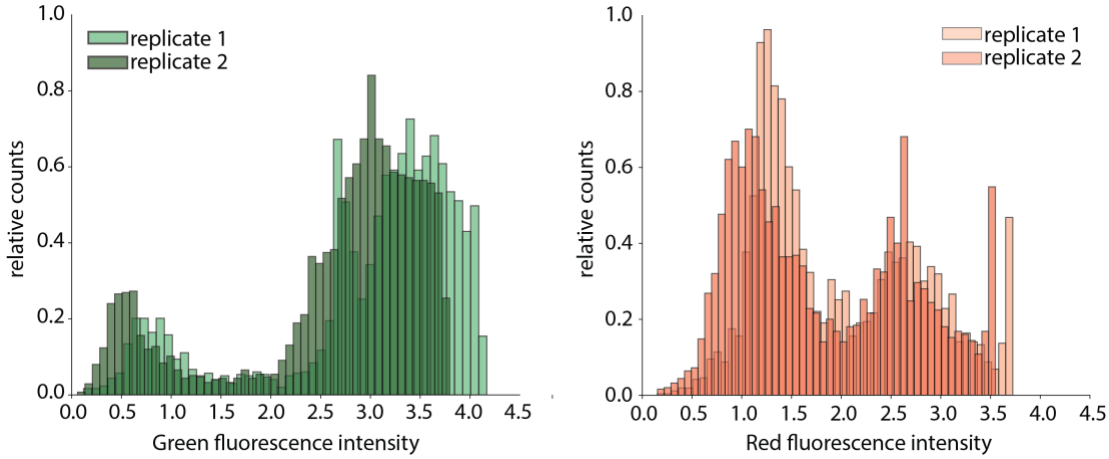


Supplementary Fig C2 GFP-like protein library construct design. The GFP-like protein library (green) was expressed as a 6x His-tagged fusion protein with EBFP2 (blue). An alpha-helical rigid linker sequence was inserted between the two fluorescent proteins to reduce FRET between them (grey). All relevant restriction enzymes for barcode mapping are labelled above the library. Grey lines indicate the location of the restriction site in the sequence. All fifteen mutations in the library are shown at their location within the sequence. At the end of the GFP library there is a double stop codon (red hexagon). The 25 bp molecular barcode is shown in magenta at the C-terminal end of the library.

Table C1 Association of amino acid substitutions with the observation of red fluorescence.

Site # (amino acid)[a]	Ancestral amino acid	Derived amino acid	P-value[b]
26	E	V	0.2
60	A	V	0.001
62	Q	H	0.000
69	T	A	0.002
74	D	H	0.067
104	T	R	0.042
105	S	N	0.000
116	Y	N	0.000
154	M	T	0.003
157	V	I	0.051
194	R	C	0.000
214	V	E	0.074
216	R	H	0.013
217	Y	deletion	0.001
219	M	V, G	0.005[1]

[a] Relative to Ancestral GFP-like protein sequence
[b] adapted from Field et al, 2009[236]; two-tailed Fisher's exact test describing the association between an amino acid substitution and the observation of red fluorescence (citation).
[1] P-value corresponds to the M219G mutation. See text for details on the valine substitution.



Supplementary Fig C3 Distribution of inferred green and red fluorescence intensities in biological replicates 1 and 2.

Supplementary Section 2: Detailed barcode extraction methods

After each sort-seq experiment, we prepared each sorted population for next generation sequencing (NovaSeq paired-end S1 300 cycle (biological replicates 1 and 2) and paired-end S4 300 cycle (biological replicate 3) as follows:

5. Sorted cells resuspended in double deionized H_2O were thawed and added to thermocycler tubes. All samples were then boiled at 95degC for 10 minutes.
6. To control for potential PCR-based errors, we prepared barcode PCR amplification reactions in triplicate by adding 10 μl of the boiled population for each sample to three new PCR tubes.
7. A Phusion polymerase (NEB catalog #: M0530S) mastermix was prepared containing the following components:

Reaction component	Volume ($\boldsymbol{\mu}$l) per 50 $\boldsymbol{\mu}$l reaction
Double deionized H_2O	23.5
5X Phusion HF bufer	10
10 mM dNTPs (NEB)	1
10 $\boldsymbol{\mu}$M $primer_{fwd}$	2.5
10 $\boldsymbol{\mu}$M $primer_{rev}$	2.5
Boiled sample	10
Phusion polymerase	0.5
Total reaction volume ($\boldsymbol{\mu}$l)	50

The $primer_{fwd}$ and $primer_{rev}$ sequences were as follows:

>$primer_{fwd}$

5'-CTAATAATCGTAGCCCGGGCAT-3'

>$primer_{rev}$

5'-AGGTCGACCCGGGGTACCG-3'

Primer oligos were obtained lyophilized from Eurofins MWG and resuspended to 100 μM with sterile double deionized H_2O. These were then diluted to a working concentration of 10 μM for PCR reactions. All primers were stored at -20degC.

40 μl of the PCR master mix was added to each PCR tube and the amplification reaction was performed with the following parameters:

Step	**# cycles**	**Temperature (°C)**	**Time**
Initial Denaturation	1	98	30 seconds
Amplification	15	98	5 seconds
		65	10 seconds
		72	2 seconds
Final Extension	1	72	5 minutes
Hold	-	12	∞

Each reaction yielded a 105 bp fragment containing the molecular barcode.

8. Post amplification, triplicate samples were pooled and samples were purified using a 1.8X bead clean-up (Omega Bio-Tek MagBind TotalPure NGS beads, SKU #: M1378-01). All samples were confirmed to have between 300 to 1000 ng total DNA after the bead purification step. The entire purified amplification

reaction was diluted to 50 μl in sterile double deionized H_2O for the following adapter ligation steps.

9. We prepared the end repair and A-tailing (ERAT) reaction according to the KAPA hyperprep kit manual. Briefly, we prepared a master mix of the following:

Component	**Volume (μl) per reaction**
ERAT buffer	7
ERAT enzyme	3
Total volume added per reaction	10 μl added to each 50 μl sample

Reactions were cycled as follows:

Step	**Temperature (°C)**	**Time**
ERAT-1	20	30 minutes
ERAT-2	65	30 minutes
HOLD	4	∞

10. Next, we prepared the adapter ligation reactions according to the KAPA Hyperprep kit. Each sorted population got a unique TruSeq-like adapter so that we could identify the origin of each sample in the de-multiplexed sequencing data. Briefly, we prepared the ligation reaction as follows:

Component	**Volume (μl) per reaction**
ddH2O	5
Ligation Buffer	30
DNA ligase	10

TruSeq-like adapter	5 (added to each reaction)
Total volume added per reaction:	45 μl added to each 60 μl ERAT sample

11. We incubated adapter ligation samples at 20degC overnight as we found that it increased the efficiency of adapter ligation.
12. Samples were purified using a 1.2X bead cleanup. Some samples required an additional bead cleanup step to remove residual adapter dimer.

Note: In our first sort-seq sequencing run (NovaSeq S1 300 cycle, biological replicates 1 and 2), we found that our PCR-free workflow led to extensive index hopping. For the third biological replicate of our sort-seq experiments, we added an additional PCR step. See the steps below for more details.

13. We prepared the following PCR reactions using the KAPA Hyperprep kit components:
14.

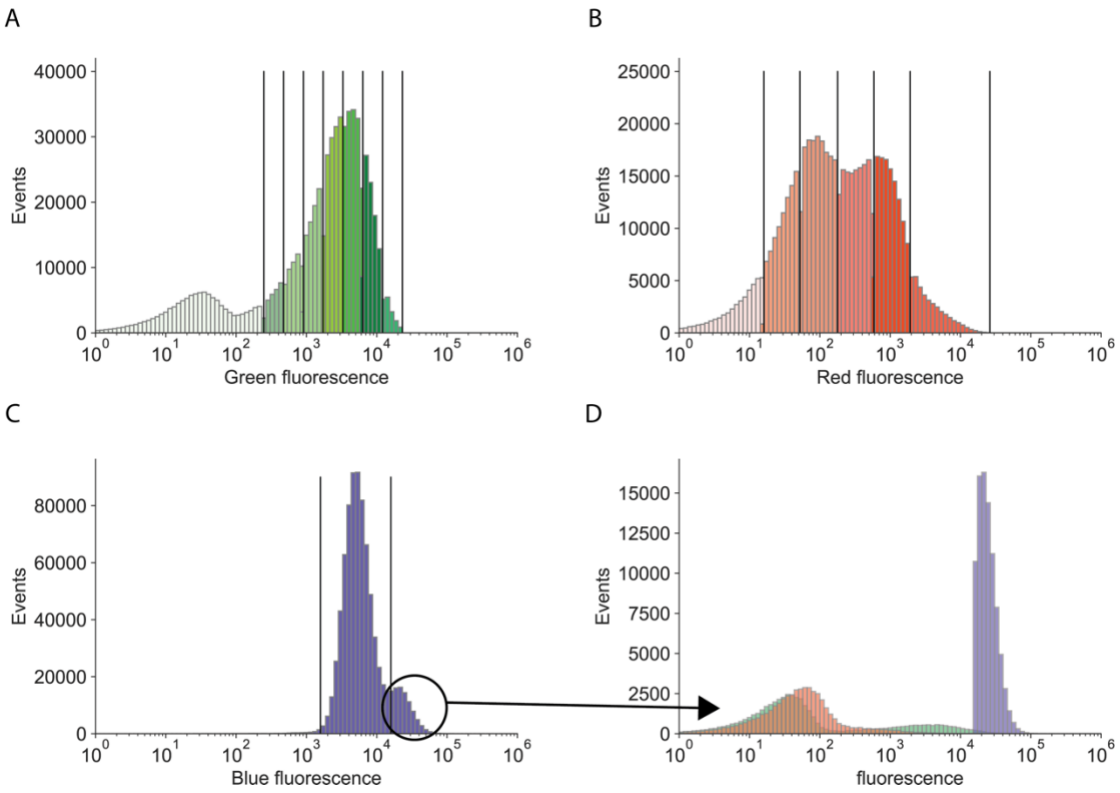
Component	**Volume (μl)**
2X KAPA HiFi HotStart Ready Mix	25
10X KAPA Library Amplification Primer Mix	5
Adapter-ligated sample	20
Total volume	50

The reaction was briefly vortexed and centrifuged. We then amplified barcodes using the following cycling protocol:

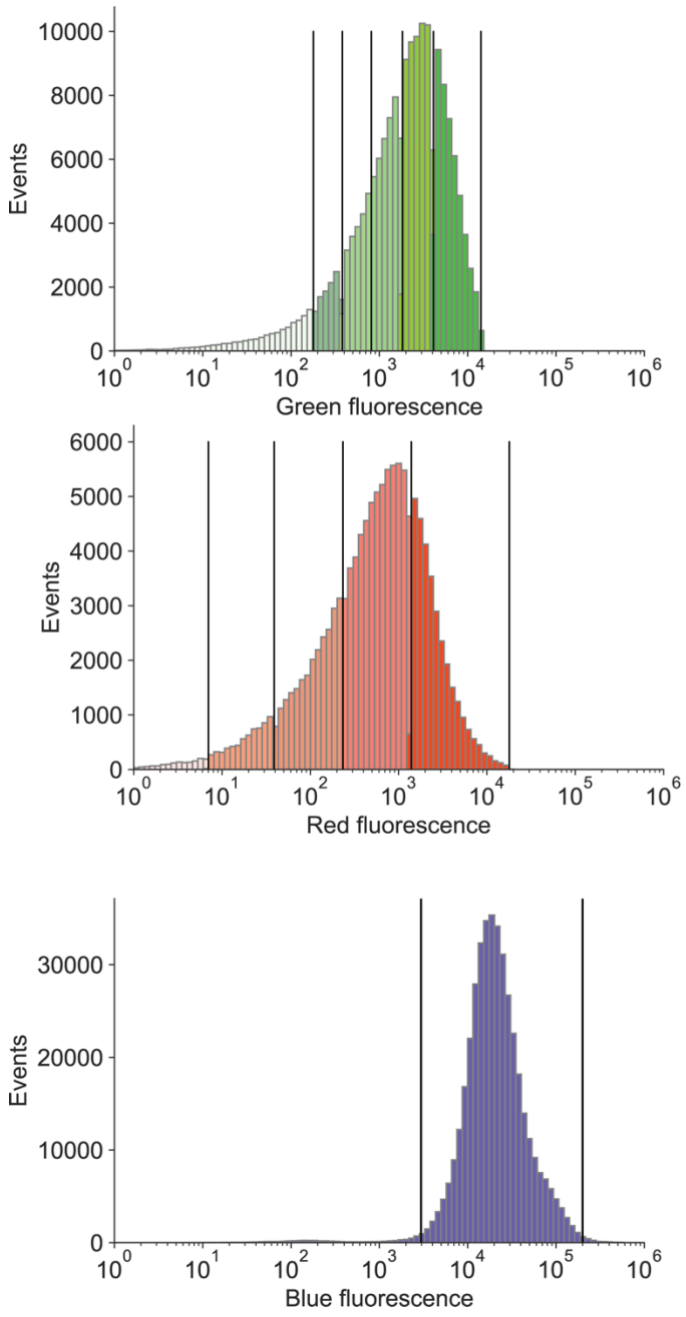
Step	Temperature (°C)	Duration	Cycles
Initial Denaturation	98	45 seconds	1
Denaturation	98	15 seconds	3
Annealing	60	30 seconds	
Extension	72	30 seconds	
Final Extension	72	1 minute	1
HOLD	4	∞	-

15. Each sample was purified using a 1X bead cleanup and eluted in elution buffer (10 mM Tris-HCl, pH 8.5, GeneJet kit). Some samples required an additional bead cleanup step.

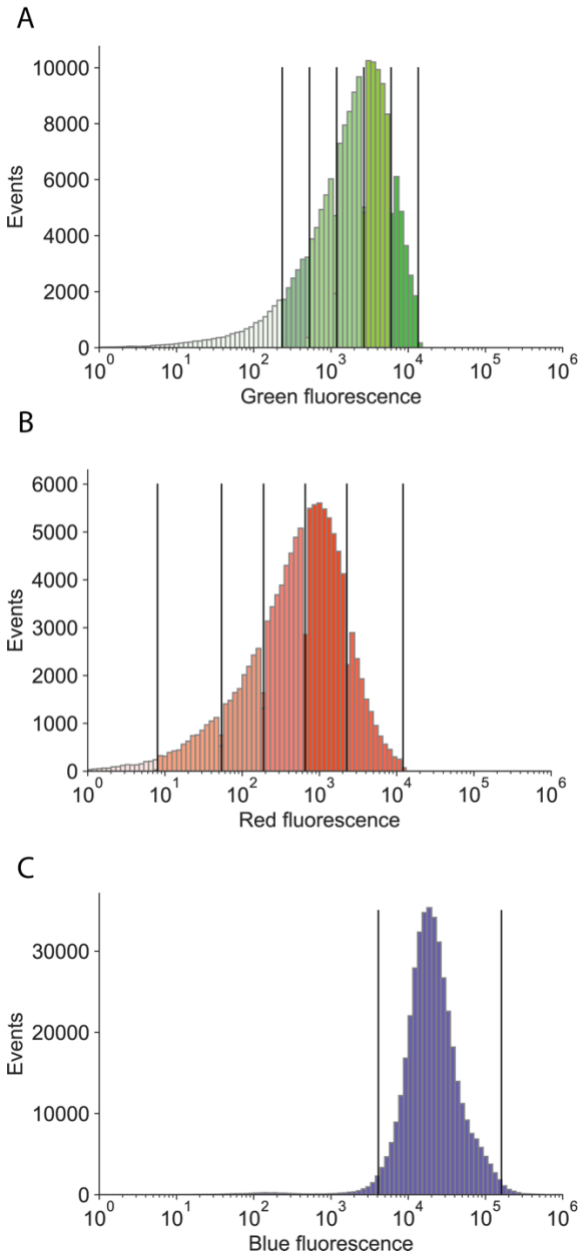
Adapter ligated samples were submitted to the University of Oregon Genomics and Cell Characterization Facility for sequencing.



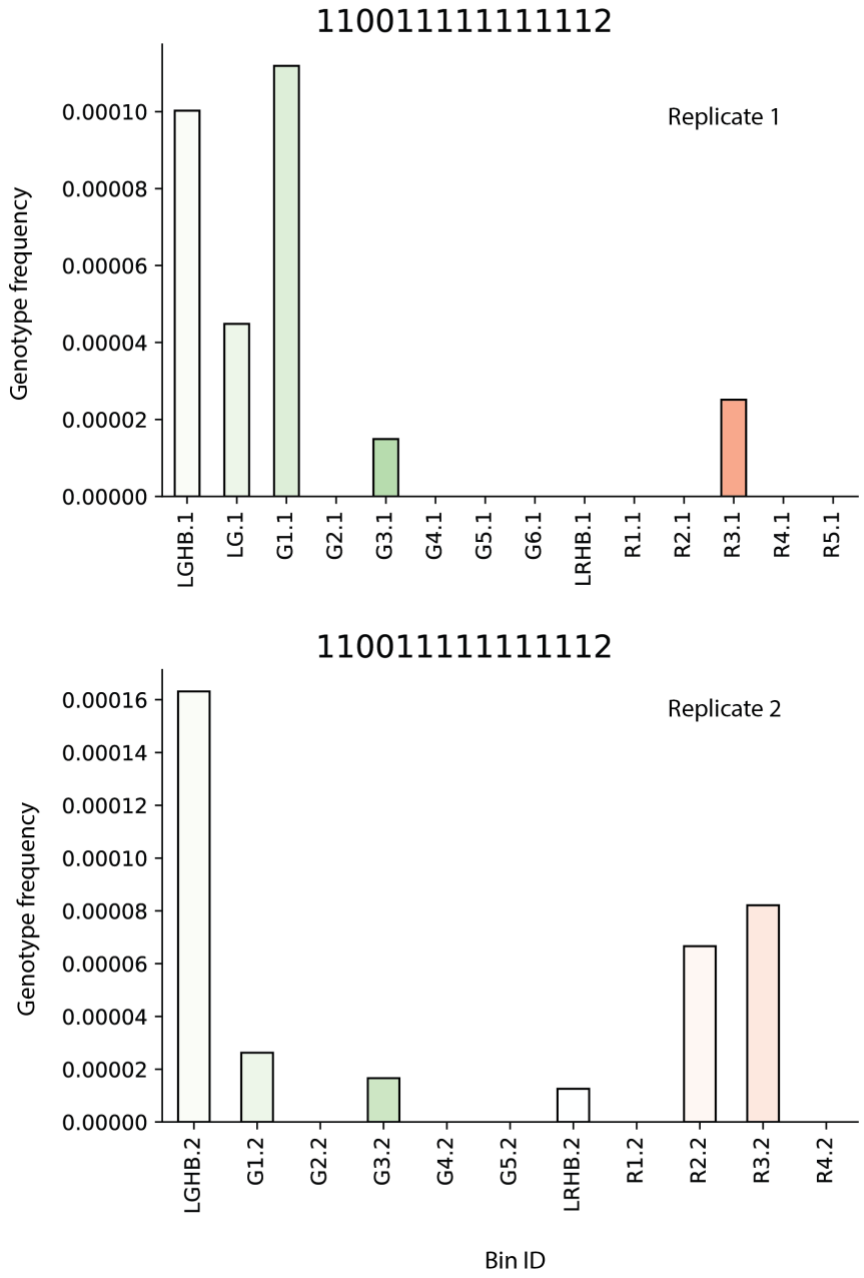
Supplementary Fig C4 Gating strategy for sort-seq experiment biological replicate #1. A) Green fluorescence intensity is shown on the x-axis with counts shown on y-axis. Black lines delineate different gates in our gating strategy. B) Red fluorescence intensity is shown on the x-axis with counts shown on y-axis. Black lines delineate different gates in our gating strategy. C) Blue fluorescence intensity is shown on the x-axis with counts shown on y-axis. Black lines delineate different gates in our gating strategy. D) Blue, red, and green fluorescence intensities for the subpopulation circled in panel C.



