PREDICTING PHENOTYPES IN SPARSELY SAMPLED GENOTYPE-PHENOTYPE MAPS

by

ZACHARY R. SAILER

A DISSERTATION

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

September 2018

DISSERTATION APPROVAL PAGE

Student: Zachary R. Sailer

Title: Predicting Phenotypes in Sparsely Sampled 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:

Jeffrey Cina	Chair
Marina Guenza	Member
John Conery	Institutional Representative
Michael Harms	Advisor

and

Janet Woodruff-Borden	Vice Provost and Dean of the Graduate School

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

Degree awarded September 2018

DISSERTATION ABSTRACT

Zachary R. Sailer

Doctor of Philosophy

Department of Chemistry and Biochemistry

September 2018

Title: Predicting Phenotypes in Sparsely Sampled Genotype-Phenotype Maps

Naturally evolving proteins must navigate a vast set of possible sequences to evolve new functions. This process depends on the genotype-phenotype map. Much effort has been directed at measuring protein genotype phenotype maps to uncover evolutionary trajectories that lead to new functions. Often, these maps are too large to comprehensively measure. Sparsely measured maps, however, are prone to missing key evolutionary trajectories. Many groups turn to computational models to infer missing phenotypes. These models treat mutations as independent perturbations to the genotype-phenotype map. A key question is how to handle non-independent effects known as epistasis. In this dissertation, we address two sources of epistasis: 1) global and 2) local epistasis. We find that incorporating global epistasis improves our predictive power, while local epistasis does not. We use our model to infer unknown phenotypes in the *Plasmodium falciparum* chloroquine transporter (PfCRT) genotype-phenotype map, a protein responsible for conferring drug resistance in malaria. From these predictions, we uncover key evolutionary trajectories that led high resistance. This dissertation includes previously published and unpublished co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Zachary R. Sailer

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
California Polytechnic State University, San Luis Obispo, CA

DEGREES AWARD:

Doctor of Philosophy, Chemistry and Biochemistry, 2018, University of Oregon
Bachelor of Science, Physics, 2018, California Polytechnic State University

GRANTS, AWARDS, AND HONORS:

Travel Award, SciPy, 2017

Travel Award, JupyterCon, 2017

Travel Award, SciPy, 2018

Art Rosen Memorial Scholar, Cal Poly SLO Physics Department, 2012

PUBLICATIONS:

Sailer, Zachary R. and Harms, Michael J., Uninterpretable interactions: epistasis as uncertainty, *bioRxiv* (2018).

Sailer, Zachary R. and Harms, Michael J., Molecular ensembles make evolution unpredictable, *PNAS* 114, 45 (2017), pp. 11938--11943.

Sailer, Zachary R. and Harms, Michael J., High-order epistasis shapes evolutionary trajectories, *PLOS Computational Biology* 13, 5 (2017), pp. e1005541.

Sailer, Zachary R. and Harms, Michael J., Detecting High-Order Epistasis in Nonlinear Genotype-Phenotype Maps, *Genetics* 205, 3 (2017), pp. 1079--1088.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

Figure Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

Over several billion years, proteins have evolved and diversified to perform various functions necessary for life. This is an incredible feat, since naturally evolving proteins must navigate a vast space of possible sequences to acquire new functions [1, 2]. While sequence space is large, it is also sparse–comprised mostly of nonfunctional or inviable sequences [3]. *How do proteins navigate sequence space and evolve new functions?*

The genotype-phenotype map is a key determinant of protein evolution. John Maynard Smith illustrated protein genotype-phenotype maps using a famous word game. [4, 5]. He imagined protein sequences as English words. To evolve between two words, evolution can only proceed through adjacent meaningful words. In the evolution from WORD to COST, there are 16 possible trajectories (Figure 1A), but only one trajectory is accessible. The distribution of meaningful words strongly shapes evolution in this space. This is analogous to how Smith imagine proteins navigate sequence space. Evolutionary trajectories are strongly shaped by the distribution of phenotype in sequence space. Ogbunugafor et al. [5] extended the word game to show how proteins might evolve a quantitative phenotype by mapping each word's frequency from various books pulled in Google Books Ngram viewer (Figure 1B). In this space, proteins accumulate mutations that incrementally improve the phenotype over evolutionary time.

The word game shows that to understand how a protein evolved, it is necessary to study the genotype-phenotype map [6, 7, 2, 8, 9, 10]. The genotype-phenotype map insight to key evolutionary questions like: why is one evolutionary trajectory taken over another? What are the biological, chemical, and physical determinants of acces-

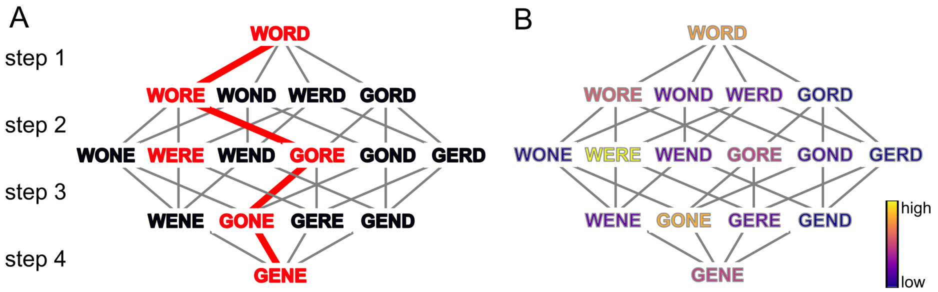


Figure 1: John Maynard Smith's wordgame. A) Panel shows the full map of English words between WORD and COST. Red word represent meaningful words English words, black words represent gibberish words. The edges represented single letter substitutions between words. The red trajectory represents the only accessible trajectory. B) Panel shows the word game with quantitative phenotypes. The colors represent the frequency of each word in Google's Ngram viewer.

sible evolutionary trajectories? Is evolution constrained? Is evolution predictable?

Measuring complete protein genotype-phenotype maps, however, is often intractable. The number of genotypes grows rapidly as the number of mutations increases. For example, in the word game (Figure 1), four mutations led to a binary genotype-phenotype with 16 genotypes. The same word game with 15 mutations (a modest number) leads to a map 32,768 genotypes. The numbers compound even quicker if we consider any amino acid at each site. For example, a protein with all 20 amino acids at 15 sites leads to a genotype-phenotype map with $> 10^{19}$ genotypes. Even modern high-throughput techniques cannot characterize maps with this many genotypes. Further, there are many cases in which high-throughput methods are not currently possible. As a result, even small genotype-phenotype maps can be experimentally challenging to characterize.

The Genotype-Phenotype Map that Led to Malarial Drug Resistance

A timely example is the chloroquine resistance transporter (PfCRT) protein that evolved to confer drug resistance in *plasmodium falciparum* [11]. Eight mutations were identified in PfCRT, that confers resistance to chloroquine (CQ). For many decades, chloroquine was the primary malaria treatment before these chloroquine-resistance parasites naturally evolved and stunted its effectiveness. This was disastrous for world health [11]. Summers et al. constructed the genotype-phenotype map from the 8 mutations. This is shown in Figure 2A [11]. Many questions immediately arose: How did the eight mutations arise? What evolutionary trajectories were the most probable (i.e. in what order did the mutations occur)? Why did this system take many decades to evolve–a stark contrast from other evolutionary systems like influenza or HIV [10].

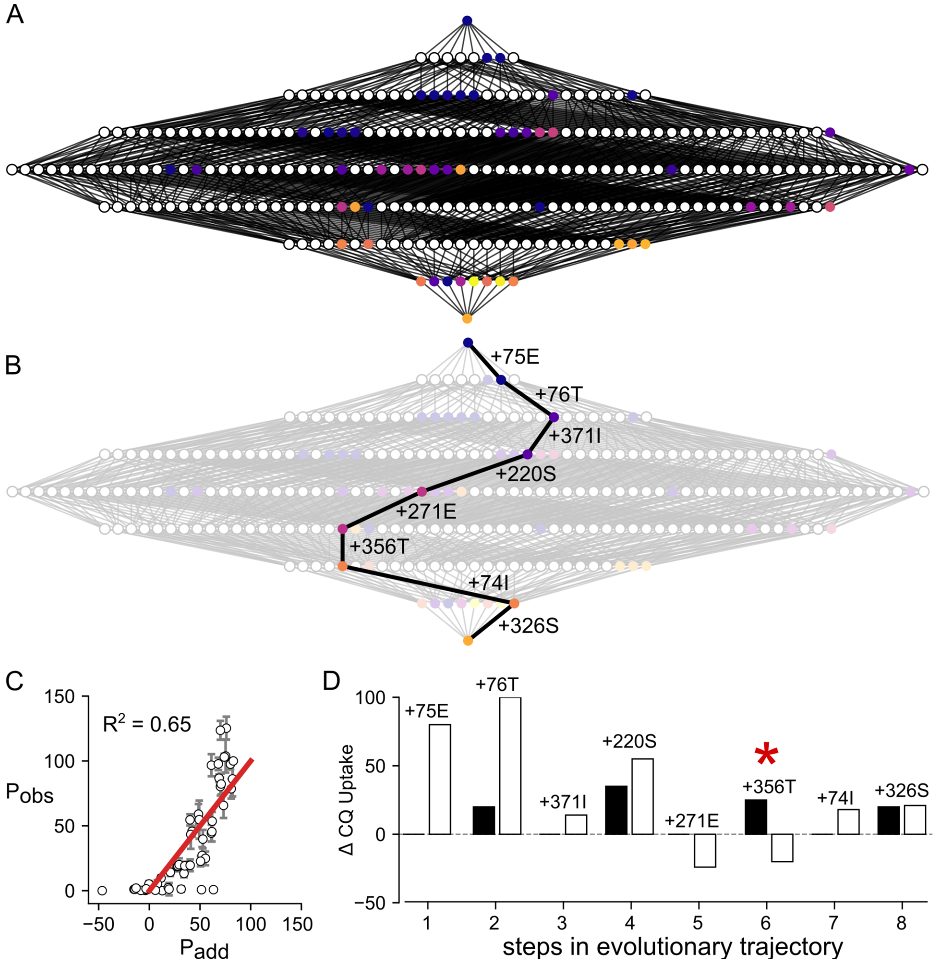


Figure 2: Incomplete PfCRT genotype-phenotype map shows epistasis. A) Panel shows known and unknown phenotypes in PfCRT's genotype-phenotype map. Nodes represent genotypes. Edges represent single substitutions. Node color represents the measured CQ uptake. White nodes represent genotypes that were not measured. B) Panel shows one possible evolutionary trajectory in PfCRT. Labels show position and mutation that occurred only each edge in the trajectory. C) Panel shows the P_{obs} vs. P_{add} curve plot for the observed phenotypes. Points represent genotypes. Error bars represent uncertainty in the measured phenotypes (two standard deviations). Red line represents the 1:1 line. D) Panel shows the quantitative effects of each mutation. Black bars represent the effect of each mutation in the trajectory. White bars represent the effect of the each mutation in the derived genotype's genetic background. Red star highlights mutation 356T which flips sign when introduced in the trajectory versus the derived background.

PfCRT plays a key role in transporting chloroquine out of the cell's digestive vacuole (DV). CQ is a diprotic base that diffuses into the DV[12]. Here, it deprotonates in the acidic environment and becomes trapped. It accumulates in high concentration, prevents the cell from detoxifying, and causes cell death [13]. To counteract the high accumulation of CQ, the parasite hijacked PfCRT to transport CQ out of the digestive vacuole. The eight mutations that arose in nature malarial strains enabled PfCRT to transport de-protonated CQ molecules across the vacuole membrane and into the intercellular space.

To understand how PfCRT evolved the ability to transport CQ, Summers et al. sought to measure portions of the genotype-phenotype map. They constructed different combinations of the 8 mutations in PfCRT (shown in color in Figure 2). They expressed these mutants on the membrane of *Xenopus laevis* oocytes cells. They then left the cells in saturating concentrations of protonated CQ and measured the CQ uptake over time. This took Summers et al. a few years to measure 52 genotypes–only a fraction of the entire map.

Using the information from the measured 52 genotypes, Summers et al. inferred a set of evolutionary trajectories that led to high-CQ transportation PfCRT [11]. One of these trajectories is shown in Figure 2B. Interestingly, they found that every trajectory required at least one neutral step–a substitution that led to no change in CQ uptake. They suggest that this may be the reason that CQ resistance took many decades to evolve.

Though a few possible trajectories were proposed, it is unclear whether these are key trajectories without knowing the complete genotype-phenotype map. It is possible that important genotypes were not measured, and therefore, key trajectories were missed. Measuring new genotypes, however, takes many weeks. At this rate, it would take over a decade to measure the complete genotype-phenotype map. On the other hand, more efficient experimental methods (such as high-throughput methods) are not currently possible.

Computational Approach to Inferring Missing Phenotypes

To alleviate the experimental challenges, this dissertation develops a computational approach to infer missing phenotypes from sparsely-sampled genotype-phenotype maps. In the PfCRT space, our model could shave off years of experimental work. We followed simple approach proposed by R.A. Fisher in 1918. He modeled each mutation as an independent perturbation to a quantitative phenotype [14, 15]. The phenotype of any genotype is simply the sum of its mutations. This is known as an additive model. We can estimate the effects of mutations in an additive model using ordinary linear regression. The regression has the form:

$$P_{obs} = P_{add} + \epsilon \tag{1}$$

where P_{obs} is the observed phenotypes, ϵ represents the regression residuals, P_{add} is the additive model. P_{add} has the form

$$P_{add} = \beta_{ref} + \beta_1 x_1 + \beta_2 x_2 + ... \tag{2}$$

where β_i represents the quantitative effect of mutation i, x_i is 0 if the mutation is present and 1 otherwise, and β_{wt} is a reference phenotype.

When we apply an additive model to the PfCRT genotype-phenotype map, it yields a fit with $R^2 = 0.65$. The error in the predictions can be seen in the correlation plot, P_{obs} versus P_{add}, shown in Figure 2C. A few immediate problems appear in the correlation plot of P_{obs} versus P_{add} that we will address throughout this dissertation.

Epistasis Impairs Prediction

The residual term in the additive model has a special name; it is called **epistasis**. Epistasis has many forms and arises from different sources [15, 16, 17, 18, 19, 20, 21, 22, 23]. Formally, epistasis is defined as a mutation having a different effect in two different genetic backgrounds. This appears as deviations from the additive model.

Epistasis plays an profound role in shaping evolutionary trajectories [2, 24, 25, 26, 16, 27, 17, 28, 29, 30, 10, 19, 31, 32, 33, 34, 20, 35, 23]. It can determine the specific order in which mutations accumulate [25, 19], stochastically open and close pathways [36, 37], and entrench mutations over evolutionary time, [38, 33], and lead to long range history contingency in a genotype-phenotype map [39, 36].

Epistasis is can be observed in PfCRT. In Figure 2D, the effect of each mutation in the trajectory is shown as black bars in Figure 2B. Summers et al. also measured the effect of these mutations in the derived genetic background, shown as white bars. We can see that each mutation has a different effect when introduced in the trajectory versus in the derived background. This observation is the result of epistasis. Interestingly, the +356T switches sign. It has a positive effect in the trajectory but a negative effect in the derived background. This is an extreme case known as sign epistasis. This non-additivity in the PfCRT genotype-phenotype map causes the additive model to yield poor results.

This dissertation addresses epistasis when predicting phenotypes. The following chapters analyze epistasis in various genotype-phenotype maps and identifies key features that improve predictive power. Finally, we apply our improved model to the PfCRT genotype-phenotype map and predict unknown phenotypes. We then collaborate with the Martin lab (Australia National University), responsible for the original PfCRT data in Summers et. al [11], to experimentally characterize 25 new genotypes. We then compare these to our predictions. In the final chapter, we layout the limitations of this approach and propose alternative avenues for future improvement.

Chapter-by-chapter Breakdown

Chapter II addresses the problem of epistasis by dissecting epistasis into two sources: global and local epistasis. We analyze various experimental genotype-phenotype maps collected from the literature. We identify global epistasis–or nonlinearity between the genotype and phenotype–in many of these maps. It explains a small fraction of the epistasis on these maps. We then analyze these maps for local epistasis. We find extensive high-order epistasis–or specific interactions between two, three, four, or even more mutations. We employ a rigorous statistical approach to confirm the observed epistasis is not due to experimental uncertainty. From this, we conclude that high-order epistasis is a ubiquitous feature of biology. The work in this chapter was published as a research article in the journal *Genetics,* and co-authored with Prof. Michael J. Harms.

Chapter III then explores the role that high-order epistasis plays in shaping evolutionary trajectories through genotype-phenotype maps. We pull various experimental genotype-phenotype maps from the literature. We enumerate evolutionary trajectories through each map and calculated their relative probabilities. We then computationally removed local epistatic interactions and observe their effects on these trajectories. In every map, we find that removing high-order epistasis drastically changed the trajectories we observed. From this, we conclude that high-order epistasis plays a significant role in evolution, and therefore, cannot be ignored. The work in this chapter was published as a research article in the journal*, PLoS Computational Biology* , and co-authored with Prof. Michael J. Harms.

Chapter IV addresses one possible source of high-order epistasis, molecular ensembles. We use a simple toy model of proteins, called protein lattice models, to explore how the protein's conformational ensemble shapes it's evolutionary trajectories. We first find that a two-state protein exhibits no high-order epistasis. On the other hand, we find that a 3+ state protein exhibits extensive high-order epistasis. This suggests that high-order epistasis in lattice proteins arises from the conformational ensemble. We examine the ensemble over the course of an evolutionary trajectory and find the cause of high-order epistasis. Because mutations affect each conformation in a slightly different way, the effect of a mutation changes over the course of an evolutionary trajectory. This leads to extensive high-order epistasis in lattice proteins. We concluded that this phenomena is likely ubiquitous in molecular systems since conformational ensembles are a common feature in biological systems. The work in this chapter was published as a research article in the journal, *Proceedings of the National Academy of Sciences,* and co-authored with Prof. Michael J. Harms.

Chapter V addresses a major problem when including local epistasis to predict missing phenotypes in sparsely-sampled genotype phenotype maps. Using both computational and experimental genotype-phenotype maps, we show that adding local epistasis yields poor predictions. The problem with fitting epistasis to sparse data is that it leads to biased estimates of epistatic coefficients. As a result, the predictions are biased. Fitting global epistasis, on the other hand, improves predictions. Thus, we conclude that the best approach to predicting phenotypes is to ignore local epistasis. The work in this chapter is currently in preparation and co-authored with Prof. Michael J. Harms.

In Chapter VI, we apply our model to infer missing phenotypes in the PfCRT genotype-phenotype map. We work in collaboration with Rowena Martin's group at Australia National University. The Martin group measured 25 genotypes not previously characterized, while we generated predictions. We address a key feature in the observed phenotype. Two mutations are necessary for CQ transport. We model this feature by including a preprocessing step that classifies genotypes as no-CQ transport or positive-CQ transport. We then fit a nonlinear additive model to positive-CQ genotypes. We predict the 204 unknown phenotype and find that many of the genotypes are nonfunctional. We then compare our predictions to the genotypes measured by the Martin group and find that our model accurately predicts with a known uncertainty. Finally, we map our predictions on the full PfCRT genotype-phenotype map and infer evolutionary trajectories that lead to high-CQ uptake. We find that many trajectories are inaccessible. The work in this chapter is currently in preparation and co-authored with Robert Summers, Sarah Shafik, Alex Joule, Prof. Rowena Martin and Prof. Michael J. Harms.

Broader Impacts

An immediate broader impact from this work is the prediction of unknown PfCRT genotypes. We were able to draw new conclusions about evolution of PfCRT. Specifically, we reveal many low-transport proteins in the genotype-phenotype map that make many evolutionary trajectories inaccessible. We also observe many neutral steps in the observed evolutionary trajectories. This could explain why CQ resistance took so many decades to evolve. It also informs future experimental work directed at understanding the evolution of drug resistance.

An even broader impact is the generalization of our model. By the end of the dissertation, we present a general model for predicting phenotypes in sparsely-sampled genotype-phenotype maps. Together, with high-throughput experimental techniques, this model has the potential to uncover massive genotype-phenotype maps that were previously inaccessible. We also address the problem of epistasis and conclude that epistasis provides a simple metric of the uncertainty. This allows us to predict large genotype-phenotype maps from very few phenotypes with known uncertainty. Further, this dissertation informs the user how many phenotypes must be measured before reaching the maximum predictive power.

All software used in this dissertation is free and open source. The prediction models can be accessed and downloaded on Github (https://github.com/harmslab).

All of the code is written in Python and built on top of core scientific python packages. Each software package includes extensive documentation and examples on how to apply to various types of experimental data, Many examples are also written in Jupyter notebooks–reproducible electronic notebooks that include code, text, and figures in a single document. This was our best effort to follow an open and reproducible approach to science.

CHAPTER II

DETECTING HIGH-ORDER EPISTASIS IN NONLINEAR GENOTYPE-PHENOTYPE MAPS

Author Contributions

Zachary Sailer (ZRS) and Michael Harms (MJH) conceptualized the paper. ZRS conducted the simulations and computational analysis. ZRS generated figures. ZRS and MJH wrote and edited the manuscript.

Abstract

High-order epistasis has been observed in many genotype-phenotype maps. These multi-way interactions between mutations may be useful for dissecting complex traits and could have profound implications for evolution. Alternatively, they could be a statistical artifact. High-order epistasis models assume the effects of mutations should add, when they could in fact multiply or combine in some other nonlinear way. A mismatch in the "scale" of the epistasis model and the scale of the underlying map would lead to spurious epistasis. In this paper, we develop an approach to estimate the nonlinear scales of arbitrary genotype-phenotype maps. We can then linearize these maps and extract high-order epistasis. We investigated seven experimental genotype-phenotype maps for which high-order epistasis had been reported previously. We find that five of the seven maps exhibited nonlinear scales. Interestingly, even after accounting for nonlinearity, we found statistically significant high-order epistasis in all seven maps. The contributions of high-order epistasis to the total variation ranged from 2.2% to 31.0%, with an average across maps of 12.7%. Our results provide

strong evidence for extensive high-order epistasis, even after nonlinear scale is taken into account. Further, we describe a simple method to estimate and account for nonlinearity in genotype-phenotype maps.

Introduction

Recent analyses of genotype-phenotype maps have revealed "high-order" epistasis—that is, interactions between three, four, and even more mutations [40, 41, 32, 42, 43, 44, 45, 46, 27, 47, 48, 49, 50, 51]. The importance of these interactions for understanding biological systems and their evolution is the subject of current debate [44, 52]. Can they be interpreted as specific, biological interactions between loci? Or are they misleading statistical correlations?

We set out to tackle one potential source of spurious epistasis: a mismatch between the "scale" of the map and the scale of the model used to dissect epistasis [14, 53, 54, 15, 16, 7]. The scale defines how to combine mutational effects. On a linear scale, the effects of individual mutations are added. On a multiplicative scale, the effects of mutations are multiplied. Other, arbitrarily complex scales, are also possible [55, 56, 57].

Application of a linear model to a nonlinear map will lead to apparent epistasis [14, 53, 54, 15, 16, 7]. Consider a map with independent, multiplicative mutations. Analysis with a multiplicative model will give no epistasis. In contrast, analysis with a linear model will give epistatic coefficients to account for the multiplicative nonlinearity [15, 16]. Epistasis arising from a mismatch in scale is mathematically valid, but obscures a key feature of the map: its scale. It is also not parsimonious, as it uses many coefficients to describe a potentially simple nonlinear function. Finally, it can be misleading because these epistatic coefficients partition global information about the nonlinear scale into (apparently) specific interactions between mutations.

Most high-order epistasis models assume a linear scale (or a multiplicative scale transformed onto a linear scale) [58, 7, 44, 52]. These models sum the independent effects of mutations to predict multi-mutation phenotypes. Epistatic coefficients account for the difference between the observed phenotypes and the phenotypes predicted by summing mutational effects. The epistatic coefficients that result are, by construction, on the same linear scale [58, 44, 52].

Because the underlying scale of genotype-phenotype maps is not known *a priori*, the interpretation of high-order epistasis extracted on a linear scale is unclear. If a nonlinear scale can be found that removes high-order epistasis, it would suggest that high-order epistasis is spurious: a highly complex description of a simple, nonlinear system. In contrast, if no such scale can be found, high-order epistasis provides a window into the profound complexity of genotype-phenotype maps.

In this paper, we set out to estimate the nonlinear scales of experimental genotype-phenotype maps. We then account for these scales in the analysis of high-order epistasis. We took our inspiration from the treatment of multiplicative maps, which can be transformed into linear maps using a log transform. Along these same lines, we set out to transform genotype-phenotype maps with arbitrary, nonlinear scales onto a linear scale for analysis of high-order epistasis. We develop our methodology using simulations and then apply it to experimentally measured genotype-phenotype maps.

Materials and Methods

Experimental data sets

We collected a set of published genotype-phenotype maps for which high-order epistasis had been reported previously. Measuring an L^{th}-order interaction requires knowing the phenotypes of all binary combinations of L mutations—that is, 2^L genotypes. The data sets we used had exhaustively covered all 2^L genotypes for five or six mutations. These data sets cover a broad spectrum of genotypes and phenotypes. Genotypes included point mutations to a single protein [25], point mutations in both members of a protein/DNA complex [32], random genomic mutations [18, 6], and binary combinations of alleles within a biosynthetic network [59]. Measured phenotypes included selection coefficients [25, 18, 6], molecular binding affinity [32], and yeast growth rate [59]. (For several data sets, the "phenotype" is a selection coefficient. We do not differentiate fitness from other properties for our analyses; therefore, for simplicity, we will refer to all maps as genotype-phenotype maps rather than specifying some as genotype-fitness maps). All data sets had a minimum of three independent measurements of the phenotype for each genotype. All data sets are available in a standardized ascii text format.

Nonlinear scale

We described nonlinearity in the genotype-phenotype map by a power transformation (see Results) [60, 61]. The independent variable for the transformation was $\vec{P}_{add}$, the predicted phenotypes of all genotypes assuming linear and additive affects for each mutation. The estimated additive phenotype of genotype i, is given by:

$$\hat{P}_{add,i} = \sum_{j=1}^{j \leq L} \langle \Delta P_j \rangle \, x_{i,j} \tag{3}$$

where $\langle \Delta P_j \rangle$ is the average effect of mutation j across all backgrounds, $x_{i,j}$ is an index that encodes whether or not mutation j is present in genotype i, and L is the number of sites. The dependent variables are the observed phenotypes $\vec{P}_{obs}$ taken from the experimental genotype-phenotype maps.

We use nonlinear least-squares regression to fit and estimate the power transformation from $\vec{P}_{add}$ to $\vec{P}_{obs}$:

$$\vec{P}_{obs} \sim \tau(\hat{\vec{P}}_{add}; \hat{\lambda}, \hat{A}, \hat{B}) + \hat{\varepsilon},$$

where ε is a residual and τ is a power transform function. This is given by:

$$\vec{P}_{obs} = \frac{(\hat{\vec{P}}_{add} + A)^{\lambda} - 1}{\lambda (GM)^{\lambda - 1}} + B,$$

where A and B are translation constants, GM is the geometric mean of $(\hat{\vec{P}}_{add} + A)$, and λ is a scaling parameter. We used standard nonlinear regression techniques to minimize d:

$$d = (\vec{P}_{scale} - \vec{P}_{obs})^2 + \varepsilon.$$

We then reversed this transformation to linearize P_{obs} using the estimated parameters $\hat{A}$, $\hat{B}$, and $\hat{\lambda}$. We did so by the back-transform:

$$P_{obs,linear} = \{\hat{\lambda}(GM)^{\lambda-1}(P_{obs} - \hat{B}) + 1\}^{1/\hat{\lambda}} - \hat{A}. \quad (4)$$

High-order epistasis model

We dissected epistasis using a linear, high-order epistasis model. These have been discussed extensively elsewhere [58, 52, 44], so we will only briefly and informally review them here.

A high-order epistasis model is a linear decomposition of a genotype-phenotype map. It yields a set of coefficients that account for all variation in phenotype. The signs and magnitudes of the epistatic coefficients quantify the effect of mutations and interactions between them. A binary map with 2^L genotypes requires 2^L epistatic coefficients and captures all interactions, up to L^{th}-order, between them. This is conveniently described in matrix notation.

$$\vec{P} = \mathbf{X}\vec{\beta}: \quad (5)$$

a vector of phenotypes $\vec{P}$ can be transformed into a vector of epistatic coefficients

$\vec{\beta}$ using a $2^L \times 2^L$ decomposition matrix that encodes which coefficients contribute to which phenotypes. If $\mathbf{X}$ is invertible, one can determine $\vec{\beta}$ from a collection of measured phenotypes by

$$\vec{\beta} = \mathbf{X}^{-1}\vec{P}. \tag{6}$$

$\mathbf{X}$ can be formulated in a variety of ways [52]. Following others in the genetics literature, we use the form derived from Walsh polynomials [58, 44, 52]. In this form, $\mathbf{X}$ is a Hadamard matrix. Conceptually, the transformation identifies the geometric center of the genotype-phenotype map and then measures the average effects of each mutation and combination of mutations in this "average" genetic background (Figure 3). To achieve this, we encoded each mutation at each site in each genotype as -1 (wildtype) or $+1$ (mutant) [58, 44, 52]. This has been called a Fourier analysis,[7, 62], global epistasis [52], or a Walsh space [58, 25]. Another common approach is to use a single wildtype genotype as a reference and encode mutations as either 0 (wildtype) or 1 (mutant) [52].

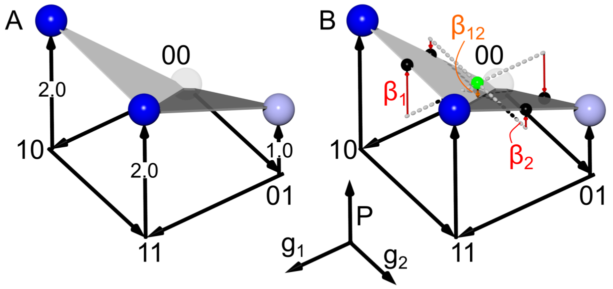


Figure 3: Epistasis can be quantified using Walsh polynomials. A) A genotype-phenotype map exhibiting negative epistasis. Axes are genotype at position 1 (g_1), genotype at position 2 (g_2), and phenotype (P). For genotypic axes, "0" denotes wildtype and "1" denotes a mutant. Phenotype is encoded both on the P-axis and as a spectrum from white to blue. The map exhibits negative epistasis: relative to wildtype, the effect of the mutations together ($P_{11} = 2$) is less than the sum of the individual effects of mutations ($P_{10} + P_{01} = 1 + 2 = 3$). B) The map can be decomposed into epistatic coefficients using a Walsh polynomial, which measures the effects of each mutation relative to the geometric center of the genotype-phenotype map (green sphere). The additive coefficients β_1 and β_2 (red arrows) are the average effect of each mutation in all backgrounds. The epistatic coefficient β_{12} (orange arrow) the variation not accounted for by β_1 and β_2. Geometrically, it is the distance between the center of the map and the "fold" given by vector connecting P_{00} and P_{11}.