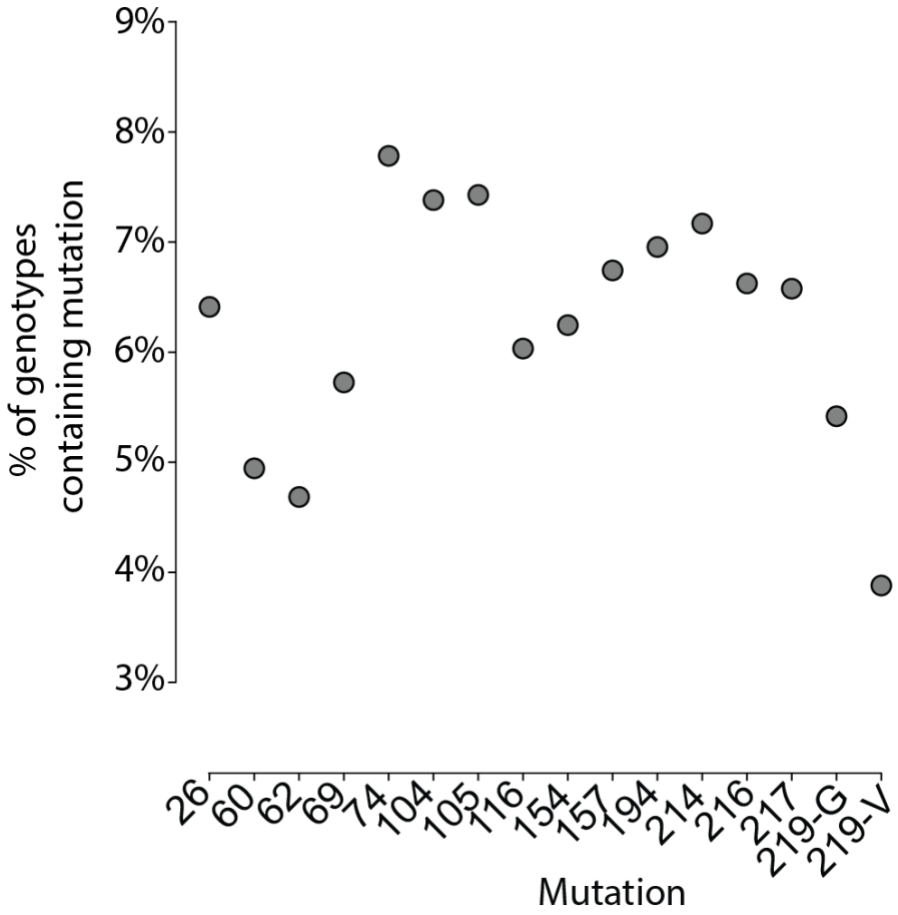
Supplementary Fig C5 Gating strategy for sort-seq experiment biological replicate #2. A) Green fluorescence intensity is shown on the x-axis with counts shown on y-axis. Black lines delineate different gates in our gating strategy. B) Red fluorescence intensity is shown on the x-axis with counts shown on y-axis. Black lines delineate different gates in our gating strategy. C) Blue fluorescence intensity is shown on the x-axis with counts shown on y-axis. Black lines delineate different gates in our gating strategy.



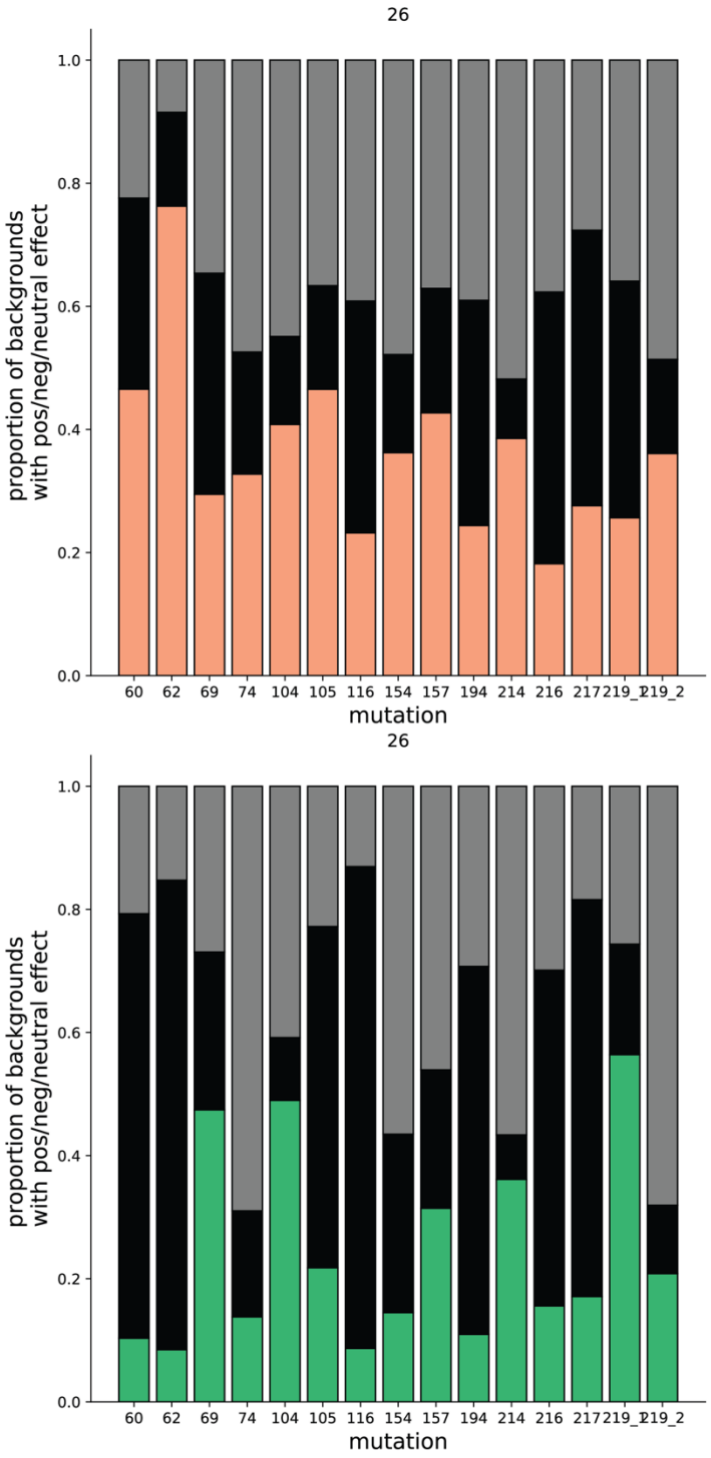
Supplementary Fig C6 Gating strategy for sort-seq experiment biological replicate #3. A) Green fluorescence intensity is shown on the x-axis with counts shown on y-axis. Black lines delineate different gates in our gating strategy. B) Red fluorescence intensity is shown on the x-axis with counts shown on y-axis. Black lines delineate different gates in our gating strategy. B) Blue fluorescence intensity is shown on the x-axis with counts shown on y-axis. Black lines delineate different gates in our gating strategy.



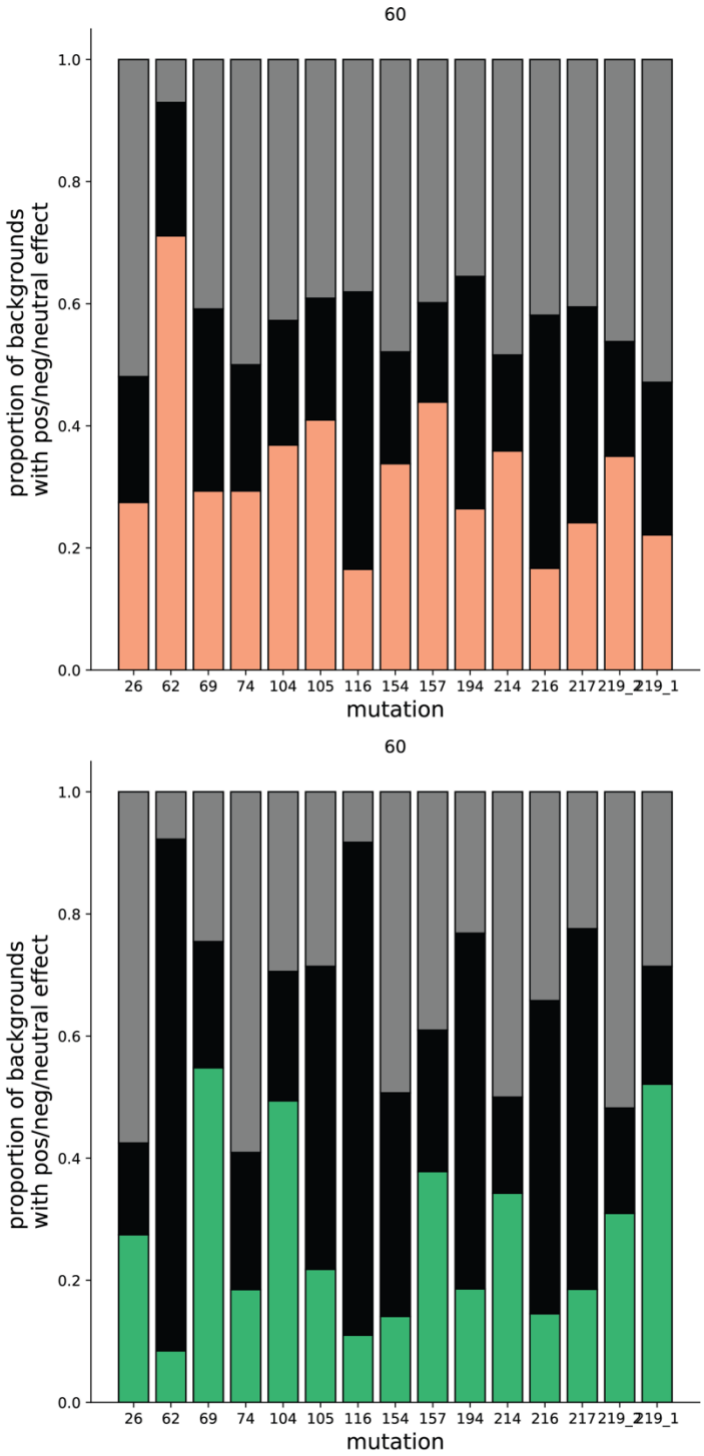
Supplementary Fig C7 Sort-seq data for high red/low green fluorescence clone lacking the Q62H mutation. The identity of each sort-seq bin is shown on the x-axis and the frequency of observations for the 110011111111112 genotype are show on the y-axis. Top panel is data for the first biological replicate. Bottom panel is data for the second biological replicate.



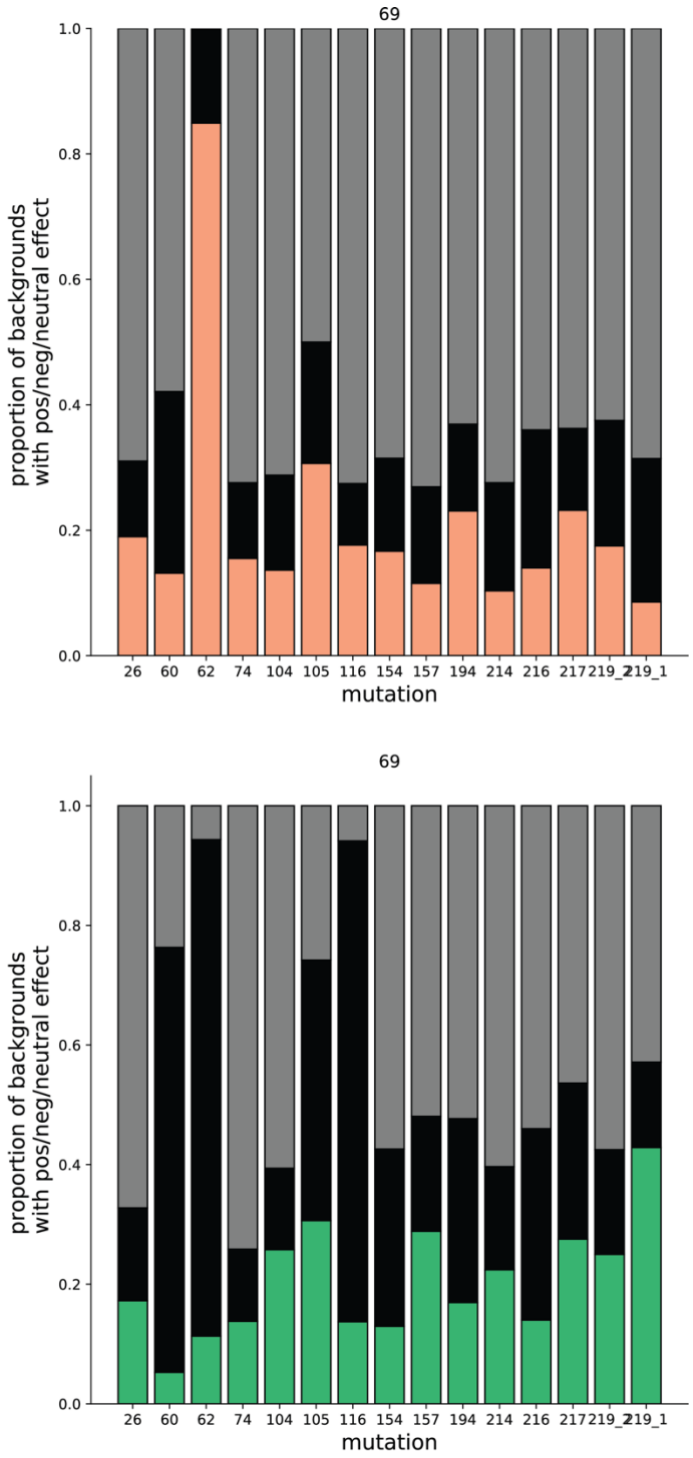
Supplementary Fig C8 Percent of genotypes that contain each mutation. The percent of pairs containing a mutation *n* (y-axis) as a function of mutation (x-axis).



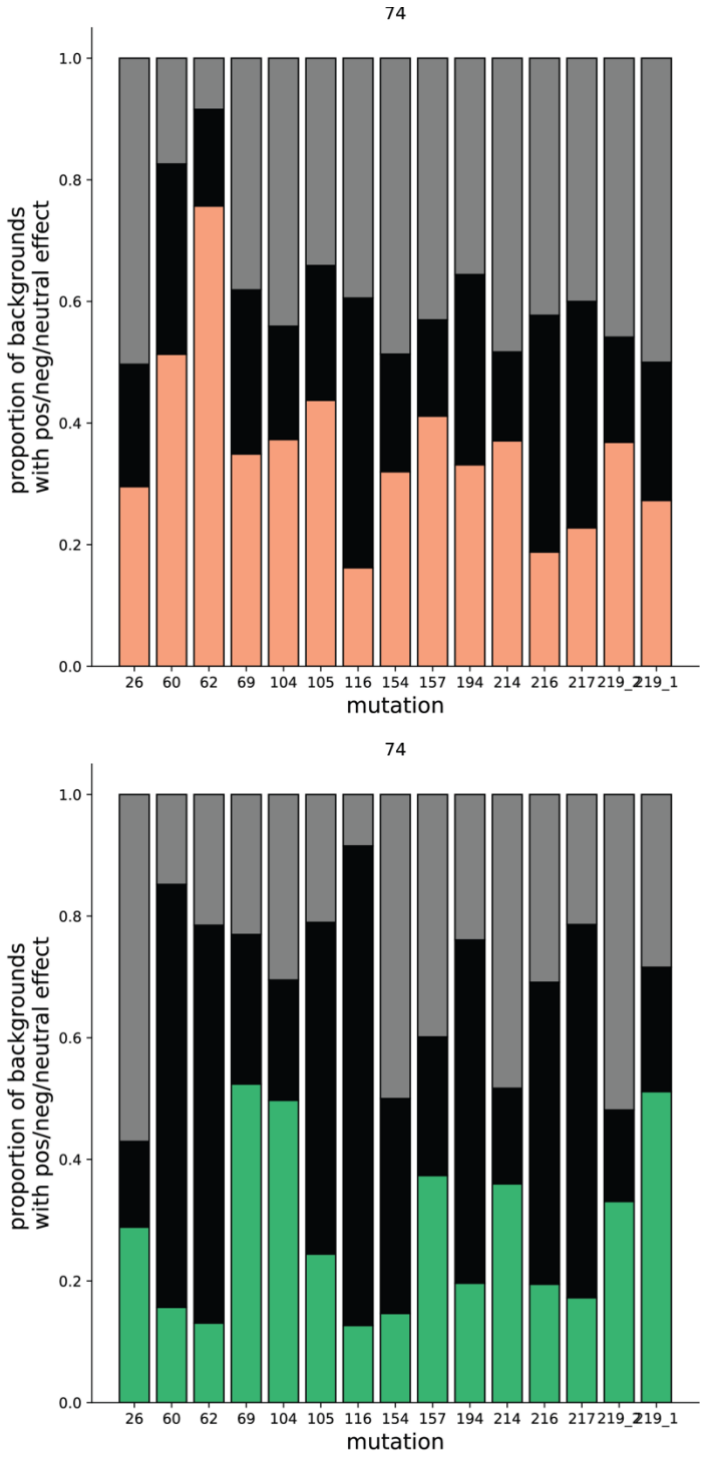
Supplementary Fig C9 Effects of each mutation in the E26V background. Top: Effect on red fluorescence, Bottom: effect on green fluorescence. Grey = neutral, black = deleterious, green = beneficial green, salmon = beneficial red.