One data set (IV, Table I) has four possible states (A, G, C and T) at two of the sites. We encoded these using the WYK tetrahedral-encoding scheme[63, 32]. Each state is encoded by a three-bit state. The wildtype state is given the bits $(1, 1, 1)$. The remaining states are encoded with bits that form corners of a tetrahedron. For example, the wildtype of site 1 is G and encoded as the $(1, 1, 1)$ state. The remaining states are encoded as follows: A is $(1, -1, -1)$, C is $(-1, 1, -1)$ and T is $(-1, -1, 1)$.

ID	genotype	phenotype	L	ref
I	scattered genomic mutations	*E. coli* fitness	5	[18]
II	chromosomes in asexual fungi	*A. niger* fitness	5	[6]
III	protein point mutants	bacterial fitness	5	[25]
IV	DNA/protein point mutants	DNA/protein binding affinity	5	[32]
V	chromosomes in asexual fungi	*A. niger* fitness	5	[6]
VI	alleles in biosynthetic network	*S. cerevisiae* haploid growth rate	6	[59]
VII	alleles in biosynthetic network	*S. cerevisiae* diploid growth rate	6	[59]

Table 1: All data sets have 2^L genotypes except the DNA/protein interaction data set (IV), which has 128 genotypes. This occurs because the data set has 2 DNA sites (each of which have 4 possible bases) and 3 protein sites (each of which has two possible amino acids).

Experimental uncertainty

We used a bootstrap approach to propagate uncertainty in measured phenotypes into uncertainty in epistatic coefficients. To do so we: 1) calculated the mean and standard deviation for each phenotype from the published experimental replicates; 2) sampled the uncertainty distribution for each phenotype to generate a pseudoreplicate vector $\vec{P}_{pseudo}$ that had one phenotype per genotype; 3) rescaled $\vec{P}_{pseudo}$ using a power-transform; and 4) determined the epistatic coefficients for $\vec{P}_{pseudo,scaled}$. We then repeated steps 2-4 until convergence. We determined the mean and variance of each epistatic coefficient after every 50 pseudoreplicates. We defined convergence as the mean and variance of every epistatic coefficient changed by < 0.1 % after addition of 50 more pseudoreplicates. On average, convergence required $\approx 100,000$ replicates per genotype-phenotype map. Finally, we used a z-score to determine if each epistatic coefficient was significantly different than zero. To account for multiple testing, we applied a Bonferroni correction to all p-values [64].

Computational methods

Our full epistasis software package—written in Python3 extended with Numpy and Scipy [65]—is available for download via github (https://harmslab.github.com/epistasis).

We used the python package *scikit-learn* for all regression [66]. Plots were generated using *matplotlib* and *jupyter* notebooks [67, 68].

Results and Discussion

Nonlinear scale induces apparent high-order epistasis

Our first goal was to understand how a nonlinear scale, if present, would affect estimates of high-order epistasis. To probe this question, we constructed a five-site binary genotype-phenotype map on a nonlinear scale, and then extracted epistasis assuming a linear scale. The nonlinear scale we chose was a saturating function:

$$P_{g,trans} = \frac{(1+K)P_g}{1+KP_g}, \tag{7}$$

where P_g is the linear phenotype of genotype g, $P_{g,trans}$ is the transformed phenotype of genotype g, and K is a scaling constant. As $K \to 0$, the map becomes linear. As K increases, mutations have systematically smaller effects when introduced into backgrounds with higher phenotypes.

We calculated P_g for all 2^L binary genotypes using the random, additive coefficients shown in Figure 4A. These coefficients included no epistasis. We then transformed P_g onto the nonlinear $P_{g,trans}$ scale using Equation 7 with the relatively shallow ($K = 2$) saturation curve shown in Figure 4B. Finally, we applied a linear epistasis model to $P_{g,trans}$ to extract epistatic coefficients.

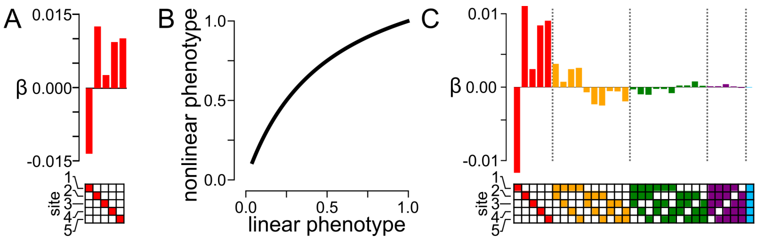


Figure 4: Nonlinearity in phenotype creates spurious high-order epistatic coefficients. A)Simulated, random, first-order epistatic coefficients. The mutated site is indicated by panel below the bar graph; bar indicates magnitude and sign of the additive coefficient. B) A nonlinear map between a linear phenotype and a saturating, nonlinear phenotype. The first-order coefficients in panel A are used to generate a linear phenotype, which is then transformed by the function shown in B. C) Epistatic coefficients extracted from the genotype-phenotype map generated in panels A and B. Bars denote coefficient magnitude and sign. Color denotes the order of the coefficient: first (β_i, red), second (β_{ij}, orange), third (β_{ijk}, green), fourth (β_{ijkl}, purple), and fifth (β_{ijklm}, blue). Filled squares in the grid below the bars indicate the identity of mutations that contribute to the coefficient.

We found that nonlinearity in the genotype-phenotype map induced extensive high-order epistasis when the nonlinearity was ignored (Figure 4C). We observed epistasis up to the fourth order, despite building the map with purely additive coefficients. This result is unsurprising: the only mechanism by which a linear model can account for variation in phenotype is through epistatic coefficients [53, 54, 15]. When given a nonlinear map, it partitions the variation arising from nonlinearity into specific interactions between mutations. This high-order epistasis is mathematically valid, but does not capture the major feature of the map—namely, saturation. Indeed, this epistasis is deceptive, as it is naturally interpreted as specific interactions between mutations. For example, this analysis identifies a specific interaction between mutations one, two, four, and five (Figure 4C, purple). But this four-way interaction is an artifact of the nonlinearity in phenotype of the map, rather than a specific interaction.

Nonlinear scale and specific epistatic interactions induce different patterns of non-additivity

Our next question was whether we could separate the effects of nonlinear scale and high-order epistasis in binary maps. One useful approach to develop intuition about epistasis is to plot the the observed phenotypes (P_{obs}) against the predicted phenotype of each genotype, assuming linear and additive mutational effects (P_{add}) [55, 56, 7]. In a linear map without epistasis, P_{obs} equals P_{add}, because each mutation would have the same, additive effect in all backgrounds. If epistasis is present, phenotypes will diverge from the $P_{obs} = P_{add}$ line.

We simulated maps including varying amounts of linear, high-order epistasis, placed them onto increasingly nonlinear scales, and then constructed P_{obs} vs. P_{add} plots. We added high-order epistasis by generating random epistatic coefficients and then calculating phenotypes using Eq. 5. We introduced nonlinearity by transforming these phenotypes with Eq. 7. For each genotype in these simulations, we calculated P_{add} as the sum of the first-order coefficients used in the generating model. P_{obs} is the observable phenotype, including both high-order epistasis and nonlinear scale.

High-order epistasis and nonlinear scale had qualitatively different effects on P_{obs} vs. P_{add} plots. Figure 5A shows plots of P_{obs} vs. P_{add} for increasing nonlinearity (left-to-right) and high-order epistasis (bottom-to-top). As nonlinearity increases, P_{obs} curves systematically relative to P_{add}. This reflects the fact that P_{add} is on a linear scale and P_{obs} is on a saturating, nonlinear scale. The shape of the curve reflects the map between the linear and saturating scale: the smallest phenotypes are underestimated and the largest phenotypes overestimated. In contrast, high-order epistasis induces random scatter away from the $P_{obs} = P_{add}$ line. This is because the epistatic coefficients used to generate the map are specific to each genotype, moving observations off the expected line, even if the scaling relationship is taken into account.

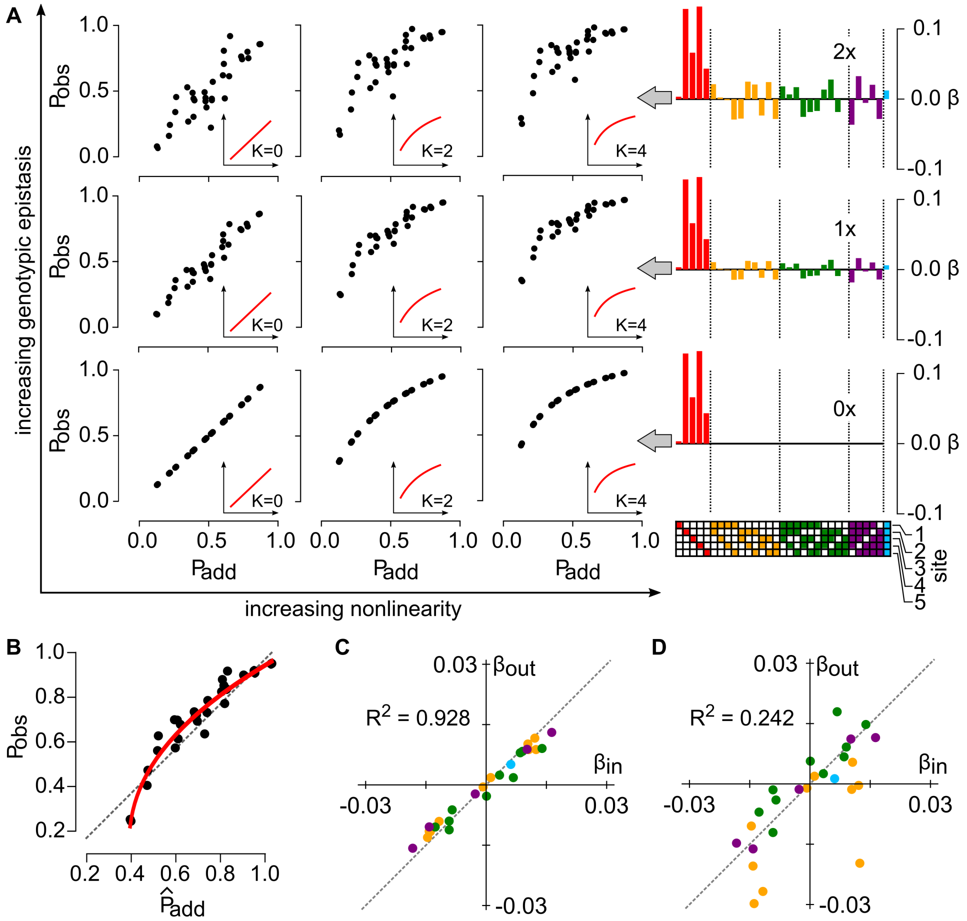


Figure 5: Epistasis and nonlinear scale induce different patterns of nonadditivity. A) Patterns of nonadditivity for increasing epistasis and nonlinear scale. Main panel shows a grid ranging from no epistasis, linear scale (bottom-left) to high epistasis, highly nonlinear scale (top-right). Insets in sub-panels show added nonlinearity. Going from left to right: $K = 0$, $K = 2$, $K = 4$. Epistatic coefficient plots to right show the magnitude of the input high-order epistasis, with colors and annotation as in Fig 2C. B) Plot of P_{obs} against $\hat{P}_{add}$ for the middle sub panel in panel A. Red line is the fit of the power transform to these data. C) Correlation between epistatic coefficients input into the simulation and extracted from the simulation after linearization by the power transform. Each point is an epistatic coefficient, colored by order. The Pearson's correlation coefficient is shown in the upper-left quadrant. D) Correlation between epistatic coefficients input into the simulation and extracted from the simulation without application of the power transform.

Nonlinearity can be separated from underlying high-order epistasis

The P_{obs} vs. P_{add} plots suggest an approach to disentangle high-order epistasis from nonlinear scale. By fitting a function to the P_{obs} vs P_{add} curve, we describe a transformation that relates the linear P_{add} scale to the (possibly nonlinear) P_{obs} scale [56, 7]. Once the form of the nonlinearity is known, we can then linearize the phenotypes so they are on an appropriate scale for epistatic analysis. Variation that remains (i.e. scatter) can then be confidently partitioned into epistatic coefficients.

In the absence of knowledge about the source of the nonlinearity, a natural choice is a power transform [60, 61], which identifies a monotonic, continuous function through P_{obs} vs. P_{add}. A key feature of this approach is that power-transformed data are normally distributed around the fit curve and thus appropriately scaled for regression of a linear epistasis model.

We tested this approach using one of our simulated data sets. One complication is that, for an experimental map, we do not know P_{add}. In the analysis above, we determined P_{add} from the additive coefficients used to generate the space. In a real map, P_{add} is not known; therefore, we had to estimate P_{add}. We did so by measuring the average effect of each mutation across all backgrounds, and then calculating $\hat{P}_{add}$ for each genotype as the sum of these average effects (Eq. 3).

We fit the power transform to P_{obs} vs. $\hat{P}_{add}$ (solid red line, Figure 5B). The curve captures the nonlinearity added in the simulation. We linearized P_{obs} using the fit model (Eq. 4), and then extracted high-order epistatic coefficients. The extracted coefficients were highly correlated with the coefficients used to generate the map ($R^2 = 0.998$) (Figure 5C). In contrast, applying the linear epistasis model to this map without first accounting for nonlinearity gives much greater scatter between the input and output coefficients ($R^2 = 0.934$) (Figure 5D). This occurs because phenotypic variation from nonlinearity is incorrectly partitioned into the linear epistatic coefficients.

Nonlinearity is a common feature of genotype-phenotype maps

Our next question was whether experimental maps exhibited nonlinear scales. We selected seven genotype-phenotype maps that had previously been reported to exhibit high-order epistasis (Table 1) and fit power transforms to each dataset (Figure 6, S1). We expected some phenotypes to be multiplicative (e.g. datasets I, II and IV were relative fitness), while we expected some to be additive (e.g. dataset IV is a free energy). Rather than rescaling the multiplicative datasets by taking logarithms of the phenotypes, we allowed our power transform to capture the appropriate scale. The power-transform identified nonlinearity in the majority of data sets. Of the seven data sets, three were less-than-additive (II, V, VI), two were greater-than-additive (III, IV), and two were approximately linear (I, VII). All data sets gave random residuals after fitting the power transform (Figure 6, S1).

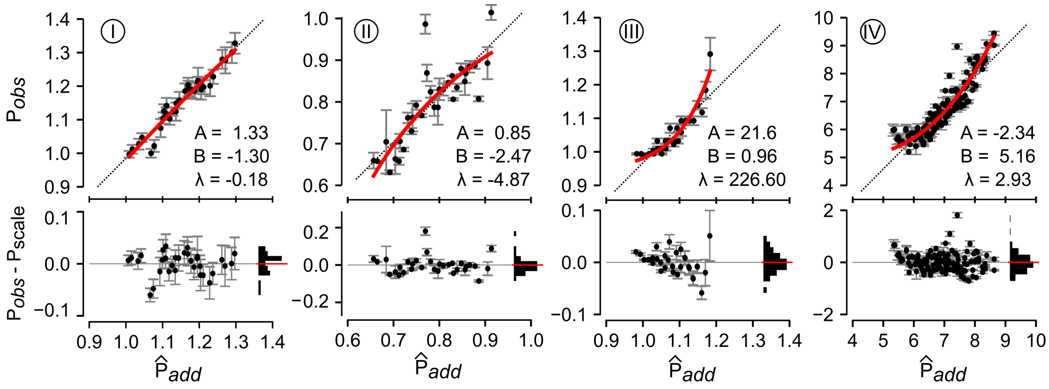


Figure 6: Experimental genotype-phenotype maps exhibit nonlinear phenotypes. Plots show observed phenotype P_{obs} plotted against $\hat{P}_{add}$ (Eq. 3) for data sets I through IV. Points are individual genotypes. Error bars are experimental standard deviations in phenotype. Red lines are the fit of the power transform to the data set. Pearson's coefficient for each fit are shown on each plot. Dashed lines are $P_{add} = P_{obs}$. Bottom panels in each plot show residuals between the observed phenotypes and the red fit line. Points are the individual residuals. Error bars are the experimental standard deviation of the phenotype. The horizontal histograms show the distribution of residuals across 10 bins. The red lines are the mean of the residuals.

High-order epistasis is a common feature of genotype-phenotype maps

With estimated scales in hand, we linearized the maps using Eq. 4 and re-measured epistasis (Figure S2). We used bootstrap sampling of uncertainty in the measured phenotypes to determine the uncertainty of each epistatic coefficient (see Methods), and then integrated these distributions to determine whether each coefficient was significantly different than zero. We then applied a Bonferroni correction to each p-value to account for multiple testing.

Despite our conservative statistical approach, we found high-order epistasis in every map studied (Figure 7A, S3). Every data set exhibited at least one statistically significant epistatic coefficient of fourth order or higher. We even detected statistically significant fifth-order epistasis (blue bar in Figure 7A, data set II). High-order coefficients were both positive and negative, often with magnitudes equal to or greater than the second-order terms. These results reveal that high-order epistasis is a robust feature of these maps, even when nonlinearity and measurement uncertainty in the genotype-phenotype map is taken into account.

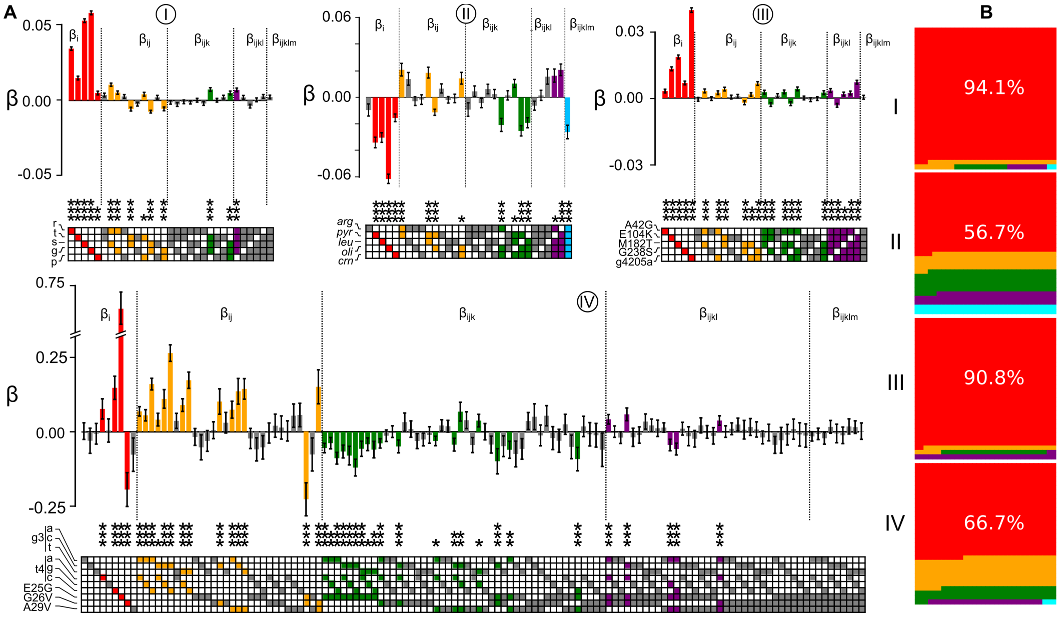


Figure 7: High-order epistasis is present in genotype-phenotype maps. A) Panels show epistatic coefficients extracted from data sets I-IV (Table 1, data set label circled above each graph). Bars denote coefficient magnitude and sign; error bars are propagated measurement uncertainty. Color denotes the order of the coefficient: first (β_i, red), second (β_{ij}, orange), third (β_{ijk}, green), fourth (β_{ijkl}, purple), and fifth (β_{ijklm}, blue). Bars are colored if the coefficient is significantly different than zero (Z-score with p-value < 0.05 after Bonferroni correction for multiple testing). Stars denote relative significance: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***). Filled squares in the grid below the bars indicate the identity of mutations that contribute to the coefficient. The names of the mutations, taken from the original publications, are indicated to the left of the grid squares. B) Sub-panels show fraction of variation accounted for by first through fifth order epistatic coefficients for data sets I-IV (colors as in panel A). Fraction described by each order is proportional to area.

We also dissected the relative contributions of each epistatic order to the remaining variation. To do so, we created truncated epistasis models: an additive model, a model containing additive and pairwise terms, a model containing additive through third-order terms, etc. We then measured how well each model accounted for variation in the phenotype using a Pearson's coefficient between the fit and the data. Finally, we asked how much the Pearson coefficient changed with addition of more epistatic coefficients. For example, to measure the contribution of pairwise epistasis, we took the difference in the correlation coefficient between the additive plus pairwise model and the purely additive model.

The contribution of epistasis to the maps was highly variable (Figure 7B, S3). For data set I, epistatic terms explained 5.9% of the variation in the data. The contributions of epistatic coefficients decayed with increasing order, with fifth-order epistasis only explaining 0.1% of the variation in the data. In contrast, for data set II, epistasis explains 43.3% of the variation in the map. Fifth-order epistasis accounts for 6.3% of the variation in the map. The other data sets had epistatic contributions somewhere between these extremes.

Accounting for nonlinear genotype-phenotype maps alters epistatic coefficients

Finally, we probed to what extent accounting for nonlinearity in phenotype altered the epistatic coefficients extracted from each space. Figure 8 and S4 show correlation plots between epistatic coefficients extracted both with and without linearization. The first-order coefficients were all highly correlated between the linear and nonlinear analyses for all data sets (Figure S5).

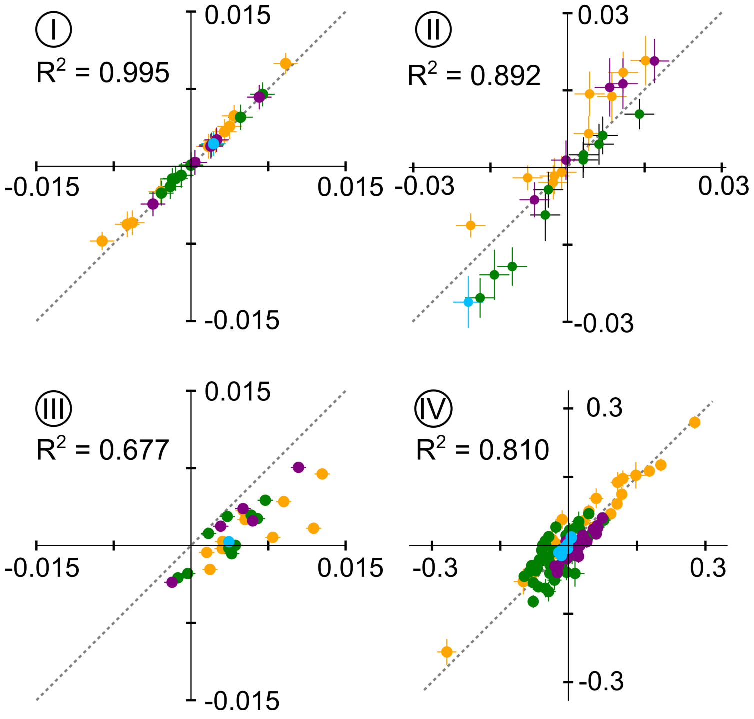


Figure 8: Nonlinear phenotypes distort measured epistatic coefficients. Sub-panels show correlation plots between epistatic coefficients extracted without accounting for nonlinearity (x-axis) and accounting for linearity (y-axis) for data sets I-IV. Each point is an epistatic coefficient, colored by order. Error bars are standard deviations from bootstrap replicates of each fitting approach.

For the epistatic coefficients, the degree of correlation depended on the degree of nonlinearity in the dataset. Data set I—which was essentially linear—had identical epistatic coefficients whether the nonlinear scale was taken into account or not. In contrast, the other data sets exhibited scatter off of the line. Data set III was particularly noteworthy. The epistatic coefficients were systematically overestimated when the nonlinear scale was ignored. Two large and favorable pairwise epistatic terms in the linear analysis became essentially zero when nonlinearity was taken into account. These interactions—M182T/g4205a and G283S/g4205a—were both noted as determinants of evolutionary trajectories in the original publication [25]; however, our results suggest the interaction is an artifact of applying a linear model to a nonlinear data set. Further $\approx 20\%$ (six of 27) epistatic coefficients flipped sign when nonlinearity was taken into account (Figure 8, III, bottom right quadrant).

Overall, we found that low-order epistatic coefficients were more robust to the linear assumption than high-order coefficients. Data set IV is a clear example of this behavior. The map exhibited noticeable nonlinearity (Figure 6). The first- and second-order terms were well correlated between the linear and nonlinear analyses (Figure 8, S4, S5). Higher-order terms, however, exhibited much poorer overall correlation. While the R^2 for second-order coefficients was 0.95, the correlation was only 0.43 for third-order. This suggests that previous analyses of nonlinear genotype-phenotype maps correctly identified the key mutations responsible for variation in the map, but incorrectly estimated the high-order epistatic effects.

Discussion

Our results reveal that both nonlinear scales and high-order epistasis play important roles in shaping experimental genotype-phenotype maps. Five of the seven data sets we investigated exhibited nonlinear scales, and all of the data sets exhibited high-order epistasis, even after accounting for nonlinearity. This suggests that both should be taken into account in analyses of genotype-phenotype maps.

Origins of nonlinear scales

We observed two basic forms of nonlinearity in these maps: saturating, less-than-additive maps and exploding, greater-than-additive maps. Many have observed less-than-additive maps in which mutations have lower effects when introduced into more optimal backgrounds [69, 17]. Such saturation has been proposed to be a key factor shaping evolutionary trajectories [69, 17, 19, 70, 71]. Further, it is intuitive that optimizing a phenotype becomes more difficult as that phenotype improves. Our nonlinear fits revealed this behavior in three different maps.

The greater-than-additive maps, in contrast, were more surprising: why would mutations have a larger effect when introduced into a more favorable background?

For the β-lactamase genotype-phenotype map (III, Figure 6), this may be an artifact of the original analysis used to generate the data set [25]. This data set describes the fitness of bacteria expressing variants of an enzyme with activity against β-lactam antibiotics. The original authors measured the minimum-inhibitory concentration (MIC) of the antibiotic against bacteria expressing each enzyme variant. They then converted their MIC values into apparent fitness by sampling from an exponential distribution of fitness values and assigning these fitness values to rank-ordered MIC values [25]. Our epistasis model extracts this original exponential distribution (Figure S6). This result demonstrates the effectiveness of our approach in extracting nonlinearity in the genotype-phenotype map.

The origins of the growth in the transcription factor/DNA binding data set are less clear (IV, Figure 6). The data set measures the binding free energy of variants of a transcription factor binding to different DNA response elements. We are aware of no physical reason for mutations to have a larger effect on free energy when introduced into a background with better binding. One possibility is that the genotype-phenotype map reflects multiple features that are simultaneously altered by mutations, giving rise to this nonlinear shape. This is a distinct possibility in this data set, where mutations are known to alter both DNA binding affinity and DNA binding cooperativity [72].

Best Practice

Because nonlinearity is a common feature of these maps, linearity should not be assumed in analyses of epistasis. Given a sufficient number of phenotypic observations, however, the appropriate scale can be estimated by construction of a P_{obs} vs. P_{add} plot and regression of a nonlinear scale model. With this scale in hand, one can then transform the genotype-phenotype map onto a linear scale appropriate for analysis using a high-order epistasis model. Our software pipeline automates this process. It takes any genotype-phenotype map in a standard text format, fits for nonlinearity, and then estimates high-order epistasis. It is freely available for download (https://harmslab.github.com/epistasis).

One important question is how to select an appropriate function to describe the nonlinear scale. By visual inspection, all of the data sets we studied were monotonic in $\hat{P}_{add}$ and could be readily captured by a power transform. Other maps may be better captured with other functions. For example, inspection of a P_{obs} vs $\hat{P}_{add}$ plot could reveal a non-monotonic scale, leading to a better fit with a polynomial than a power-transform. Another possibility is that external biological knowledge motivates scale choice [56].

The choice of model determines what fraction of the variation is assigned to "scale" versus "epistasis." The more complicated the function chosen, the more variation in the data is shifted from epistasis and into scale. One could, for example, fit a completely uninformative Lth-order polynomial, which would capture all of variation as scale and none as epistasis. Scale estimation should be governed by the well-established principles of model regression: find the simplest function that captures the maximum amount of variation in the data set without fitting stochastic noise. Because epistasis is scatter off the scale line (noise), model-selection approaches like the F-test, Akaike Information Criterion, and inspection of fit residuals are a natural strategy for partitioning variation between scale and epistasis.

Interpretation

Another powerful aspect of this approach is that it allows explicit separation of two distinct origins of non-additivity in genotype-phenotype maps.

This can be illustrated with a simple, conceptual, example. Imagine mutations to an enzyme, expressed in bacteria, that have a less-than-additive effect on bacterial growth rate. To a first approximation, this epistasis could have two origins. The first is at the level of the enzyme: maybe the mutations have a specific, negative chemical interactions that alter enzyme rate. The second is at the level of the whole cell: maybe, above a certain activity, the enzyme is fast enough that some other part of the cell starts limiting growth. Mutations continue to improve enzyme activity, but growth rate does not reflect this. These two origins of less-than-additive behavior will have different effects in a P_{add} vs. P_{obs} plot: saturation of growth rate will appear as nonlinearity, interactions between mutations at the enzyme level will appear as linear epistasis. Our analysis would reveal this pattern and set up further experiments to tease apart these possibilities.

This may also provide important evolutionary insights. One important question is to what extent evolutionary paths are shaped by global constraints versus specific interactions that lead to specific historical contingencies [36, 33, 19]. For example, recent work has shown specific epistatic interactions lead to sequence-level unpredictability, while a globally less-than-additive scale leads to predictable phenotypes in evolution [19]. Our analysis approach naturally distinguishes these origins of non-additivity, and thus these evolutionary possibilities. Prevailing magnitude epistasis [6], global epistasis [19], and diminishing-returns epistasis [17, 69, 71, 70] will all appear as nonlinear scales. In contrast, specific interactions will appear in specific coefficients in the linear epistasis model. Our detection of nonlinearity and high-order epistasis in most datasets suggests that both forms of non-additivity will be in play over evolutionary time.

High-order epistasis

Finally, our work reveals that high-order epistasis is, indeed, a common feature of genotype-phenotype maps. Our study could be viewed as an attempt to "explain away" previously observed high-order epistasis. To do so, we both accounted for nonlinearity in the map and propagated experimental uncertainty to the epistatic coefficients. Surprisingly—to the authors, at least—high-order epistasis was robust to these corrections.