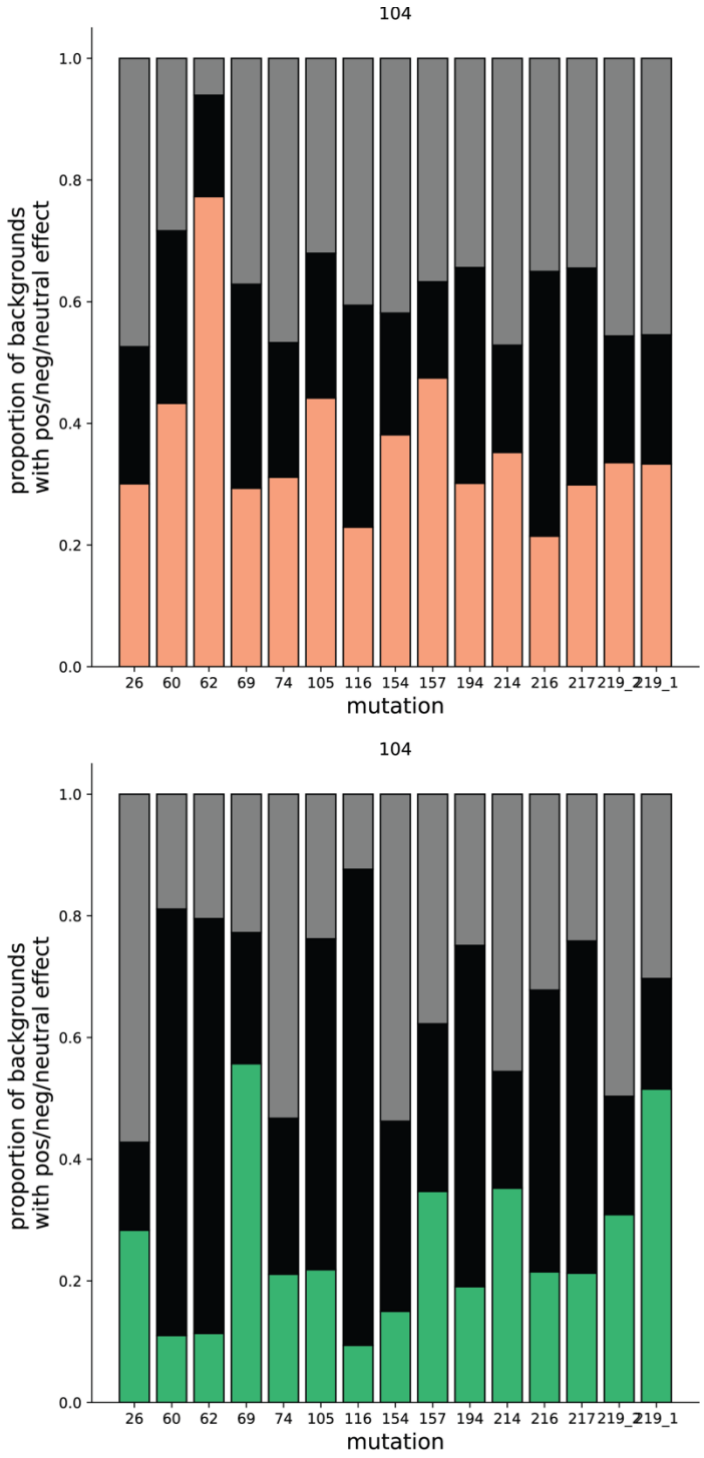
Supplementary Fig C10 Effects of each mutation in the A60V background. Top: Effect on red fluorescence, Bottom: effect on green fluorescence. Grey = neutral, black = deleterious, green = beneficial green, salmon = beneficial red.



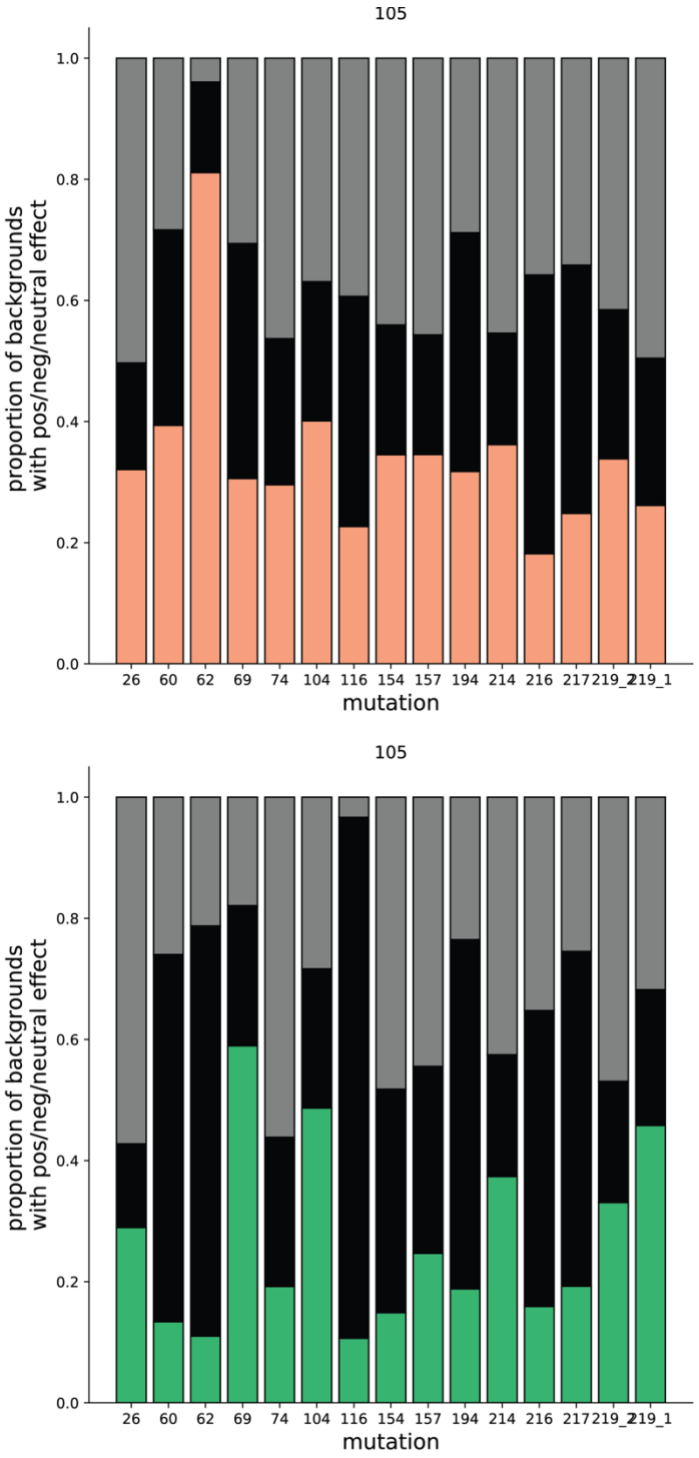
Supplementary Fig C11 Effects of each mutation in the T69A background. Top: Effect on red fluorescence, Bottom: effect on green fluorescence. Grey = neutral, black = deleterious, green = beneficial green, salmon = beneficial red.



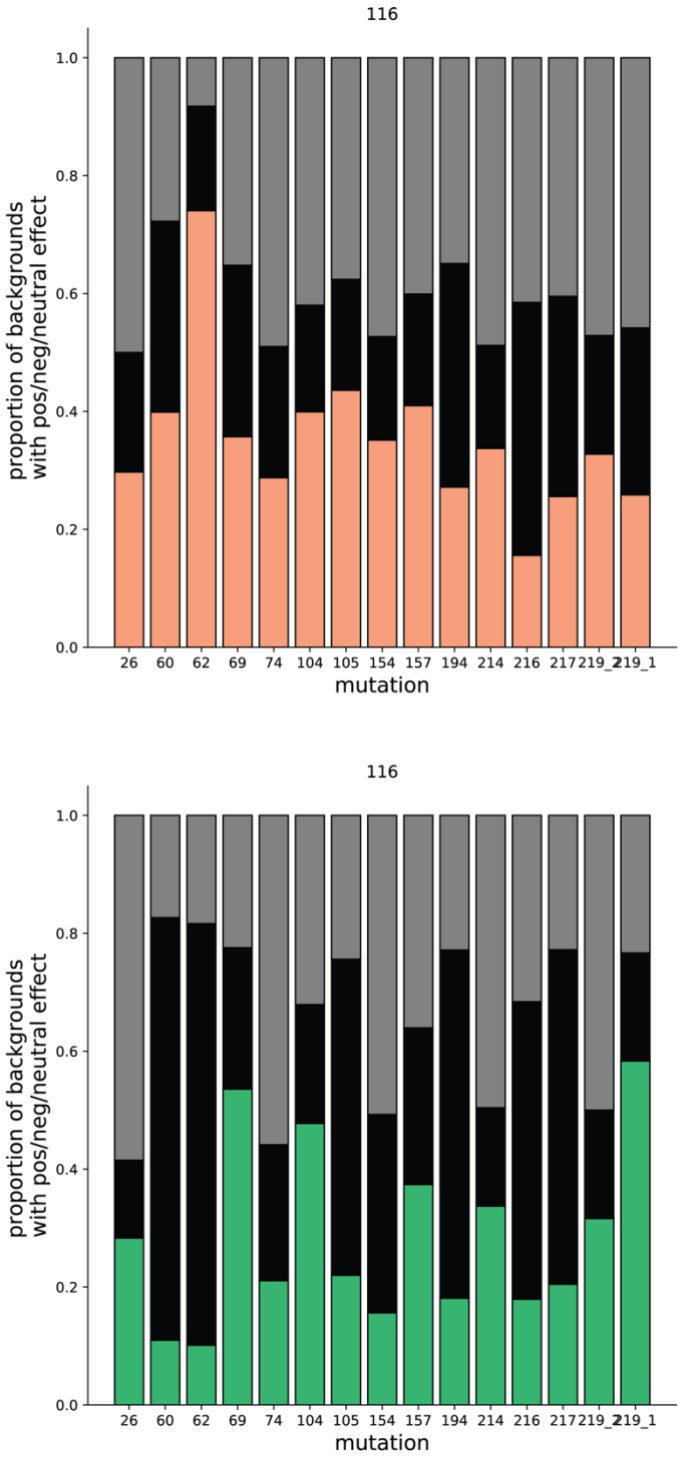
Supplementary Fig C12 Effects of each mutation in the D74H background. Top: Effect on red fluorescence, Bottom: effect on green fluorescence. Grey = neutral, black = deleterious, green = beneficial green, salmon = beneficial red.



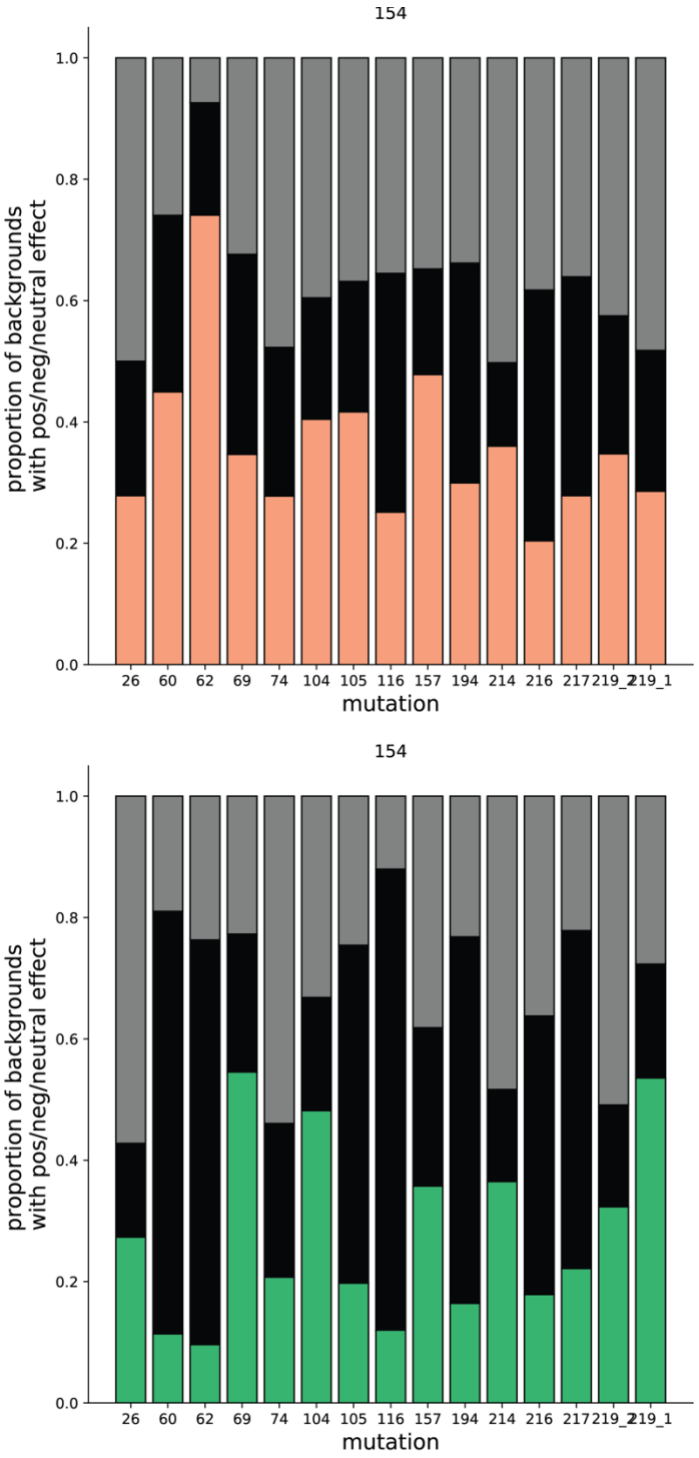
Supplementary Fig C13 Effects of each mutation in the T104R background. Top: Effect on red fluorescence, Bottom: effect on green fluorescence. Grey = neutral, black = deleterious, green = beneficial green, salmon = beneficial red.



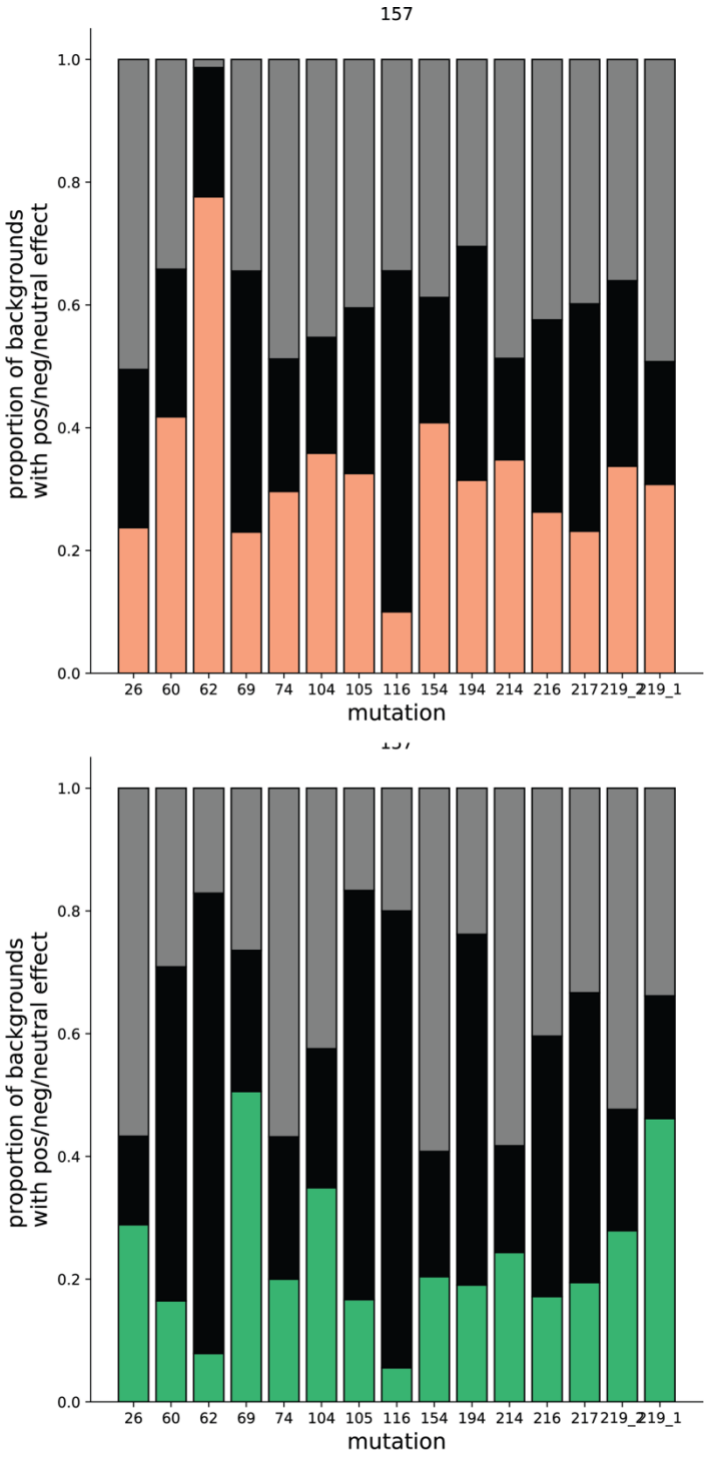
Supplementary Fig C14 Effects of each mutation in the S105N background. Top: Effect on red fluorescence, Bottom: effect on green fluorescence. Grey = neutral, black = deleterious, green = beneficial green, salmon = beneficial red.



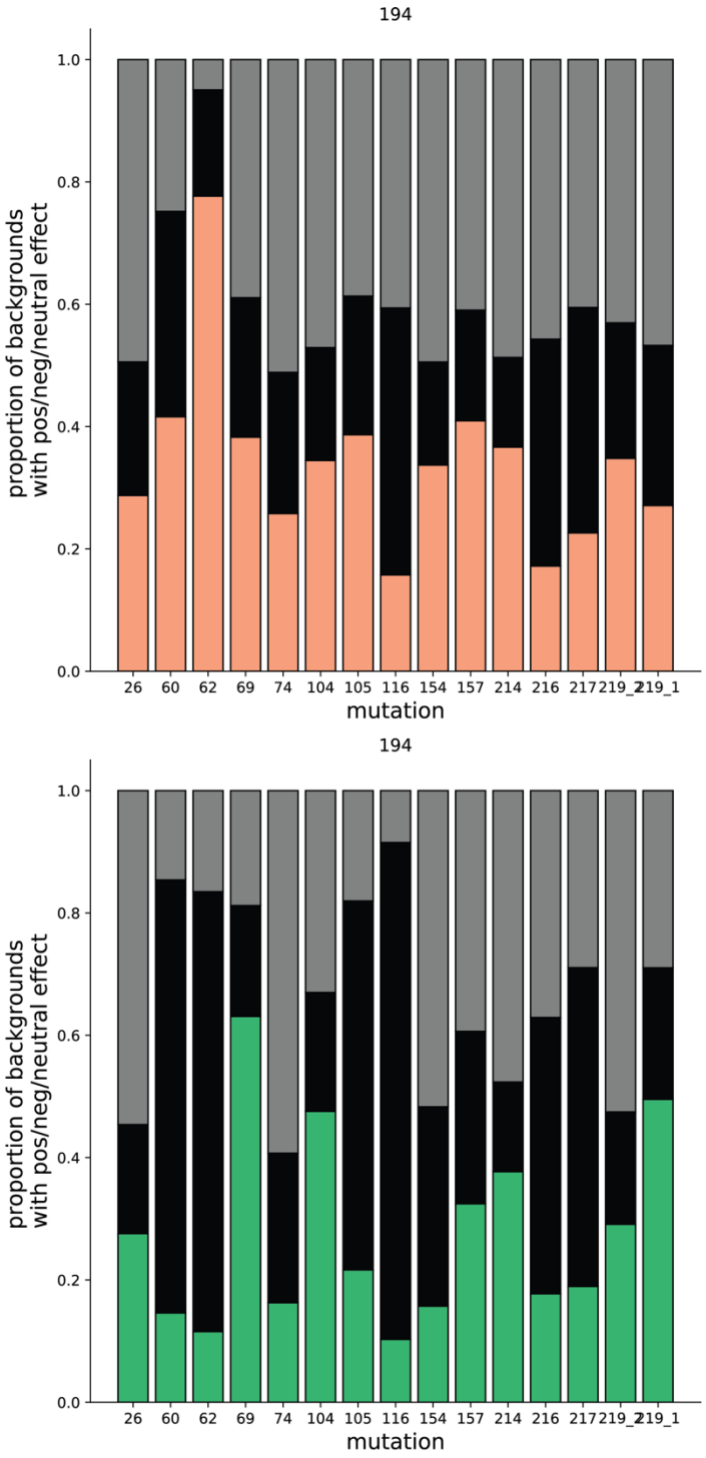
Supplementary Fig C15 Effects of each mutation in the Y116N background. Top: Effect on red fluorescence, Bottom: effect on green fluorescence. Grey = neutral, black = deleterious, green = beneficial green, salmon = beneficial red.



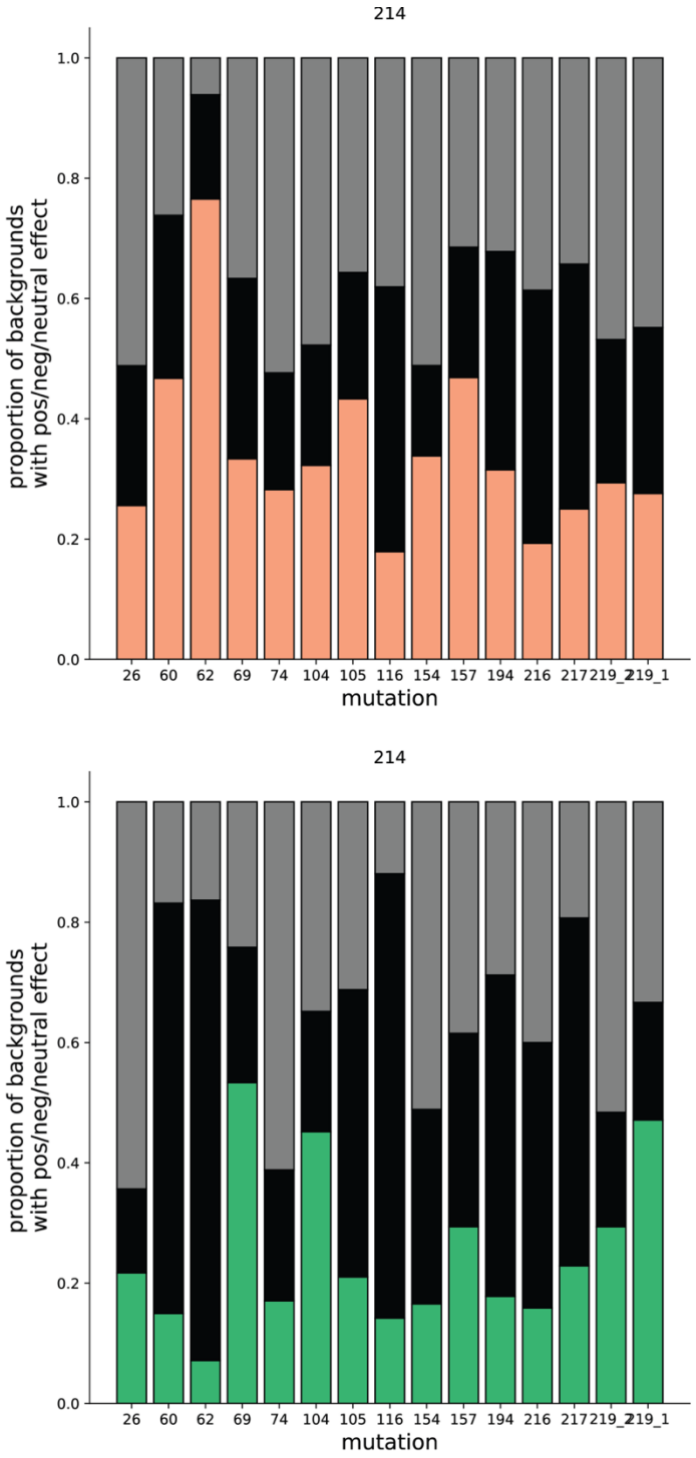
Supplementary Fig C16 Effects of each mutation in the M154T background. Top: Effect on red fluorescence, Bottom: effect on green fluorescence. Grey = neutral, black = deleterious, green = beneficial green, salmon = beneficial red.



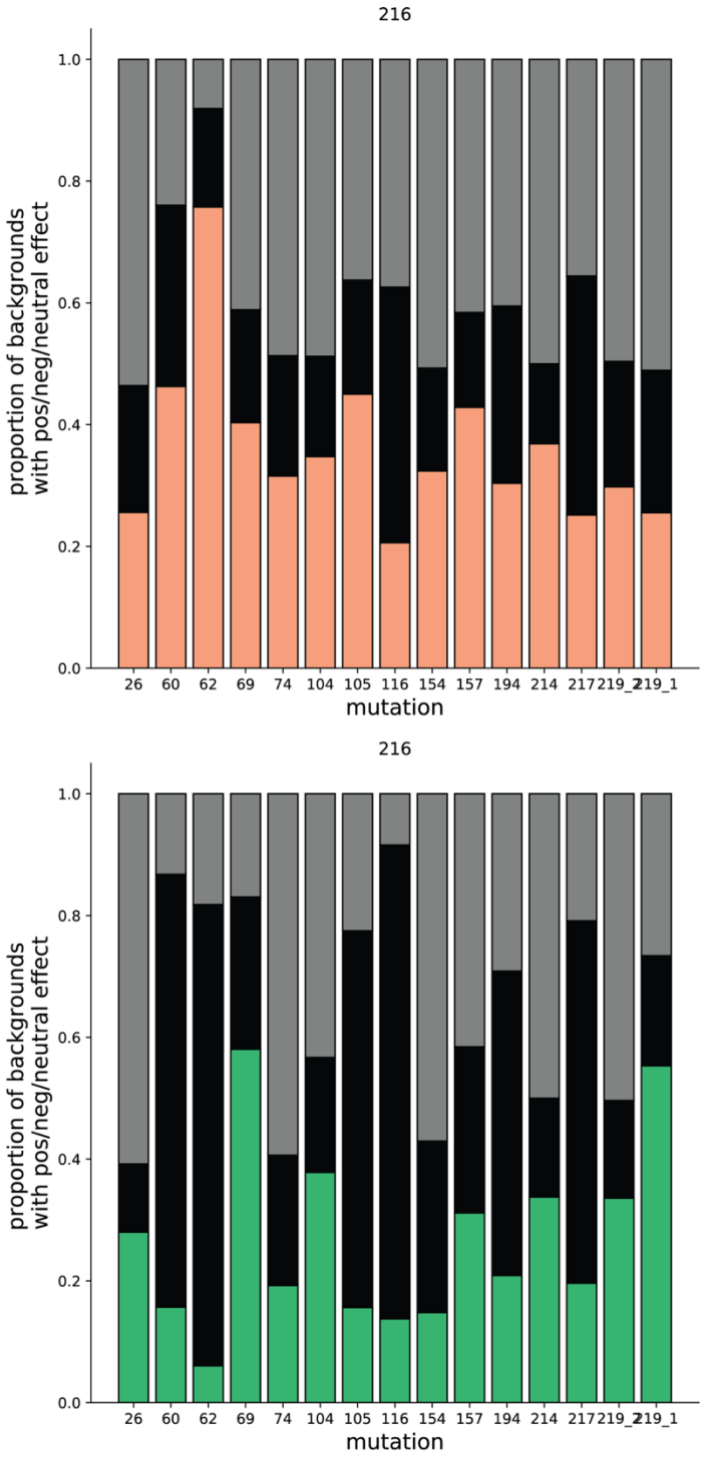
Supplementary Fig C17 Effects of each mutation in the V157I background. Top: Effect on red fluorescence, Bottom: effect on green fluorescence. Grey = neutral, black = deleterious, green = beneficial green, salmon = beneficial red.



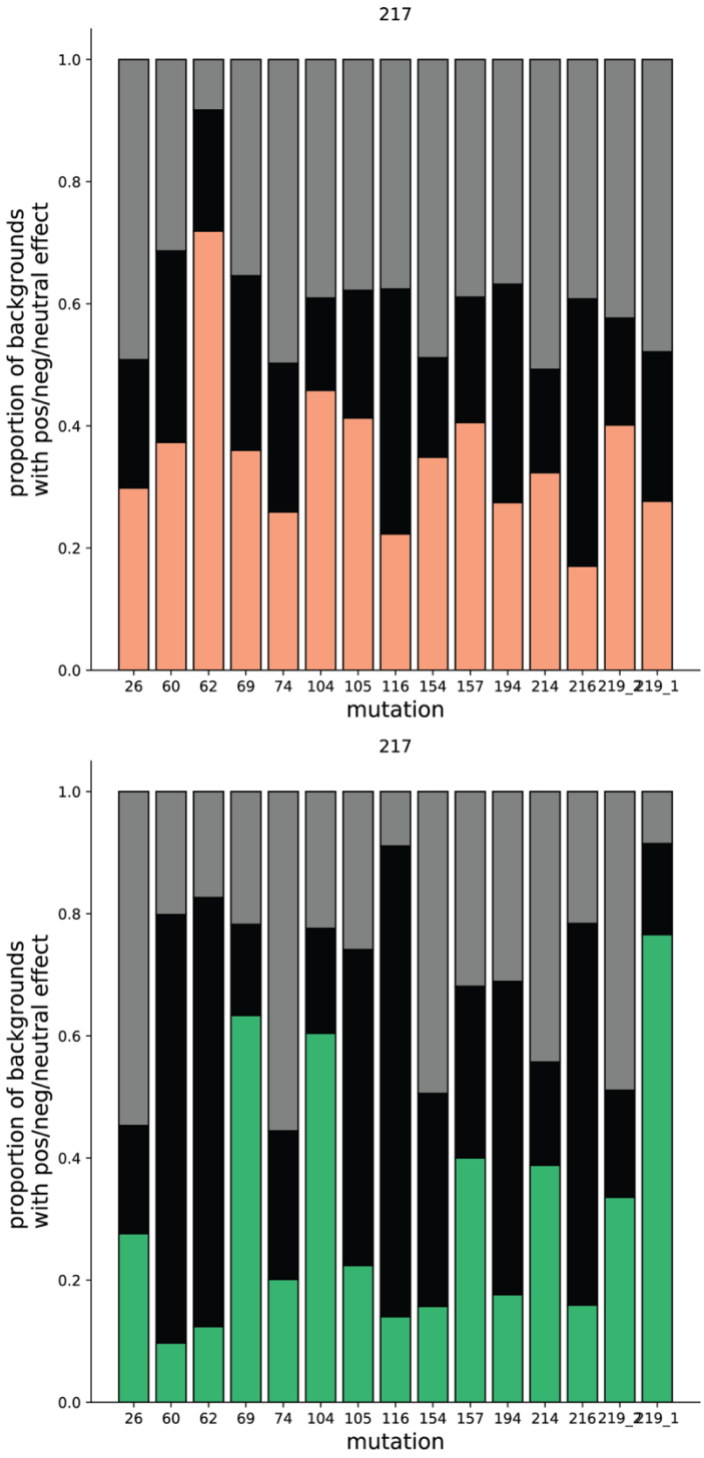
Supplementary Fig C18 Effects of each mutation in the R194C background. Top: Effect on red fluorescence, Bottom: effect on green fluorescence. Grey = neutral, black = deleterious, green = beneficial green, salmon = beneficial red.



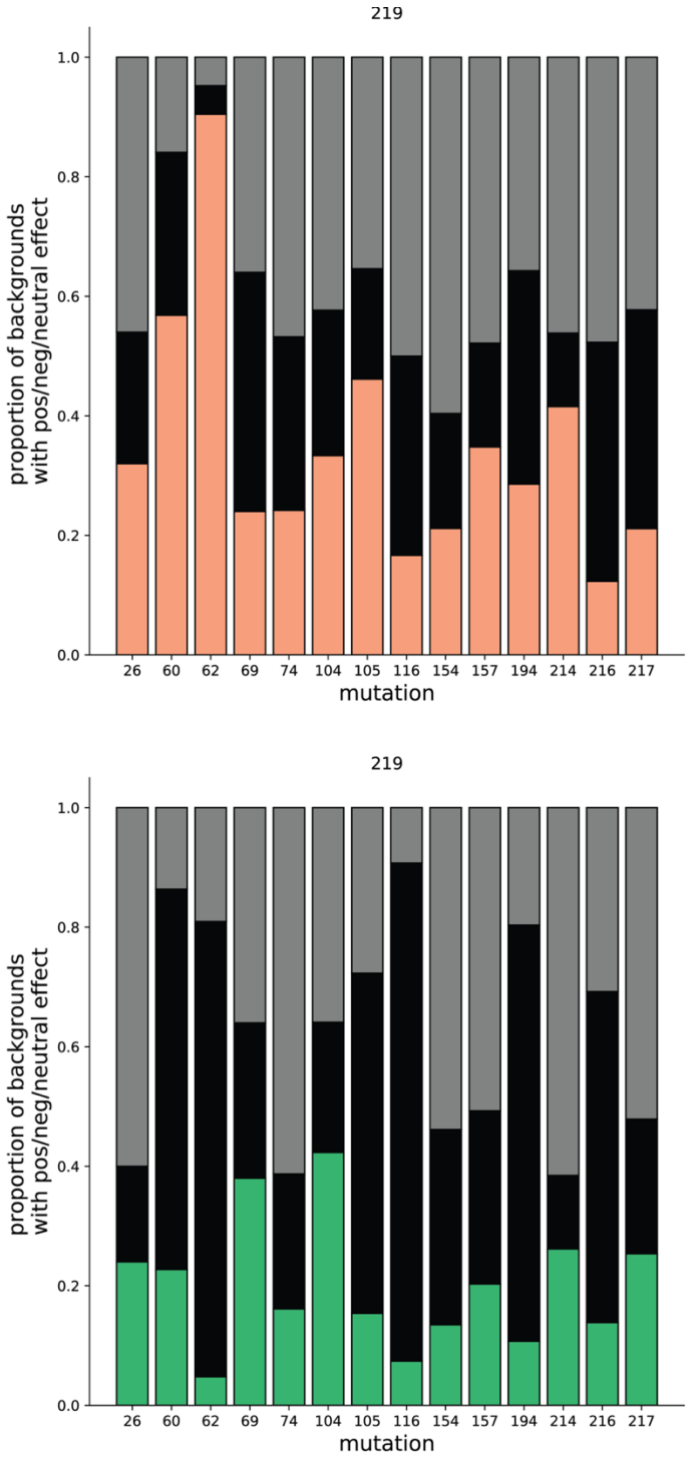
Supplementary Fig C19 Effects of each mutation in the V214E background. Top: Effect on red fluorescence, Bottom: effect on green fluorescence. Grey = neutral, black = deleterious, green = beneficial green, salmon = beneficial red.



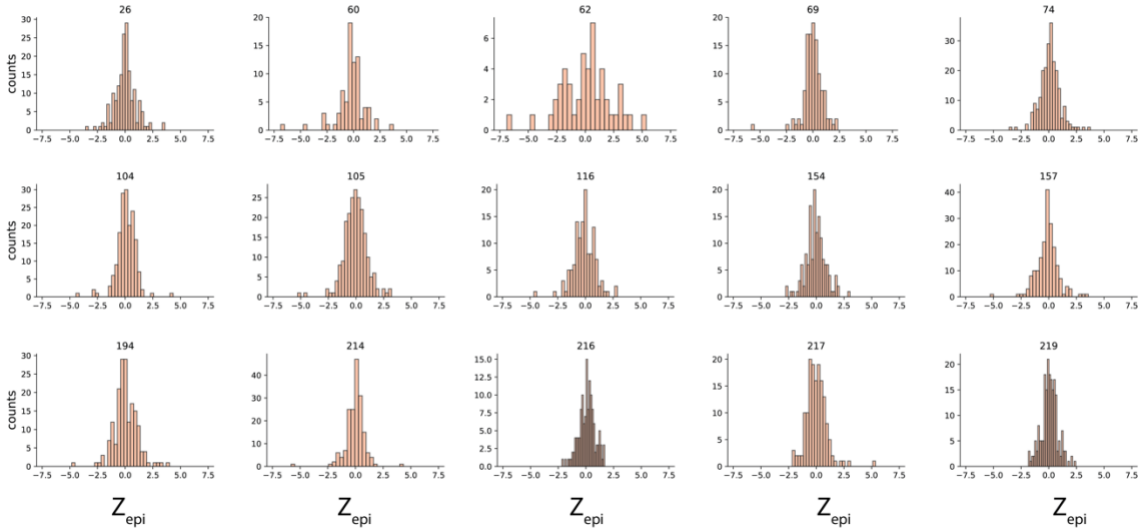
Supplementary Fig C20 Effects of each mutation in the R216H background. Top: Effect on red fluorescence, Bottom: effect on green fluorescence. Grey = neutral, black = deleterious, green = beneficial green, salmon = beneficial red.