High-order epistasis can make huge contributions to genotype-phenotype maps. In data set II, third-order and higher epistasis accounts for fully 31.0% of the variation in the map. The average contribution, across maps, is 12.7%. We also do not see a consistent decay in the contribution of epistasis with increasing order. In data sets II, V and VI, third-order epistasis contributes more variation to the map than second-order epistasis. This suggests that epistasis could go to even higher orders in larger genotype-phenotype maps.

The generality of these results across all genotype-phenotype maps is unclear. The maps we analyzed were measured and published because they were “interesting,” either from a mechanistic or evolutionary perspective. Further, most of the maps have a single, maximum phenotype peak. The nonlinearity and high-order epistasis we observed may be common for collections of mutations that, together, optimize a function, but less common in “flatter” or more random genotype-phenotype maps. This can only be determined by characterization of genotype-phenotype maps with different structural features.

The observation of this epistasis also raises important questions: What are the origins of third, fourth, and even fifth-order correlations in these data sets? What, mechanistically, leads to a five-way interaction between mutations? Does neglecting high-order epistasis bias estimates of low-order epistasis [73]? What can this epistasis tell us about the biological underpinning of these maps [74, 75, 43, 76]?

The evolutionary implications are also potentially fascinating. Epistasis creates temporal dependency between mutations: the effect of a mutation depends strongly on specific mutations that fixed earlier in time [77, 78, 36, 33] How does this play out for high-order epistasis, which introduces long-range correlations across genotype-phenotype maps? Do these low magnitude interactions matter for evolutionary outcomes or dynamics? These, and questions like them, are challenging and fascinating future avenues for further research.

CHAPTER III

HIGH-ORDER EPISTASIS SHAPES EVOLUTIONARY TRAJECTORIES

Author Contributions

Zachary Sailer (ZRS) and Michael Harms (MJH) conceptualized the paper. ZRS conducted the simulations and computational analysis. ZRS generated figures. ZRS and MJH wrote and edited the manuscript.

Abstract

An important question in evolutionary biology is the extent to which past events shape later evolution. High-order epistasis—where the effect of a mutation is determined by interactions with two or more other mutations—provides insight into this link between past and present. If high-order epistasis determines evolutionary trajectories, it reveals that the effect of a mutation depends on multiple, past substitutions: the past can strongly shape the present. High-order epistasis makes small, but detectable, contributions to genotype-fitness maps, but its evolutionary consequences remain poorly understood. To determine the effect of high-order epistasis on evolutionary trajectories, we computationally removed high-order epistasis from experimental genotype-fitness maps and then compared trajectories through maps both with and without high-order epistasis. Here we show that high-order epistasis, despite its small magnitude, strongly shapes the accessibility and probability of evolutionary trajectories. This suggests that evolutionary outcomes are profoundly dependent on multiple, past events.

Author Summary

A key goal for evolutionary biologists is understanding why one evolutionary trajectory is taken rather than others. This requires understanding how individual mutations, as well as interactions between them, determine the availability of evolutionary pathways. We used a robust statistical analysis to reveal interactions between up to five mutations in published datasets, meaning that the effect of a mutation can depend on the presence or absence of four other mutations. Simulations reveal that these interactions strongly shape evolutionary trajectories, and that mutations which occur early in a trajectory continue to exert profound effects on later evolution. These multi-way interactions are ubiquitous in biological datasets, suggesting that past events strongly determine future outcomes.

Introduction

Epistasis creates historical contingency, as it means that the effect of a mutation depends on previous substitutions [25, 16, 36, 19, 9, 37]. Interactions between pairs of mutations can cause mutations to accumulate in a specific order [25, 19], stochastically open and close pathways [36, 37], and make evolution irreversible [38, 33].

The effects of high-order epistasis—interactions between three or more mutations—on evolution are less well understood. Statistically-significant high-order epistasis has been observed in multiple genotype-phenotype maps [40, 44, 42, 52, 32, 41, 22, 37, 79], even when steps are taken to minimize its contribution to epistasis models [22]. Its magnitude is generally lower than the individual and pairwise epistatic effects of mutations [22]. Several studies have suggested that it can alter evolutionary outcomes [44, 37, 79], but its overall importance for evolution is not well understood. Does high-order epistasis alter evolutionary outcomes? Or are trajectories primarily shaped by the additive and pairwise epistatic effects of mutations?

We set out to assess the effect of high-order epistasis on evolutionary trajectories through experimentally measured genotype-fitness maps. We decomposed these maps into contributions from nonlinear scale, additive effects, and epistasis at different orders ranging from second to fifth. We then calculated "truncated" maps with different orders of epistasis deleted. By comparing the fitness values and probabilities of individual evolutionary trajectories through the truncated maps, we can reveal the extent to which high-order epistasis determines evolutionary outcomes.

Materials and Methods

Removal of high-order epistasis

We used the following protocol to remove specific orders of epistasis from genotype-fitness maps. The steps correspond directly to the pipeline shown in Fig S1, which is described in detail in Sailer et al. [22].

1. We identified an appropriate, possibly nonlinear, scale for the map by fitting a power transformation to the genotype-fitness map:

$$\vec{F}_{experimental} = \frac{(\hat{\vec{F}}_{add} + A)^{\lambda} - 1}{\lambda (GM)^{\lambda - 1}} + B, \tag{8}$$

   where $\vec{F}_{experimental}$ is the vector of the observed fitness values, $\hat{\vec{F}}_{add}$ is the fitness of each genotype assuming each mutation has the same, average effect in all backgrounds, A and B are translation constants, GM is the geometric mean of $(\hat{\vec{F}}_{add} + A)$, and λ is a scaling parameter. $\hat{\vec{F}}_{add}$ is given by:

$$F_{add,i} = \sum_{j=1}^{j \leq L} \langle \Delta F_j \rangle \, x_{i,j}$$

   where $\langle \Delta F_j \rangle$ is the average effect of mutation j across all backgrounds, $x_{i,j}$ is an index that encodes whether or not mutation j is present in genotype i, and

L is the number of sites. We first regressed $\hat{F}_{add}$, and then regressed the power transform.

2. We linearized each map by transforming each element in $\vec{F}_{experimental}$ with the nonlinear scale and coefficients determined in step 1. For each element in $\vec{F}_{experimental}$, we performed:

$$F_{linear} = \{\hat{\lambda}(GM)^{\lambda-1}(F_{experimental} - \hat{B}) + 1\}^{1/\hat{\lambda}} - \hat{A}.$$

3. We decomposed the variation in fitness into epistatic coefficients using a linear decomposition of the form:

$$\vec{\beta} = \mathbf{X}^{-1}\vec{F}_{linear},$$

where $\vec{\beta}$ is a collection of epistatic coefficients (ranging from 0^{th} to L^{th} order) and $\mathbf{X}$ is a design matrix that indicates which coefficients contribute to fitness in which genotype. For most of the work described, we used a Hadamard matrix for $\mathbf{X}$, which uses the geometric center of the genotype-fitness map as a reference state. [58, 44, 52, 22]. To construct this matrix, we encoded each mutation within each genotype as -1 (wildtype) or +1 (mutant) [52, 22]. For the final section, we use a "local" matrix for $\mathbf{X}$, which measures the effect of each mutation relative to a defined reference phenotype. To construct this matrix, we encoded each mutation within each genotype as 0 (wildtype) or 1 (mutant). These to forms of $\mathbf{X}$ can be readily inter-converted [52].

4. We truncated epistasis from the linearized map by setting the epistatic coefficients from orders of interest to 0, creating $\vec{\beta}_{trunc}$.

5. We recalculated the linearized fitness values, with truncated epistasis by:

$$\vec{F}_{linear,trunc} = \mathbf{X}\vec{\beta}_{trunc}.$$

6. We transformed the $\vec{F}_{linear,trunc}$ onto the original, nonlinear scale using Eq. 8, with $\vec{F}_{linear,trunc}$ in place of $\vec{F}_{add}$.

7. We used the final $\vec{F}_{trunc}$ values to construct a genotype-fitness map in which orders of epistasis were selectively removed, leaving the global, nonlinear scale intact.

We quantified the contribution of epistasis to each map (ϕ) by determining the difference in the variation explained by the i^{th} and $(i-1)^{th}$ orders. $\phi = \rho_i^2 - \rho_{i-1}^2$, where ρ_x^2 is the squared Pearson coefficient between linear fitness values in a model truncated to order x ($\vec{F}_{linear,trunc-to-x}$) and linear fitness values determined from the original map ($\vec{F}_{linear}$).

Evolutionary Trajectories

We calculated the probability of a given evolutionary trajectory as series of independent, sequential fixation events. We assumed that the time to fixation for each mutation was much less than the time between mutations (the so-called strong selection/weak mutation regime) [80, 25, 18]. The relative probability of an evolutionary trajectory i is the product of its required fixation events relative to all possible trajectories:

$$p_i = \frac{\prod_{x \in S_i} \pi_{x \to x+1}}{\sum_{j \in T} \prod_{x \in S_j} \pi_{x \to x+1}},$$

where $\pi_{x \to x+1}$ is the fixation probability for genotype $x+1$ in the x background, S_i is the set of steps that compose trajectory i, and T is the set of all forward trajectories. The model assumes the mutation rate is the same for all sites, and that population size and mutation rates are fixed over the evolutionary trajectory [81, 82, 18, 25, 83]. We calculated $\pi_{x \to x+1}$ for each step using the Gillespie model [84]

$$\pi_{x \to x+1} = \frac{1 - e^{-s_{x \to x+1}}}{1 - e^{-Ns_{x \to x+1}}} = \frac{1 - e^{-(1 - w_{x+1}/w_x)}}{1 - e^{-N(1 - w_{x+1}/w_x)}},$$

where N is population size, s is the selection coefficient and w_x and w_{x+1} are the relative fitnesses of the x and $x+1$ genotypes visited over the trajectory.

To determine the difference between sets of trajectories in maps with and without high-order epistasis, we measured the magnitude of the difference in probability for all $L!$ forward trajectories through each space. We did so by:

$$\theta = \frac{\sum_{i=1}^{i=L!} \left| p_i^{experimental} - p_i^{trunc} \right|}{2},$$

where $p_i^{experimental}$ is the probability of the i^{th} trajectory within the experimental map and p_i^{trunc} is the probability of that same trajectory in a truncated map, with high-order epistasis removed.

Software

We implemented the epistasis and trajectory models using Python 3 extended with the *numpy* and *scipy* packages [65]. We used the python package *scikit-learn* to perform linear regression with truncated forms of these models [66]. Plots were generated using *matplotlib* and *jupyter* notebooks [67, 68]. Our full software package is available in the *epistasis* package via github (https://harmslab.github.com/epistasis).

Results

High-order epistasis is common in all maps

Our first goal was to determine the contributions of each order of epistasis to fitness in six experimentally measured genotype-fitness maps (Table 2). Each map consisted of all possible combinations of 5 mutations ($2^5 = 32$ genotypes) in a haploid genome. The mutations in datasets I and IV arose during adaptive, experimental evolution of *E. coli*, and occur throughout the genome [18, 85]. The mutations in datasets II and VI each occur in single genes that confer drug resistance in *E. coli* and HIV, respectively [25, 27]. The mutations in datasets III and V were introduced randomly into the *A. niger* genome [6]. Previous workers characterized components of fitness for each genotype under defined experimental conditions. For four of the datasets (I, III, IV, and V), the authors measured relative fitness using competition assays. In dataset II, the authors measured minimum inhibitory concentration in the presence of an antibiotic, and from this estimated relative fitness [25]. In dataset VI, the authors measured HIV infectivity in an ex vivo assay, then treated this activity as a proxy for fitness [27]. All datasets exhibited a single fitness peak.

ID	genotype	organism	reference
I	scattered genomic mutations	*E. coli*	[18]
II	β-lactamase enzyme point mutations	*E. coli*	[25]
III	chromosomes combinations	*A. niger*	[6]
IV	scattered genomic mutations	*E. coli*	[85]
V	chromosomes combinations	*A. niger*	[6]
VI	envelope glycoprotein point mutations	HIV-1	[27]

Table 2: Experimental genotype-fitness maps

We previously analyzed four of these datasets, finding small magnitude, but statistically-significant, high-order epistasis in each map [22]. We used this same approach to characterize epistasis in the remaining two maps (Figure S1, Materials & Methods). We sought to account for confounding effects that could lead to spurious epistasis, which would, in turn, lead to spurious effects on evolutionary trajectories. The most important confounding effect is the scale of the map. Models of high-order epistasis sum the effects of mutations and then account for deviation from this expectation by epistasis [58, 52]. But there is no *a priori* reason to assume mutational effects should add: they may multiply or combine on some other nonlinear scale [15, 9, 52, 22]. To account for this, we empirically determined a nonlinear scale for each map using a power-transform, and then used this to linearize each map [22].

We then decomposed the linearized maps into epistatic coefficients using Walsh polynomials [58, 44, 52]. This approach uses the geometric center of the genotype-fitness map as reference state and reveals global correlations in the effects of mutations across the map. Each order of epistasis accounts for variation that is not explained by the sum of all lower-order contributions. For example, third-order coefficients account for any "leftover" variation in the fitness of triple mutants after the first-order (additive) and second-order (pairwise) effects of those mutations are taken into account.

We determined the contribution of each order of epistasis to the total variation in fitness for each dataset by sequentially setting fifth-, fourth-, third-, and second-order epistatic coefficients to zero. We recalculated the fitness of each genotype using each "truncated" model. This is directly analogous to decomposing a sound wave into a sum of frequencies using a Fourier transform [52]. After decomposition, the original sound wave can be approximated by a sum of principal frequencies, followed by a reverse Fourier transform. By selectively including frequencies, one can identify those that contribute most to the final sound wave. Our analysis follows the same logic, approximating fitness (the sound wave) using a collection of epistatic coefficients (sound frequencies).

We quantified the contribution of each epistatic order by measuring the change in fitness when the i^{th} order of epistasis was included in the model. As a metric, we used $\phi = \rho_i^2 - \rho_{i-1}^2$, where ρ_x^2 is the squared Pearson's coefficient between the measured fitness of each genotype and its fitness calculated for a model truncated to the x^{th} order. ϕ ranges from -1 to 1. Figure 9A shows this calculation for dataset I. As epistatic orders are added, ρ^2 between the truncated model and measured fitness values improves. This allows determination of ϕ for each order: first-order coefficients (additive effects) account for 94.0% of variation in fitness; second-order (pairwise epistasis) for 3.8%; third for 1.2%, fourth for 0.9%, and fifth for 0.1%.

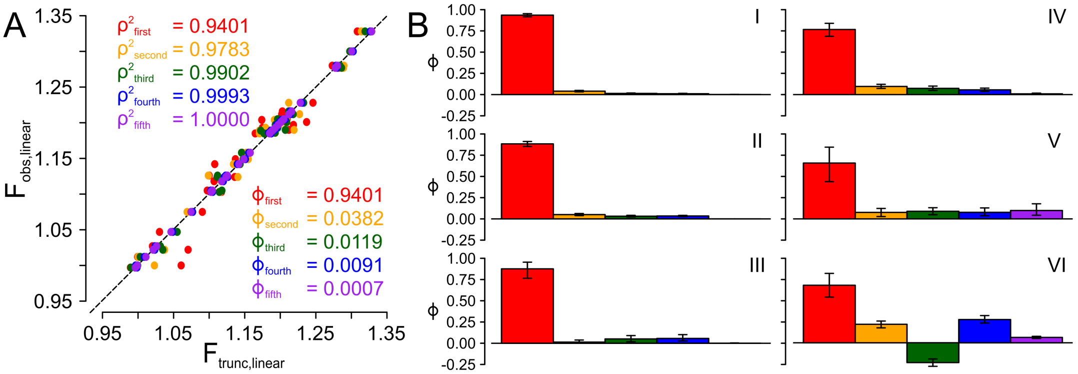


Figure 9: Contributions of epistasis to variation in fitness. Panel A: Correlation between observed (linearized) fitness and fitness calculated for truncated epistasis models for dataset I. Each point on the plot is a single genotype-fitness pair; the dashed line is a 1:1 line. Colors correspond to the truncation order: to first (red), second (orange), third (green), fourth (blue), and fifth (purple). ρ^2 and ϕ for each order are shown on the plot, colored by order. Panel B summarizes the contributions of each order of epistasis to the variation in fitness for all six datasets. Dataset is indicated with roman numeral. Colors follow panel A.

We then applied this analysis to all six datasets. Figure 9B summarizes these results. The total contribution of epistasis to variation in fitness ranged from 6.0% (dataset I) to 32.2% (dataset VI). Other datasets exhibited intermediate levels of epistasis, comparable in magnitude to high-order epistasis observed in similar datasets [22, 44, 79]. In all datasets, the first-order (additive) effects of mutations made the largest contribution to variation in fitness. Outside of this, there was no simple pattern in the relative contributions of the different orders. In dataset I, II and IV, the contribution of epistasis to variation decayed with increasing order. In dataset V, epistasis does not decay. In dataset VI, the addition third-order epistasis (without fourth-order epistasis) actually does a worse job of predicting fitness than second-order alone. The quantitative and qualitative differences in the contribution of epistasis across datasets allow us to study how altering epistasis alters evolutionary trajectories.

Epistasis alters evolutionary trajectories

Our next question was how each order of epistasis altered evolutionary trajectories. We first back-transformed our truncated, linearized maps onto the original scale. This creates a genotype-fitness map without specific epistatic interactions, but on the original, possibly nonlinear, scale of the map. We calculated the relative probabilities of all $L!$ forward trajectories through these maps, starting from the ancestral state and ending at the derived state [25, 6, 18]. Because the maps describe fitnesses of asexual organisms with large population sizes, we modeled trajectories as a series of sequential fixation events captured by a Gillespie model for haploid organisms with large population size (Materials & Methods). [80, 25, 18]. In this scheme, the probability of a trajectory is the product of the probabilities of its individual fixation events, normalized across all trajectories.

We visualized these trajectories by overlaying them on the genotype-fitness map weighted by their relative probabilities. Higher probability mutations have thicker lines connecting them. Figure 10 shows this analysis for dataset I. We started with a purely additive map (top left). All trajectories are accessible with similar probabilities because, in this map, all mutations are individually favorable. We then added successive orders of epistasis and recalculated trajectories through each new map. The addition of second-order epistasis altered the availabilities of trajectories. The changes are most readily evident in the lower row in Figure 10, which shows the change in the probability of each edge and node in the map. The left side of the map is red (indicating loss of probability), while the right side of the map is blue (indicating gain of probability). Addition of each new order, moving left to right across Figure 10, alters the probability of trajectories through the map.

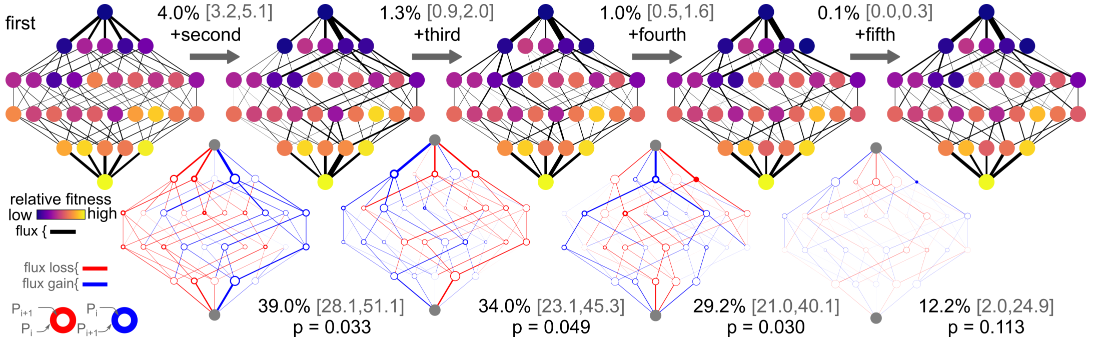


Figure 10: Epistasis alters evolutionary trajectories through genotype-fitness maps. Figure show the effects of increasing orders of epistasis on evolutionary trajectories for dataset I. **Top Row:** Genotype-fitness maps with increasing amounts of epistasis included, increasing from none (far left) to fifth-order (far right). Networks show all 2^5 genotypes, arranged from ancestral (top) to derived (bottom), colored by relative fitness from low (purple) to high (yellow). Edges show the probability of a given mutation in a given background from low (thin) to high (heavy). Mutations with no probability have no edge. The numbers above the arrows are ϕ, with 95% confidence intervals in brackets. **Bottom row:** Change in trajectory probability as each order of epistasis is added. Edges reveal loss of probability (red) or gain of probability (blue). The weight of the edge is directly proportional to the change in probability. Mutations whose probability do not change have no edge. For each node, the thickness of the ring reveals the change in probability that this genotype is visited. For genotypes whose probability goes down, the red area indicates the loss in probability. For genotypes whose probability goes up, the blue area indicates the increase in probability. The numbers below each network are θ, with 95% confidence intervals in brackets. The p-value measures whether the observed value of θ would be expected from fitting experimental noise (Fig 3).

To quantify differences in the sets of trajectories with increasing epistasis, we calculated the change in the probabilities of all 120 forward trajectories through maps with different amounts of epistasis included (θ). A θ of 0.0 indicates that the set of trajectories through the spaces are identical, while a θ of 1.0 means the sets of trajectories do not overlap at all (Materials & Methods). Intermediate values indicate that some fraction of the trajectory probability density is shared between the maps. In dataset I, trajectories through the additive and second-order epistatic maps have $\theta = 0.390$. Put another way, the addition of pairwise epistasis to the additive map shifts 39.0% of the trajectory probability density. Addition of each new order of epistasis has a smaller effect on trajectory probability: $\theta_{2\rightarrow 3} = 0.340$, $\theta_{3\rightarrow 4} = 0.292$, and $\theta_{4\rightarrow5} = 0.122$.

To determine confidence intervals on our estimates of ϕ and θ, we sampled from the fitness measurement uncertainty for each genotype, generating a collection of pseudoreplicate genotype-fitness maps (Figure S2A). We then decomposed each pseudoreplicate map into epistatic coefficients (including refitting the scale) and remeasured ϕ and θ for each epistatic order. Figure 11A shows this calculation for dataset I. From these distributions, we can determine 95% confidence intervals for ϕ and θ (shown as gray, bracketed values in Figure 10).

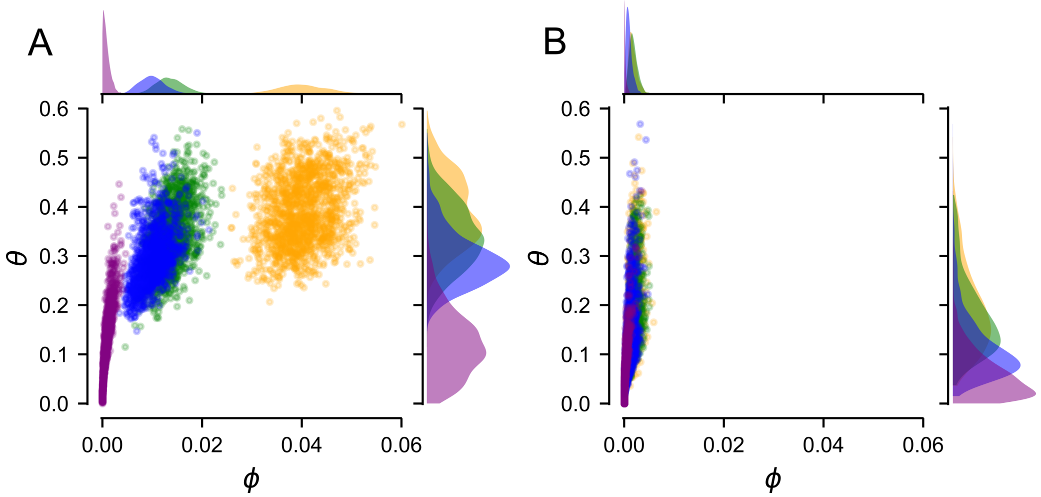


Figure 11: Changes in trajectories are not the result of experimental uncertainty. Data in panel A and B are for dataset I. Panel A shows the distribution of ϕ and θ for 10,000 pseudoreplicates generated by sampling uncertainty in each the fitness of each genotype (Fig S2A). Colors denote order of epistasis, as in panel A. Panel B shows the epistasis extracted from datasets without epistasis, but experimental uncertainty (Fig S2B).

We next asked whether the observed epistasis and its effect on trajectories could be the result of uncertainty in the fitness values. An epistasis model accounts for random noise as leftover variation, and thus as apparent epistasis [22]. We thus posed the following question: if the epistasis at a given order resulted only from noise, what effect would it have on ϕ and θ? To ask this question, we constructed "null" maps with truncated epistasis, but noisy fitness values (Figure S2B). We took our truncated maps at each order and then assigned each fitness the same variance that was measured for the original, un-truncated fitness values. We sampled from this uncertainty to generate pseudoreplicates, extracted apparent epistasis—in this case, arising from noise—and then calculated ϕ and θ for the pseudoreplicate. This allows us to construct distributions of ϕ and θ for epistasis arising purely from experimental noise.

We show this calculation for dataset I in Figure 10B. Unlike the experimental distributions, which spread out in ϕ, the distributions arising from random noise cluster at low values of ϕ. The ϕ/θ distributions of second-, third-, and fourth-order epistasis minimally overlap in Figure 10A versus Figure 10B. This indicates that the signal for epistasis in the datasets is greater than expected from noise in the measured fitness values. In contrast, the ϕ/θ distribution for fifth-order epistasis overlaps between Figure 10A and 10B: the effect of fifth-order epistasis cannot be distinguished from noise. Because we are interested in the effect of epistasis on trajectories (θ), we determined a p-value for each θ. We took the mode of θ at each order from Figure 10A, and determined its percentile on the corresponding null distribution in Figure 10B. For second-, third-, and fourth-order epistasis, this yields a p-value < 0.05. In contrast, the p-value for fifth order was 0.12.

With these quantification tools in hand, we next studied the relationship between epistasis and evolutionary trajectories for the increasing levels of epistasis exhibited by the remaining five datasets. Figure S3-S8 summarize our analyses for all six datasets.

It is helpful to compare dataset I (Figure 10) and dataset V (Figure 12). While epistasis accounts for 6.0% of variation in fitness for dataset I, it accounts for 32% of the variation in fitness for dataset V. The large amount of epistasis in dataset V means that epistasis at all orders has a massive effect on evolutionary trajectories through this space. The addition of fourth-order epistasis is particularly striking. With only third-order epistasis and down, there are multiple paths through the space. With the addition of fourth-order epistasis, all paths but two become inaccessible. The addition of fifth-order epistasis opens the space up again, but to a different set of trajectories than what existed in the third-order space.

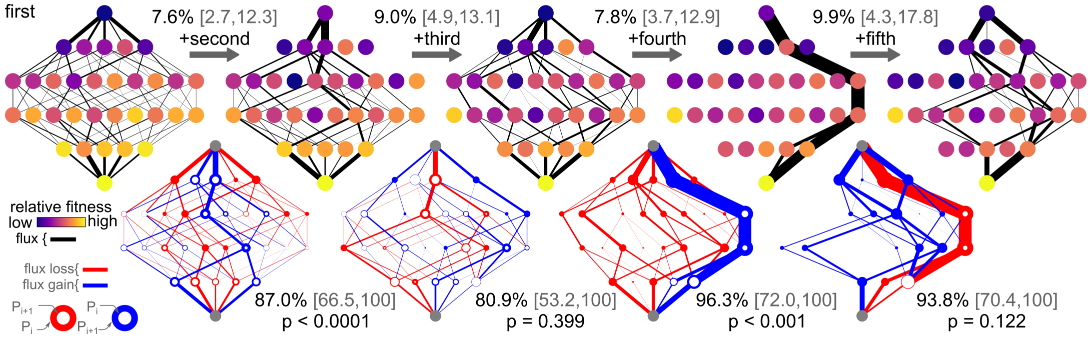


Figure 12: Epistasis alters trajectories in dataset V. Altered trajectories in dataset V with increasing epistasis. Colors, panel layouts, and statistics are as in Fig 2.

We next asked whether magnitude of epistasis or the order of epistasis was a stronger predictor of its effect on evolutionary trajectories. We plotted ϕ versus θ for each order for each dataset on a single plot (Figure 13A). This reveals a correlation between the magnitude of the epistasis and its effect on trajectories. In contrast, we see no correlation between the order of epistasis and its effect on evolutionary trajectories (Figure 13B). When epistasis contributes more than ≈5% of the variation in fitness, regardless of order, the divergence in trajectory probabilities with and without the epistasis is 40% or greater. The magnitude of epistasis—not its order—predicts its effect.

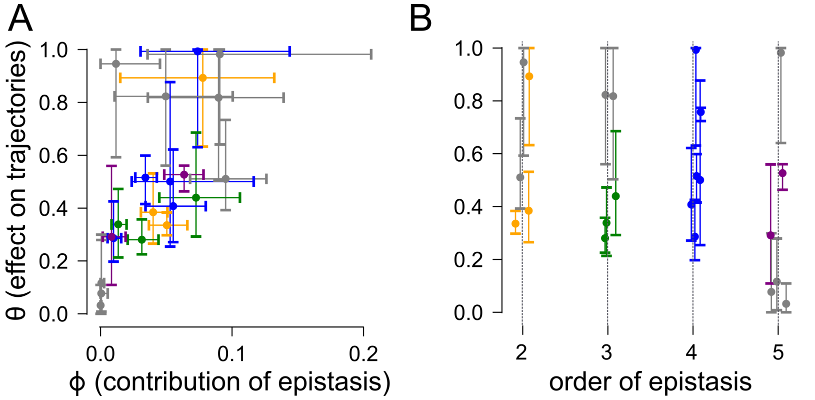


Figure 13: The magnitude of epistasis, not its order, predicts its effects on trajectories. Plot shows θ graphed against ϕ for all datasets. Points are colored by the order of epistasis: second (orange), third (green), fourth (blue), and fifth (purple). Error bars are 95% confidence intervals. Gray points are orders that could not be distinguished from experimental uncertainty ($p < 0.05$).

High-order epistasis limits evolutionary predictability

Our next question was more practical: how important is epistasis for predicting evolutionary trajectories in these datasets? We imagined an experiment in which we measured the effects of all mutations in the ancestral genotype. We then asked if we could take these individually measured mutational effects and predict evolutionary trajectories.

To ask this question, we re-analyzed the epistasis present in all six datasets, this time using the ancestral genotype as the reference state. In this formulation, the first-order coefficients are the effect of each mutation by itself in the ancestral background, the second-order coefficients are the difference in the effects of mutations introduced in pairs versus separately, and third-order coefficients are the difference in the fitness of genotypes combining three mutations versus two mutations that cannot be explained by the first- and second-order coefficients. (This has been called the “biochemical” or “local” model of high-order epistasis [52].) We describe this further in the Materials & Method section.

To characterize the effect of epistasis on our ability to predict evolutionary trajectories of increasing length, we calculated the probability of all possible forward trajectories of a defined number of steps starting from the ancestral genotype, and then repeated this probability calculation using maps truncated to various orders of epistasis. The difference in the actual and truncated map trajectory probability distributions measures our predictive power for evolutionary trajectories. We show these results in Figure 14 for all six datasets. In each panel, we plot inclusion of increasing orders of epistasis left-to-right (starting from additive and going to fifth-order) and increasing trajectory length bottom-to-top (starting from one-step and going to five-step). The overlap between the trajectory distribution for the truncated and real map for each epistasis/trajectory-length is shown as a color ranging from white (perfect prediction) to red (poor prediction).

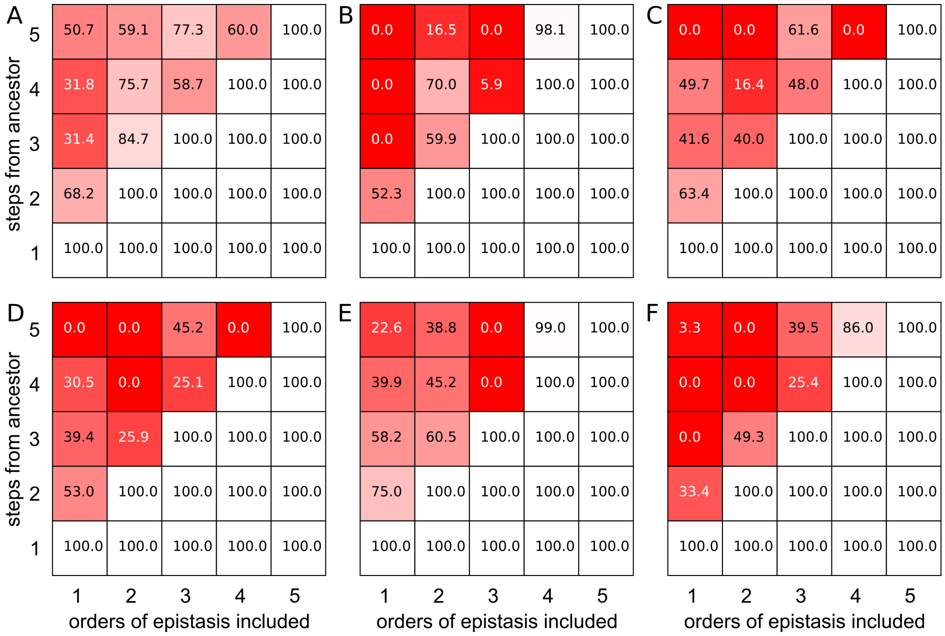


Figure 14: Epistasis complicates predicting trajectories from the ancestral genotype. Panels show trajectory prediction accuracy (color) for different amounts of epistasis included in the model (x-axis) and for different length trajectories away from the ancestral genotype (y-axis). Accuracy is measured as the difference in the probability distributions for trajectories through the truncated and original maps, ranging from 0.0% (red, poor accuracy) to 100% (white, perfect accuracy). Panels A-F correspond to datasets I-VI.

We found that all orders of epistasis were important for predicting evolutionary trajectories. Dataset IV (panel D) illustrates behavior seen across all datasets, so we will use it as a specific example. In this dataset, additive coefficients are inadequate to capture even two-step trajectories: the trajectory probability distribution for two-step mutations only overlaps by 53.0% for the truncated and real maps. The prediction gets worse for longer trajectories, dropping to 39.4% for three steps, 30.5% for four steps, and 0.0% for five steps. The overlap for the final step is 0.0% because the additive model does not predict that the five-mutation genotype will be more fit than the four-mutation genotype. Trajectories in the additive map therefore do not proceed to this final genotype.

Adding pairwise epistasis to the model allows perfect "prediction" of the two-step trajectories, as we have perfect knowledge of the fitness values of all possible single and double mutants. But the three-step and four-step trajectories are predicted worse with pairwise epistasis included than with the additive map. The three-step overlap is 25.9%, while the four-step and five-step trajectory overlap is 0.0. The four-mutation and five-mutation genotypes are predicted to have low fitness. Adding third-order epistasis—now imagining that we characterized all possible single, double, and triple mutants in the ancestral genotype—allows us to "predict" trajectories up to three steps long; however, it fails for four- and five-step trajectories. The overlap is 25.1% and 45.2% respectively. Even the addition of fourth-order epistasis is insufficient to capture the five-step trajectories: the overlap for five-step trajectories is 0.0%.

Dataset IV is a particularly clean example, but all six datasets exhibit similar behavior (Figure 14). Neglecting epistasis leads to poor predictions of trajectories starting from the ancestral genotype. The lower-order the truncation, the worse the prediction as more mutations accumulate. Third- and fourth-order epistasis had an appreciable effect on all datasets. Fifth-order epistasis had an effect in four of the six datasets. Like the analysis using the global model above, high-order epistasis relative to the ancestral genotype potently alters evolutionary trajectories.

Discussion

Our analysis reveals that high-order epistasis can strongly shape evolutionary trajectories. Removal of three-, four-, and five-way interactions between mutations significantly alters the probabilities of trajectories through genotype-fitness maps (Figure 10, Figure 12). This result is robust to uncertainty in the measured fitness values (Figure 11) and appears to be a general pattern in many maps (Figure 13). Finally, neglecting high-order epistasis leads to poor predictions of evolutionary trajectories through these maps (Figure 14).

In the majority of datasets, low-order models provide useful estimates of fitness. For datasets I-IV, ignoring three-way and higher-interactions yields fitness values within 15% of the actual map (Figure 9B). Dataset I would be particularly close, yielding fitness values within 2.5% of the actual map. This is consistent with other analyses of high-order epistasis in other datasets, which suggest that additive and pairwise epistatic effects can often provide sufficient information to predict multi-mutation fitness values to within 5-10% [40, 44, 42, 52, 32, 41, 22, 37].

While low-order models can often describe fitness with some degree of precision, low-order models are inadequate to describe evolutionary trajectories in any of the datasets. Even in dataset I, third- and fourth-order interactions potently shape evolutionary trajectories. The probability distributions of trajectories with and without fourth-order epistasis differ by 29.2%. And, as the magnitude of epistasis increases, its effect on trajectories grows (Figure 13A). In some instances, addition of high-order interactions completely shifts the set of trajectories available (Figure 12).

The effect of high-order epistasis on evolutionary trajectories is profound. We can build this intuition by imagining predicting evolutionary trajectories. If we start with knowledge of the individual effects of mutations in the ancestral background we can predict the first move perfectly, but not the second move. Pairwise epistasis means the effect of the second mutation is modulated by the presence of the first. We might try to overcome this difficulty by measuring the effect of each mutation and each pair of mutations, thereby accounting for pairwise epistasis. But our results reveal this is still insufficient to predict trajectories past the second step. There are three-way interactions that alter the effect of the third mutation, even after accounting for the first- and second-order effects of mutations. This continues all the way to fifth-order in these five-site datasets.

This has two implications. First, this adds to the growing recognition of extensive contingency in evolution [86, 36, 19, 33]. The effect of an event today is contingent on a whole collection of previous events. Remarkably, we found that this contingency is mediated by epistasis at all orders, including up to five-way interactions between mutations. Second, this work implies that measuring the individual effects of many mutations in a single genetic background, despite revealing a local fitness landscape [34, 37, 20], will be of limited utility for understanding evolution past the first few moves.

We expect the effect of high-order epistasis on trajectories will be amplified in larger maps that have more mutations. In a larger map, more mutations compete for fixation—each modulated by high-order interactions with previous substitutions—leading to even greater contingency on specific substitutions that occurred in the past. Further, the small maps we studied artificially limit the effects of high-order epistasis, as larger maps could, potentially, have even higher-order interactions. But even if no epistasis above fifth-order is present, trajectories will have more steps in a larger map; therefore, a fifth-order interaction could alter the relative probabilities of many more future moves in a larger space.

One open question is the effect of recombination on this radical contingency. We studied trajectories in which mutations fixed sequentially. This means our results are directly applicable to asexual organisms and loci in tight linkage, such as mutations to individual genes. Once recombination comes into play, other dynamics become possible. While recombination can completely overcome pairwise epistasis [87], it is unclear whether this result will apply to higher-order interactions.

High-order epistasis appears to be a ubiquitous feature of experimental genotype-fitness (and genotype-phenotype) maps [40, 44, 42, 52, 32, 41, 22, 37]. The origins of this epistasis remain unknown. Further, epistasis may go to much higher-order than yet observed, leading to extremely long-term memory in evolution. The observation of cryptic epistasis between genetic backgrounds that appear similar, but in which mutations have radically different effects, may point to high-order epistasis between mutations in diverging backgrounds[88, 34]. Whatever the origins or order may be, our work reveals that combinations of early substitutions continue to have an effect as future mutations accumulate: the past continues to press upon the present.

CHAPTER IV

MOLECULAR ENSEMBLES MAKE EVOLUTION UNPREDICTABLE

Author Contributions

Zachary Sailer (ZRS) and Michael Harms (MJH) conceptualized the paper. ZRS conducted the simulations and computational analysis. ZRS generated figures. ZRS and MJH wrote and edited the manuscript.

Abstract

Evolutionary prediction is of deep practical and philosophical importance. Here we show, using a simple computational protein model, that protein evolution remains unpredictable even if one knows the effects of all mutations in an ancestral protein background. We performed a virtual deep mutational scan—revealing the individual and pairwise epistatic effects of every mutation to our model protein—and then used this information to predict evolutionary trajectories. Our predictions were poor. This is a consequence of statistical thermodynamics. Proteins exist as ensembles of similar conformations. The effect of a mutation depends on the relative probabilities of conformations in the ensemble, which in turn depend on the exact amino acid sequence of the protein. Accumulating substitutions alter the relative probabilities of conformations, thereby changing the effects of future mutations. This manifests itself as subtle, but pervasive, high-order epistasis. Uncertainty in the effect of each mutation accumulates and undermines prediction. Because conformational ensembles are an inevitable feature of proteins, this is likely universal.

Significance

A long-standing goal in evolutionary biology is predicting evolution. Here we show that the architecture of macromolecules fundamentally limits evolutionary predictability. Under physiological conditions, macromolecules like proteins flip between multiple structures, forming an ensemble of structures. A mutation affects all of these structures in slightly different ways, redistributing the relative probabilities of structures in the ensemble. As a result, mutations that follow the first mutation have a different effect than they would if introduced before. This implies that knowing the effects of every mutation in an ancestor would be insufficient to predict evolutionary trajectories past the first few steps, leading to profound unpredictability in evolution. We therefore conclude that detailed evolutionary predictions are not possible given the chemistry of macromolecules.

Introduction

Is evolution predictable? This is a fundamental question in evolutionary biology, both for philosophical [89, 90, 91, 36] and practical reasons [34, 92]. Deep mutational scanning experiments provide an intriguing new avenue to think about evolutionary prediction. These experiments reveal the effects of huge numbers of mutations and thus provide rich information about local adaptive landscapes [1, 93, 20]. This leads to a simple question: if we know the effect of every mutation in an ancestral genotype, can we predict future evolutionary trajectories? If not, what limits our ability to make predictions?

To pose this question, we attempted to predict the evolution of simple physical protein model given a virtual deep mutational scan. In this context, prediction is knowing which mutations would accumulate, in what order, given knowledge of the effects of the mutations in the ancestral background.We attempted an “easy” prediction, reasoning that it could act as a starting point for more difficult scenarios involving more complex evolutionary processes. To maximize predictive success, we studied adaptive trajectories in which the environment was stable, there was consistent directional selection, and mutations were fixed by selection rather than drift.

We studied the evolution of improved thermodynamic stability—a shared feature of all folded proteins and a target of natural selection in many contexts [94, 95, 96]. This is a useful phenotype for a number of reasons. First, because stability is a thermodynamic quantity, studying it may reveal features common to the evolution of other thermodynamic properties such as allostery and ligand binding. Second, biological systems—from molecules to ecosystems—are ultimately physical; therefore, insights at the physical level may provide insights for higher levels of biological organization [39, 97].

Surprisingly, we found that our predictions were quite poor. We even added all pairwise epistatic effects of mutations to our predictive model, requiring a massive virtual deep mutational scan of all possible pairs of mutations. Even this did not allow robust predictions of evolutionary trajectories. We find that the unpredictability arises directly from the thermodynamic ensemble of conformations populated by macromolecules—revealing a profound link between protein physics and the evolutionary process.

Materials and Methods

All of our analyses are are contained in Python scripts and Jupyter notebooks available on Github (https://github.com/harmslab/notebooks-epistasis-ensembles). Full details are given in the SI Appendix.

For the protein lattice model simulations, we extended the *latticeproteins* package originally written by Prof. Jesse Bloom (https://github.com/harmslab/latticeproteins) [98] using Miyazawa and Jernigan contact energies (from Table V in their paper) and reduced temperature units [99]. We randomly generated 12-site protein sequences and then evolved each sequence until the fraction folded was ≈ 0.7. We calculated the probability of a given evolutionary trajectory as a series of independent, sequential fixation events using a strong-selection, weak-mutation model [80]. In this model, the fixation probability for going from genotype x to $x+1$ is:

$$\pi_{x\rightarrow x+1} = 1 - e^{-(w_{x+1}/w_x - 1)},$$

where w_x and w_{x+1} are the relative fitnesses of the x and $x+1$ genotypes (SI Appendix).

To quantify the difference between the predicted and actual trajectories, we calculated the magnitude of the difference in probability of all observed trajectories through each space [35]

We predicted the ΔG°_{N} for each genotype and then used this to predict evolutionary trajectories. For the additive model, the predicted stability of a genotype with a set of $\{M\}$ mutations was:

$$\Delta\hat{G}^{\circ}_{N,\{M\}} = \Delta G^{\circ}_{N,anc} + \sum_{i\in\{M\}} \Delta\Delta G^{\circ}_{N,i}$$

where $\Delta G^{\circ}_{N,anc}$ is the stability of the ancestor and $\Delta\Delta G^{\circ}_{N,i}$ is the effect of the i^{th} mutation in the ancestral background. For the pairwise epistatic model, we added epistatic coefficients:

$$\Delta\hat{G}^{\circ}_{N,\{M\}} = \Delta G^{\circ}_{N,anc} + \sum_{i\in\{M\}} \Delta\Delta G^{\circ}_{N,i} + \sum_{i<j\in\{M\}} \Delta\Delta\Delta G^{\circ}_{N,ij}$$

where $\Delta\Delta\Delta G^{\circ}_{N,ij}$ accounts for any pairwise epistasis between the mutations i and j. We quantified this epistasis by introducing mutations i, j, and then i and j together. We then took the difference:

$$\Delta\Delta\Delta G^{\circ}_{N,ij} = \Delta G^{\circ}_{N,ij} - \left(\Delta G^{\circ}_{N,anc} + \Delta\Delta G^{\circ}_{N,i} + \Delta\Delta G^{\circ}_{N,j}\right)$$

where $\Delta\Delta G^{\circ}_{N,i}$ and $\Delta\Delta G^{\circ}_{N,j}$ are the individual effects of mutations i and j in the ancestral background and and $\Delta G^{\circ}_{N,ij}$ is the stability of the ij double mutant.

To extract high-order epistasis, we used the *epistasis* package (https://harmslab.github.com/epistasis [22]. A genotype with L mutations is described by 2^L hierarchical epistatic coefficients [58, 44, 52, 22]. We generated all 2^6 binary combinations of the substitutions that accumulated between the ancestor and most probable final sequence, calculating ΔG°_{N} for all 64 mutants. Because we were doing predictions starting from the ancestral state, we used the so-called "biochemical model" that uses the ancestral genotype as the reference state **?**]. See SI Appendix for further details.

Software

All of our analyses are are contained in Python scripts and Jupyter notebooks available on Github (https://github.com/harmslab/notebooks-epistasis-ensembles). For the protein lattice model simulations, we extended the *latticeproteins* package originally written by Prof. Jesse Bloom (https://github.com/harmslab/latticeproteins) [98]. All protein lattice simulations use the effective interresidue contact energies estimated by Miyazawa and Jernigan and reduced temperature units [99]. We used the *epistasis* Python package for our analysis of high-order epistasis in binary lattice protein maps. (https://github.com/harmslab/epistasis) [22].

Generating ancestral sequences

We randomly generated 12-site protein sequences and then evolved each sequence so it was adjacent to a fitness peak, but not at the peak, allowing for future evolution. Starting from the random sequence, we calculated the effect of all possible mutations and then chose one randomly, weighted by its effect on stability. We continued this until the fraction folded was ≈ 0.7.

Evolutionary trajectories

We calculated the probability of a given evolutionary trajectory as a series of independent, sequential fixation events. We described the fitness of each genotype as proportional to the fraction of the molecules in the native conformation. This is given by:

$$w = \frac{1}{1 + e^{\Delta G^{\circ}_{N}}}.$$

We assumed that the time to fixation for each mutation was much less than the time between mutations (the so-called strong selection/weak mutation regime) [80]. The relative probability of an evolutionary trajectory i is the product of its required fixation events relative to all possible trajectories:

$$p_i = \frac{\prod_{x \in \{S_i\}} \pi_{x \to x+1}}{\sum_{j \in \{T\}} \prod_{x \in \{S_j\}} \pi_{x \to x+1}},$$

where $\pi_{x \to x+1}$ is the fixation probability for genotype $x + 1$ in the x background, $\{S_i\}$ is the set of steps that compose trajectory i, and $\{T\}$ is the set of all trajectories. The model assumes the mutation rate is the same for all sites, that the population size is large, and that mutation rates are fixed over the evolutionary trajectory [80]. We calculated $\pi_{x \to x+1}$ for each step using the Gillespie model for infinite populations [80]:

$$\pi_{x \to x+1} = 1 - e^{-(w_{x+1}/w_x - 1)},$$

where w_x and w_{x+1} are the relative fitnesses of the x and $x + 1$ genotypes visited over the trajectory.

To quantify the difference between between the predicted and actual trajectories, we calculated the magnitude of the difference in probability of all observed trajectories through each space [22]. We did so by:

$$\theta = \sum_{i \in \{T\}} \frac{|p_{pred,i} - p_{actual,i}|}{2},$$

where $\{T\}$ is the set of all trajectories observed in both maps, $p_{pred,i}$ is the probability of the i^{th} trajectory in the predicted map, and $p_{actual,i}$ is the probability of that same trajectory in the actual map. This ranges from 0 (perfect overlap) to 1 (no overlap).

Epistasis analysis

A genotype with L mutations is described by 2^L hierarchical epistatic coefficients [58, 44, 52, 22]. We generated all 2^6 binary combinations of the substitutions that accumulated between the ancestor and most probable final sequence, calculating ΔG_N° for all 64 mutants. Because we were doing predictions starting from the ancestral state, we used the so-called "biochemical model" that uses the ancestral genotype as the reference state [52]. First-order coefficients describe the effects of individual mutations ($\Delta\Delta G_N^\circ$). Second-order coefficients capture the difference between the individual effects of mutations and their effect together ($\Delta\Delta\Delta G_N^\circ$). Third-order coefficients capture leftover variation in phenotype after first- and second-order effects are accounted for ($\Delta\Delta\Delta\Delta G_N^\circ$). This continues to the L^{th} order. Because we are using ΔG_N° as our phenotype, it is already on a defined linear scale and can be analyzed without empirically identifying the scale [52, 22].

Results

We set out to predict trajectories that increased the stability of a lattice protein. Lattice models have been used extensively in studies of protein folding and evolution [100, 101, 3, 98, 97]. A lattice model captures the fact that the weak interactions that define the structure of a protein stochastically break and form under cellular conditions, causing proteins to fluctuate between multiple conformations [102, 103, 104]. These structural ensembles are critical to functions such as allostery [105, 104], enzyme activity [103], complex assembly [106], and regulation [107].

A lattice model describes a protein ensemble as a collection of conformations on a grid. Some conformations will be favored, others disfavored. The favorability of each conformation is quantified by its internal energy (E_c), which depends on the contacts between amino acids in that conformation. Conformations with more favorable contacts are more likely than those with fewer contacts. The overall stability of the protein is described by the free energy of the native conformation (ΔG_N°), which quantifies the population of the native conformation relative to all other conformations in the ensemble (Figure 15A). Using reduced temperature units, this is given by:

$$\Delta G_N^\circ = E_N + ln\left(-e^{-E_N} + \sum_{i=1}^{C} e^{-E_i}\right), \tag{9}$$

where E_N is the internal energy of the native conformation and the sum on the right goes over all C conformations in the ensemble.

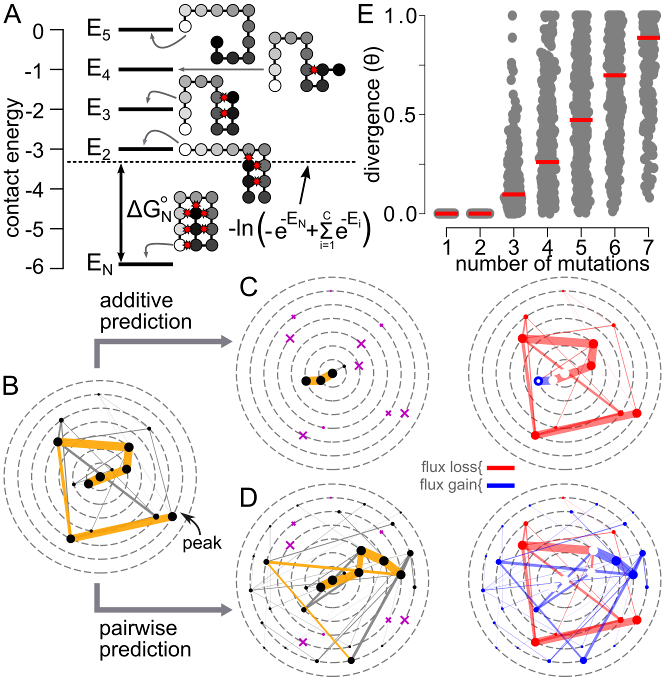


Figure 15: Evolution is unpredictable in protein lattice models. Panel A shows the meaning of ΔG°_{N} using five lattice model conformations of the many thousands possible. Each conformation is a single, non-intersecting chain on a 2D grid. Amino acids are shown as circles colored by position in the sequence from black to white. Peptide bonds are dark bars. Non-covalent interactions are red stars. The strength of each noncovalent interaction depends on the identities of the interacting amino acids. The contact energy of each conformation is shown as a dark line; the Boltzmann-weighted average of the non-native conformations is shown as a dashed line. B) Relative probabilities of evolutionary trajectories starting from an ancestral genotype (center). Circles indicate genotypes; lines indicate mutational steps. The size of each circle and width of each line indicates its probability. Dashed gray lines indicate increasing number of sequence differences from the ancestor. The orange trajectory is the highest probability trajectory. C) Predicted trajectories using an additive predictive model. The left panel is colored as B. The genotypes visited in the actual trajectories but not the predicted trajectories are shown as purple "×" with sizes proportional to their probability. The right panel shows the difference between the predicted and actual trajectories. Red lines indicate trajectories missed in the prediction; blue lines indicate trajectories incorrectly added by the prediction. D) Predicted trajectories using a pairwise epistatic predictive model, with colors as in C. E) Divergence between predicted and actual trajectories for 1,000 starting genotypes as a function of number of mutations. Gray points are individual simulations. Red bars are the means for all genotypes after that number of steps.

We studied evolution in a strong-selection, weak-mutation regime [84]. This assumes the population size is large and that mutations fix sequentially—a reasonable assumption for a single gene for which recombination is rare. This also removes uncertainty due to drift. We defined fitness as proportional to the fraction of the molecules in the native conformation: $w = 1/[1 + exp(\Delta G^{\circ}_{N})]$, as the fraction of molecules folded—not the free energy—is under selection in most biological contexts [95].

We generated a random 12 amino acid protein sequence with $w \approx 0.7$ as our starting point (see Materials & Methods). We calculated w for all genotypes differing by a single mutation relative to the ancestor and then determined the relative fixation probability for all point mutants. We then stepped out to all accessible genotypes and repeated the protocol. Any mutation with a non-zero fixation probability was considered accessible. Iterating this procedure generates a branching set of trajectories that improve the stability of the original native conformation. By comparing the relative fixation probabilities for each mutation along each trajectory, we can calculate the total probability flux through each possible trajectory (see Materials & Methods).

We started by calculating ground-truth evolutionary trajectories against which to compare our predictions (Figure 15B). For the sequence in Figure 15B, we found one main evolutionary trajectory (shown in orange) leading to a fitness peak six mutations away. There were also several lower probability trajectories accessible (shown in gray).

We next set out to predict these trajectories using information extracted from a virtual deep mutational scanning experiment. We calculated the change in ΔG°_{N} for all 228 possible point mutants to the ancestral genotype. Using this information, we could then predict ΔG°_{N} for any genotype as the free energy of the ancestor plus the sum of the effects of all mutations in the genotype. Finally, we could use these predicted ΔG°_{N} values to calculate probable evolutionary trajectories.

These predictions were quite poor (Figure 15C). While the first move is correctly identified, the next move is incorrect. For the sequence shown in Figure 15C, our predicted trajectories are limited to a peak directly adjacent to the ancestral genotype rather than the actual peak six mutations away.

This result is unsurprising: we did not include any epistasis in our predictions. Like real proteins, residues in a lattice model form direct contacts with each other. We would therefore expect to see pairwise epistasis—a difference in ΔG_N° when mutations are introduced together versus separately. We therefore re-ran our predictions of trajectories accounting for both the individual effects of mutations and pairwise epistasis between them. Practically, this involved another, larger virtual deep-mutational scanning experiment: we calculated ΔG_N° for all 228 possible single mutants and all $22,836$ possible double mutants. By comparing the effects of each mutation together and in pairs, we could build a more sophisticated prediction model that accounts for pairwise interactions between mutations (see Materials & Methods).

Addition of pairwise epistasis improved our predictions relative to the additive model, but we still performed quite poorly (Figure 15D). Although the addition of pairwise epistasis allows the trajectories to escape the local region of the ancestral genotype, the predicted and actual trajectories diverge after the second step. Many genotypes are visited that were not seen in the actual trajectories, while many genotypes in the actual trajectories were missed (purple crosses). This includes the actual fitness peak.

To verify that this was a robust feature of lattice proteins, we then repeated our pairwise epistasis predictions for 1,000 different random starting sequences. To characterize the quality of our predictions, we calculated the difference in the probabilities of matched trajectories between our predicted and actual maps (θ). This metric ranges from 0.0 (no difference) to 1.0 (complete difference) (see Materials & Methods). We then calculated θ as a function of number of steps from the ancestral genotype for all 1,000 spaces (Figure 15E). In all spaces, we correctly predict the first two moves. (This is because we built the prediction model using the fitnesses of these genotypes—hardly a difficult prediction.) Addition of the third mutation, however, causes immediate divergence between the predicted and actual trajectories. This divergence continues until, by the sixth step, the predicted and actual maps have an average divergence of 0.9.

Ensembles induce epistasis

Our predictions accounted for all pairwise epistasis, yet still failed. It follows that there must be three-way (or greater) epistatic interactions between mutations. This is surprising, as lattice models are built from pairwise contacts alone. What is the source of this "high-order" epistasis? We know that these interactions must be indirect at a structural level, as the only direct interactions are pairwise. We therefore searched for mechanisms that would lead to indirect interactions between mutations.

Allostery, where binding at one site indirectly affects activity at a distant site, is a useful analog of this problem. One way that allostery can arise is through a conformational ensemble [105, 104]. Binding at one site perturbs the relative populations of different structures in the conformational ensemble. This can, indirectly, change activity at another site. We hypothesized that a similar phenomenon was leading to evolutionary unpredictability.

Figure 16 shows a highly simplified lattice model that illustrates indirect, "ensemble-induced" epistasis between a pair of mutations. (The logic can be extended to indirect multi-way interactions between mutations, but visualization of the phenomenon is much easier for pairwise epistasis). In this example, we have a six amino acid protein where each site can be either hydrophobic (H) or polar (P). We will consider introducing two mutations that do not contact one another in the the structure: $H2P$ (orange) and $H4P$ (purple). We start with the non-epistatic case (panel A). This ensemble has only two conformations: A and B. ΔG°_{A} is simply the difference in the contact energy between conformation A and conformation B (Figure 16A, Table 3). $H2P$ destablizes conformation A, $H4P$ indirectly stabilizes conformation A by destabilizing conformation B. If we sum the effects of the mutations, we obtain the correct stability of the double mutant.

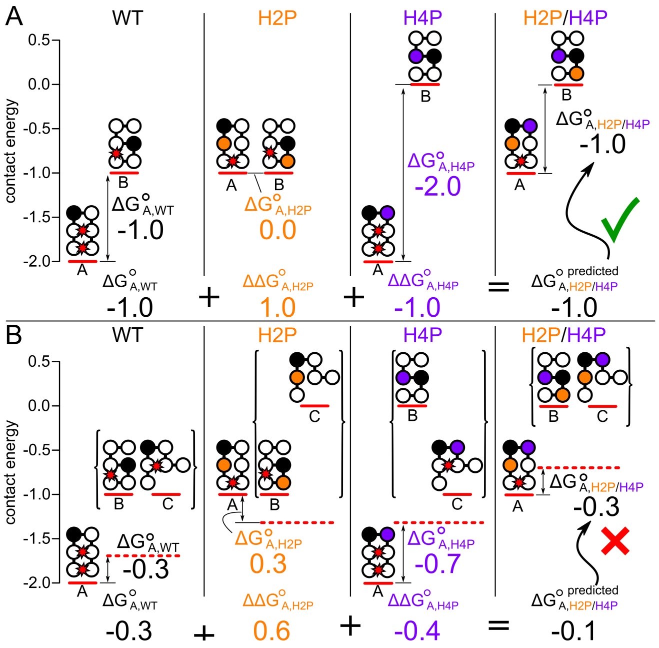


Figure 16: Conformational ensembles induce epistasis. Panels show how epistasis arises from the conformational ensemble of a simple, six-residue lattice protein. In this model, residues can be either hydrophobic (H, white circles) or polar (P, filled circles). Favorable H-H contacts are worth -1 and are denoted by a red star. Colors denote mutations: H2P (orange) and H4P (purple). Solid red lines indicate the contact energies of each conformation. Genotypes are denoted above each sub-panel. The thermodynamic stability of state A is shown for each genotype (for example, $\Delta G^{\circ}_{A,WT}$). The information used to calculate $\Delta G^{\circ predicted}_{A,H2P/H4P}$ is shown along the bottom of panels A and C. **Panel A**: In the two-state system, the effects of H2P and H4P sum in the H2P/H4P mutant; therefore, $\Delta G^{\circ predicted}_{A,H2P/H4P}$ is correct (green check mark). In **panel** B, we see that addition of a third state in the ensemble leads to epistasis. ΔG°_{A} for each genotype is now the difference in the contact energy of conformation A and the Boltzmann-weighted sum of the contact energies of conformations B and C (dashed red line). Because of this nonlinearity, $\Delta G^{\circ}_{A,WT} + \Delta\Delta G^{\circ}_{A,H2P} + \Delta\Delta G^{\circ}_{A,H4P} \neq \Delta G^{\circ}_{A,H2P/H4P}$ (red "×").