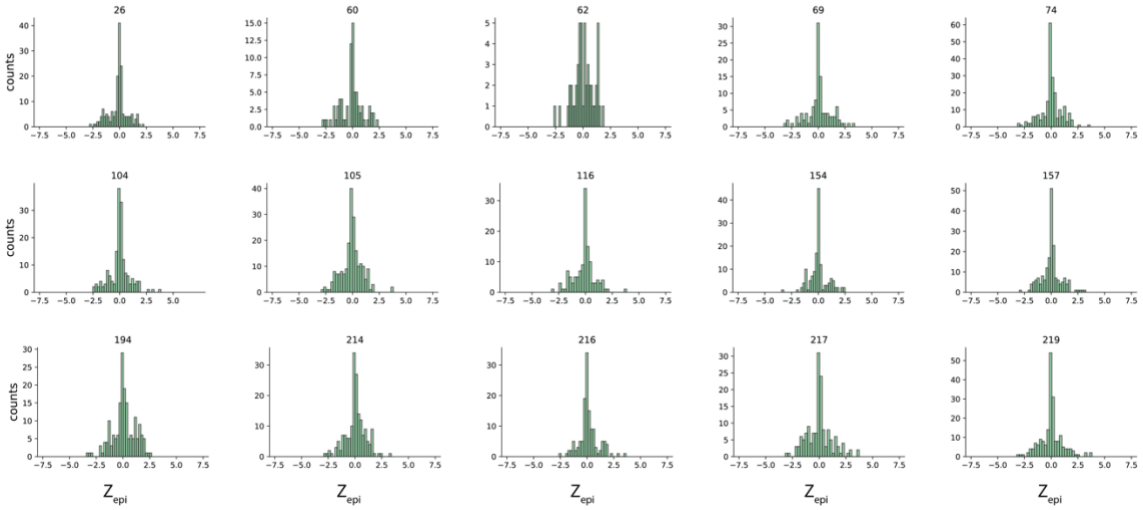
Supplementary Fig C21 Effects of each mutation in the Y217Del background. Top: Effect on red fluorescence, Bottom: effect on green fluorescence. Grey = neutral, black = deleterious, green = beneficial green, salmon = beneficial red.



Supplementary Fig C22 Effects of each mutation in the M219 background. Top: Effect on red fluorescence, Bottom: effect on green fluorescence. Grey = neutral, black = deleterious, green = beneficial green, salmon = beneficial red.



Supplementary Fig C23 Distributions of Z-scores for pairwise epistasis in red fluorescence for all substitutions. Z_{epi} scores for red fluorescence is shown on the x-axis and counts are shown on the y-axis. Each plot is a different genotype, which is indicated at the top center of each plot.



Supplementary Fig C24 Distributions of Z-scores for pairwise epistasis in green fluorescence for all substitutions. Z_{epi} scores for green fluorescence is shown on the x-axis and counts are shown on the y-axis. Each plot is a different genotype, which is indicated at the top center of each plot.

REFERENCES CITED

(1) Bloom, J. D.; Gong, L. I.; Baltimore, D. Permissive Secondary Mutations Enable the Evolution of Influenza Oseltamivir Resistance. *Science* **2010**, *328* (5983), 1272–1275. https://doi.org/10.1126/science.1187816.

(2) Roe, A. M.; Shur, N. From New Screens to Discovered Genes: The Successful Past and Promising Present of Single Gene Disorders. *Am. J. Med. Genet. C Semin. Med. Genet.* **2007**, *145C* (1), 77–86. https://doi.org/10.1002/ajmg.c.30121.

(3) Maynard Smith, J. Natural Selection and the Concept of a Protein Space. *Nature* **1970**, *225* (5232), 563–564. https://doi.org/10.1038/225563a0.

(4) Currin, A.; Swainston, N.; J. Day, P.; B. Kell, D. Synthetic Biology for the Directed Evolution of Protein Biocatalysts: Navigating Sequence Space Intelligently. *Chem. Soc. Rev.* **2015**, *44* (5), 1172–1239. https://doi.org/10.1039/C4CS00351A.

(5) de Visser, J. A. G. M.; Krug, J. Empirical Fitness Landscapes and the Predictability of Evolution. *Nat. Rev. Genet.* **2014**, *15* (7), 480–490. https://doi.org/10.1038/nrg3744.

(6) Fowler, D. M.; Fields, S. Deep Mutational Scanning: A New Style of Protein Science. *Nat. Methods* **2014**, *11* (8), 801–807. https://doi.org/10.1038/nmeth.3027.

(7) O'Maille, P. E.; Malone, A.; Dellas, N.; Andes Hess, B.; Smentek, L.; Sheehan, I.; Greenhagen, B. T.; Chappell, J.; Manning, G.; Noel, J. P. Quantitative Exploration of the Catalytic Landscape Separating Divergent Plant Sesquiterpene Synthases. *Nat. Chem. Biol.* **2008**, *4* (10), 617–623. https://doi.org/10.1038/nchembio.113.

(8) Szendro, I. G.; Schenk, M. F.; Franke, J.; Krug, J.; Visser, J. A. G. M. de. Quantitative Analyses of Empirical Fitness Landscapes. *J. Stat. Mech. Theory Exp.* **2013**, *2013* (01), P01005. https://doi.org/10.1088/1742-5468/2013/01/P01005.

(9) Weinreich, D. M.; Watson, R. A.; Chao, L. Perspective: Sign Epistasis and Genetic Costraint on Evolutionary Trajectories. *Evolution* **2005**, *59* (6), 1165–1174. https://doi.org/10.1111/j.0014-3820.2005.tb01768.x.

(10) Weinreich, D. M.; Lan, Y.; Wylie, C. S.; Heckendorn, R. B. Should Evolutionary Geneticists Worry about Higher-Order Epistasis? *Curr. Opin. Genet. Dev.* **2013**, *23* (6), 700–707. https://doi.org/10.1016/j.gde.2013.10.007.

(11) Breen, M. S.; Kemena, C.; Vlasov, P. K.; Notredame, C.; Kondrashov, F. A. Epistasis as the Primary Factor in Molecular Evolution. *Nature* **2012**, *490* (7421), 535–538. https://doi.org/10.1038/nature11510.

(12) Gong, L. I.; Suchard, M. A.; Bloom, J. D. Stability-Mediated Epistasis Constrains the Evolution of an Influenza Protein. *eLife* **2013**, *2*, e00631. https://doi.org/10.7554/eLife.00631.

(13) Lunzer, M.; Golding, G. B.; Dean, A. M. Pervasive Cryptic Epistasis in Molecular Evolution. *PLOS Genet.* **2010**, *6* (10), e1001162. https://doi.org/10.1371/journal.pgen.1001162.

(14) Bershtein, S.; Segal, M.; Bekerman, R.; Tokuriki, N.; Tawfik, D. S. Robustness–Epistasis Link Shapes the Fitness Landscape of a Randomly Drifting Protein. *Nature* **2006**, *444* (7121), 929–932. https://doi.org/10.1038/nature05385.

(15) Anderson, D. W.; McKeown, A. N.; Thornton, J. W. Intermolecular Epistasis Shaped the Function and Evolution of an Ancient Transcription Factor and Its DNA Binding Sites. *eLife* **2015**, *4*, e07864. https://doi.org/10.7554/eLife.07864.

(16) Kvitek, D. J.; Sherlock, G. Reciprocal Sign Epistasis between Frequently Experimentally Evolved Adaptive Mutations Causes a Rugged Fitness Landscape. *PLoS Genet.* **2011**, *7* (4), e1002056. https://doi.org/10.1371/journal.pgen.1002056.

(17) Hall, D. W.; Agan, M.; Pope, S. C. Fitness Epistasis among 6 Biosynthetic Loci in the Budding Yeast Saccharomyces Cerevisiae. *J. Hered.* **2010**, *101* (suppl_1), S75–S84. https://doi.org/10.1093/jhered/esq007.

(18) Weinreich, D. M.; Delaney, N. F.; DePristo, M. A.; Hartl, D. L. Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science* **2006**, *312* (5770), 111–114. https://doi.org/10.1126/science.1123539.

(19) Poelwijk, F. J.; Kiviet, D. J.; Weinreich, D. M.; Tans, S. J. Empirical Fitness Landscapes Reveal Accessible Evolutionary Paths. *Nature* **2007**, *445* (7126), 383–386. https://doi.org/10.1038/nature05451.

(20) Poelwijk, F. J.; Krishna, V.; Ranganathan, R. The Context-Dependence of Mutations: A Linkage of Formalisms. *PLOS Comput. Biol.* **2016**, *12* (6), e1004771. https://doi.org/10.1371/journal.pcbi.1004771.

(21) Cordell, H. J. Epistasis: What It Means, What It Doesn't Mean, and Statistical Methods to Detect It in Humans. *Hum. Mol. Genet.* **2002**, *11* (20), 2463–2468. https://doi.org/10.1093/hmg/11.20.2463.

(22) Phillips, P. C. Epistasis — the Essential Role of Gene Interactions in the Structure and Evolution of Genetic Systems. *Nat. Rev. Genet.* **2008**, *9* (11), 855–867. https://doi.org/10.1038/nrg2452.

(23) Sailer, Z. R.; Harms, M. J. High-Order Epistasis Shapes Evolutionary Trajectories. *PLOS Comput. Biol.* **2017**, *13* (5), e1005541. https://doi.org/10.1371/journal.pcbi.1005541.

(24) Starr, T. N.; Thornton, J. W. Epistasis in Protein Evolution. *Protein Sci.* **2016**, *25* (7), 1204–1218. https://doi.org/10.1002/pro.2897.

(25) DePristo, M. A.; Weinreich, D. M.; Hartl, D. L. Missense Meanderings in Sequence Space: A Biophysical View of Protein Evolution. *Nat. Rev. Genet.* **2005**, *6* (9), 678–687. https://doi.org/10.1038/nrg1672.

(26) Weinreich, D. M.; Watson, R. A.; Chao, L. Perspective: Sign Epistasis and Genetic Costraint on Evolutionary Trajectories. *Evolution* **2005**, *59* (6), 1165–1174. https://doi.org/10.1111/j.0014-3820.2005.tb01768.x.

(27) Chou, H.-H.; Chiu, H.-C.; Delaney, N. F.; Segrè, D.; Marx, C. J. Diminishing Returns Epistasis Among Beneficial Mutations Decelerates Adaptation. *Science* **2011**.

(28) Kryazhimskiy, S.; Rice, D. P.; Jerison, E. R.; Desai, M. M. Global Epistasis Makes Adaptation Predictable despite Sequence-Level Stochasticity. *Science* **2014**, *344* (6191), 1519–1522. https://doi.org/10.1126/science.1250939.

(29) Salverda, M. L. M.; Dellus, E.; Gorter, F. A.; Debets, A. J. M.; van der Oost, J.; Hoekstra, R. F.; Tawfik, D. S.; de Visser, J. A. G. M. Initial Mutations Direct Alternative Pathways of Protein Evolution. *PLoS Genet.* **2011**, *7* (3), e1001321. https://doi.org/10.1371/journal.pgen.1001321.

(30) Tufts, D. M.; Natarajan, C.; Revsbech, I. G.; Projecto-Garcia, J.; Hoffmann, F. G.; Weber, R. E.; Fago, A.; Moriyama, H.; Storz, J. F. Epistasis Constrains Mutational Pathways of Hemoglobin Adaptation in High-Altitude Pikas. *Mol. Biol. Evol.* **2015**, *32* (2), 287–298. https://doi.org/10.1093/molbev/msu311.

(31) Shah, P.; McCandlish, D. M.; Plotkin, J. B. Contingency and Entrenchment in Protein Evolution under Purifying Selection. *Proc. Natl. Acad. Sci.* **2015**, *112* (25), E3226–E3235. https://doi.org/10.1073/pnas.1412933112.

(32) Miton, C. M.; Tokuriki, N. How Mutational Epistasis Impairs Predictability in Protein Evolution and Design. *Protein Sci.* **2016**, *25* (7), 1260–1272. https://doi.org/10.1002/pro.2876.

(33) Sarkisyan, K. S.; Bolotin, D. A.; Meer, M. V.; Usmanova, D. R.; Mishin, A. S.; Sharonov, G. V.; Ivankov, D. N.; Bozhanova, N. G.; Baranov, M. S.; Soylemez, O.; Bogatyreva, N. S.; Vlasov, P. K.; Egorov, E. S.; Logacheva, M. D.; Kondrashov, A. S.; Chudakov, D. M.; Putintseva, E. V.; Mamedov, I. Z.; Tawfik, D. S.; Lukyanov, K. A.; Kondrashov, F. A. Local Fitness Landscape of the Green Fluorescent Protein. *Nature* **2016**, *533* (7603), 397–401. https://doi.org/10.1038/nature17995.

(34) Sailer, Z. R.; Harms, M. J. Detecting High-Order Epistasis in Nonlinear Genotype-Phenotype Maps. *Genetics* **2017**, *205* (3), 1079–1088. https://doi.org/10.1534/genetics.116.195214.

(35) Bateson, W. Heredity and Variation in Modern Lights. *Darwin Mod. Sci.* **1909**.

(36) Fisher, R. A. *The Correlation between Relatives on the Supposition of Mendelian Inheritance*; Trans R. Soc. Edin., 1918; Vol. 52.

(37) Miller, C.; Davlieva, M.; Wilson, C.; White, K. I.; Couñago, R.; Wu, G.; Myers, J. C.; Wittung-Stafshede, P.; Shamoo, Y. Experimental Evolution of Adenylate Kinase Reveals Contrasting Strategies toward Protein Thermostability. *Biophys. J.* **2010**, *99* (3), 887–896. https://doi.org/10.1016/j.bpj.2010.04.076.

(38) Dellus-Gur, E.; Elias, M.; Caselli, E.; Prati, F.; Salverda, M. L. M.; de Visser, J. A. G. M.; Fraser, J. S.; Tawfik, D. S. Negative Epistasis and Evolvability in TEM-1 β-Lactamase—The Thin Line between an Enzyme's Conformational Freedom and Disorder. *J. Mol. Biol.* **2015**, *427* (14), 2396–2409. https://doi.org/10.1016/j.jmb.2015.05.011.

(39) Otwinowski, J. Biophysical Inference of Epistasis and the Effects of Mutations on Protein Stability and Function. *Mol. Biol. Evol.* **2018**, *35* (10), 2345–2354. https://doi.org/10.1093/molbev/msy141.

(40) Adams, R. M.; Kinney, J. B.; Walczak, A. M.; Mora, T. Epistasis in a Fitness Landscape Defined by Antibody-Antigen Binding Free Energy. *Cell Syst.* **2019**, *8* (1), 86-93.e3. https://doi.org/10.1016/j.cels.2018.12.004.

(41) da Silva, J.; Coetzer, M.; Nedellec, R.; Pastore, C.; Mosier, D. E. Fitness Epistasis and Constraints on Adaptation in a Human Immunodeficiency Virus Type 1 Protein Region. *Genetics* **2010**, *185* (1), 293–303. https://doi.org/10.1534/genetics.109.112458.

(42) Bridgham, J. T.; Ortlund, E. A.; Thornton, J. W. An Epistatic Ratchet Constrains the Direction of Glucocorticoid Receptor Evolution. *Nature* **2009**, *461* (7263), 515–519. https://doi.org/10.1038/nature08249.

(43) Harms, M. J.; Thornton, J. W. Historical Contingency and Its Biophysical Basis in Glucocorticoid Receptor Evolution. *Nature* **2014**, *512* (7513), 203–207. https://doi.org/10.1038/nature13410.

(44) Ortlund, E. A.; Bridgham, J. T.; Redinbo, M. R.; Thornton, J. W. Crystal Structure of an Ancient Protein: Evolution by Conformational Epistasis. *Science* **2007**, *317* (5844), 1544–1548. https://doi.org/10.1126/science.1142819.

(45) Poelwijk, F. J.; Kiviet, D. J.; Weinreich, D. M.; Tans, S. J. Empirical Fitness Landscapes Reveal Accessible Evolutionary Paths. *Nature* **2007**, *445* (7126), 383–386. https://doi.org/10.1038/nature05451.

(46) Giger, L.; Caner, S.; Obexer, R.; Kast, P.; Baker, D.; Ban, N.; Hilvert, D. Evolution of a Designed Retro-Aldolase Leads to Complete Active Site Remodeling. *Nat. Chem. Biol.* **2013**, *9* (8), 494–498. https://doi.org/10.1038/nchembio.1276.

(47) Sailer, Z. R.; Harms, M. J. Molecular Ensembles Make Evolution Unpredictable. *Proc. Natl. Acad. Sci.* **2017**, *114* (45), 11938–11943. https://doi.org/10.1073/pnas.1711927114.

(48) Sykora, J.; Brezovsky, J.; Koudelakova, T.; Lahoda, M.; Fortova, A.; Chernovets, T.; Chaloupkova, R.; Stepankova, V.; Prokop, Z.; Smatanova, I. K.; Hof, M.; Damborsky, J. Dynamics and Hydration Explain Failed Functional Transformation in Dehalogenase Design. *Nat. Chem. Biol.* **2014**, *10* (6), 428–430. https://doi.org/10.1038/nchembio.1502.

(49) Harms, M. J.; Thornton, J. W. Evolutionary Biochemistry: Revealing the Historical and Physical Causes of Protein Properties. *Nat. Rev. Genet.* **2013**, *14* (8), 559–571. https://doi.org/10.1038/nrg3540.

(50) Dickinson, B. C.; Leconte, A. M.; Allen, B.; Esvelt, K. M.; Liu, D. R. Experimental Interrogation of the Path Dependence and Stochasticity of Protein Evolution Using Phage-Assisted Continuous Evolution. *Proc. Natl. Acad. Sci.* **2013**, *110* (22), 9007–9012. https://doi.org/10.1073/pnas.1220670110.

(51) Kaltenbach, M.; Jackson, C. J.; Campbell, E. C.; Hollfelder, F.; Tokuriki, N. Reverse Evolution Leads to Genotypic Incompatibility despite Functional and Active Site Convergence. *eLife* **2015**, *4*, e06492. https://doi.org/10.7554/eLife.06492.