What if we add a third conformation (C) to the thermodynamic ensemble? This is shown in panel C. ΔG°_{A} is now the difference between the contact energy of A and the log of the Boltzmann-weighted sum of the contact energies of B and C (Figure 16B, Table 3). The mutations now no longer behave additively. In the wild type background, *H2P* is destabilizing by 0.6 (reduced energy units). In the *H4P* background, *H2P* is destabilizing by only 0.4. This is indirect, ensemble-induced epistasis. In the wildtype background, $H2P$ destabilizes conformation A such that it has the same contact energy as conformation B. These states strongly compete with one another, causing $H2P$ to have a relatively large effect (0.6). $H4P$ alters this effect by destabilizing conformation B. This means that, when $H2P$ is introduced, conformation B does not compete effectively with conformation A. As a result, $H2P$ has a more mild effect (0.4).

ensemble complexity	Phenotype
two-state	$E_N - E_U$
three-state	$E_N + ln\left(e^{-E_U} + e^{-E_{U'}}\right)$
full ensemble	$E_N + ln\left(-e^{-E_N} + \sum_{i=1}^{C} e^{-E_i}\right)$

Table 3: Mathematical formulation of multi-state ensembles in lattice proteins.

The presence of the third state in the ensemble leads to epistasis between these mutations and an incorrect prediction of the double-mutant stability. Predicting the effect of a mutation in a future genetic background therefore requires knowing its effect on every member of the ensemble, not simply its aggregate effect on the entire population of states. This is, in practice, impossible to measure. Ensemble-induced epistasis is directly analogous to ensemble-induced allostery [105, 104]; however, we have now substituted mutations for binding events and binding sites.

Evolutionary trajectories exhibit extensive ensemble-induced epistasis

From our reasoning above, we would predict that ensemble-induced epistasis would arise for any conformational ensemble with more than two states, and that it could lead to epistasis of any order. We therefore set out to quantify the epistasis present in our evolutionary trajectories. Quantitatively, epistasis accounts for variation not accounted for by lower-ordered effects of mutations. Pairwise epistasis is the difference in the effects of two mutations introduced together versus separately. Three-way epistasis is the difference in ΔG°_{N} for a triple mutant versus the predicted ΔG°_{N} from the individual and pairwise epistatic coefficients. In thermodynamic terms, individual effects are $\Delta\Delta G^{\circ}_{N}$, pairwise interactions are $\Delta\Delta\Delta G^{\circ}_{N}$, and three-way interactions are $\Delta\Delta\Delta\Delta G^{\circ}_{N}$. This can be extended to any order of interaction [58, 52, 22].

Measuring an L^{th}-order interaction requires characterizing 2^L combinations of mutations. To access a collection of 2^L genotypes, we constructed binary genotype-phenotype maps containing all possible combinations of the mutations between the ancestral genotypes and their highest probability genotype six mutations away (e.g. between the ancestor and the peak on Figure 15B). We decomposed epistasis in ΔG°_{N} using a linear model capturing the individual effects of mutations and any interactions between them [58, 52, 22] using the ancestral genotype as the reference state [52].

We detected high-order epistasis in every calculated trajectory. Figure 17A shows the average magnitude of epistatic coefficients of increasing order for all spaces. The magnitude decays with increasing order, but is still detectable up to sixth order in all spaces.

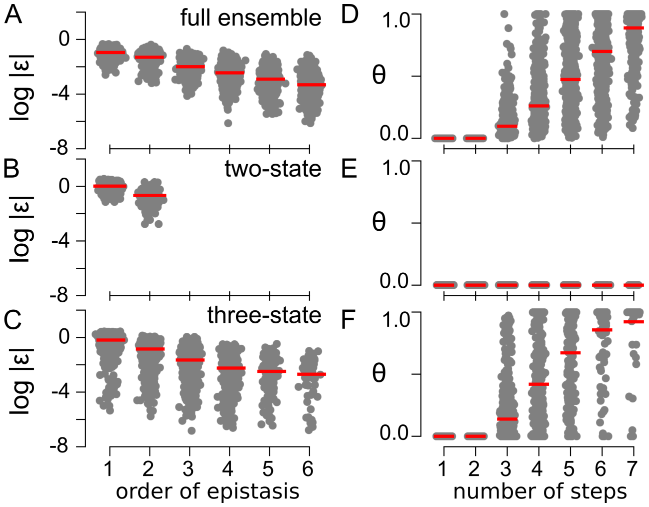


Figure 17: Ensemble-induced epistasis leads to unpredictibility. Panels A-C are jitter plots that show the average magnitude of epistasis $|\varepsilon|$ observed for increasing orders of epistasis in 1,000 maps generated from full ensemble (A), two-state (B), and three-state (C) ensembles. Gray points represent the average magnitude of the epistastic coefficients at a given order for a single map. Red bars indicate the means. Panels D-F show divergence between the "true" trajectories and predicted trajectories for full (D), two-state (E), and three state maps (F). For clarity, panel 3D reproduces the data in panel 1E.

If the ensemble is, indeed, the source of this epistasis, we predicted that it would disappear if we removed the ensemble. We therefore generated truncated lattice models that had only two or three conformations in their ensemble. The two-state ensemble had the native state and the lowest energy non-native conformation (N and U). The three state ensemble had the native state and the two lowest energy non-native conformations (N, U and U').

We predicted that the two-state model would exhibit only pairwise epistasis, while the three-state model would exhibit higher-ordered epistasis. This can be understood from the energy function for each ensemble. ΔG°_{N} for the two-state model is linear with respect to contact energy (Table 3). The only epistasis that arises is via direct interactions encoded in the contact energy. In a three-state (or higher) ensemble, ΔG°_{N} no longer reduces to a linear difference in contact energies (Table 3). This means that mutations have nonlinear effects on the probabilities of conformations within the ensemble, leading ensemble-induced epistasis (Figure 16).

To investigate epistasis in these reduced ensembles, we generated binary maps as before: we used either a two-state or three-state ensemble to calculate ΔG°_{N} for each genotype, calculated evolutionary trajectories, and then built binary maps between the ancestor and most probable final genotype. We then decomposed these maps to extract epistasis. As predicted, the two-state ensembles exhibited only pairwise epistasis (Figure 17B). In contrast, the three-state ensemble exhibited extensive high-order epistasis (Figure 17C)—just like the full ensemble (Figure 17A).

We next asked whether reducing the ensembles altered predictability. If the unpredictability we initially observed arises from epistasis induced by the ensemble, we would predict high predictability for two-state ensemble, but poor predictability for the three-state ensemble. We observed precisely this pattern. A pairwise prediction model was able to perfectly predict evolutionary trajectories for two-state ensemble (Figure 17E), but failed to predict evolutionary trajectories for the three-state ensemble (Figure 17F). The unpredictability observed for the three-state ensemble is directly comparable to the unpredictability observed for the full ensemble (Figure 17D).

Predictions fail even when high-order epistasis is included

If molecular ensembles lead to epistasis which undermines evolutionary prediction, an obvious solution is to characterize even higher orders of epistasis. What if we knew all three-way interactions? Or four-way interactions? Is there some order of epistasis that, when characterized, allows long-range evolutionary predictions?

We cannot ask this question for open-ended evolutionary trajectories as it rapidly becomes intractable. (Even for our twelve-site lattice protein, quantifying all possible three-way interactions would require characterizing 1,533,034 genotypes). Instead, we chose to try to predict trajectories from ancestral to derived genotype through the binary maps above, each containing 2^6 genotypes. We built increasingly complex epistatic models, ranging from first-order (constructed from characterization of six point mutants in the ancestral background) to sixth-order (constructed from characterization of all combinations of point mutants in the ancestral background).

We found that incorporation of high-order epistasis led to little improvement in our predictions. Figure 18 shows the deviation between our predicted and actual trajectories for models incorporating increasingly higher orders of epistasis. Panel A shows predictions using the full ensemble. The additive model begins to deviate from the actual trajectories after the first step; the pairwise after the second; the three-way after the third. As soon as the model has to make a prediction beyond the phenotypes that were used to build the model, trajectories begin to deviate. Even our fifth-order model—which required knowing the phenotypes of 63 of the 64 genotypes in the space—does not always correctly predict the final step (purple curve).

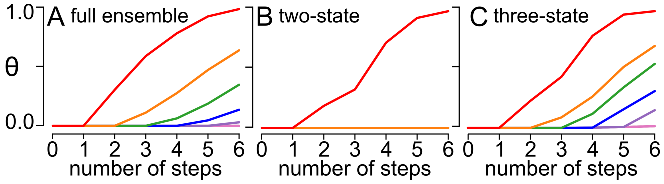


Figure 18: Addition of high-order epistasis does not lead to predictability. Panels show the deviation between predicted and actual trajectories through binary genotype-phenotype maps using predictive models with increasing orders of epistasis: additive (red), pairwise (orange), three-way (green), four-way (blue), five-way (purple) and six-way (pink). Panels correspond maps using a full-ensemble (A), two-state ensemble (B) and three-state ensemble (C) maps. Each curve is averaged over 1,000 maps.

This unpredictability arises from the thermodynamic ensemble. Panels B and C show the same analysis for the two-state and three-state ensembles respectively. For the two-state ensemble, we were able to predict trajectories perfectly with the addition of pairwise epistasis. For the three-state model, we see similar behavior to the full ensemble: inclusion of high-order epistasis does not improve predictions. This is because the epistasis does not capture specific interactions, but instead reveals that the ensemble is changing quantitatively and nonlinearly as mutations accumulate. No matter what order of epistasis is characterized, the future remains obscure.

Discussion

Our work demonstrates that the physical properties of proteins can lead to profound evolutionary unpredictability. Because each mutation alters the relative probabilities of all conformations of a protein, the quantitative effect of a mutation is different in every genetic background. As a result, the effect of a mutation early in a trajectory does not predict its effect later and evolutionary trajectories become unpredictable. Because thermodynamic ensembles are a natural aspect of molecular architecture and ubiquitous for function, we expect this is a universal link between the biochemistry of macromolecules and their evolution.

A key point from our work is that unpredictability can arise even in this extraordinary simple system. The problem of predicting evolution will only become harder as the complexity and realism of the models increases. Using a larger protein, for example, would increase the number of possible options and degeneracy of trajectories, making predictions more challenging. Likewise, constructing a more realistic evolutionary model—incorporating drift, for example—increases the number of available trajectories and makes evolutionary prediction more challenging than the strong selection case (SI Appendix).

Ensemble-induced epistasis is likely common

Our work suggests that any macromolecule that populates three or more conformations can exhibit ensemble-induced epistasis. This is an extraordinarily common set of conditions, as most macromolecular functions require populating multiple states [102, 105, 103, 104]. For example, consider an allosterically inhibited enzyme that takes two conformations E (inactive) and E^* (active). An inhibitor I binds to E, shifting the population from E^* to E. The fraction of the enzyme in the active form given $[I]$ is:

$$E \cdot I \rightleftarrows I + E \rightleftarrows E^*$$
$$f_{active} = \frac{[E^*]}{[E^*] + [E] + [E \cdot I]}.$$

This protein has three distinct states—E^*, E and $E \cdot I$—that have different structures and would thus respond differentially to mutations. We would therefore expect ensemble-induced epistasis in f_{active}.

In addition to theoretical considerations, there is experimental evidence that ensemble-induced epistasis shapes evolution [108, 109]. The most direct is an engineered evolutionary trajectory that converts a protein from one fold to another [108, 97]. Midway through the trajectory, a single mutation switches the fold. If introduced earlier in the trajectory, the mutation does not have the same effect. The fold-switching mutation has its singular effect because other mutations have pre-stabilized the alternate fold. This is ensemble-induced epistasis: mutations perturb the relative stability of a non-native conformation, opening up a new evolutionary trajectory.

Another observation is the presence of high-order epistasis in every combinatorial protein genotype-phenotype studied [44, 22]. The phenotypes studied are diverse, including binding affinity, spectroscopic properties, and enzyme activity. Although there is no direct evidence that the high-order epistasis in these maps arises from underlying ensembles, ensemble-induced epistasis provides a simple, universal explanation that unites these disparate observations of high-order epistasis.

One question that remains is how our observations in lattice models map quantitatively to real proteins. What is the magnitude of ensemble-induced epistasis in real systems? How many steps can be predicted before real evolutionary predictions diverge? This will depend on the details of the sequence and its associated ensemble. Our results suggest, however, that ensemble-induced epistasis will eventually lead to divergence between predicted and actual trajectories in any protein genotype-phenotype map.

It also remains to be seen if something analogous to ensemble-induced epistasis exists on a larger scale, such as in a signaling network. Such networks do exhibit ensemble-like behavior, populating a collection of different configurations that rearrange in response to stimuli, sometimes even exhibiting stable three-state character [110, 111] We might therefore be able to explain high-order epistasis in such systems using an ensemble framework [44, 22].

Interpreting epistasis

Our analysis also sheds light on the question of the origin and interpretation of high-order epistasis. First, our work shows that it is relatively easy to create a system with irreducible high-order epistasis, even with a very simple lattice protein. There is no simple scale that reduces the epistasis [15, 22]: it is an integral part of the system.

Second, our work shows there is no mechanistic interpretation for epistatic coefficients that arise by such a process. The epistasis is fundamentally statistical rather than biological [15]. The ensemble effectively encrypts the interactions that give rise to the epistasis. A three-way interaction cannot be interpreted in a direct physical manner, nor in a way to predict which conformations changed. It quantifies the effect of the mutation, integrated over its effect on all conformations in the ensemble. In our view, the best interpretation of epistasis in macroevolutionary trajectories is as a means to quantify uncertainty in future predictions—not necessarily as a way to gain mechanistic insight into the system.

Evolution is unpredictable

Epistasis makes a relatively small contribution to variation in our lattice models (Figure 17A), as well as real datasets [44, 22]. This allows prediction of phenotypes with relatively high accuracy, as has been noted before in lattice models [98]. Approximate phenotypes are, however, insufficient for predicting trajectories. Because evolutionary trajectories are a contingent series of steps, small uncertainties in phenotypes are amplified into large uncertainties in trajectories [35]. Practically, this means you can predict a multi-mutation phenotype from a deep mutational scanning experiment, but likely not its evolutionary accessibility from the ancestral state.

Many previous discussions of unpredictability have revolved around robustness of trajectories to external factors such as environmental perturbation [90], genetic drift [36], or a change in the nature of selection [112]. The unpredictability we observe arises from the architecture of protein systems themselves. Our work indicates that the physical architecture of biomolecules naturally leads to ensemble-induced epistasis. Accumulating mutations thus alter the effects of future mutations, making evolution unpredictable given information about the effects of mutations in the ancestral state.

CHAPTER V

UNINTERPRETABLE INTERACTIONS: EPISTASIS AS UNCERTAINTY

Author Contributions

Zachary Sailer (ZRS) and Michael Harms (MJH) conceptualized the paper. ZRS conducted the simulations and computational analysis. ZRS generated figures. ZRS and MJH wrote the chapter.

Introduction

Epistasis—that is, non-additivity between mutations—is a ubiquitous feature of genotype-phenotype maps [1, 113, 44, 114, 32, 79, 115, 116, 117, 118, 119, 22, 120, 121, 122]. Epistasis can provide mechanistic insight into the determinants of phenotypes [123, 124, 40, 51]; however, it also complicates predicting unmeasured phenotypes [9, 34, 23, 125], as the effect of a mutation changes depending on the presence or absence of other mutations. Despite a century of work [14], epistasis remains challenging to analyze and interpret [15, 16, 87, 44, 122].

One approach is to decompose epistasis into specific pairwise and high-order interactions between mutations [58, 44, 52, 22, 126, 122]. This is often done by treating each coefficient as a linear and independent perturbation to the additive phenotype [58, 52]. Such an approach is a direct extension of classic approaches in quantitative genetics and biochemistry. In a genetics context, one might measure the effect of a mutation in two genetic backgrounds to dissect metabolic and regulatory pathways [40, 51]. Likewise, mutant cycles are a mainstay of biochemistry. Introducing mutations individually and together allows one to infer the nature of physical interactions between residues in macromolecules [123, 124].

Although linear epistasis models are very commonly used [44, 114, 32, 79, 120, 121], two recent observations raise questions about their utility. The first is that regression can lead to biased estimates of linear epistatic coefficients, and thus poor predictive power of epistatic models [73]. The second is that one can generate maps with extensive pairwise and high-order epistasis using a toy model of proteins that do not explicitly include such interactions [23]. This indicates that there may be no simple way to relate linear epistatic coefficients back to underlying biology, thus undermining their utility as indicators of biological mechanism.

Motivated by these concerns, we set out to systematically investigate linear epistatic models constructed from twelve published genotype-phenotype maps. We focused on two criteria for utility: the ability of such models to predict unmeasured phenotypes and the ability of such coefficients to provide mechanistic insight into the map. We studied maps for which all 2^L combinations of L mutations were measured. Because these maps have the same number of observations as coefficients in a high-order epistatic model, they can be readily decomposed into epistatic coefficients from second to L^{th}-order. Further, the selected maps cover many different classes of genotypes, phenotypes, and total magnitudes of epistasis.

We find that the epistatic coefficients we extract by regression from such maps are quite poor at predicting unmeasured phenotypes. This arises from bias in the regressed coefficients—exactly as predicted by Otwinowski and Plotkin [73]. Further, we find we can generate epistatic coefficients similar to experimental coefficients by simply using randomly assigned phenotypes. This suggests that the pairwise and high-order interactions we extract are likely decompositions of random noise. We therefore propose that we should not decompose genotype-phenotype maps into specific interactions between mutations using linear models. Rather, in the context of a whole genotype-phenotype map, epistasis is best interpreted as a global metric capturing roughness [7, 127]. This translates directly to a measure of uncertainty on predicted phenotypes, as well as an indication that an improved mechanistic model is required.

Materials and Methods

Linear epistasis models

We used a linear epistasis model to decompose genotype-phenotype maps into up to L^{th}-order epistatic coefficients. The model is linear in that it consists of a collection of independent epistatic coefficients that are summed to describe each phenotype [14, 52]. (The assumption of linearity contrasts with other models, such as a Potts model, in which mutations sum in a nonlinear fashion [118]). There are two common formulations a linear epistasis model, the Hadamard model (sometimes called a Walsh or Fourier model) and the biochemical model [52]. The approaches differ in their choice of coordinate origin. Each model has been described in detail elsewhere [58, 44, 52]. The two models are related by a simple set of linear transformations [52]. Throughout the text, we describe our results using the Hadamard model, but our conclusions are robust to the choice of model (see supplemental figures referenced throughout the text).

The Hadamard model uses the geometric center of the map as the coordinate origin [58, 44, 52, 22]. Each genotype is made up of L sites. In a binary genotype-phenotype map, the sites have two possible states: "wildtype" or "derived". Both states have equal effects but opposite signs. Each mutation is treated as a linear perturbation away from the origin of the map,

$$P = \beta_{\text{origin}} + \sum_{i}^{L} \beta_i x_i \tag{10}$$

where β_{origin} is the origin of the genotype-phenotype map, β_i is the effect of site i, and x_i is 1 if site i is "wildtype" and -1 if "derived". We can then add linear coefficients to describe interactions between mutations to Eq. 10. For pairwise interactions, this has the form:

$$P = \beta_{\text{origin}} + \sum_{i}^{L} \beta_i x_i + \sum_{j<i}^{L} \beta_{ij} x_i x_j \tag{11}$$

where β_{ij} is a pairwise epistatic coefficient. For the high-order model, the expansion continues:

$$P = \beta_{\text{origin}} + \sum_{i}^{L} \beta_i x_i + \sum_{j<i}^{L} \beta_{ij} x_i x_j + \sum_{k<j<i}^{L} \beta_{ijk} x_i x_j x_k + ... \tag{12}$$

The model can be expanded all the way to L^{th}-order interactions.

Linearizing experimental genotype-phenotype maps

Prior to extracting epistatic coefficients from experimental genotype-phenotype maps, we corrected each map for *global* epistasis, which arises when mutations combine on some scale other than an additive scale [17, 70, 56, 22, 21]. This violates the assumption of linearity inherent in the epistasis models [14, 15, 22]. Global epistasis manifests as a non-normal distribution of the residuals between the $\vec{P}_{\text{obs}}$ (the vector of observed phenotypes) and $\vec{P}_{\text{add}}$ (the vector phenotypes calculated using an additive model) [22, 21]. Such epistasis can be minimized by identifying a nonlinear function T that captures global curvature in the relationship between $\vec{P}_{\text{obs}}$ and $\vec{P}_{\text{add}}$, yielding normally distributed fit residuals [60, 7, 22, 21]:

$$\vec{P}_{\text{obs}} = T(\vec{P}_{\text{add}}) + \vec{\varepsilon}. \tag{13}$$

where $\vec{\varepsilon}$ are the fit residuals. We linearized all experimental maps by fitting a second-order spline to the $\vec{P}_{\text{obs}}$ vs. $\vec{P}_{\text{add}}$ curve for each map prior to extracting linear epistatic coefficients [21].

Epistasis and linear regression

We used linear regression to regress epistasis models against experimental and simulated genotype-phenotype maps. We formulated the problem as follows:

$$\vec{P}_{\text{obs}} = \mathbf{X}\vec{\beta} + \vec{\varepsilon} \tag{14}$$

where $\vec{P}_{\text{obs}}$ is a vector of observed phenotypes (corrected for global epistasis), $\vec{\beta}$ is a vector of epistatic coefficients, $\mathbf{X}$ is a matrix that encodes the sign of each coefficient according to Eq. 12, and $\vec{\varepsilon}$ is a vector of residuals. The goal was to estimate coefficients in $\vec{\beta}$ that minimized the magnitudes of the values in $\vec{\varepsilon}$.

We used three different regression approaches: ordinary least-squares, lasso, and ridge. The number of coefficients in these maps grows rapidly with the number of sites. For a binary map with L sites, there are 2^L possible fit coefficients. Lasso and ridge regression are strategies to identify only those coefficients that contribute significantly to the variation in the data. These strategies have been used previously to dissect linear epistatic models [73, 126]. Throughout the text, we describe results using lasso regression, but our conclusions are robust to the choice of regression strategy (see supplemental figures referenced throughout the text).

Simulating epistatic genotype-phenotype maps

We constructed genotype-phenotype maps using Equations 10 and 12. First, we set the additive coefficients to random values drawn from a normal distribution. We then added all 2^{nd}- through L^{th}-order epistatic coefficients. We set the values of the coefficients to random values drawn from a different normal distribution. The widths of the additive and epistatic distributions were tuned to match the relative magnitudes of epistatic coefficients extracted from experimental maps. Further, we could tune the fraction of epistasis in a simulated genotype-phenotype map by changing the relative widths of the additive and epistatic distributions with respect to one another.

Software

We implemented the epistasis models using Python 3 extended with *numpy*, *scipy*, and *pandas* [65, 128]. We used the Python package *scikit-learn* to perform ordinary-, lasso-, and ridge- regression [66]. We used the Python package *lmfit* to perform nonlinear-least squares regression [129]. Plots were generated using *matplotlib* and *Jupyter* notebooks [67, 68]. Our full software packages are available in the *gpmap* (https://harmslab.github.com/gpmap) and *epistasis* (https://harmslab.github.com/epistasis) packages on Github.

Results

Regression yields biased estimates of epistatic coefficients

We started with a straightforward question: What fraction of a genotype-phenotype map must we observe to resolve a linear epistatic model that predicts unmeasured phenotypes? We simulated a genotype-phenotype map consisting of all 2^8 binary combinations of 8 mutations. We then assigned random epistatic coefficients using an 8th-order Hadamard matrix, such that epistasis accounted for 20% of the variation in phenotype (see methods). The epistatic coefficients were similar in magnitude and sign to those extracted from experimental genotype-phenotype maps (Fig S1).

To test our ability to predict phenotypes, we masked a fraction of the genotypes, fit linear epistatic models to the unmasked genotypes, and attempted to predict the masked genotypes. We then calculated the correlation between the model and unmasked observations (ρ^2_{train}) and the model and masked observations (ρ^2_{test}). We repeated this for 1,000 pseudo-replicate training and test sets.

As a starting point, we fit the additive model (Eq. 10). We found that the additive model converged on $\rho^2_{train} = \rho^2_{test} = 0.8$ when $\gtrapprox 30\%$ of the map was used for the fit (red lines, Fig 19A). The model converges once each mutation has been observed across a sufficient number of genetic backgrounds to average out the epistatic perturbations to the phenotype. Because, by construction, 20% of the variation in the map is due to epistasis, the best the additive model can do is explain 80% of the variation in phenotype.

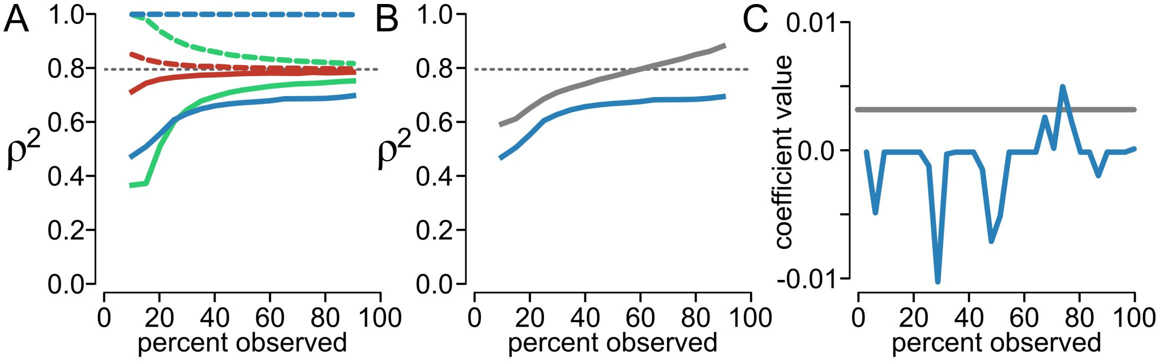


Figure 19: Linear epistatic coefficients cannot be estimated from an incomplete, simulated genotype-phenotype map. A) Fit scores versus the percent of the genotypes in the map used to train the model, from 10% to 90%. The dashed gray line indicates the amount of additive variation in the map (80%). Colors indicate model order: additive (red), pairwise epistasis (green), and high-order epistasis (blue). Dashed lines indicate ρ^2_{train} and solid lines indicate ρ^2_{test}. B) Fit scores versus the fraction of map used to train the model. Blue curve uses regressed coefficients (reproduced from panel A). Gray curve shows ρ^2_{test} if we use the coefficients used to generate the map. C) Value of a pairwise epistatic coefficient as we add data to the fit. Gray line indicates the value of the coefficient used to generate the map.

We next tried to improve our predictive power by adding coefficients describing either pairwise interactions between mutations (Eq. 11) or all interactions (up to eighth-order) (Eq. 12). Because Eq. 12 is the model we used to generate the map, this model should, in principle, be able to explain all variation in the map.

We found that neither the pairwise nor high-order models performed as well as the additive model (green and blue lines, Fig 19A). Even when 90% of the genotypes were included in the training set, the pairwise and high-order models had ρ^2_{test} of 0.73 and 0.62—much less than the value of 0.80 achieved by the additive model. Worse, this failure to predict the test set was accompanied by much higher ρ^2_{train} values. The high-order model, in particular, had a correlation of 1.0 with the training set (Fig 19A, dashed blue line), even while test set correlation languished around 0.6 (Fig 19A, solid blue line).

This result arises because regression yields biased estimates of the epistatic coefficients [73]. We know that high-order epistasis is present, because we used the same high-order model we are now fitting to generate the underlying map. The fit coefficients, however, do not accurately capture this variation. This can be seen in Fig 19B. The blue line reproduces ρ^2_{test} for the high-order model from Fig 19A. The gray curve shows values of ρ^2_{test} calculated using the epistatic coefficients used to generate the map. The divergence between these curves indicates that the regression fails to extract the correct values for the epistatic coefficients.

This can also be seen by examining the values of the extracted epistatic coefficients. The blue curve in Fig 19C shows the estimated value of a single, pairwise epistatic coefficient within the high-order model as data are added to the training set. The gray line shows the coefficient used to generate the map. Rather than monotonically converging to the true value, the estimated coefficient fluctuates in both magnitude and sign as data are added. This was common for all coefficients. We found that, on average, 65% of the pairwise coefficients flipped signs as data was added to our model.

These observations were robust to the choice of epistatic model and regression method. We used the Hadamard epistatic model with lasso regression for the results shown, but obtained identical results for all combinations of the Hadamard and biochemical epistatic models with ordinary, lasso, or ridge regression (see methods, Fig S2).

Predictive epistatic models cannot be extracted from experimental genotype-phenotype maps

We next asked whether experimental genotype-phenotype maps exhibited similar bias in their regressed epistatic coefficients. We analyzed 12 experimentally characterized genotype-phenotype maps (Table 4). All maps contained all combinations of L mutations, ranging in size from 32 to 128 genotypes. The maps consisted of very different classes of genotypes: collections of point mutations within a single gene, scattered genomic point mutations, or alternate alleles of genes in a metabolic network. The measured phenotypes are also diverse: competitive fitness, binding affinity, and parameters like growth rate and sporulation efficiency.

ID	genotype	phenotype	L	reference
I	genomic mutations	*E. coli* fitness	5	[18]
II	point mutants	bacterial fitness	5	[25]
III	chromosomes	*A. niger* fitness	5	[6]
IV	point mutants	binding affinity	5	[32]
V	alleles in network	*S. cerevisiae* growth rate	6	[59]
VI	alleles in network	*S. cerevisiae* growth rate	6	[59]
VII	genomic mutations	*E. coli* fitness	5	[85]
VIII	genomic mutations	*E. coli* fitness	5	[85]
IX	chromosomes	*A. niger* fitness	5	[6]
X	alleles in network	*S. cerevisiae* sporulation	6	[59]
XI	alleles in network	*S. cerevisiae* mating	6	[59]
XII	genomic mutations	*E. coli* fitness	6	[79]

Table 4: Published experimental genotype-phenotype maps.

We started by linearizing the experimental genotype-phenotype maps (see methods) [22, 21]. We then dissected each map into linear epistatic coefficients. Because all genotype-phenotype pairs in these maps have been measured, we have the same number of observations as epistatic coefficients. We can therefore decompose epistasis into linear coefficients using a matrix transformation [58, 52], avoiding complications arising from regression. We found that all of these maps exhibited statistically significant pairwise and high-order epistatic interactions. Epistasis contributed from 6% to 79% of the variation in these maps (Fig 20A).