(52) Alexander, P. A.; He, Y.; Chen, Y.; Orban, J.; Bryan, P. N. A Minimal Sequence Code for Switching Protein Structure and Function. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (50), 21149–21154. https://doi.org/10.1073/pnas.0906408106.

(53) Field, S. F.; Matz, M. V. Retracing Evolution of Red Fluorescence in GFP-Like Proteins from Faviina Corals. *Mol. Biol. Evol.* **2010**, *27* (2), 225–233. https://doi.org/10.1093/molbev/msp230.

(54) Halabi, N.; Rivoire, O.; Leibler, S.; Ranganathan, R. Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell* **2009**, *138* (4), 774–786. https://doi.org/10.1016/j.cell.2009.07.038.

(55) Khan, A. I.; Dinh, D. M.; Schneider, D.; Lenski, R. E.; Cooper, T. F. Negative Epistasis Between Beneficial Mutations in an Evolving Bacterial Population. *Science* **2011**, *332* (6034), 1193–1196. https://doi.org/10.1126/science.1203801.

(56) Lunzer, M.; Miller, S. P.; Felsheim, R.; Dean, A. M. The Biochemical Architecture of an Ancient Adaptive Landscape. *Science* **2005**, *310* (5747), 499–501. https://doi.org/10.1126/science.1115649.

(57) McKeown, A. N.; Bridgham, J. T.; Anderson, D. W.; Murphy, M. N.; Ortlund, E. A.; Thornton, J. W. Evolution of DNA Specificity in a Transcription Factor Family Produced a New Gene Regulatory Module. *Cell* **2014**, *159* (1), 58–68. https://doi.org/10.1016/j.cell.2014.09.003.

(58) Tokuriki, N.; Tawfik, D. S. Chaperonin Overexpression Promotes Genetic Variation and Enzyme Evolution. *Nature* **2009**, *459* (7247), 668–673. https://doi.org/10.1038/nature08009.

(59) Bloom, J. D.; Labthavikul, S. T.; Otey, C. R.; Arnold, F. H. Protein Stability Promotes Evolvability. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (15), 5869–5874. https://doi.org/10.1073/pnas.0510098103.

(60) Bloom, J. D.; Silberg, J. J.; Wilke, C. O.; Drummond, D. A.; Adami, C.; Arnold, F. H. Thermodynamic Prediction of Protein Neutrality. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (3), 606–611. https://doi.org/10.1073/pnas.0406744102.

(61) Sideraki, V.; Huang, W.; Palzkill, T.; Gilbert, H. F. A Secondary Drug Resistance Mutation of TEM-1 Beta-Lactamase That Suppresses Misfolding and Aggregation. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98* (1), 283–288. https://doi.org/10.1073/pnas.011454198.

(62) Beadle, B. M.; Shoichet, B. K. Structural Bases of Stability-Function Tradeoffs in Enzymes. *J. Mol. Biol.* **2002**, *321* (2), 285–296. https://doi.org/10.1016/s0022-2836(02)00599-5.

(63) Shoichet, B. K.; Baase, W. A.; Kuroki, R.; Matthews, B. W. A Relationship between Protein Stability and Protein Function. *Proc. Natl. Acad. Sci. U. S. A.* **1995**, *92* (2), 452–456. https://doi.org/10.1073/pnas.92.2.452.

(64) Bloom, J. D.; Arnold, F. H.; Wilke, C. O. Breaking Proteins with Mutations: Threads and Thresholds in Evolution. *Mol. Syst. Biol.* **2007**, *3*, 76. https://doi.org/10.1038/msb4100119.

(65) Novais, A.; Comas, I.; Baquero, F.; Cantón, R.; Coque, T. M.; Moya, A.; González-Candelas, F.; Galán, J.-C. Evolutionary Trajectories of Beta-Lactamase CTX-M-1 Cluster Enzymes: Predicting Antibiotic Resistance. *PLoS Pathog.* **2010**, *6* (1), e1000735. https://doi.org/10.1371/journal.ppat.1000735.

(66) Smock, R. G.; Gierasch, L. M. Sending Signals Dynamically. *Science* **2009**, *324* (5924), 198–203. https://doi.org/10.1126/science.1169377.

(67) Boehr, D. D.; McElheny, D.; Dyson, H. J.; Wright, P. E. The Dynamic Energy Landscape of Dihydrofolate Reductase Catalysis. *Science* **2006**, *313* (5793), 1638. https://doi.org/10.1126/science.1130258.

(68) del Sol, A.; Tsai, C.-J.; Ma, B.; Nussinov, R. The Origin of Allosteric Functional Modulation: Multiple Pre-Existing Pathways. *Structure* **2009**, *17* (8), 1042–1050. https://doi.org/10.1016/j.str.2009.06.008.

(69) Ma, B.; Nussinov, R. Enzyme Dynamics Point to Stepwise Conformational Selection in Catalysis. *Nanotechnol. MiniaturizationMechanisms* **2010**, *14* (5), 652–659. https://doi.org/10.1016/j.cbpa.2010.08.012.

(70) Motlagh, H. N.; Wrabl, J. O.; Li, J.; Hilser, V. J. The Ensemble Nature of Allostery. *Nature* **2014**, *508* (7496), 331–339. https://doi.org/10.1038/nature13001.

(71) Bunzel, H. A.; Anderson, J. L. R.; Hilvert, D.; Arcus, V. L.; van der Kamp, M. W.; Mulholland, A. J. Evolution of Dynamical Networks Enhances Catalysis in a Designer Enzyme. *Nat. Chem.* **2021**, 1–6. https://doi.org/10.1038/s41557-021-00763-6.

(72) James, L. C.; Roversi, P.; Tawfik, D. S. Antibody Multispecificity Mediated by Conformational Diversity. *Science* **2003**, *299* (5611), 1362. https://doi.org/10.1126/science.1079731.

(73) Yogurtcu, O. N.; Bora Erdemli, S.; Nussinov, R.; Turkay, M.; Keskin, O. Restricted Mobility of Conserved Residues in Protein-Protein Interfaces in Molecular Simulations. *Biophys. J.* **2008**, *94* (9), 3475–3485. https://doi.org/10.1529/biophysj.107.114835.

(74) Hilser, V. J. An Ensemble View of Allostery. *Science* **2010**, *327* (5966), 653–654. https://doi.org/10.1126/science.1186121.

(75) Franke, J.; Klözer, A.; Visser, J. A. G. M. de; Krug, J. Evolutionary Accessibility of Mutational Pathways. *PLOS Comput. Biol.* **2011**, *7* (8), e1002134. https://doi.org/10.1371/journal.pcbi.1002134.

(76) Yokoyama, S.; Xing, J.; Liu, Y.; Faggionato, D.; Altun, A.; Starmer, W. T. Epistatic Adaptive Evolution of Human Color Vision. *PLOS Genet* **2014**, *10* (12), e1004884. https://doi.org/10.1371/journal.pgen.1004884.

(77) Brown, K. M.; Costanzo, M. S.; Xu, W.; Roy, S.; Lozovsky, E. R.; Hartl, D. L. Compensatory Mutations Restore Fitness during the Evolution of Dihydrofolate Reductase. *Mol. Biol. Evol.* **2010**, *27* (12), 2682–2690. https://doi.org/10.1093/molbev/msq160.

(78) da Silva, J.; Coetzer, M.; Nedellec, R.; Pastore, C.; Mosier, D. E. Fitness Epistasis and Constraints on Adaptation in a Human Immunodeficiency Virus Type 1 Protein Region. *Genetics* **2010**, *185* (1), 293–303. https://doi.org/10.1534/genetics.109.112458.

(79) Gavrilets, S. *Fitness Landscapes and the Origin of Species*; Princeton University Press: Princeton, New Jersey, 2004.

(80) Gavrilets, S. Evolution and Speciation on Holey Adaptive Landscapes. *Trends Ecol. Evol.* **1997**, *12* (8), 307–312. https://doi.org/10.1016/S0169-5347(97)01098-7.

(81) de Visser, J. A. G. M.; Krug, J. Empirical Fitness Landscapes and the Predictability of Evolution. *Nat. Rev. Genet.* **2014**, *15* (7), 480–490. https://doi.org/10.1038/nrg3744.

(82) Lobkovsky, A. E.; Wolf, Y. I.; Koonin, E. V. Predictability of Evolutionary Trajectories in Fitness Landscapes. *PLOS Comput. Biol.* **2011**, *7* (12), e1002302. https://doi.org/10.1371/journal.pcbi.1002302.

(83) Palmer, A. C.; Toprak, E.; Baym, M.; Kim, S.; Veres, A.; Bershtein, S.; Kishony, R. Delayed Commitment to Evolutionary Fate in Antibiotic Resistance Fitness Landscapes. *Nat. Commun.* **2015**, *6*, 7385. https://doi.org/10.1038/ncomms8385.

(84) Zagorski, M.; Burda, Z.; Waclaw, B. Beyond the Hypercube: Evolutionary Accessibility of Fitness Landscapes with Realistic Mutational Networks. *PLOS Comput. Biol.* **2016**, *12* (12), e1005218. https://doi.org/10.1371/journal.pcbi.1005218.

(85) Franke, J.; Klözer, A.; Visser, J. A. G. M. de; Krug, J. Evolutionary Accessibility of Mutational Pathways. *PLOS Comput Biol* **2011**, *7* (8), e1002134. https://doi.org/10.1371/journal.pcbi.1002134.

(86) Wu, N. C.; Dai, L.; Olson, C. A.; Lloyd-Smith, J. O.; Sun, R. Adaptation in Protein Fitness Landscapes Is Facilitated by Indirect Paths. *eLife* **2016**, *5*, e16965. https://doi.org/10.7554/eLife.16965.

(87) Greenbury, S. F.; Schaper, S.; Ahnert, S. E.; Louis, A. A. Genetic Correlations Greatly Increase Mutational Robustness and Can Both Reduce and Enhance Evolvability. *PLOS Comput. Biol.* **2016**, *12* (3), e1004773. https://doi.org/10.1371/journal.pcbi.1004773.

(88) Wagner, A. Neutralism and Selectionism: A Network-Based Reconciliation. *Nat. Rev. Genet.* **2008**, *9* (12), 965–974. https://doi.org/10.1038/nrg2473.

(89) Draghi, J. A.; Parsons, T. L.; Wagner, G. P.; Plotkin, J. B. Mutational Robustness Can Facilitate Adaptation. *Nature* **2010**, *463* (7279), 353–355. https://doi.org/10.1038/nature08694.

(90) Bloom, J. D.; Labthavikul, S. T.; Otey, C. R.; Arnold, F. H. Protein Stability Promotes Evolvability. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (15), 5869–5874. https://doi.org/10.1073/pnas.0510098103.

(91) Gruner, W.; Giegerich, R.; Strothmann, D.; Reidys, C.; Weber, J.; Hofacker, I. L.; Stadler, P. F.; Schuster, P. *Analysis of RNA Sequence Structure Maps by Exhaustive Enumeration*; Working Paper 95-10–099; Santa Fe Institute, 1995.

(92) Martinez, M. A.; Pezo, V.; Marlière, P.; Wain-Hobson, S. Exploring the Functional Robustness of an Enzyme by in Vitro Evolution. *EMBO J.* **1996**, *15* (6), 1203–1210.

(93) Schuster, P.; Fontana, W.; Stadler, P. F.; Hofacker, I. L. From Sequences to Shapes and Back: A Case Study in RNA Secondary Structures. *Proc. Biol. Sci.* **1994**, *255* (1344), 279–284. https://doi.org/10.1098/rspb.1994.0040.

(94) Ferrada, E.; Wagner, A. A Comparison of Genotype-Phenotype Maps for RNA and Proteins. *Biophys. J.* **2012**, *102* (8), 1916–1925. https://doi.org/10.1016/j.bpj.2012.01.047.

(95) Bornberg-Bauer, E. How Are Model Protein Structures Distributed in Sequence Space? *Biophys. J.* **1997**, *73* (5), 2393–2403. https://doi.org/10.1016/S0006-3495(97)78268-7.

(96) Rost, B. Protein Structures Sustain Evolutionary Drift. *Fold. Des.* **1997**, *2, Supplement 1*, S19–S24. https://doi.org/10.1016/S1359-0278(97)00059-X.

(97) Aguirre, J.; Buldú, J. M.; Stich, M.; Manrubia, S. C. Topological Structure of the Space of Phenotypes: The Case of RNA Neutral Networks. *PLOS ONE* **2011**, *6* (10), e26324. https://doi.org/10.1371/journal.pone.0026324.

(98) Ugalde, J. A.; Chang, B. S. W.; Matz, M. V. Evolution of Coral Pigments Recreated. *Science* **2004**, *305* (5689), 1433–1433. https://doi.org/10.1126/science.1099597.

(99) Alexander, P. A.; He, Y.; Chen, Y.; Orban, J.; Bryan, P. N. A Minimal Sequence Code for Switching Protein Structure and Function. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (50), 21149–21154. https://doi.org/10.1073/pnas.0906408106.

(100) Baier, F.; Hong, N.; Yang, G.; Pabis, A.; Miton, C. M.; Barrozo, A.; Carr, P. D.; Kamerlin, S. C.; Jackson, C. J.; Tokuriki, N. Cryptic Genetic Variation Shapes the Adaptive Evolutionary Potential of Enzymes. *eLife* **2019**, *8*, e40789. https://doi.org/10.7554/eLife.40789.

(101) Maisnier-Patin, S.; Paulander, W.; Pennhag, A.; Andersson, D. I. Compensatory Evolution Reveals Functional Interactions between Ribosomal Proteins S12, L14 and L19. *J. Mol. Biol.* **2007**, *366* (1), 207–215. https://doi.org/10.1016/j.jmb.2006.11.047.

(102) Yang, G.; Anderson, D. W.; Baier, F.; Dohmen, E.; Hong, N.; Carr, P. D.; Kamerlin, S. C. L.; Jackson, C. J.; Bornberg-Bauer, E.; Tokuriki, N. Higher-Order Epistasis Shapes the Fitness Landscape of a Xenobiotic-Degrading Enzyme. *Nat. Chem. Biol.* **2019**, *15* (11), 1120–1128. https://doi.org/10.1038/s41589-019-0386-3.

(103) Yokoyama, S.; Xing, J.; Liu, Y.; Faggionato, D.; Altun, A.; Starmer, W. T. Epistatic Adaptive Evolution of Human Color Vision. *PLOS Genet.* **2014**, *10* (12), e1004884. https://doi.org/10.1371/journal.pgen.1004884.

(104) Ancel, L. W.; Fontana, W. Plasticity, Evolvability, and Modularity in RNA. *J. Exp. Zool.* **2000**, *288* (3), 242–283. https://doi.org/10.1002/1097-010X(20001015)288:3<242::AID-JEZ5>3.0.CO;2-O.

(105) Sailer, Z. R.; Harms, M. J. Molecular Ensembles Make Evolution Unpredictable. *Proc. Natl. Acad. Sci.* **2017**, *114* (45), 11938–11943. https://doi.org/10.1073/pnas.1711927114.

(106) Tsai, C.-J.; Nussinov, R. A Unified View of "How Allostery Works." *PLOS Comput. Biol.* **2014**, *10* (2), e1003394. https://doi.org/10.1371/journal.pcbi.1003394.

(107) Wei, G.; Xi, W.; Nussinov, R.; Ma, B. Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell. *Chem. Rev.* **2016**, *116* (11), 6516–6551. https://doi.org/10.1021/acs.chemrev.5b00562.

(108) Ecsédi, P.; Kiss, B.; Gógl, G.; Radnai, L.; Buday, L.; Koprivanacz, K.; Liliom, K.; Leveles, I.; Vértessy, B.; Jeszenői, N.; Hetényi, C.; Schlosser, G.; Katona, G.; Nyitray, L. Regulation of the Equilibrium between Closed and Open Conformations of Annexin A2 by N-Terminal Phosphorylation and S100A4-Binding. *Structure* **2017**, *25* (8), 1195-1207.e5. https://doi.org/10.1016/j.str.2017.06.001.

(109) Malashkevich, V. N.; Varney, K. M.; Garrett, S. C.; Wilder, P. T.; Knight, D.; Charpentier, T. H.; Ramagopal, U. A.; Almo, S. C.; Weber, D. J.; Bresnick, A. R. Structure of Ca2+-Bound S100A4 and Its Interaction with Peptides Derived from Nonmuscle Myosin-IIA. *Biochemistry* **2008**, *47* (18), 5111–5126. https://doi.org/10.1021/bi702537s.

(110) Vallely, K. M.; Rustandi, R. R.; Ellis, K. C.; Varlamova, O.; Bresnick, A. R.; Weber, D. J. Solution Structure of Human Mts1 (S100A4) As Determined by NMR Spectroscopy. *Biochemistry* **2002**, *41* (42), 12670–12680. https://doi.org/10.1021/bi020365r.

(111) Garrett, S. C.; Hodgson, L.; Rybin, A.; Toutchkine, A.; Hahn, K. M.; Lawrence, D. S.; Bresnick, A. R. A Biosensor of S100A4 Metastasis Factor Activation: Inhibitor Screening and Cellular Activation Dynamics. *Biochemistry* **2008**, *47* (3), 986–996. https://doi.org/10.1021/bi7021624.

(112) Dellus-Gur, E.; Elias, M.; Caselli, E.; Prati, F.; Salverda, M. L. M.; de Visser, J. A. G. M.; Fraser, J. S.; Tawfik, D. S. Negative Epistasis and Evolvability in TEM-1 β-Lactamase—The Thin Line between an Enzyme's Conformational Freedom and Disorder. *J. Mol. Biol.* **2015**, *427* (14), 2396–2409. https://doi.org/10.1016/j.jmb.2015.05.011.

(113) Seeliger, M. A.; Nagar, B.; Frank, F.; Cao, X.; Henderson, M. N.; Kuriyan, J. C-Src Binds to the Cancer Drug Imatinib with an Inactive Abl/c-Kit Conformation and a Distributed Thermodynamic Penalty. *Structure* **2007**, *15* (3), 299–311. https://doi.org/10.1016/j.str.2007.01.015.

(114) Wilson, C.; Agafonov, R. V.; Hoemberger, M.; Kutter, S.; Zorba, A.; Halpin, J.; Buosi, V.; Otten, R.; Waterman, D.; Theobald, D. L.; Kern, D. Using Ancient Protein Kinases to Unravel a Modern Cancer Drug's Mechanism. *Science* **2015**, *347* (6224), 882–886. https://doi.org/10.1126/science.aaa1823.