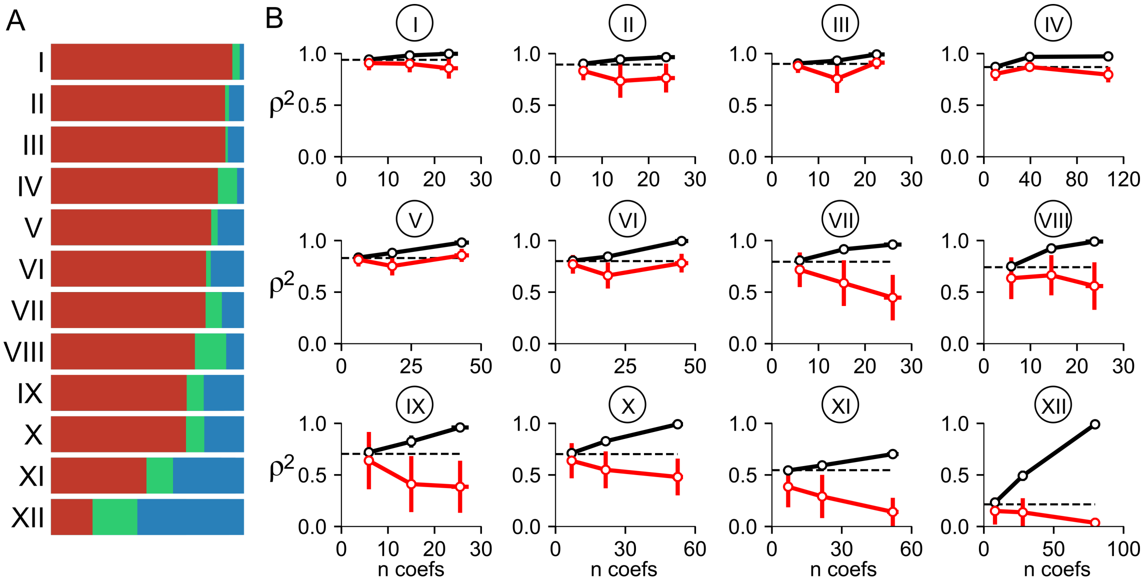


Figure 20: Predictive epistatic coefficients cannot be resolved from experimental genotype-phenotype maps. A) Bars show the fraction of variation in phenotype explained by additive effects (red), pairwise epistasis (green), or any order of high-order epistasis (blue). Each bar is for one of the twelve experimental genotype-phenotype maps. B) Each sub-panel shows ρ^2_{train} (black) and ρ^2_{test} (red) for the map indicated above the graph as epistatic orders are added to the model. The x-axis is the number of parameters used in the fit. Points are, from left to right: additive, pairwise, and high-order epistasis. Points and lines indicate the mean of 1,000 pseudoreplicate samples. Error bars are standard deviation of pseudoreplicate results. The dashed lines indicate the fraction of the variation in the map explained by the additive model. These fits used the Hadamard model with lasso regression. See Fig S3 for other epistatic models and regression strategies.

We next probed our ability to extract predictive epistatic coefficients from the linearized maps. We created a training set consisting of 80% of the genotype-phenotype pairs in each map, regressed models against this set of observations, and then predicted the phenotypes of the remaining 20% of the genotypes. As above, we fit the additive, pairwise and high-order models. We then repeated this for 1,000 pseudoreplicate training and test sets on each map.

As with our simulations, we found we could not reliably extract predictive epistatic coefficients (Fig 20B). In 11 of 12 maps, the additive model performed better than any other model. In seven of the twelve maps (I, VII, VIII, IX, X, XI, and XII), ρ^2_{test} consistently decreased with each addition of epistatic coefficients. In four of the maps (II, III, V, and VI) the addition of pairwise epistasis led to a drop in ρ^2_{test} that was partially offset by the addition of high-order coefficients. Ultimately, however, the high-order model did no better than the additive model in these maps. Map IV was the the only map in which adding epistatic coefficients had any positive effect: the addition of pairwise epistasis led to a small increase in ρ^2_{test} (from 0.80 to 0.87). This is achieved, however, by increasing the number of fit parameters from 10 to 40, implying that each epistatic coefficient contributed very little to the overall model. As with the simulated maps, these observations were robust to the choice of epistatic model and regression strategy (Fig S3).

Experimental epistatic coefficients cannot be distinguished from a random model

These results indicate that predictive, linear epistatic coefficients cannot be estimated by regression in these genotype-phenotype maps. We must characterize essentially every phenotype in a genotype-phenotype map to resolve the epistatic coefficients that describe the map. But, if we have measured every phenotype, there are no more phenotypes to predict. One might conclude that understanding epistasis requires measuring *every* genotype-phenotype pair in a map.

Given the effort required to measure every phenotype, we posed another question: is it worth exhaustively characterizing a map just to extract epistatic coefficients? Or, put differently, are the epistatic coefficients one can decompose from a complete map informative? We approached this question by comparing the epistatic coefficients extracted from an experimental genotype-phenotype map to those extracted from a null model. Our null model was a random map: we generated phenotypes with an additive model and then perturbed each phenotype by a random value drawn from a normal distribution centered at zero. This is an appropriate null model because the generating model has no mechanistic interactions at all; any correlations between mutations arise from noise. Such a map consists entirely of "statistical" epistasis [15].

We decomposed the epistasis in Map VIII using all 32 measured phenotypes and compared the resulting epistasis to our null model. Fig 21A-C shows the epistasis extracted from the experimental map. In this map 26% of the variation in phenotype is due to epistasis (Fig 21B). The residuals between the additive model and the observed phenotypes are normally distributed (Fig 21B). When we decompose the epistasis, we find that pairwise coefficients capture 16.2% and high-order coefficients capture 9.2% of the variation in phenotype.

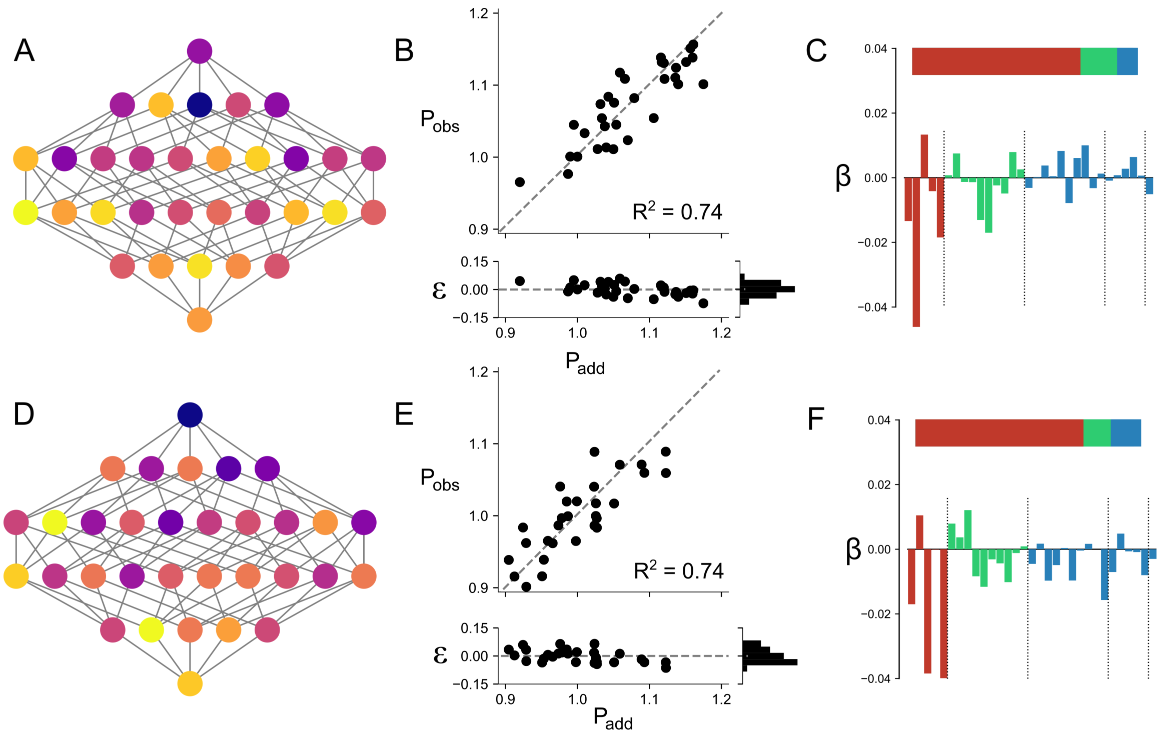


Figure 21: Experimental maps resemble random maps. Panels A and D show genotype-phenotype maps. Each node is a genotype; each edge is a single point mutation. Colors indicate quantitative phenotype. Panels B and E show the correlation between the observed phenotypes and the additive model, with fit residuals shown below the plot. Panels C and F indicate the magnitude of epistasis in each map as in Fig 20A (top subpanel) and the values of all model coefficients (bottom subpanel). Colors indicate additive components (red), pairwise components (green), and high-order components (blue). Each bar shows the value of a single model coefficient: the red bars correspond to the 5 additive coefficients, the green bars to the 10 pairwise coefficients, and the blue bars to the 17 high-order coefficients. Panels A-C are for experimental map VIII; panels D-F are for a simulated map with random epistasis.

We next constructed our null map. We generated a collection of random additive coefficients and calculated P_{add} for each genotype. We then added random perturbations to each phenotype, drawn from a normal distribution with a mean of 0 and a standard deviation selected to yield a total magnitude of epistasis similar to the experimental map. This sampling procedure gave the P_{obs} vs. P_{add} curve shown in Fig 21E. As with the experimental map, epistasis accounted for 26% of the total variation in the map. We then decomposed this random epistasis with a high-order epistasis model (Fig 21F).

The overall structure of epistasis is indistinguishable between the experimental and a random map, even though the values of the specific epistatic coefficients are different (Fig 21C vs. F). If we generate many random maps—effectively, sampling over the possible configurations of epistatic coefficients that arise from a random variation in phenotype—we cannot distinguish the experimental map from among the decoys (Fig S4). This suggests that the linear epistatic coefficients extracted from this map should be viewed as decompositions of random noise, unless this can be shown otherwise.

Using an additive model to treat epistasis

Our results speak against decomposing epistasis into collections of linear interaction terms. So how should we treat epistasis? We will touch on nonlinear treatments in the discussion, but before doing so, we will explore our top-performing epistasis model from above: the additive model.

The additive model treats epistasis as residual variation not explicitly accounted for by the model. If we measure the phenotypes of a set of combinatorial genotypes, we observe the effect of each mutation in a large number of genetic backgrounds (Fig 22A). We can describe the effect of mutation i with two numbers, its average effect $\langle \beta_i \rangle$ and the variance of its effect σ_i^2. This same logic applies at the level of whole genotypes. If we have linearized the genotype-phenotype map [7, 22, 21], the residuals between $\vec{P}_{\text{obs}}$ and $\vec{P}_{\text{add}}$ will be normally distributed (Fig 21B). As a result, the phenotype of a genotype g is given by:

$$P_{\text{obs,g}} = P_{\text{add,g}} \pm \xi \tag{15}$$

where ξ is the standard deviation of the residuals between $\vec{P}_{\text{obs}}$ and $\vec{P}_{\text{add}}$. This is the basic definition of epistasis given by Fisher [14], applied across the whole map.

This view is particularly useful for predicting unmeasured phenotypes. First: it means each predicted phenotype has a known, normally distributed uncertainty. Even if a large amount of variation remains unexplained by the additive model, it is safely partitioned into a random normal distribution. Put another way, ξ acts as a prediction interval. Second: because the additive model has few terms, we can train it using a very small amount of data.

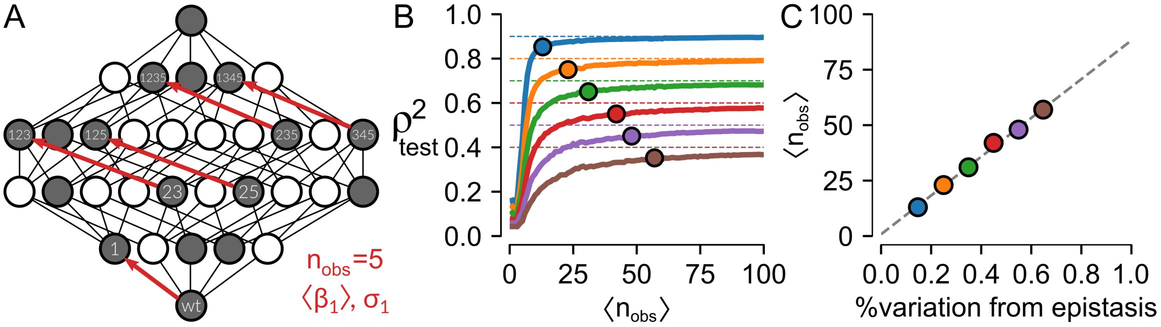


Figure 22: Epistasis as uncertainty. A) A partially characterized map. Circles represent genotypes, some of which have been measured (filled), some of which have not (unfilled). Lines represent single point mutations. Given these observations, we measure the effect of mutation 1 in five different backgrounds (red arrows) and can thus calculate the mean and variance in its effect across the map ($\langle\beta_1\rangle$ and σ_1). B) ρ^2_{test} versus the average number of times each mutation is seen in randomly sampled genotype-phenotype maps with epistasis responsible for 10% (blue) to 60% (brown) of the variation in the maps. Points indicate where ρ^2_{test} is within 5% of the maximum predictive power of the additive model. C) A calibration curve indicating how many times, on average, one must observe each mutation in map to resolve the additive coefficients in a map with different fractions of epistasis.

Following this line of reasoning, we asked how many phenotypes we would have to measure to construct a maximally predictive additive model. We constructed additive maps with different alphabet sizes (ranging from 2 to 5) and numbers of mutations (ranging from 6 to 8). We then injected random epistasis ranging in magnitude from 10% to 60% of the variation in the phenotype. We simulated experiments where we measured one random genotype at a time, added it to our observations, and predicted the phenotypes of the remaining genotypes. We then plotted ρ^2_{test} as a function of the average number of times we saw each individual mutation across all genetic backgrounds ($\langle n_{obs} \rangle$).

When plotted as a function of $\langle n_{obs} \rangle$, ρ^2_{test} rapidly rises and then saturates at the magnitude of the epistasis in the map, independent of alphabet size and number of mutations (Fig 22B). We next asked, as a function of the magnitude of the epistasis in the map, when our predictions would be within 0.05 of the best achievable ρ^2_{test}. This is indicated by the points on Fig 22B. We plotted these values as a function of the magnitude of the epistasis in the map. This reveals a linear relationship between the average number of times we need to see each mutation and the total epistasis in the map (Fig 22C).

We set out to test this approach using a partially sampled, experimental genotype-phenotype map characterizing the binding specificity of dCas9 to 23-base-pair oligonucleotides (Fig 23A). The published experiment sampled $59,394$ of the 7×10^{13} (4^{23}) possible oligonucleotides. Although all bases were sampled at all positions, there was significant bias towards a specific base at each position in the library (Fig 23A). The map exhibited a highly non-linear relationship between $\vec{P}_{\text{obs}}$ and $\vec{P}_{\text{add}}$ (Fig 23B), so we linearized the map with a 5th-order spline (Eq. 13), yielding normal residuals between $\vec{P}_{\text{obs,linearized}}$ and $\vec{P}_{\text{add}}$ (Fig 23C). We then assessed the predictive power of the map: we added genotypes individually to a training set and evaluated our ability to predict the test set. We found that we were able to fit a model using $\approx 4,000$ genotypes to predict the remaining $\approx 55,000$ measurements. Because of the biased sampling of genotypes in the map, it took $4,000$ genotypes to observe each individual mutation a sufficient number of times to resolve the additive effects of all mutations (Fig 23D). Our prediction curve saturated after we had seen each mutation at least 39 times. This is in good agreement with our calibration curve on simulated data, which indicated we would need to observe each mutation an average of 40 times (with random sampling) to saturate an additive model in which epistasis was responsible for 38% of the variation in the map.

The predictive power of this model is quite good considering its simplicity: we are able to predict any phenotype to ±38% given we only sampled one, one-billionth of a percent of the map. Extensive epistasis remains, but it follows a normal distribution with a known standard deviation. While there are certainly more sophisticated models, an additive model provides significant predictive power for this map.

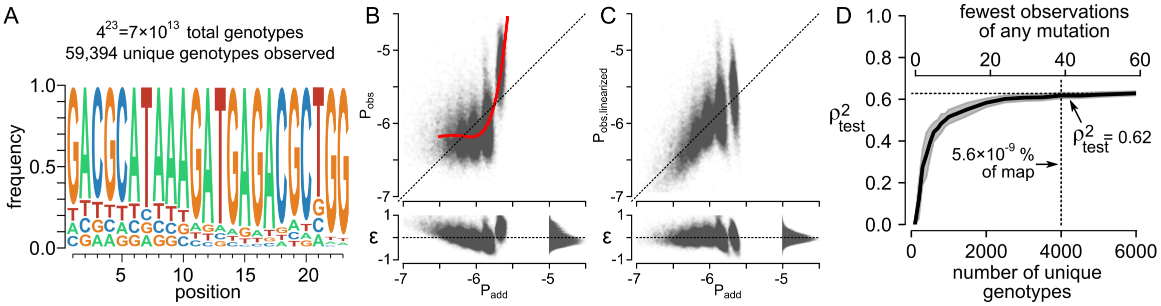


Figure 23: A predictive, additive model can be trained on a large genotype-phenotype map. A) Summary of the genotype-phenotype map reported in [117]. Map consists of 23 sites, each with four bases with the frequency at each site shown in the sequence logo. The total map has 7×10^{13} genotypes; the publication reports measured phenotypes for 59, 394 genotypes. B) Raw P_{obs} vs. P_{add} plot for the map. Each point is a genotype. The fit residuals are shown below the main plot. We fit an 5th-order spline to linearize the map (red curve). C) The linearized form of the map, with epistasis removed using the spline shown in panel B. D) A predictive model can be trained using ≈ 4, 000 genotypes. The bottom x-axis shows the number of unique genotypes used to train the model (sampled randomly); the top x-axis shows the fewest number of times any mutation was seen in that sample given the bias in the frequencies of the input mutations. ρ^2_{test} was measured against the remaining 50, 000+ genotypes not used to train the model.

Discussion

Our results suggest that a linear model should not be used to extract pairwise and multi-way interactions between mutations in a genotype-phenotype map. Regressed epistatic coefficients are biased (Fig 19B), unstable to the addition of new data (Fig 19C), and not useful for predicting unmeasured phenotypes (Fig 20A). Far from being an anomaly, this appears to be a shared feature of a collection of a dozen high-precision, combinatorially-complete genotype-phenotype maps (Fig 20B). Further, we can generate epistatic coefficients very similar to those observed in these maps using a simulated map in which we added random, normally distributed noise to each phenotype (Fig 21). This argues for viewing epistatic coefficients as uninterpretable decompositions of random variation, unless shown otherwise.

Viewed mechanistically, this is unsurprising. The epistatic models under investigation assume linearity, but biology is nonlinear [26, 130, 17, 70, 56, 22, 21]. There is no reason to believe that a linear model will capture a complicated nonlinear system in a predictive and interpretable way. For example, we showed recently that we could generate high-order epistasis using a toy thermodynamic model of proteins with only explicit pairwise interactions [23]. The epistasis arise because mutations have a nonlinear effect on the relative populations of individual protein conformations. As a result, epistatic coefficients cannot be interpreted mechanistically—they are purely "statistical" [15].

Further, our results indicate that the signs and magnitudes of specific epistatic interactions extracted from genotype-phenotype maps have no universal meaning. For example, in Fig 19C, the selected pairwise coefficient flips between positive, zero, and negative. If different genotypes of the map are characterized, we obtain different values for the pairwise coefficient, and thus a different interpretation for the effect of epistasis on the phenotype.

Treating epistasis with an additive model

A simple way to treat epistasis is as the residual variation after fitting an additive model (Eq. 15). Despite its simplicity, this is a useful perspective. It can be used to predict unmeasured phenotypes in a genotype-phenotype map with known uncertainty. This is because deviation from the additive model is determined by the magnitude of epistasis in the map (Fig 22A). Further, the simplicity of the model means we can characterize an extremely sparse sample of combinations of mutations across a genotype-phenotype map and still predict missing phenotypes (Fig 22C).

This suggests that sparsely sampling combinatorial genotypes, rather than aiming to exhaustively characterize point mutants, may be a powerful way to understand and predict genotype-phenotype maps. As long as each mutation is seen across a sufficiently large number of genetic backgrounds, we can resolve its average effect across a volume of the genotype-phenotype map. In contrast, exhaustively sampling point mutations in a single background—such as a deep mutational scan—will yield mutational effects specific to whatever genetic background is used. Epistasis is not averaged out, meaning such coefficients should not provide high predictive power when mutations are combined.

Interpretation of epistasis as a prediction interval only holds when the fit residuals are normally distributed about zero. Curvature between $\vec{P}_{\text{obs}}$ and $\vec{P}_{\text{add}}$ will lead to non-normal residuals and, thus, a distorted picture of the uncertainty [22, 21]. Multiple methods exist for linearizing genotype-phenotype maps, including taking the log of phenotypes [15], power transforms [22], splines [21], and even mechanistic models [56, 131]. This is one area for improvement, as better global models will decrease the amount of variation that must be explained by the additive model. For example, in our analysis of the dCas9 binding specificity, there is still structure in the residuals, with some clustering along P_{add} despite linearization using an 5th-order spline (Fig 23C). A model that captures such variation could improve predictive power. Further improvement of global models will thus be an important area of investigation.

Moving away from linear models

We see three promising ways forward. The first is to view epistasis in terms of its consequences for evolutionary trajectories. This includes metrics like the number of accessible trajectories, number of fitness peaks, and summary statistics such as the roughness to slope ratio [7, 127, 132]. These metrics generally do not allow prediction of unmeasured phenotypes nor mechanistic understanding of the map, but can provide useful insights into evolutionary trajectories and outcomes without the poor behavior observed in linear epistasis models.

The second is to use non-biological, nonlinear models to extract information from each map. These include tools such as Potts models [133, 118], variational auto encoders [134, 135], and neural networks [136, 137]. Such approaches can yield predictive models of genotype-phenotype maps, and will no doubt continue to grow in popularity and sophistication. One downside to these models is a requirement for massive amounts of training data—which may not always be feasible, even in the modern high-throughput era. Further, it may be difficult to link such models to an underlying biological mechanism.

The third is to attempt to model the underlying mechanistic process that leads to the map [70, 56, 131, 138]. Rather than taking a "top-down" approach, in which one dissects epistasis into statistical interactions that are hopefully meaningful, one can instead take a "bottom-up" approach, in which one calculates phenotypes from genotypes using a mechanistic biological model. This model can then be trained against measured phenotypes. This provides a predictive model for unmeasured phenotypes, as well as providing mechanistic insight into map between genotype and phenotype. A good example is that of Schenk et al, who dissected a genotype-phenotype map by explicitly modeling the effect of each mutation on protein stability and enzymatic activity [56]. This model captured extensive variation in the map that could not be described with a linear model, while also providing mechanistic insight into the protein under investigation.

Conclusion

Epistasis was described by Fisher as residual variation left over after fitting an additive model [14]. While it may sometimes be productive to separate these residuals into specific statistical coefficients, a better approach is to build better model. In our view, the long-term goal should not be interpreting epistatic interactions between mutations; rather, the long-term goal should be building mechanistic models that fit experimental observations and, ultimately, make epistasis disappear.

CHAPTER VI

PREDICTING DRUG RESISTANCE IN MALARIA VIA THE PFCRT GENOTYPE-PHENOTYPE MAP

Author Contributions

Zachary Sailer (ZRS) and Michael Harms (MJH) conceptualized the paper. ZRS conducted the computational analysis. ZRS generated figures. Robert Summers, Sarah Shafik, Alex Joule, and Prof. Rowena Martin collected the experimental data. ZRS wrote the chapter.

Introduction

Genotype-phenotype maps strongly shape evolution [4, 80, 26, 25, 8, 30, 7, 36, 9, 32, 37, 35, 23, 120]. The distribution and connectivity of phenotypes in a map determines the accessibility of adaptive evolutionary trajectories [6, 7, 2, 8, 9, 10], tunes evolutionary dynamics [139, 140, 18, 30], and alters population structure [141, 142]. Studying such maps can, however, be challenging as the size of a map rapidly increases as the number of mutations increases. For example, a map with four sites, each of which can can be in one of two states, has only 16 genotypes (2^4). In contrast, a map with 15 such sites has 32,768 genotypes (2^{15}). Because of the experimental challenge of exhaustively characterizing every genotype, researchers often only have access to portions of a map of interest [143, 27, 11, 37]. Being able to infer the phenotypes of missing portions of a genotype-genotype map could therefore be extremely useful.

In this work, we set out to infer unmeasured phenotypes in a map of intermediate size, containing $2^8 = 256$ genotypes. This regime is particularly important, as the evolution of features like drug or pesticide resistance often involve 5-10 mutations (32 to 1,024 genotypes) [25, 144, 55, 145, 11, 79]. Many important questions in such maps require knowledge of the phenotypes of all (or most) genotypes: Were there many or few evolutionary trajectories between the ancestral and modern protein? Were the pathway(s) adaptive, or were functionally neutral steps required? Why does resistance sometime evolve quickly [146, 147, 148], while taking decades in other scenarios [149, 150]?

A map with hundreds of combinatorial genotypes is large enough it may be difficult to characterize exhaustively, particularly for phenotypes that are difficult to characterize by high-throughput methods. Unfortunately, such a map is also small enough that it is not readily analyzed using sophisticated, data-hungry machine learning models. To address this problem, we have developed a straightforward approach to infer missing phenotypes from a small number of measured phenotypes. Our goal is to use combinatorial samples covering $\approx 20\%$ of a map to infer the remaining $\approx 80\%$, with well-characterized uncertainty in our predictions. Such knowledge would allow robust and statistically-informed analysis of evolutionary trajectories through a modeled genotype-phenotype map.

As a model system, we used the map for the evolution of the *Plasmodium falciparum* Chloroquine Resistance Transporter (PfCRT). PfCRT gives the malarial parasite resistance to chloroquine (CQ), a front-line malarial treatment for many decades [151, 150]. CQ is a diprotic base that diffuses into the *P. falciparum* food vacuole where it protonates, accumulates, and causes cell death by interfering with heme metabolism [152, 153] (Fig 24A). Eight mutations gave PfCRT the ability to transport protonated CQ out of the food vacuole [151, 154, 11] (Fig 24A).

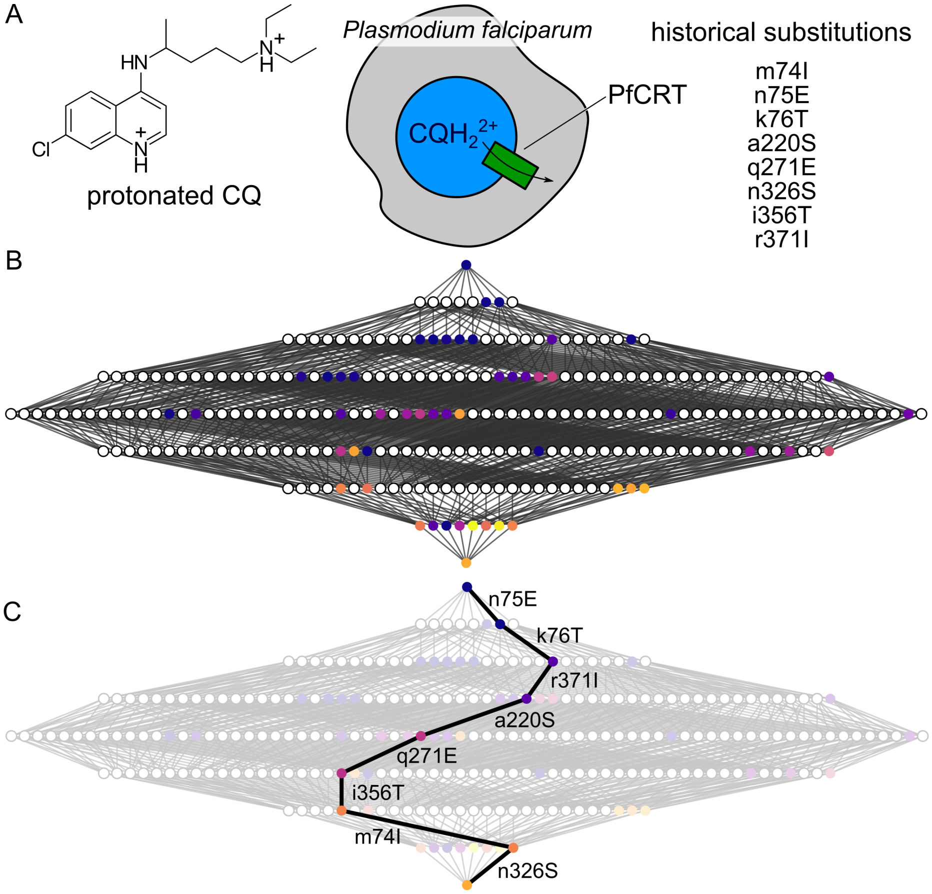


Figure 24: We have a sparse sample of phenotypes in the transition from low to high PfCRT CQ transportation activity. A) Cartoon representation of the mechanism of PfCRT-induced CQ resistance. A list of the eight mutations that accumulated between the ancestral and derived protein is shown on the right, with the ancestral amino acid in lowercase and the derived amino acid in uppercase. B) Network shows the complete PfCRT genotype-phenotype map. Nodes represent genotypes; edges represent single mutations. Node color indicates the previously measured level of CQ transport [11]. White nodes represent genotypes that have not been measured. C) One possible evolutionary trajectory from the ancestral protein to the modern PfCRT. Only six of the eight steps increase CQ transport.

The PfCRT map is an excellent system for developing a predictive model. The Martin lab previously characterized 52 of the 256 possible combinations of the 8 mutations in PfCRT [11]. They expressed these mutants in *Xenopus laevis* oocytes and then quantified CQ uptake by these cells, allowing them to fill in 20% of the PfCRT genotype-phenotype map (Fig 24B). From this, they were able to infer possible evolutionary trajectories that led to high CQ transport in PfCRT [11]. One of these trajectories is shown in Fig 24C. Interestingly, they found that every trajectory required at least one step in which the mutation did not alter CQ uptake. The lack of adaptive trajectories may explain why CQ resistance took many decades to evolve.