(115) Bessonnard, S.; De Mot, L.; Gonze, D.; Barriol, M.; Dennis, C.; Goldbeter, A.; Dupont, G.; Chazaud, C. Gata6, Nanog and Erk Signaling Control Cell Fate in the Inner Cell Mass through a Tristable Regulatory Network. *Development* **2014**, *141* (19), 3637–3648. https://doi.org/10.1242/dev.109678.

(116) Hameri, T.; Boldi, M.-O.; Hatzimanikatis, V. Statistical Inference in Ensemble Modeling of Cellular Metabolism. *PLOS Comput. Biol.* **2019**, *15* (12), e1007536. https://doi.org/10.1371/journal.pcbi.1007536.

(117) Khazaei, T.; McGuigan, A. P.; Mahadevan, R. Ensemble Modeling of Cancer Metabolism. *Front. Physiol.* **2012**, *3*. https://doi.org/10.3389/fphys.2012.00135.

(118) Lu, M.; Jolly, M. K.; Gomoto, R.; Huang, B.; Onuchic, J.; Ben-Jacob, E. Tristability in Cancer-Associated MicroRNA-TF Chimera Toggle Switch. *J. Phys. Chem. B* **2013**, *117* (42), 13164–13174. https://doi.org/10.1021/jp403156m.

(119) Tran, L. M.; Rizk, M. L.; Liao, J. C. Ensemble Modeling of Metabolic Networks. *Biophys. J.* **2008**, *95* (12), 5606–5617. https://doi.org/10.1529/biophysj.108.135442.

(120) Venturi, V.; Kerényi, Á.; Reiz, B.; Bihary, D.; Pongor, S. Locality versus Globality in Bacterial Signalling: Can Local Communication Stabilize Bacterial Communities? *Biol. Direct* **2010**, *5* (1), 30. https://doi.org/10.1186/1745-6150-5-30.

(121) Bershtein, S.; Segal, M.; Bekerman, R.; Tokuriki, N.; Tawfik, D. S. Robustness–Epistasis Link Shapes the Fitness Landscape of a Randomly Drifting Protein. *Nature* **2006**, *444* (7121), 929–932. https://doi.org/10.1038/nature05385.

(122) Gong, L. I.; Suchard, M. A.; Bloom, J. D. Stability-Mediated Epistasis Constrains the Evolution of an Influenza Protein. *eLife* **2013**, *2*, e00631. https://doi.org/10.7554/eLife.00631.

(123) Kumar, A.; Natarajan, C.; Moriyama, H.; Witt, C. C.; Weber, R. E.; Fago, A.; Storz, J. F. Stability-Mediated Epistasis Restricts Accessible Mutational Pathways in the Functional Evolution of Avian Hemoglobin. *Mol. Biol. Evol.* **2017**, *34* (5), 1240–1251. https://doi.org/10.1093/molbev/msx085.

(124) Petrović, D.; Risso, V. A.; Kamerlin, S. C. L.; Sanchez-Ruiz, J. M. Conformational Dynamics and Enzyme Evolution. *J. R. Soc. Interface* **2018**, *15* (144), 20180330. https://doi.org/10.1098/rsif.2018.0330.

(125) Otwinowski, J.; McCandlish, D. M.; Plotkin, J. B. Inferring the Shape of Global Epistasis. *Proc. Natl. Acad. Sci.* **2018**, *115* (32), E7550–E7558. https://doi.org/10.1073/pnas.1804015115.

(126) Bridgham, J. T.; Carroll, S. M.; Thornton, J. W. Evolution of Hormone-Receptor Complexity by Molecular Exploitation. *Science* **2006**, *312* (5770), 97–101. https://doi.org/10.1126/science.1123348.

(127) Chiotti, K. E.; Kvitek, D. J.; Schmidt, K. H.; Koniges, G.; Schwartz, K.; Donckels, E. A.; Rosenzweig, F.; Sherlock, G. The Valley-of-Death: Reciprocal Sign Epistasis Constrains Adaptive Trajectories in a Constant, Nutrient Limiting Environment. *Genomics* **2014**, *104* (6, Part A), 431–437. https://doi.org/10.1016/j.ygeno.2014.10.011.

(128) Palmer, A. C.; Toprak, E.; Baym, M.; Kim, S.; Veres, A.; Bershtein, S.; Kishony, R. Delayed Commitment to Evolutionary Fate in Antibiotic Resistance Fitness Landscapes. *Nat. Commun.* **2015**, *6* (1), 1–8. https://doi.org/10.1038/ncomms8385.

(129) Poelwijk, F. J.; Tănase-Nicola, S.; Kiviet, D. J.; Tans, S. J. Reciprocal Sign Epistasis Is a Necessary Condition for Multi-Peaked Fitness Landscapes. *J. Theor. Biol.* **2011**, *272* (1), 141–144. https://doi.org/10.1016/j.jtbi.2010.12.015.

(130) Salverda, M. L. M.; Dellus, E.; Gorter, F. A.; Debets, A. J. M.; van der Oost, J.; Hoekstra, R. F.; Tawfik, D. S.; de Visser, J. A. G. M. Initial Mutations Direct Alternative Pathways of Protein Evolution. *PLoS Genet.* **2011**, *7* (3), e1001321. https://doi.org/10.1371/journal.pgen.1001321.

(131) Weinreich, D. M. Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science* **2006**, *312* (5770), 111–114. https://doi.org/10.1126/science.1123539.

(132) Weinreich, D. M.; Lan, Y.; Wylie, C. S.; Heckendorn, R. B. Should Evolutionary Geneticists Worry about Higher-Order Epistasis? *Curr. Opin. Genet. Dev.* **2013**, *23* (6), 700–707. https://doi.org/10.1016/j.gde.2013.10.007.

(133) Wu, N. C.; Dai, L.; Olson, C. A.; Lloyd-Smith, J. O.; Sun, R. Adaptation in Protein Fitness Landscapes Is Facilitated by Indirect Paths. *eLife* **2016**, *5*, e16965. https://doi.org/10.7554/eLife.16965.

(134) Motlagh, H. N.; Hilser, V. J. Agonism/Antagonism Switching in Allosteric Ensembles. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (11), 4134–4139.

(135) Barker, B.; Xu, L.; Gu, Z. Dynamic Epistasis under Varying Environmental Perturbations. *PloS One* **2015**, *10* (1), e0114911. https://doi.org/10.1371/journal.pone.0114911.

(136) Flynn, K. M.; Cooper, T. F.; Moore, F. B.-G.; Cooper, V. S. The Environment Affects Epistatic Interactions to Alter the Topology of an Empirical Fitness Landscape. *PLOS Genet.* **2013**, *9* (4), e1003426. https://doi.org/10.1371/journal.pgen.1003426.

(137) Guerrero, R. F.; Scarpino, S. V.; Rodrigues, J. V.; Hartl, D. L.; Ogbunugafor, C. B. Proteostasis Environment Shapes Higher-Order Epistasis Operating on Antibiotic Resistance. *Genetics* **2019**, *212* (2), 565–575. https://doi.org/10.1534/genetics.119.302138.

(138) Joshi, C. J.; Prasad, A. Epistatic Interactions among Metabolic Genes Depend upon Environmental Conditions. *Mol. Biosyst.* **2014**, *10* (10), 2578–2589. https://doi.org/10.1039/C4MB00181H.

(139) Nosil, P.; Villoutreix, R.; de Carvalho, C. F.; Feder, J. L.; Parchman, T. L.; Gompert, Z. Ecology Shapes Epistasis in a Genotype–Phenotype–Fitness Map for Stick Insect Colour. *Nat. Ecol. Evol.* **2020**. https://doi.org/10.1038/s41559-020-01305-y.

(140) Remold, S. K.; Lenski, R. E. Pervasive Joint Influence of Epistasis and Plasticity on Mutational Effects in Escherichia Coli. *Nat. Genet.* **2004**, *36* (4), 423–426. https://doi.org/10.1038/ng1324.

(141) Samir, P.; Rahul; Slaughter, J. C.; Link, A. J. Environmental Interactions and Epistasis Are Revealed in the Proteomic Responses to Complex Stimuli. *PLOS ONE* **2015**, *10* (8), e0134099. https://doi.org/10.1371/journal.pone.0134099.

(142) Dill, K. A. Dominant Forces in Protein Folding. *Biochemistry* **1990**, *29* (31), 7133–7155. https://doi.org/10.1021/bi00483a001.

(143) Park, H.; Bradley, P.; Greisen, P.; Liu, Y.; Mulligan, V. K.; Kim, D. E.; Baker, D.; DiMaio, F. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J. Chem. Theory Comput.* **2016**, *12* (12), 6201–6212. https://doi.org/10.1021/acs.jctc.6b00819.

(144) Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; Labonte, J. W.; Pacella, M. S.; Bonneau, R.; Bradley, P.; Dunbrack, R. L.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme, T.; Gray, J. J. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **2017**, *13* (6), 3031–3048. https://doi.org/10.1021/acs.jctc.7b00125.

(145) Anderson, D. W.; McKeown, A. N.; Thornton, J. W. Intermolecular Epistasis Shaped the Function and Evolution of an Ancient Transcription Factor and Its DNA Binding Sites. *eLife* **2015**, *4*, e07864. https://doi.org/10.7554/eLife.07864.

(146) Bridgham, J. T.; Carroll, S. M.; Thornton, J. W. Evolution of Hormone-Receptor Complexity by Molecular Exploitation. *Science* **2006**, *312* (5770), 97–101. https://doi.org/10.1126/science.1123348.

(147) D'Souza, G.; Waschina, S.; Kaleta, C.; Kost, C. Plasticity and Epistasis Strongly Affect Bacterial Fitness after Losing Multiple Metabolic Genes. *Evol. Int. J. Org. Evol.* **2015**, *69* (5), 1244–1254. https://doi.org/10.1111/evo.12640.

(148) Kachanovsky, D. E.; Filler, S.; Isaacson, T.; Hirschberg, J. Epistasis in Tomato Color Mutations Involves Regulation of Phytoene Synthase 1 Expression by Cis-Carotenoids. *Proc. Natl. Acad. Sci.* **2012**, *109* (46), 19021–19026. https://doi.org/10.1073/pnas.1214808109.

(149) Lunzer, M.; Golding, G. B.; Dean, A. M. Pervasive Cryptic Epistasis in Molecular Evolution. *PLOS Genet.* **2010**, *6* (10), e1001162. https://doi.org/10.1371/journal.pgen.1001162.

(150) McCandlish, D. M.; Otwinowski, J.; Plotkin, J. B. Detecting Epistasis from an Ensemble of Adapting Populations. *Evol. Int. J. Org. Evol.* **2015**, *69* (9), 2359–2370. https://doi.org/10.1111/evo.12735.

(151) Musso, G.; Costanzo, M.; Huangfu, M.; Smith, A. M.; Paw, J.; Luis, B.-J. S.; Boone, C.; Giaever, G.; Nislow, C.; Emili, A.; Zhang, Z. The Extensive and Condition-Dependent Nature of Epistasis among Whole-Genome Duplicates in Yeast. *Genome Res.* **2008**, *18* (7), 1092–1099. https://doi.org/10.1101/gr.076174.108.

(152) Morrison, A. J.; Wonderlick, D. R.; Harms, M. J. Ensemble Epistasis: Thermodynamic Origins of Nonadditivity between Mutations. *Genetics* **2021**, No. iyab105. https://doi.org/10.1093/genetics/iyab105.

(153) Tsai, C.-J.; Ma, B.; Nussinov, R. Folding and Binding Cascades: Shifts in Energy Landscapes. *Proc. Natl. Acad. Sci.* **1999**, *96* (18), 9970–9972. https://doi.org/10.1073/pnas.96.18.9970.

(154) Tsai, C.-J.; Nussinov, R. A Unified View of "How Allostery Works." *PLOS Comput. Biol.* **2014**, *10* (2), e1003394. https://doi.org/10.1371/journal.pcbi.1003394.

(155) Bell, C. E.; Lewis, M. A Closer View of the Conformation of the Lac Repressor Bound to Operator. *Nat. Struct. Biol.* **2000**, *7* (3), 209–214. https://doi.org/10.1038/73317.

(156) Chuprina, V. P.; Rullmann, J. A. C.; Lamerichs, R. M. J. N.; van Boom, J. H.; Boelens, R.; Kaptein, R. Structure of the Complex of Lac Repressor Headpiece and an 11 Base-Pair Half-Operator Determined by Nuclear Magnetic Resonance Spectroscopy and Restrained Molecular Dynamics. *J. Mol. Biol.* **1993**, *234* (2), 446–462. https://doi.org/10.1006/jmbi.1993.1598.

(157) Friedman, A. M.; Fischmann, T. O.; Steitz, T. A. Crystal Structure of Lac Repressor Core Tetramer and Its Implications for DNA Looping. *Science* **1995**, *268* (5218), 1721–1727. https://doi.org/10.1126/science.7792597.

(158) Gilbert, W.; Müller-Hill, B. Isolation of the Lac Repressor. *Proc. Natl. Acad. Sci.* **1966**, *56* (6), 1891–1898. https://doi.org/10.1073/pnas.56.6.1891.

(159) Jacob, F.; Monod, J. Genetic Regulatory Mechanisms in the Synthesis of Proteins. *J. Mol. Biol.* **1961**, *3* (3), 318–356. https://doi.org/10.1016/S0022-2836(61)80072-7.

(160) Lewis, M. The Lac Repressor. *C. R. Biol.* **2005**, *328* (6), 521–548. https://doi.org/10.1016/j.crvi.2005.04.004.

(161) Lewis, M.; Chang, G.; Horton, N. C.; Kercher, M. A.; Pace, H. C.; Schumacher, M. A.; Brennan, R. G.; Lu, P. Crystal Structure of the Lactose Operon Repressor and Its Complexes with DNA and Inducer. *Science* **1996**, *271* (5253), 1247–1254. https://doi.org/10.1126/science.271.5253.1247.

(162) Monod, J.; Wyman, J.; Changeux, J.-P. On the Nature of Allosteric Transitions: A Plausible Model. *J. Mol. Biol.* **1965**, *12* (1), 88–118. https://doi.org/10.1016/S0022-2836(65)80285-6.

(163) Müller-Hill, B.; Rickenberg, H. V.; Wallenfels, K. Specificity of the Induction of the Enzymes of the Lac Operon in Escherichia Coli. *J. Mol. Biol.* **1964**, *10* (2), 303–318. https://doi.org/10.1016/S0022-2836(64)80049-8.

(164) Perez, P. J.; Clauvelin, N.; Tam, G.; Olson, W. K. Conformational Changes in the Lac Repressor Protein Effect DNA Loop Energetics and Topology. *Biophys. J.* **2014**, *106* (2), 71a. https://doi.org/10.1016/j.bpj.2013.11.467.

(165) Slijper, M.; Bonvin, A. M. J. J.; Boelens, R.; Kaptein, R. Refined Structure of Lac Repressor Headpiece (1-56) Determined by Relaxation Matrix Calculations from 2D and 3D NOE Data: Change of Tertiary Structure upon Binding to ThelacOperator. *J. Mol. Biol.* **1996**, *259* (4), 761–773. https://doi.org/10.1006/jmbi.1996.0356.

(166) Taraban, M.; Zhan, H.; Whitten, A. E.; Langley, D. B.; Matthews, K. S.; Swint-Kruse, L.; Trewhella, J. Ligand-Induced Conformational Changes and Conformational Dynamics in the Solution Structure of the Lactose Repressor Protein. *J. Mol. Biol.* **2008**, *376* (2), 466–481. https://doi.org/10.1016/j.jmb.2007.11.067.

(167) Vos, M. G. J. de; Poelwijk, F. J.; Battich, N.; Ndika, J. D. T.; Tans, S. J. Environmental Dependence of Genetic Constraint. *PLOS Genet.* **2013**, *9* (6), e1003580. https://doi.org/10.1371/journal.pgen.1003580.

(168) Sochor, M. A. In Vitro Transcription Accurately Predicts Lac Repressor Phenotype in Vivo in Escherichia Coli. *PeerJ* **2014**, *2*, e498. https://doi.org/10.7717/peerj.498.

(169) Barry, J. K.; Matthews, K. S. Substitutions at Histidine 74 and Aspartate 278 Alter Ligand Binding and Allostery in Lactose Repressor Protein. *Biochemistry* **1999**, *38* (12), 3579–3590. https://doi.org/10.1021/bi982577n.

(170) Meinhardt, S.; Manley, M. W., Jr; Becker, N. A.; Hessman, J. A.; Maher, L. J., III; Swint-Kruse, L. Novel Insights from Hybrid LacI/GalR Proteins: Family-Wide Functional Attributes and Biologically Significant Variation in Transcription Repression. *Nucleic Acids Res.* **2012**, *40* (21), 11139–11154. https://doi.org/10.1093/nar/gks806.

(171) Swint-Kruse, L.; Elam, C. R.; Lin, J. W.; Wycuff, D. R.; Matthews, K. S. Plasticity of Quaternary Structure: Twenty-Two Ways to Form a LacI Dimer. *Protein Sci. Publ. Protein Soc.* **2001**, *10* (2), 262–276.

(172) Swint-Kruse, L.; Zhan, H.; Matthews, K. S. Integrated Insights from Simulation, Experiment, and Mutational Analysis Yield New Details of LacI Function. *Biochemistry* **2005**, *44* (33), 11201–11213. https://doi.org/10.1021/bi050404+.

(173) Flynn, T. C.; Swint-Kruse, L.; Kong, Y.; Booth, C.; Matthews, K. S.; Ma, J. Allosteric Transition Pathways in the Lactose Repressor Protein Core Domains: Asymmetric Motions in a Homodimer. *Protein Sci. Publ. Protein Soc.* **2003**, *12* (11), 2523–2541.

(174) Bell, C. E.; Barry, J.; Matthews, K. S.; Lewis, M. Structure of a Variant of Lac Repressor with Increased Thermostability and Decreased Affinity for Operator11Edited by D. Rees. *J. Mol. Biol.* **2001**, *313* (1), 99–109. https://doi.org/10.1006/jmbi.2001.5041.

(175) Taute, K. M.; Gude, S.; Nghe, P.; Tans, S. J. Evolutionary Constraints in Variable Environments, from Proteins to Networks. *Trends Genet.* **2014**, *30* (5), 192–198. https://doi.org/10.1016/j.tig.2014.04.003.

(176) Anderson, D. W.; Baier, F.; Yang, G.; Tokuriki, N. The Adaptive Landscape of a Metallo-Enzyme Is Shaped by Environment-Dependent Epistasis. *Nat. Commun.* **2021**, *12* (1), 3867. https://doi.org/10.1038/s41467-021-23943-x.

(177) Hayden, E. J.; Ferrada, E.; Wagner, A. Cryptic Genetic Variation Promotes Rapid Evolutionary Adaptation in an RNA Enzyme. *Nature* **2011**, *474* (7349), 92–95. https://doi.org/10.1038/nature10083.