Although a few possible trajectories were revealed, it is unclear whether there are other accessible trajectories without knowing the complete genotype-phenotype map. There are $8! = 40,320$ possible forward trajectories through this map. The measured phenotypes allow us to assess the accessibility of 428 of these trajectories. This leaves $39,892$ trajectories—98.9%—for which we are missing one or more mutational steps. Measuring new phenotypes, however, is quite labor intensive, making it difficult to completely characterize the map.

We therefore set out to build a predictive model of the PfCRT genotype-phenotype map. The model incorporates the additive effects of mutations, a nonlinear scale, and logistic classifier. Further, we study the prediction error in the model, allowing for us to account for prediction uncertainty in evolutionary analyses. We have released our implementation of the model as an open source Python software package (GPSEER; https:github.com/harmslab/gpseer). The approach we describe should be general to many genotype-phenotype maps.

Results

Model development

We started with a linear, additive model that treats each mutation as an independent perturbation to the quantitative phenotype of CQ transport. We then added non-additive features to this model, as needed, to better describe the experimental observations.

We started by describing the phenotype of any genotype as the sum of the effects of its mutations [15, 155]:

$$P_{obs} \sim P_{model}$$

where P_{obs} is the observed phenotype and P_{model} is a linear model with the form:

$$P_{model} \sim \beta_{ref} + \beta_1 x_1 + \beta_2 x_2 + ... \tag{16}$$

β_{ref} is the phenotype of the reference genotype, β_i represents the quantitative effect of mutation i, and x_i is 0 if the mutation is present and 1 otherwise. We can estimate the effects of mutations in an additive model using linear regression:

$$P_{obs} = P_{model} + \epsilon, \tag{17}$$

where ϵ is the regression residual.

We fit this model to the 52 previously measured phenotypes of PfCRT. This yielded an R^2_{train} value of 0.65 between the model and the training data. In an effort to improve the model, we then inspected a graph of observed phenotype (P_{obs}) versus the predicted phenotype (P_{model}) (Fig 25A). This is an extremely powerful plot that can reveal mismatches between the linear, additive model and the statistical process that generated the data [22, 21].

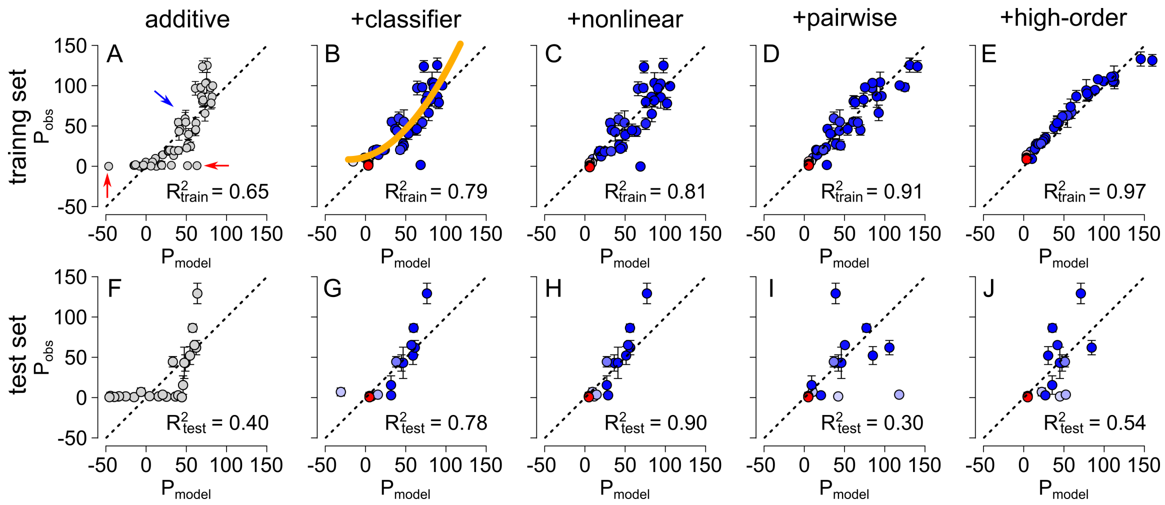


Figure 25: We can train models on measured phenotypes to predict previously unmeasured phenotypes. Panels show observed phenotype (P_{obs}) vs. predicted phenotype (P_{model}) for different input data and models. Vertical error-bars represent experimental uncertainty (two standard deviations) in the observed phenotypes. The dashed line is the 1:1 line (perfect agreement); the R^2 for is shown on each plot. The top row (panels A-E) shows fit quality for the training set (52 previously published phenotypes); the bottom row (panels F-J) shows fit quality for the test set (24 newly measured phenotypes). Columns, from left to right, are increasing model sophistication: additive (A,F), additive+classifier (B,G), additive+classifier+nonlinear (C,H), additive+classifier+nonlinear+pairwise epistasis (D,I), and additive+classifier+nonlinear+all epistatic orders (E,J).

The first observation we made from this graph was that genotypes appear to fall into two distinct classes. One class consisted of genotypes for which P_{obs} remained close to zero, even as P_{model} changes (red arrows, 25A). The other class consisted of genotypes for which P_{obs} increased monotonically with P_{model} (blue arrow, 25A). This suggested two, different, underlying processes. The linear model ends up making a poor compromise between these two classes.

We addressed this issue by adding a classifier to the model. We defined a threshold of 5% CQ uptake, corresponding to the minimum detection threshold of the CQ-transport assay [11]. We labeled genotypes with $P_{obs} \geq 5\%$ as *transporters* and $P_{obs} < 5\%$ as *non-transporters.* We then constructed a logistic classifier that predicts the class of any genotype given its sequence (see Materials & Methods). We trained this model on the PfCRT dataset and colored the points on the P_{obs} vs. P_{model} curve according to the likelihood the genotype belonged to the *transporter* or *non-transporter* class (Fig 25B). Note that the spread of points indicated by the red arrow in Fig 25A has collapsed to a set of overlapping red points in Fig 25B. We could then train the linear model against members of *transporter* class. The resulting fit is shown in Fig 25B, with an R^2_{train} of 0.79.

After we applied the classifier, we noticed a second feature of the P_{obs} vs. P_{model} plot: nonlinearity between P_{obs} and P_{model} (yellow line in Fig 25B). Such curvature arises when mutations do not add, but instead combine on a nonlinear scale [22, 21]. We therefore set out to improve the model by linearizing our data by 2^{nd}-order spline interpolation [21]. The spline we fit is shown as the yellow line in Fig 25B. We used the spline to linearize the data and then re-fit the additive model. The resulting fit is shown in Fig 25C; the addition of the nonlinear spline improved R^2_{train} to 0.81.

The additive model, classifier, and nonlinear correction leave 19% of the quantitative variation unaccounted for (the scatter off the line in Fig 25C). One way to treat the remaining scatter is with epistatic interactions between specific mutations [58, 52, 44, 22, 126]. We added all pairwise interactions to the model and then fit the model parameters using lasso regression [156, 126]. This approach uses L1 regularization to penalize the addition of unneeded parameters, thereby minimizing the number of new parameters added. This approach added 24 parameters, while increasing the R^2_{train} to 0.91 (Fig 25D).

Finally, it has been noted previously that under-specifying epistasis—that is, ignoring multi-way interactions—can lead to biased estimates of the low-order terms [73, 157]. We therefore also built a complete, high-order model that has all pairwise through eight-way interactions between mutations [58, 44, 52, 22]. As before, we used lasso regression to discard parameters that did not contribute to the fit. This added 70 more parameters than the pairwise model, increasing the R^2 to 0.97 (Fig 25E).

Testing the model with newly measured phenotypes

While the Harms lab developed the model using the original 52 measured PfCRT genotypes [11], the Martin lab measured 24 new phenotypes to test the predictive power of the model (Table 1). Genotypes were selected to sample across genotypes scattered throughout the map, as well as to different magnitudes of predicted phenotypes. CQ uptake was measured relative to the derived, 8-mutation PfCRT variant, where 100% indicates transport activity identical to this protein. The CQ uptake of the newly measured genotypes ranged from 0% to 129.2 ± 6.4% (Table 1). 14 of the 24 newly measured genotypes exhibited no transport above the detection threshold (5% CQ transport).

Table 1: Measured CQ transport phenotypes.

	% CQ uptake		
genotype	mean	std err	n
INKAENTR	0.0	0.8	5
MNTAQSTR	6.4	1.0	6
MNTSQSTI	0.8	0.8	4
METAESTR	42.5	2.3	5
MNTSENTR	0.5	0.3	4
MNTAENTR	1.2	0.6	4
MEKSENTR	0.7	0.2	4
IETAQNTR	15.0	2.6	5
IETAESII	51.8	2.8	4
METAENTI	42.8	4.9	4
INTSESII	44.0	3.3	3
IETSQSIR	61.7	4.5	3
IETSQNIR	64.9	0.6	3
METSQSTI	86.2	2.5	3
INKAQNIR	0.7	0.5	4
MNKAENIR	1.0	0.9	4
MNKAQSIR	1.5	1.1	4
MNKSQNIR	0.9	1.0	3
MNKAQNTR	0.1	0.2	5
MNKAQNII	-0.1	0.2	5
INTSQSII	3.0	0.9	10
METSESII	129.2	6.4	10
IETAQNII	2.2	1.7	9
IEKSESII	-0.1	0.3	14

We then used the 24 newly measured phenotypes to test the predictions of the models we had trained on the 52 published PfCRT genotypes. We predicted the phenotypes of all 24 new phenotypes and then compared the predictions to the measurements (Fig 25F-J). The addition of the classifier and nonlinear spline both improve the fit to the test data, with the R^2_{test} increasing from 0.40, to 0.78, to 0.90 (Fig 25F-H). The addition of pairwise epistasis, however, dramatically undermined the predictive power of the model. While R^2_{train} jumps from 0.81 to 0.91 with the addition of pairwise epistasis (Fig 25D), R^2_{test} fell from 0.90 to 0.30. (25I). The high-order epistatic model exhibited identical behavior (compare Fig 25E and J). This suggests that the epistatic interactions between mutations are not meaningful, but instead arise from over-fitting.

These results are summarized in Fig 26A. While R^2_{train} climbs monotonically as the model is made more complicated, R^2_{test} climbs with the addition of the classifier and nonlinear terms, but then drops precipitously upon addition of epistasis. To verify that this was not due to something specific to our training or test set, we repeated the analysis using k-fold cross validation. We combined the 76 phenotypes (the original 52 plus the newly measured 24) and then sampled from this set to create arbitrary pseudoreplicate 52-phenotype training sets and 24-phenotype test sets. These results replicate what we saw on the initial test and training set: fitting epistatic interactions leads to poor predictive power (Fig 26B).

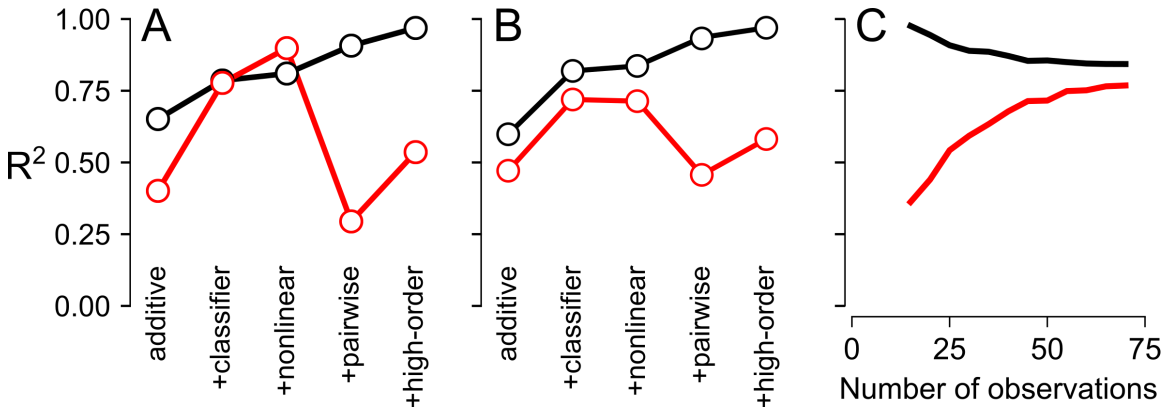


Figure 26: The additive+classifier+nonlinear model captures the most variation without over fitting. In all panels, lines denote R^2_{train} (black) and R^2_{test} (red). In Panels A and B, model complexity increases from left to right: additive model, add the classifier, add a nonlinear spline, add pairwise epistasis, and add high-order epistasis. A) R^2_{train} calculated from the 52 phenotypes used to train the model; R^2_{test} calculated from the 24 newly measured phenotypes. B) R^2_{train} calculated from 50-genotype pseudoreplicate samples from the 76 total genotypes; R^2_{test} from the matched 26-genotype pseudoreplicate test set. C) The mean of R^2_{train} and R^2_{test}, calculated for the additive+classifier+nonlinear model applied to pseudoreplicate samples, converge as the number of observations in the training set is increased.

We next set out to identify how many genotypes we would need to measure to fully train our best model (additive+classifier+nonlinear). We again took a pseudoreplicate cross-validation approach. We sampled the 76 measured phenotypes to create training sets ranging from 10 to 66 genotypes. We then trained our model on each pseudoreplicate training set and asked how well that model did, both at reproducing its training set and in predicting its matched test set. The results are shown in Fig 26C. R^2_{train} starts high and then decays. This is because we start with almost as many parameters as observations, and can identify a model that fits those few observations extremely well. As the number of observations increases, R^2_{train} decreases because more variation not accounted for by the model is encountered. In contrast, the predictive power of the model, measured by R^2_{test}, steadily increases until it plateaus at $\approx$ 60 genotypes. This indicates that measuring more than 60 genotypes provides sufficient data to train the model.

Model error

We next set out to quantify the error in our predictions. This model has two main sources of error: incorrect classification of genotypes as *non-transporter* or *transporter*, and quantitative uncertainty in the predicted phenotypes from the additive and non-linear portion of the model.

We estimated the false-positive and false-negative rate of the classifier using a pseudoreplicate approach. We took the 76 measured genotypes, broke them into 5,000 arbitrary 61-genotype training sets paired with 15-genotype test sets. We then retrained the classifier on these pseudoreplicate sets and measured our false positive and false negative rates. The average false positive and negative rates were 6.2% and 5.2%, respectively.

The uncertainty on our quantitative predictions is given by the fit residuals (scatter off the 1:1 line in Fig 25C,H) [157]. The standard deviation on the residuals is 19%, meaning that we can predict each quantitative phenotype (i.e. the phenotype of each member of the *transporter* class) to ±19% CQ uptake. This is much higher than the average experimental uncertainty on each phenotype (2.4% CQ uptake); we therefore treated the residual uncertainty as the quantitative uncertainty on each predicted phenotype.

The PfCRT map has a small number of viable genotypes

Using the predictive model, we can now reconstruct the entire map. We retrained the additive+classifier+nonlinear model on the 76 total experimental measurements. We then used the model to predict the remaining 180 genotypes, yielding an estimate of the complete PfCRT genotype-phenotype map (Fig 27A).

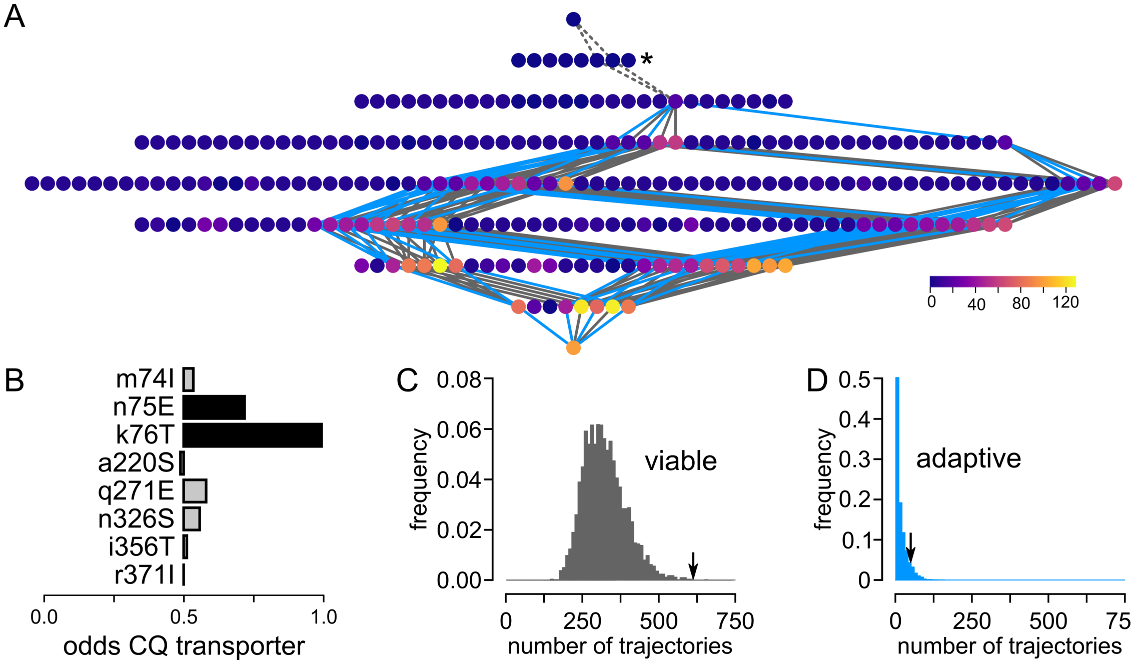


Figure 27: Only a few pathways to high CQ-transport proteins are accessible. A) Complete genotype-phenotype map with 76 measured and 178 predicted phenotypes. Nodes are genotypes. Colors are as in Fig 1. Star and dashed lines indicate first mutation, which does not exhibit CQ transport. Viable trajectories (which, after the first mutation, only pass through genotypes with detectable CQ transport) are shown in gray; adaptive trajectories (which, after the first mutation, continuously improve CQ transport) are shown in blue. B) Contribution of mutations to classifier, where the x-axis is the odds that the mutation is found in genotypes that transport CQ. C) The estimated number of viable trajectories, integrated over experimental uncertainty in the measured phenotypes and prediction uncertainty in the model. The arrow indicates the number of trajectories in the maximum likelihood model. D) The estimated number of adaptive trajectories, integrated over uncertainty as in panel C. The arrow indicates the number of trajectories in the maximum likelihood model.

We found that the majority of the map—163 genotypes—exhibited no detectable CQ transport. Strikingly, none of the single mutants exhibit detectable activity—meaning that the first evolutionary step in this transition had very little effect on CQ transport. Note that this was a prediction of the original, 52-genotype model; we validated this result by experimentally measuring all six previously unmeasured single-mutants: they did, indeed, exhibit no activity (Table 1).

The reason for the low number of CQ transport genotypes is revealed by the classifier. We counted how often each mutation appears in genotypes that the classifier places in the *transporter* class. If a mutation does not contribute to the classifier, it will appear in both classes equally (odds = 0.5). We found that the only the n75E and k76T mutations are strongly enriched in *transporter*s (27B), consistent with their proposed critical roles in CQ transport [11]. Only 64 genotypes have both of these mutations, thus strongly constraining the map.

PfCRT evolution is highly constrained

We next set out to understand the nature of evolutionary trajectories through this estimated map. We calculated the number of "viable" pathways—defined as trajectories in which every genotype after the single-mutant exhibited transport activity. We found that there were 612 trajectories that met this criterion through the estimated map (gray trajectories in Fig 27A). To account for uncertainty in the classifier, we generated 5,000 replicate maps in which we flipped the the classification of genotypes between *transporter* and *non-transporter* according to the estimated false negative and false negative rates. This yielded the distribution seen in Fig 27C. Within the accuracy of the predictive model, we conclude there are between 150 and 650 trajectories that pass only through genotypes that transport CQ. This corresponds to 0.4% to 1.6% of the possible forward trajectories.

This is likely an overestimate of the number of evolutionarily realistic trajectories; however, as many of these trajectories require transitions from higher to lower CQ transport. We therefore calculated the set of trajectories that continuously improve CQ transport. As before, we allowed the first mutation to have no effect, as there are no single mutants with measurable CQ transport. We then considered a trajectory accessible if each subsequent mutational step improved or had no effect on CQ transport. This yields the 51 blue trajectories shown in Fig 27A. To account for uncertainty in the experimental measurements and model, we again sampled from the error distributions. As before, we started by sampling the classifier false positive and false negative rates. We then added an additional layer of sampling, perturbing each the phenotype of each genotype in the *transporter* class by sampling from experimental or prediction uncertainty for that phenotype.

This yields the distribution shown in Fig 27D. 50% of the re-sampled maps yield no adaptive trajectories; 95% of the samples have 75 or fewer adaptive trajectories. This suggests that only $< 0.2\%$ of forward paths through this map are adaptive with respect to CQ transport. Recall also that we allowed an (apparently) non-adaptive step for the first mutation and allowed any step that did not decrease fitness, even if it's effect was neutral. We can therefore conclude that there are no trajectories through the map that continuously improve CQ transport within our detection limit, and that there are very few that do so after the first non-adaptive step.

Discussion

We have measured only 76 of 256 genotypes—29.7% of the map—but can now make robust conclusions about the distribution of CQ-transport activity across the PfCRT map and how this may shape the evolutionary process. The model we employ is lightweight, fast and should be applicable to other, potentially very large, genotype-phenotype maps.

Global features, not local epistasis, are useful for prediction

One key conclusion from our work is that estimating global features—a classifier and nonlinear scale—was much more powerful than estimating individual epistatic interactions between residues. The classifier identified a step-function in the map, from non-functional to functional, mediated by the two mutations n75E and k76T. Once the genotypes were classified, the nonlinear function then identified the appropriate scale on which to sum mutational effects. Both the classifier and scale capture a large amount of the variation in a phenotype with few, meaningful, parameters to allow prediction of unmeasured phenotypes.

In contrast, modeling specific epistatic interactions between residues gave no predictive power (Fig 26A). The values of pairwise epistatic coefficients cannot be extracted reliably from sparse samples of maps that contain high-order epistasis [157], because the high-order epistasis biases the low-order terms. Since every map we have investigated exhibits high-order epistasis [44, 35, 122, 121], it follows that the addition of epistatic coefficients describing specific interactions will not provide predictive information [157]. The PfCRT dataset demonstrates this quite powerfully: even the addition of pairwise epistatic coefficients strongly undermines prediction (Fig 26A). This suggests that the scatter off the linear model (Fig 25C) is caused by a mechanism besides specific interactions between mutations.

Another powerful aspect of using global features, rather than specific epistasis, is that we can fit the model with relatively few measured phenotypes. We recently estimated that seeing each mutation across ≈ 40 combinatorial backgrounds was sufficient to estimate its additive effect [157]. Our results are in line with this: our predictions only marginally improve after we have observed 20% of the map (Fig 26C), which corresponds to seeing each mutation at least 26 times. This implies that the number of genotypes one must characterize to resolve the map increases linearly with the number of mutations, even as the number of genotypes in the whole map scales exponentially.

Because the number of model terms increases linearly, our implementation should be effective even for massive genotype-phenotype maps. The primary requirement is that the training data consists of combinatorial genotypes, rather than different mutations introduced individually into the same genetic background. This is because the additive coefficients will be much more effectively resolved if one can average over background-dependent effects [157].

Future improvements

There are a two main ways to improve the model. The first would be defining a more general framework for the classifier. For the PfCRT map, a logistic classifier readily identified the two classes of genotypes visually apparent in the data. In other maps, a different classifier may be much more effective. This could be a Gaussian process classifier [158, 159] (which we have, in fact, implemented in our software package), a variant of a principle-component analysis [160], or any number of other unsupervised classifier algorithms [66].

The second—and probably more important—way to improve the approach would be through a better description of the underlying nonlinearity that gives rise to the map itself. We currently fit the nonlinear scale with a spline [21]. This is a powerful way to identify the curvature in data, but the spline is an empirical parameter that has no intrinsic, mechanistic meaning. In principle, it would be much more powerful to define a model that describes the mechanism that underlies the nonlinearity in the map [56, 23, 21]. With such a model in hand, one could describe the global shape of the model without resorting to the ad hoc approach of fitting a nonlinear scale. Further, such a mechanistic model could potentially remove apparently random epistatic interactions between the mutations, thus yielding a highly predictive model.

PfCRT evolution

Filling in the PfCRT map revealed that there are, in fact, very few viable trajectories connecting the ancestral genotype to the modern, high-resistance genotype. Several observations from the initial study have been borne out and extended with this analysis. First, across the whole map, CQ activity strictly requires both E75 and T76. This lends further credence to the idea that both of these residues are directly involved in transporting the positively charged from of CQ. As has been noted previously, n75E added a negative charge, while k76T removed a positive charge—strongly suggesting that these mutations operate by an electrostatic mechanism [154].

Second, because *both* n75E and k76T are required, the first step in any single-step trajectory starting at the ancestor likely did not alter CQ transport. This means that selection for CQ resistance could not drive the first mutation to a high frequency. This may indicate why CQ resistance took so long to evolve in PfCRT, relative to the evolution of other forms of antibiotic resistance. One of several low-probability scenarios were likely required: 1) The first mutation rose to a high frequency by drift or selection for some other trait, followed by fixation of the second mutation by selection; 2) The second mutation occurred by chance in a low-frequency member of the population that already had the first mutation; or 3) The two mutations were brought together by recombination across the gene.

Finally, the requirement for both n75E and k76T dramatically lowers the number of possible evolutionary trajectories through the map. It is conceivable, within the uncertainty of the model and experimental measurements, that there is in fact only one path from ancestral to derived—suggesting that adaptive evolution can indeed be highly constrained yet still reach high-fitness peaks [25, 145, 127].

Conclusion

These results show that a simple model can be quite effective in filling in unmeasured phenotypes in a genotype-phenotype map. Treating global properties—such as classifying genotypes with shared behavior and fitting nonlinear scale—allowed us to apply an additive model that explained 81% of the variation in an experimental genotype-phenotype map. Adding complexity via specific epistasis, far from helping, dramatically undermined the prediction of phenotypes. This implies that one need not measure a massive number of genotypes to infer the global properties of the map.

Materials and Methods

Linear epistasis models

We used a linear epistasis model to decompose the PfCRT genotype-phenotype map into up to 8^{th}-order epistatic coefficients. We used the Hadamard model which uses the geometric center of the map as the coordinate origin. Each genotype is made up of L sites. Each site has two possible states: "wildtype" or "derived" which are treated as a linear perturbations away from the origin of the map,

$$P = \beta_{\text{origin}} + \sum_{i}^{L} \beta_i x_i \tag{18}$$

where β_{origin} is the origin of the genotype-phenotype map, β_i is the effect of site i, and x_i is 1 if site i is "wildtype" and -1 if "derived".

We can add linear coefficients to describe interactions between mutations to Eq. 18. For pairwise interactions, this has the form:

$$P = \beta_{\text{origin}} + \sum_{i}^{L} \beta_i x_i + \sum_{j<i}^{L} \beta_{ij} x_i x_j \tag{19}$$

where β_{ij} is a pairwise epistatic coefficient. For the high-order model, the expansion continues:

$$P = \beta_{\text{origin}} + \sum_{i}^{L} \beta_i x_i + \sum_{j<i}^{L} \beta_{ij} x_i x_j + \sum_{k<j<i}^{L} \beta_{ijk} x_i x_j x_k + ... \tag{20}$$

The model can be expanded all the way to 8^{th}-order interactions.

Logistic classifier model

We used a logistic classifier to classify genotypes in the P_{obs} vs. P_{add} plot as active or inactive. A genotype is active if it's observed phenotype P_{obs} is greater than a defined threshold. The threshold represents the detection limit in our experiment. We labelled the observed phenotypes P_{obs} as 0 (inactive) or 1 (active):

$$P_{obs,class} = \begin{cases} 1 & \text{if } P_{obs} > \text{ threshold} \\ 0 & \text{else} \end{cases}$$

We then construct and train a logistic model to classify the additive phenotypes P_{add} given the observe phenotype classes $P_{obs,class}$:

$$P_{obs,class} = P_{add,class} = \begin{cases} 1 & \text{if } \alpha_0\beta_0 + \sum_i^L \alpha_i(\beta_i x_i) + \xi > 0 \\ 0 & \text{else} \end{cases}$$

where β_i are the additive coefficients determined in 18, α_i are the odds that genotypes with mutation i are inactive, and ξ is the error in our model (which follows a logistic distribution) [161, 162].

<u>Nonlinear model</u>

We addressed global epistasis in the genotype-phenotype map by identifying a nonlinear function T that captures global curvature in the relationship between $\vec{P}_{\text{obs}}$ and $\vec{P}_{\text{add}}$,

$$\vec{P}_{\text{obs}} = T(\vec{P}_{\text{add}}) + \vec{\varepsilon}. \tag{21}$$

where $\vec{\varepsilon}$ are the fit residuals. We fit a 2^{nd}-order spline to the P_{obs} vs. P_{add} curve [21] in the PfCRT map. We then used this transform to linearize the map and extract linear epistatic coefficients.

Software

All software is free and open source. The prediction models can be accessed and downloaded on Github (https://github.com/harmslab/gpseer). All of the code is written in Python and built on top of core scientific python packages [65, 128, 129, 67]. We have written extensive documentation and examples of how to apply to various types of experimental data (https://gpseer.readthedocs.io). The software takes a list of genotypes with their measured phenotypes and, from that, trains the model. The code is written with a modular application programming interface, allowing users to try each layer of the model—classifier, nonlinear fit, and epistasis—with simple Python scripts (or integrated into a programming environment such as Jupyter). Because it is built on the existing epistasis package (https://github.com/harmslab/epistasis) [22], it can also handle arbitrary genotype alphabets (it builds the alphabet from the input data). Finally, all statistical tests and cross-validation are implemented within the software, allowing for any researcher to fit and select between different predictive models of the genotype-phenotype map.

CHAPTER VII

CONCLUSIONS AND FUTURE DIRECTIONS

This dissertation addresses a key question when predicting genotype-phenotype maps. How should we handle non-additivity–or epistasis? We identified two sources of epistasis: 1) global epistasis and 2) local epistasis. Each source must be treated separately. First, global epistasis is important feature to capture. If a global feature, like nonlinearity, can be resolved from a sparsely-sampled genotype phenotype map, it can inform prediction. This dissertation shows various ways to infer global epistasis. Local epistasis, on the other hand, should be treated as uncertainty. In general, specific epistatic interactions cannot be accurately estimated from a sparsely-sampled genotype-phenotype map. As a consequence, using local epistasis yields biased predictions.

A major limitation of this approach is when local epistasis is a large contributor to the variation in a genotype-phenotype map. In this scenario, the model will not predict phenotypes accurately. However, this model will estimate how much local epistasis is present within a heavy computational cost. We can apply this model and know immediately if we will be able to predict phenotypes accurately.