(178) Hayden, E. J.; Wagner, A. Environmental Change Exposes Beneficial Epistatic Interactions in a Catalytic RNA. *Proc. R. Soc. B Biol. Sci.* **2012**, *279* (1742), 3418–3425. https://doi.org/10.1098/rspb.2012.0956.

(179) Lindsey, H. A.; Gallie, J.; Taylor, S.; Kerr, B. Evolutionary Rescue from Extinction Is Contingent on a Lower Rate of Environmental Change. *Nature* **2013**, *494* (7438), 463–467. https://doi.org/10.1038/nature11879.

(180) Ogbunugafor, C. B.; Wylie, C. S.; Diakite, I.; Weinreich, D. M.; Hartl, D. L. Adaptive Landscape by Environment Interactions Dictate Evolutionary Dynamics in Models of Drug Resistance. *PLOS Comput. Biol.* **2016**, *12* (1), e1004710. https://doi.org/10.1371/journal.pcbi.1004710.

(181) Li, C.; Zhang, J. Multi-Environment Fitness Landscapes of a TRNA Gene. *Nat. Ecol. Evol.* **2018**, *2* (6), 1025–1032. https://doi.org/10.1038/s41559-018-0549-8.

(182) Lagator, M.; Igler, C.; Moreno, A. B.; Guet, C. C.; Bollback, J. P. Epistatic Interactions in the Arabinose Cis-Regulatory Element. *Mol. Biol. Evol.* **2016**, *33* (3), 761–769. https://doi.org/10.1093/molbev/msv269.

(183) Lagator, M.; Paixão, T.; Barton, N. H.; Bollback, J. P.; Guet, C. C. On the Mechanistic Nature of Epistasis in a Canonical Cis-Regulatory Element. *eLife* **2017**, *6*, e25192. https://doi.org/10.7554/eLife.25192.

(184) Shultzaberger, R. K.; Malashock, D. S.; Kirsch, J. F.; Eisen, M. B. The Fitness Landscapes of Cis-Acting Binding Sites in Different Promoter and Environmental Contexts. *PLOS Genet.* **2010**, *6* (7), e1001042. https://doi.org/10.1371/journal.pgen.1001042.

(185) Nghe, P.; Kogenaru, M.; Tans, S. J. Sign Epistasis Caused by Hierarchy within Signalling Cascades. *Nat. Commun.* **2018**, *9* (1), 1451. https://doi.org/10.1038/s41467-018-03644-8.

(186) Caudle, S. B.; Miller, C. R.; Rokyta, D. R. Environment Determines Epistatic Patterns for a SsDNA Virus. *Genetics* **2014**, *196* (1), 267–279. https://doi.org/10.1534/genetics.113.158154.

(187) Filteau, M.; Hamel, V.; Pouliot, M.-C.; Gagnon-Arsenault, I.; Dubé, A. K.; Landry, C. R. Evolutionary Rescue by Compensatory Mutations Is Constrained by Genomic and Environmental Backgrounds. *Mol. Syst. Biol.* **2015**, *11* (10), 832. https://doi.org/10.15252/msb.20156444.

(188) James, L. C.; Tawfik, D. S. Conformational Diversity and Protein Evolution – a 60-Year-Old Hypothesis Revisited. *Trends Biochem. Sci.* **2003**, *28* (7), 361–368. https://doi.org/10.1016/S0968-0004(03)00135-X.

(189) Tokuriki, N.; Tawfik, D. S. Protein Dynamism and Evolvability. *Science* **2009**, *324* (5924), 203–207. https://doi.org/10.1126/science.1169375.

(190) Jolly, M. K.; Kulkarni, P.; Weninger, K.; Orban, J.; Levine, H. Phenotypic Plasticity, Bet-Hedging, and Androgen Independence in Prostate Cancer: Role of Non-Genetic Heterogeneity. *Front. Oncol.* **2018**, *8*, 50. https://doi.org/10.3389/fonc.2018.00050.

(191) Pigliucci, M.; Murren, C. J.; Schlichting, C. D. Phenotypic Plasticity and Evolution by Genetic Assimilation. *J. Exp. Biol.* **2006**, *209* (12), 2362–2367. https://doi.org/10.1242/jeb.02070.

(192) Javier Zea, D.; Miguel Monzon, A.; Fornasari, M. S.; Marino-Buslje, C.; Parisi, G. Protein Conformational Diversity Correlates with Evolutionary Rate. *Mol. Biol. Evol.* **2013**, *30* (7), 1500–1503. https://doi.org/10.1093/molbev/mst065.

(193) Tóth-Petróczy, Á.; Tawfik, D. S. The Robustness and Innovability of Protein Folds. *Curr. Opin. Struct. Biol.* **2014**, *26*, 131–138. https://doi.org/10.1016/j.sbi.2014.06.007.

(194) Charon, J.; Barra, A.; Walter, J.; Millot, P.; Hébrard, E.; Moury, B.; Michon, T. First Experimental Assessment of Protein Intrinsic Disorder Involvement in an RNA Virus Natural Adaptive Process. *Mol. Biol. Evol.* **2018**, *35* (1), 38–49. https://doi.org/10.1093/molbev/msx249.

(195) Broom, A.; Rakotoharisoa, R. V.; Thompson, M. C.; Zarifi, N.; Nguyen, E.; Mukhametzhanov, N.; Liu, L.; Fraser, J. S.; Chica, R. A. Ensemble-Based Enzyme Design Can Recapitulate the Effects of Laboratory Directed Evolution in Silico. *Nat. Commun.* **2020**, *11* (1), 4808. https://doi.org/10.1038/s41467-020-18619-x.

(196) Adams, R. M.; Mora, T.; Walczak, A. M.; Kinney, J. B. Measuring the Sequence-Affinity Landscape of Antibodies with Massively Parallel Titration Curves. *eLife* **2016**, *5*, e23156. https://doi.org/10.7554/eLife.23156.

(197) Aditham, A. K.; Markin, C. J.; Mokhtari, D. A.; DelRosso, N.; Fordyce, P. M. High-Throughput Affinity Measurements of Transcription Factor and DNA Mutations Reveal Affinity and Specificity Determinants. *Cell Syst.* **2021**, *12* (2), 112-127.e11. https://doi.org/10.1016/j.cels.2020.11.012.

(198) Rocklin, G. J.; Chidyausiku, T. M.; Goreshnik, I.; Ford, A.; Houliston, S.; Lemak, A.; Carter, L.; Ravichandran, R.; Mulligan, V. K.; Chevalier, A.; Arrowsmith, C. H.; Baker, D. Global Analysis of Protein Folding Using Massively Parallel Design, Synthesis, and Testing. *Science* **2017**, *357* (6347), 168. https://doi.org/10.1126/science.aan0693.

(199) Tack, D. S.; Tonner, P. D.; Pressman, A.; Olson, N. D.; Levy, S. F.; Romantseva, E. F.; Alperovich, N.; Vasilyeva, O.; Ross, D. The Genotype-Phenotype Landscape of an Allosteric Protein. *Mol. Syst. Biol.* **2021**, *17* (3), e10179. https://doi.org/10.15252/msb.202010179.

(200) Cohen, S. E.; Erb, M. L.; Selimkhanov, J.; Dong, G.; Hasty, J.; Pogliano, J.; Golden, S. S. Dynamic Localization of the Cyanobacterial Circadian Clock Proteins. *Curr. Biol.* **2014**, *24* (16), 1836–1844. https://doi.org/10.1016/j.cub.2014.07.036.

(201) Arai, R.; Ueda, H.; Kitayama, A.; Kamiya, N.; Nagamune, T. Design of the Linkers Which Effectively Separate Domains of a Bifunctional Fusion Protein. *Protein Eng. Des. Sel.* **2001**, *14* (8), 529–532. https://doi.org/10.1093/protein/14.8.529.

(202) Wycuff, D. R.; Matthews, K. S. Generation of an AraC-AraBAD Promoter-Regulated T7 Expression System. *Anal. Biochem.* **2000**, *277* (1), 67–73. https://doi.org/10.1006/abio.1999.4385.

(203) Kubitschek, H. E.; Friske, J. A. Determination of Bacterial Cell Volume with the Coulter Counter. *J. Bacteriol.* **1986**, *168* (3), 1466–1467. https://doi.org/10.1128/jb.168.3.1466-1467.1986.

(204) Hoffmann, S. A.; Kruse, S. M.; Arndt, K. M. Long-Range Transcriptional Interference in E. Coli Used to Construct a Dual Positive Selection System for Genetic Switches. *Nucleic Acids Res.* **2016**, *44* (10), e95. https://doi.org/10.1093/nar/gkw125.

(205) Poelwijk, F. J.; de Vos, M. G. J.; Tans, S. J. Tradeoffs and Optimality in the Evolution of Gene Regulation. *Cell* **2011**, *146* (3), 462–470. https://doi.org/10.1016/j.cell.2011.06.035.

(206) Foreman-Mackey, D.; Hogg, D. W.; Lang, D.; Goodman, J. Emcee: The MCMC Hammer. *Publ. Astron. Soc. Pac.* **2013**, *125* (925), 306. https://doi.org/10.1086/670067.

(207) Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M. H.; Brett, M.; Haldane, A.; del Río, J. F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T. E. Array Programming with NumPy. *Nature* **2020**, *585* (7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2.

(208) Reback, J.; jbrockmendel; McKinney, W.; Bossche, J. V. den; Augspurger, T.; Cloud, P.; Hawkins, S.; gfyoung; Sinhrks; Roeschke, M.; Klein, A.; Petersen, T.; Tratner, J.; She, C.; Ayd, W.; Hoefler, P.; Naveh, S.; Garcia, M.; Schendel, J.; Hayden, A.; Saxton, D.; Shadrach, R.; Gorelli, M. E.; Li, F.; Jancauskas, V.; attack68; McMaster, A.; Battiston, P.; Seabold, S.; Dong, K. *Pandas-Dev/Pandas: Pandas 1.3.2*; Zenodo, 2021. https://doi.org/10.5281/zenodo.5203279.

(209) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

(210) Maynard Smith, J. Natural Selection and the Concept of a Protein Space. *Nature* **1970**, *225* (5232), 563–564. https://doi.org/10.1038/225563a0.

(211) Malcolm, B. A.; Wilson, K. P.; Matthews, B. W.; Kirsch, J. F.; Wilson, A. C. Ancestral Lysozymes Reconstructed, Neutrality Tested, and Thermostability Linked to Hydrocarbon Packing. *Nature* **1990**, *345* (6270), 86–89. https://doi.org/10.1038/345086a0.

(212) Tan, L.; Serene, S.; Chao, H. X.; Gore, J. Hidden Randomness between Fitness Landscapes Limits Reverse Evolution. *Phys. Rev. Lett.* **2011**, *106* (19), 198102. https://doi.org/10.1103/PhysRevLett.106.198102.

(213) Jiang, P.-P.; Corbett-Detig, R. B.; Hartl, D. L.; Lozovsky, E. R. Accessible Mutational Trajectories for the Evolution of Pyrimethamine Resistance in the Malaria Parasite Plasmodium Vivax. *J. Mol. Evol.* **2013**, *77* (3), 81–91. https://doi.org/10.1007/s00239-013-9582-z.

(214) Brown, K. M.; Costanzo, M. S.; Xu, W.; Roy, S.; Lozovsky, E. R.; Hartl, D. L. Compensatory Mutations Restore Fitness during the Evolution of Dihydrofolate Reductase. *Mol. Biol. Evol.* **2010**, *27* (12), 2682–2690. https://doi.org/10.1093/molbev/msq160.

(215) Costanzo, M. S.; Brown, K. M.; Hartl, D. L. Fitness Trade-Offs in the Evolution of Dihydrofolate Reductase and Drug Resistance in Plasmodium Falciparum. *PLOS ONE* **2011**, *6* (5), e19636. https://doi.org/10.1371/journal.pone.0019636.

(216) Aita, T.; Iwakura, M.; Husimi, Y. A Cross-Section of the Fitness Landscape of Dihydrofolate Reductase. *Protein Eng. Des. Sel.* **2001**, *14* (9), 633–638. https://doi.org/10.1093/protein/14.9.633.

(217) Szendro, I. G.; Schenk, M. F.; Franke, J.; Krug, J.; Visser, J. A. G. M. de. Quantitative Analyses of Empirical Fitness Landscapes. *J. Stat. Mech. Theory Exp.* **2013**, *2013* (01), P01005. https://doi.org/10.1088/1742-5468/2013/01/P01005.

(218) Lozovsky, E. R.; Chookajorn, T.; Brown, K. M.; Imwong, M.; Shaw, P. J.; Kamchonwongpaisan, S.; Neafsey, D. E.; Weinreich, D. M.; Hartl, D. L. Stepwise Acquisition of Pyrimethamine Resistance in the Malaria Parasite. *Proc. Natl. Acad. Sci.* **2009**, *106* (29), 12025–12030. https://doi.org/10.1073/pnas.0905922106.

(219) de Visser, J. A. G. M.; Krug, J. Empirical Fitness Landscapes and the Predictability of Evolution. *Nat. Rev. Genet.* **2014**, *15* (7), 480–490. https://doi.org/10.1038/nrg3744.

(220) Weinreich, D. M.; Lan, Y.; Jaffe, J.; Heckendorn, R. B. The Influence of Higher-Order Epistasis on Biological Fitness Landscape Topography. *bioRxiv* **2017**, 164798. https://doi.org/10.1101/164798.

(221) Yokoyama, S.; Yang, H.; Starmer, W. T. Molecular Basis of Spectral Tuning in the Red- and Green-Sensitive (M/LWS) Pigments in Vertebrates. *Genetics* **2008**, *179* (4), 2037–2043. https://doi.org/10.1534/genetics.108.090449.

(222) Gavrilets, S. Evolution and Speciation on Holey Adaptive Landscapes. *Trends Ecol. Evol.* **1997**, *12* (8), 307–312. https://doi.org/10.1016/S0169-5347(97)01098-7.

(223) Cariani, P. A. Extradimensional Bypass. *Biosystems* **2002**, *64* (1), 47–53. https://doi.org/10.1016/S0303-2647(01)00174-5.

(224) Zagorski, M.; Burda, Z.; Waclaw, B. Beyond the Hypercube: Evolutionary Accessibility of Fitness Landscapes with Realistic Mutational Networks. *PLOS Comput. Biol.* **2016**, *12* (12), e1005218. https://doi.org/10.1371/journal.pcbi.1005218.

(225) Aguirre, J.; Buldú, J. M.; Stich, M.; Manrubia, S. C. Topological Structure of the Space of Phenotypes: The Case of RNA Neutral Networks. *PLOS ONE* **2011**, *6* (10), e26324. https://doi.org/10.1371/journal.pone.0026324.

(226) Babajide, A.; Hofacker, I. L.; Sippl, M. J.; Stadler, P. F. Neutral Networks in Protein Space: A Computational Study Based on Knowledge-Based Potentials of Mean Force. *Fold. Des.* **1997**, *2* (5), 261–269. https://doi.org/10.1016/S1359-0278(97)00037-0.

(227) Wagner, A. Neutralism and Selectionism: A Network-Based Reconciliation. *Nat. Rev. Genet.* **2008**, *9* (12), 965–974. https://doi.org/10.1038/nrg2473.

(228) Field, S. F.; Matz, M. V. Retracing Evolution of Red Fluorescence in GFP-Like Proteins from Faviina Corals. *Mol. Biol. Evol.* **2010**, *27* (2), 225–233. https://doi.org/10.1093/molbev/msp230.

(229) Sarkisyan, K. S.; Bolotin, D. A.; Meer, M. V.; Usmanova, D. R.; Mishin, A. S.; Sharonov, G. V.; Ivankov, D. N.; Bozhanova, N. G.; Baranov, M. S.; Soylemez, O.; Bogatyreva, N. S.; Vlasov, P. K.; Egorov, E. S.; Logacheva, M. D.; Kondrashov, A. S.; Chudakov, D. M.; Putintseva, E. V.; Mamedov, I. Z.; Tawfik, D. S.; Lukyanov, K. A.; Kondrashov, F. A. Local Fitness Landscape of the Green Fluorescent Protein. *Nature* **2016**, *533* (7603), 397–401. https://doi.org/10.1038/nature17995.

(230) Zhao, L.; Liu, Z.; Levy, S. F.; Wu, S. Bartender: A Fast and Accurate Clustering Algorithm to Count Barcode Reads. *Bioinformatics* **2018**, *34* (5), 739–747. https://doi.org/10.1093/bioinformatics/btx655.

(231) Peterman, N.; Levine, E. Sort-Seq under the Hood: Implications of Design Choices on Large-Scale Characterization of Sequence-Function Relations. *BMC Genomics* **2016**, *17* (1), 206. https://doi.org/10.1186/s12864-016-2533-5.

(232) Kim, H.; Zou, T.; Modi, C.; Dörner, K.; Grunkemeyer, T. J.; Chen, L.; Fromme, R.; Matz, M. V.; Ozkan, S. B.; Wachter, R. M. A Hinge Migration Mechanism Unlocks the Evolution of Green-to-Red Photoconversion in GFP-like Proteins. *Struct. Lond. Engl. 1993* **2015**, *23* (1), 34–43. https://doi.org/10.1016/j.str.2014.11.011.

(233) Wachter, R. M. Photoconvertible Fluorescent Proteins and the Role of Dynamics in Protein Evolution. *Int. J. Mol. Sci.* **2017**, *18* (8), 1792. https://doi.org/10.3390/ijms18081792.

(234) Lukyanov, K. A.; Chudakov, D. M.; Lukyanov, S.; Verkhusha, V. V. Photoactivatable Fluorescent Proteins. *Nat. Rev. Mol. Cell Biol.* **2005**, *6* (11), 885–890. https://doi.org/10.1038/nrm1741.

(235) Mizuno, H.; Mal, T. K.; Tong, K. I.; Ando, R.; Furuta, T.; Ikura, M.; Miyawaki, A. Photo-Induced Peptide Cleavage in the Green-to-Red Conversion of a Fluorescent Protein. *Mol. Cell* **2003**, *12* (4), 1051–1058. https://doi.org/10.1016/S1097-2765(03)00393-9.

(236) Matz, M. V.; Fradkov, A. F.; Labas, Y. A.; Savitsky, A. P.; Zaraisky, A. G.; Markelov, M. L.; Lukyanov, S. A. Fluorescent Proteins from Nonbioluminescent Anthozoa Species. *Nat. Biotechnol.* **1999**, *17* (10), 969–973. https://doi.org/10.1038/13657.