A major benefit of this approach is that it can predict massive genotype-phenotype maps from very sparse data. The model is extremely lightweight and computationally optimized. It also generalizes to predict any arbitrary genotype-phenotype map. Further, the interpretation of the model parameters are intuitive. Each mutation has an average quantitative effect on the phenotype. We can resolve which mutations are key to function and use this information direct further analysis.

A key follow up to this dissertation is: *can we do better than a nonlinear additive* *model?* Recently, new approaches have been proposed that pull information from outside the genotype-phenotype map. For example, some groups are exploring variational auto-encoders (VAE) as a viable approach. VAEs use a deep neural network to decompose a large set of aligned genotypes, identify latent key features, and recompose these features to predict unknown phenotypes. The interpretation of such models are less clear, since the latent variable is not interpretable. However, they offer a promising future for phenotypic prediction.

CHAPTER VIII

APPENDIX

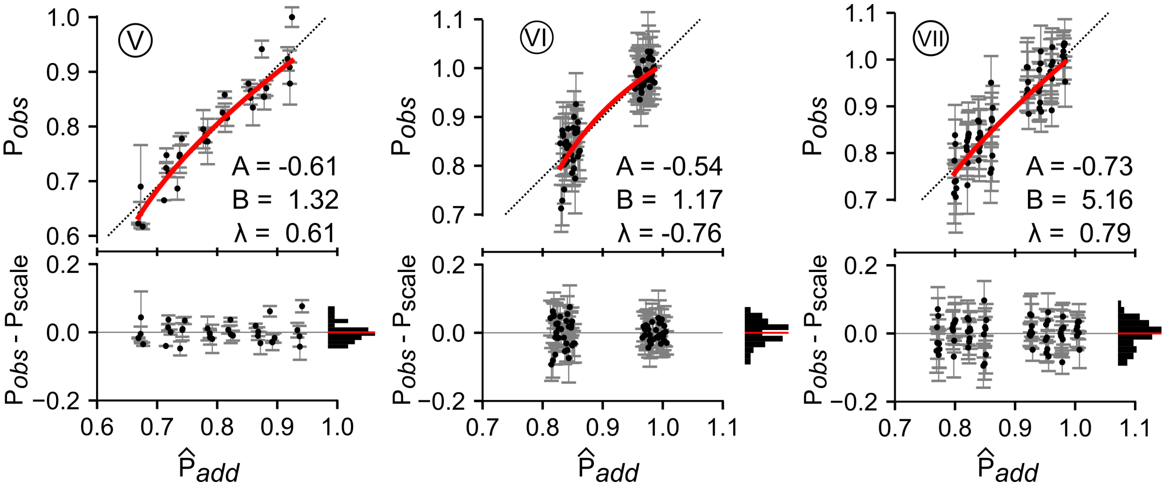


Figure 28: Experimental genotype-phenotype maps exhibit nonlinear phenotypes. Plots show observed phenotype P_{obs} plotted against $\hat{P}_{add}$ (Eq. 3) for data sets V through VII. Points are individual genotypes. Error bars are experimental standard deviations in phenotype. Red lines are the fit of the power transform to the data set. Pearson's coefficient for each fit are shown on each plot. Dashed lines are $P_{add} = P_{obs}$. Bottom panels in each plot show residuals between the observed phenotypes and the red fit line. Points are the individual residuals. Errorbars are the experimental standard deviation of the phenotype. The horizontal histograms show the distribution of residuals across 10 bins. The red lines are the mean of the residuals. Datasets VI and VII form distinct clusters because each map has a single, large-effect mutation. The two clusters correspond to genotypes with and without the large-effect mutation.

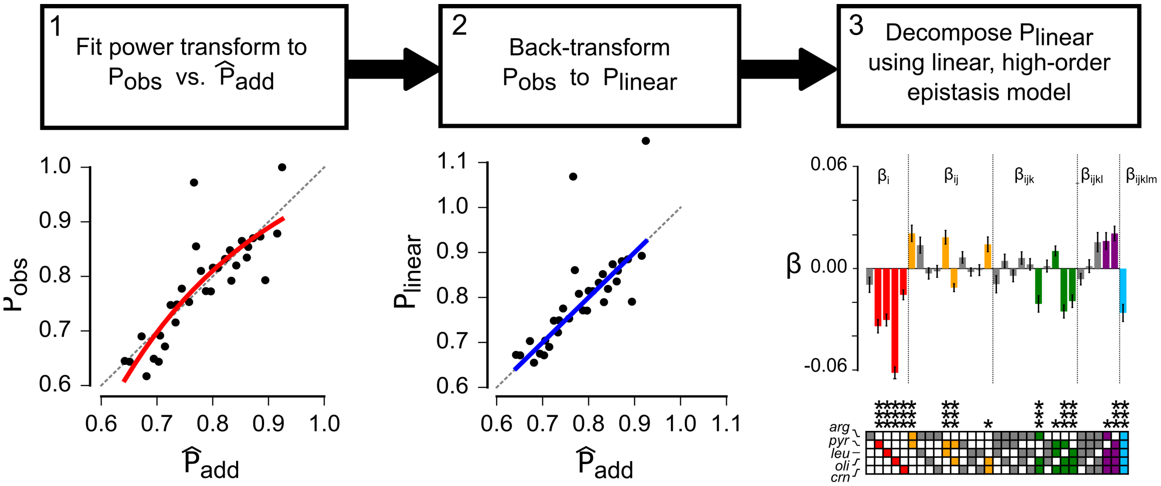


Figure 29: Nonlinear phenotypes can be transformed to linear scale to estimate high-order epistasis. Flowchart shows the steps for estimating high-order epistasis in nonlinear genotype-phenotype maps. The plots beneath the chart show this pipeline for data set II. In step 1, a power transform function is used to fit the P_{obs} versus $\hat{P}_{add}$ plot and estimate the map's scale. In step 2, the inverse of the fitted transform is used to back-transform P_{obs} to a linear scale, P_{linear}. In step 3, a linear, high-order epistasis model is used to fit the variation in P_{linear}. In the left plot, points are individual genotypes, red line is the resulting fit and dashed line is the $P_{add} = P_{obs}$. In the middle plot, the blue line is the new scale of P_{obs} after back transforming. In the right plot, bars represent additive and epistatic coefficients extracted from the linear phenotypes. Error bars are propagated measurement uncertainty. Color denotes the order of the coefficient: first (β_i, red), second (β_{ij}, orange), third (β_{ijk}, green), fourth (β_{ijkl}, purple), and fifth (β_{ijklm}, blue). Bars are colored if the coefficient is significantly different than zero (Z-score with p-value <0.05 after Bonferroni correction for multiple testing). Stars denote relative significance: p < 0.05 (*), p < 0.01 (**), p < 0.001 (***). Filled squares in the grid below the bars indicate the identity of mutations that contribute to the coefficient.

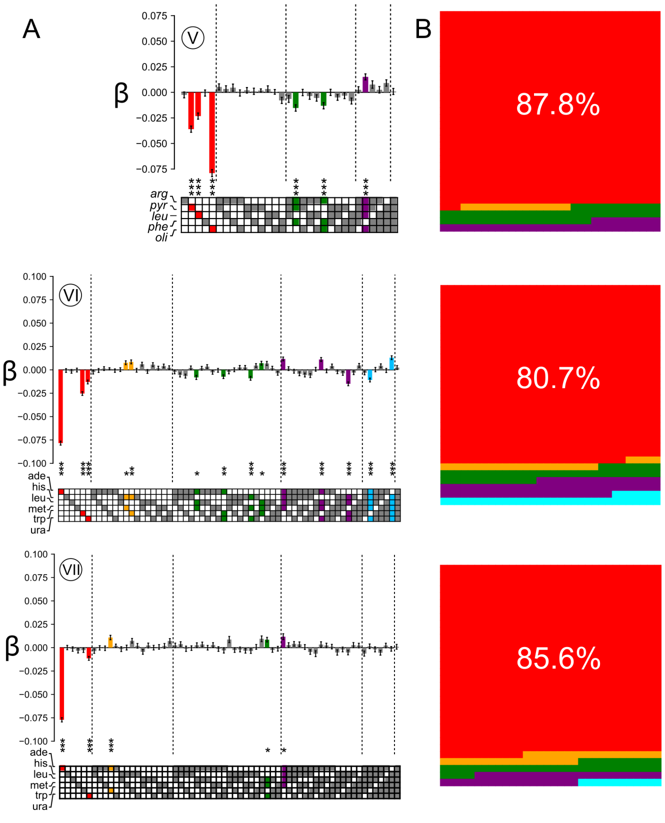


Figure 30: High-order epistasis is present in genotype-phenotype maps. A) Panels show epistatic coefficients extracted from data sets V-VII (Table 1, data set label circled above each graph). Bars denote coefficient magnitude and sign; error bars are propagated measurement uncertainty. Color denotes the order of the coefficient: first (β_i, red), second (β_{ij}, orange), third (β_{ijk}, green), fourth (β_{ijkl}, purple), and fifth (β_{ijklm}, blue). Bars are colored if the coefficient is significantly different than zero (Z-score with p-value <0.05 after Bonferroni correction for multiple testing). Stars denote relative significance: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***). Filled squares in the grid below the bars indicate the identity of mutations that contribute to the coefficient. The names of the mutations, taken from the original publications, are indicated to the left of the grid squares. B) Sub-panels show fraction of variation accounted for by first through fifth order epistatic coefficients for data sets I-IV (colors as in panel A). Fraction described by each order is proportional to area.

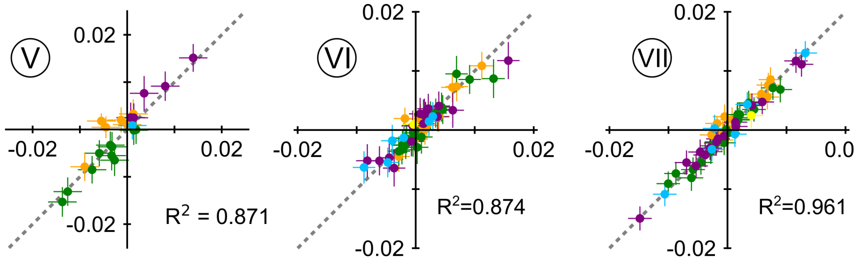


Figure 31: Nonlinear phenotypes distort measured epistatic coefficients. Sub-panels show correlation plots between epistatic coefficients extracted without accounting for nonlinearity (x-axis) and accounting for linearity (y -axis) for data sets V-VII. Each point is an epistatic coefficient, colored by order.

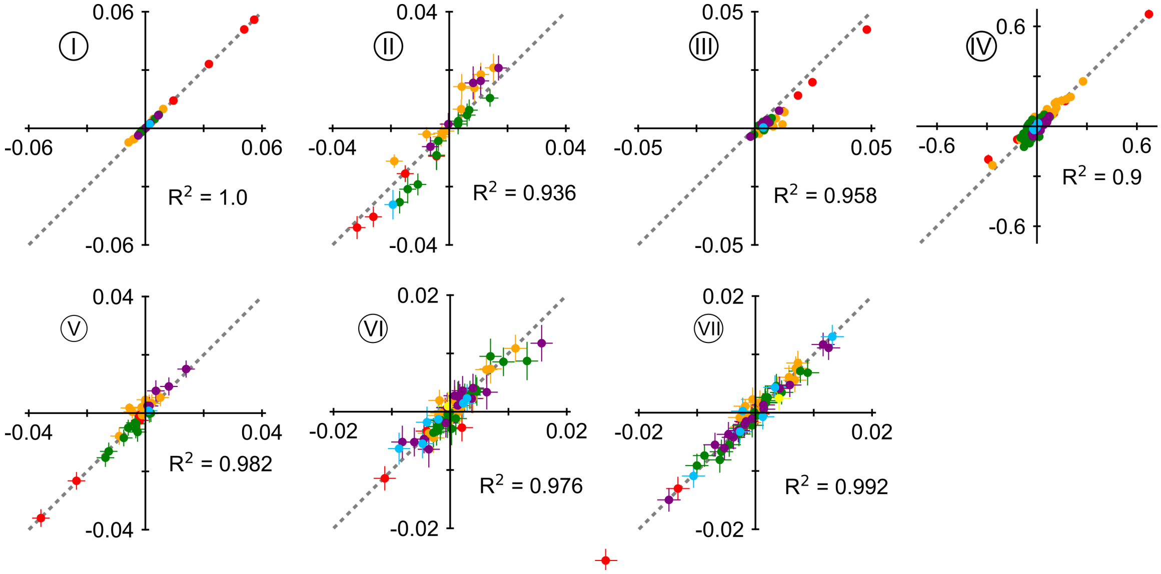


Figure 32: Additive coefficients are well estimated, even when nonlinearity is neglected. Sub-panels show correlation plots between both additive and epistatic coefficients extracted without accounting for nonlinearity (x-axis) and accounting for linearity (y-axis) for data sets I-VII. Each point is an epistatic coefficient, colored by order. Error bars are standard deviations from bootstrap replicates of each fitting approach.

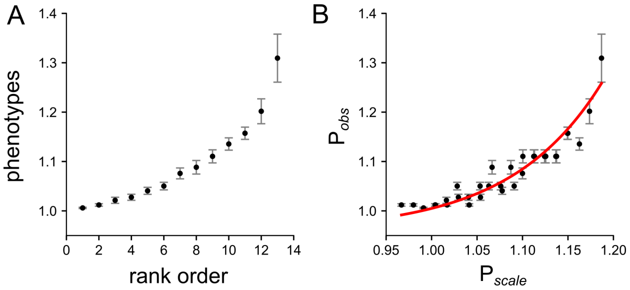


Figure 33: Exponential fitness model leads to global nonlinearity in the β-lactamase data set (III). A) A recapitulation of the map used in the original publication [**?**]. We first rank-ordered the genotypes according to the measured property (the minimum inhibitory concentration of a β-lactam antibiotic against a clonal population of bacteria expressing that protein). This gave us 13 classes of genotypes, as some genotypes had equivalent MIC values. We then drew 3,000 random fitness values from the distribution $W = 1 + x$, where x is an exponential distribution centered around $\bar{x} = 0.1$. We took the top 13 values from this distribution and assigned them, in value order, to each of the 32 β-lactamase genotypes. Panel A shows the average and standard deviation of the fitness values W assigned to each of these ranks if we repeat the protocol above 1,000 times. B) Best fit for the power-transform for data set III. Solid red line denotes the best fit (nonlinear). This fit successfully pulls out the original distribution of W.

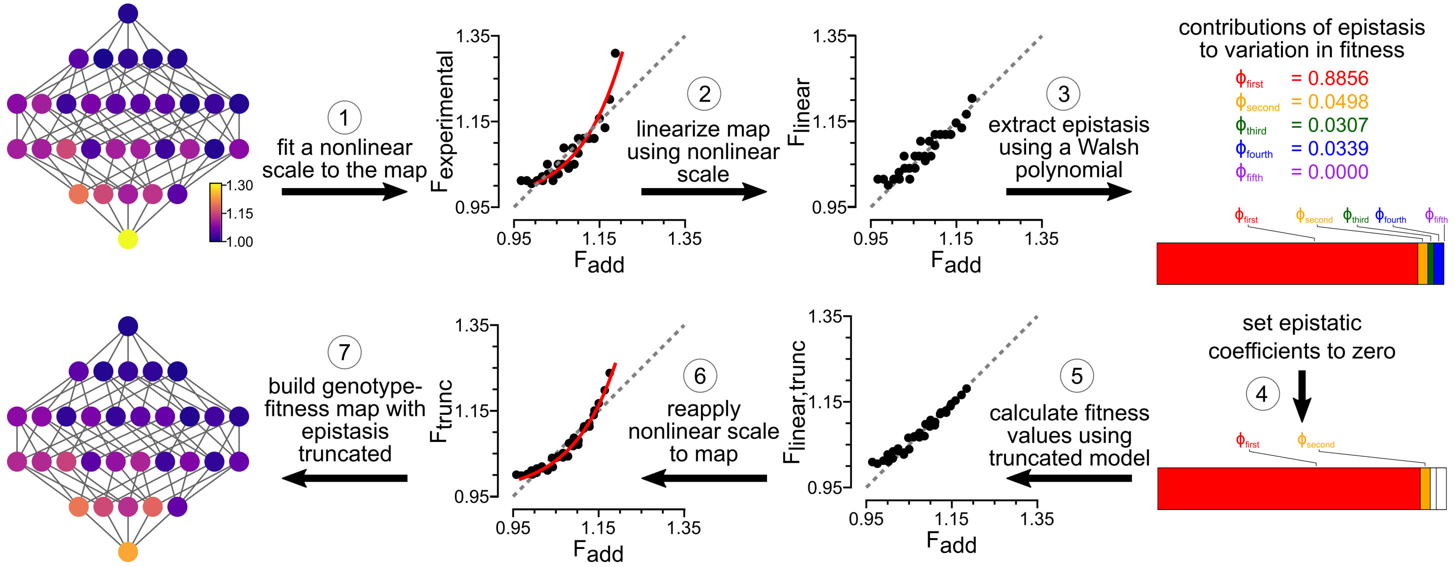
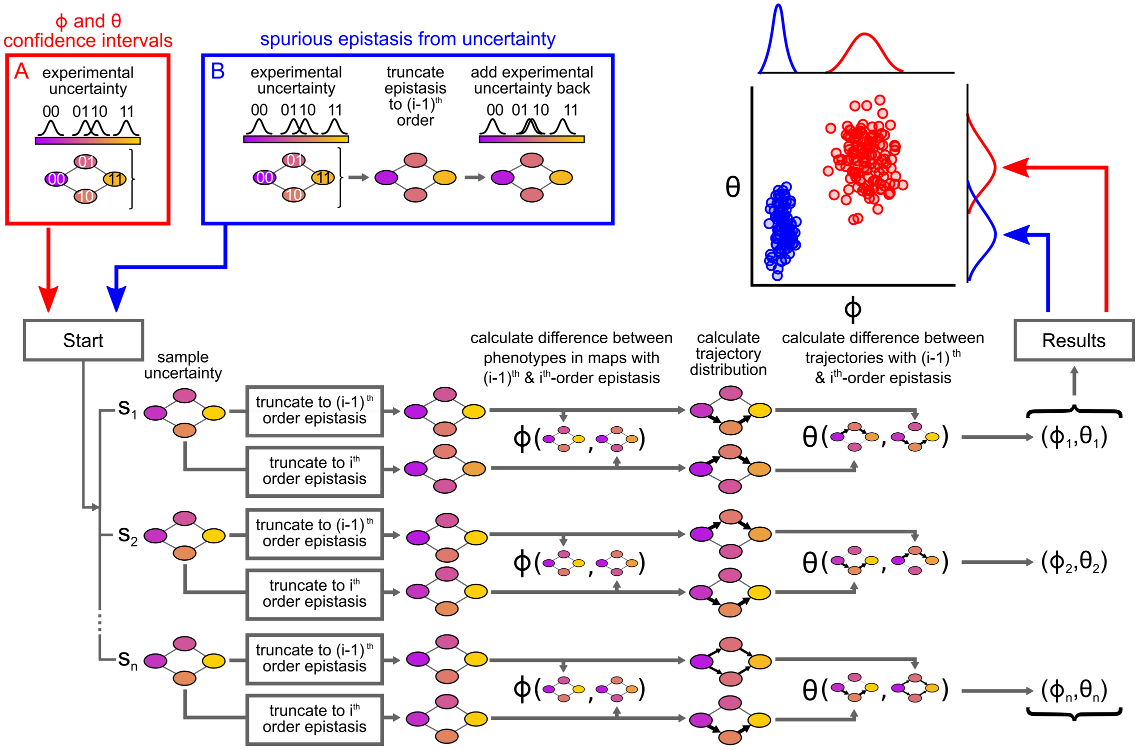
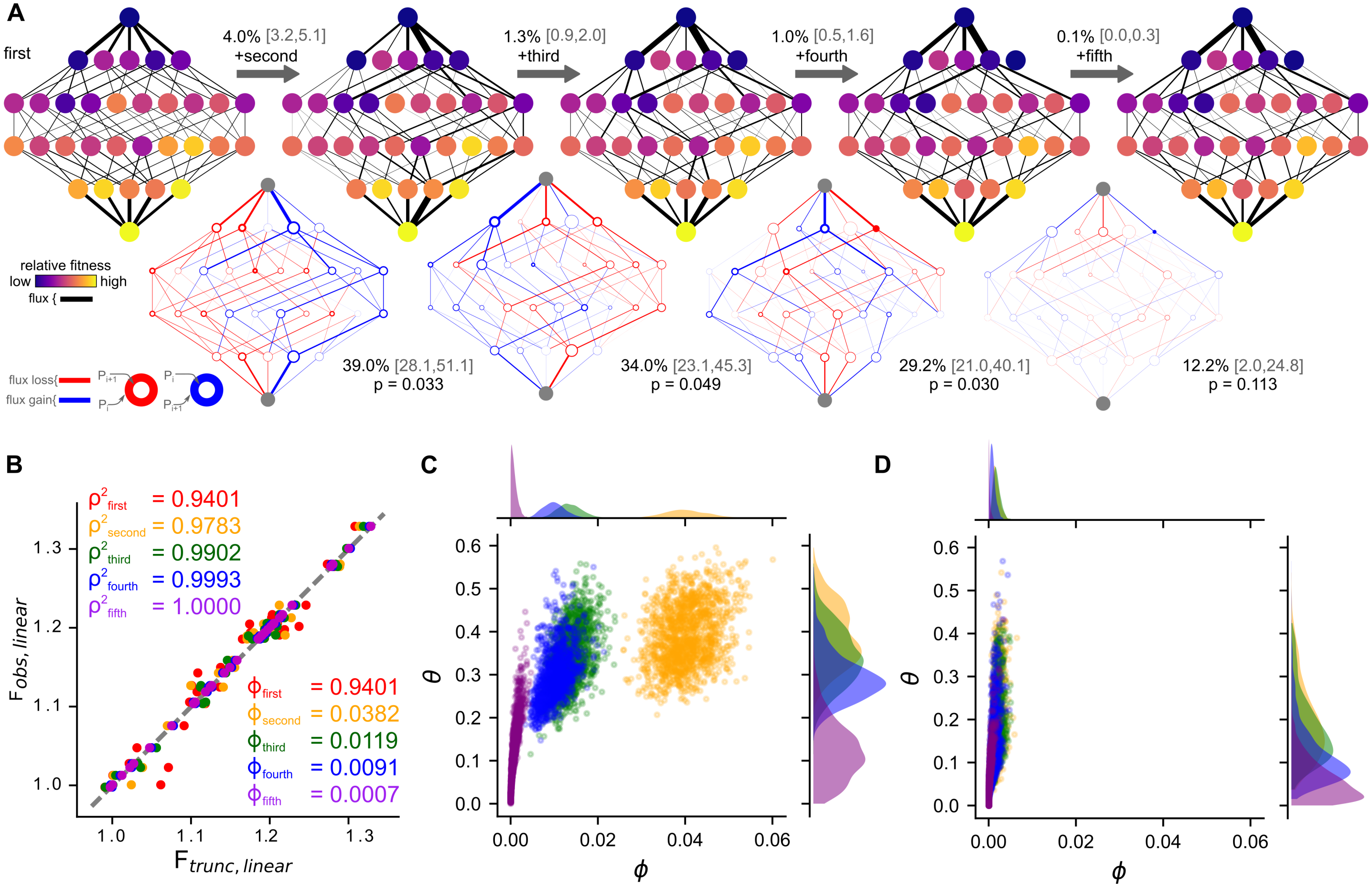

1.30
1.15
1.00
1
fit a nonlinear scale to the map
F experimental
1.35
1.15
0.95
0.95
1.15
1.35
F add
2
linearize map using nonlinear scale
F linear
3
extract epistasis using a Walsh polynomial
contributions of epistasis to variation in fitness
ϕ first = 0.8856
ϕ second = 0.0498
ϕ third = 0.0307
ϕ fourth = 0.0339
ϕ fifth = 0.0000
set epistatic coefficients to zero
4
5
calculate fitness values using truncated model
F linear,trunc
6
reapply nonlinear scale to map
F trunc
7
build genotype-fitness map with epistasis truncated

Figure 34: Flow chart for removing epistasis in a genotype-fitness map. This chart describes the pipeline we used to truncate epistasis from genotype-fitness maps. The data shown are for dataset II. Networks (left) show all 2^5 genotypes, arranged from ancestral (top) to derived (bottom), colored by relative fitness from 1.0 (purple) to 1.30 (yellow). The correlation plots (middle) show the fitness of each genotype plotted against the fitness of that genotype assuming each mutation has a linear, additive effect on fitness (F_{add}). Y-axes correspond to: the experimentally measured fitness ($F_{experimental}$, panel 2); the experimentally measured fitness linearized using the red scale in panel 2 (F_{linear}, panel 3); fitness values with third-, fourth- and fifth-order epistasis removed, on the linear scale from panel 3 ($F_{linear,trunc}$, panel 5); and fitness values with truncated epistasis on the red nonlinear scale from panel 2 (F_{trunc}, panel 6). The right-most panels show the fraction of variation explained by first- (red), second- (orange), third- (green), and fourth-order (purple) epistatic coefficients. The area occupied by each color indicates its contribution to the fitness on the linear scale.



ϕ and θ
confidence intervals
A
experimental uncertainty
00 01 10 11
01
00
11
10
spurious epistasis from uncertainty
B
experimental uncertainty
truncate epistasis to (i-1)th order
add experimental uncertainty back
θ
ϕ
Start
Results
sample uncertainty
s1
s2
sn
truncate to (i-1)th order epistasis
truncate to ith order epistasis
calculate difference between phenotypes in maps with (i-1)th & ith-order epistasis
calculate trajectory distribution
calculate difference between trajectories with (i-1)th & ith-order epistasis
ϕ(,)
θ(,)
(ϕ1,θ1)
(ϕ2,θ2)
(ϕn,θn)

Figure 35: Schematic of our resampling protocol. Two-mutation maps are shown throughout, colored by fitness from low (purple) to high (yellow). We sampled from two maps: the original map with uncertainty (A, red) and a "null" map in which epistasis was removed, but experimental uncertainty maintained (B, blue). We used the same sampling protocal on each ("Start"). We generated pseudoreplicates $(s_1, s_2, ...s_n)$ from uncertainty (Gaussian curves above the color spectrum in A and B). We then truncated the pseudoreplicate to i^{th} and $(i-1)^{th}$ order epistasis and calculated ϕ and θ for each pseudoreplicate: $\{(\phi_1, \theta_1), (\phi_2, \theta_2), ...(\phi_n, \theta_n)$. We can then plot and compare these distributions on ϕ/θ axes.



dd

Figure 36: Epistasis alters trajectories in dataset I. A) Colors, panel layouts, and statistics are as in Fig 2. B) Colors, panel layouts, and statistics are as in Fig 1A. C-D): Colors and panel layouts are as in Fig 3.

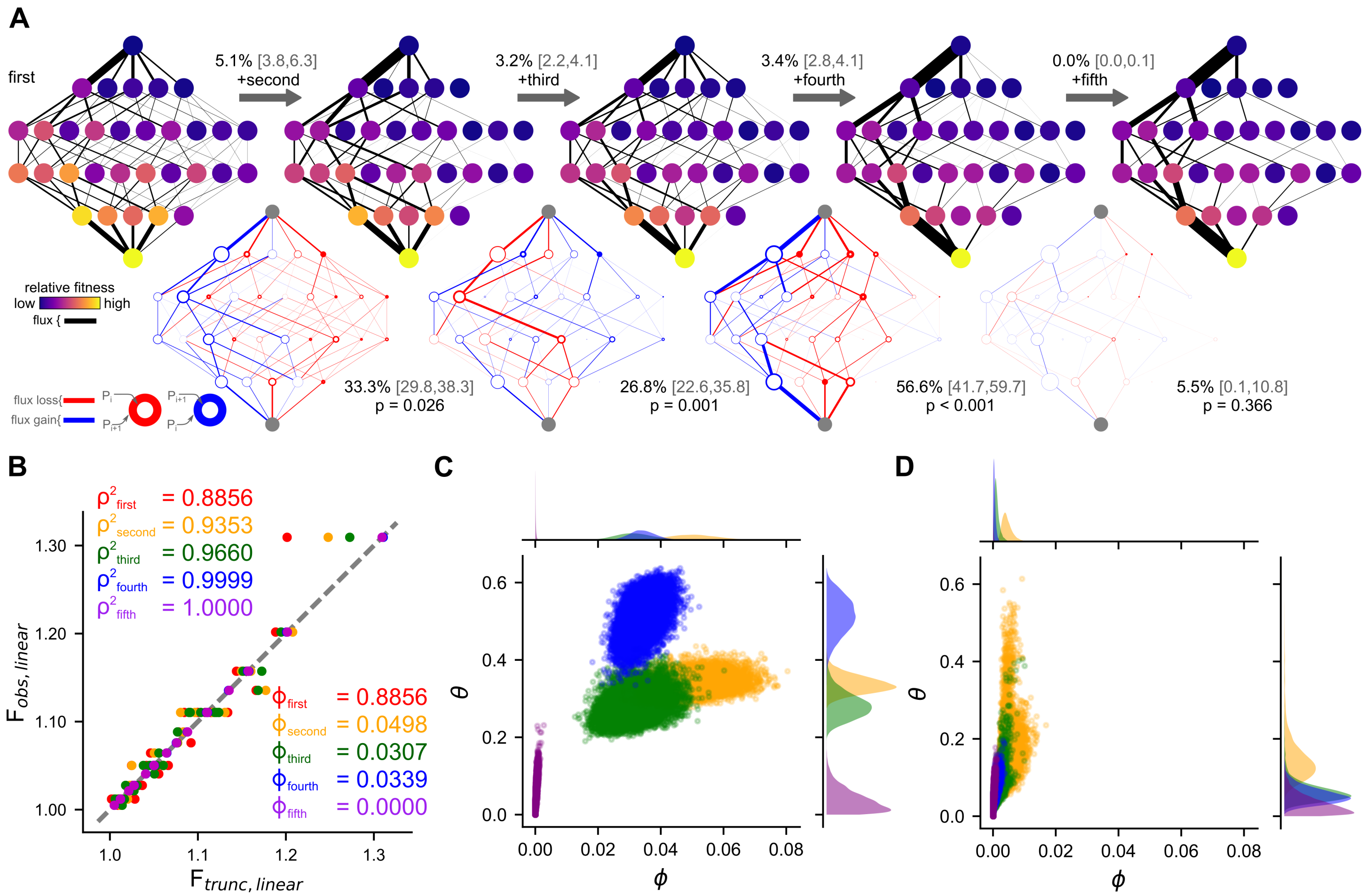


Figure 37: Epistasis alters trajectories in dataset II. A) Colors, panel layouts, and statistics are as in Fig 2. B) Colors, panel layouts, and statistics are as in Fig 1A. C-D): Colors and panel layouts are as in Fig 3.

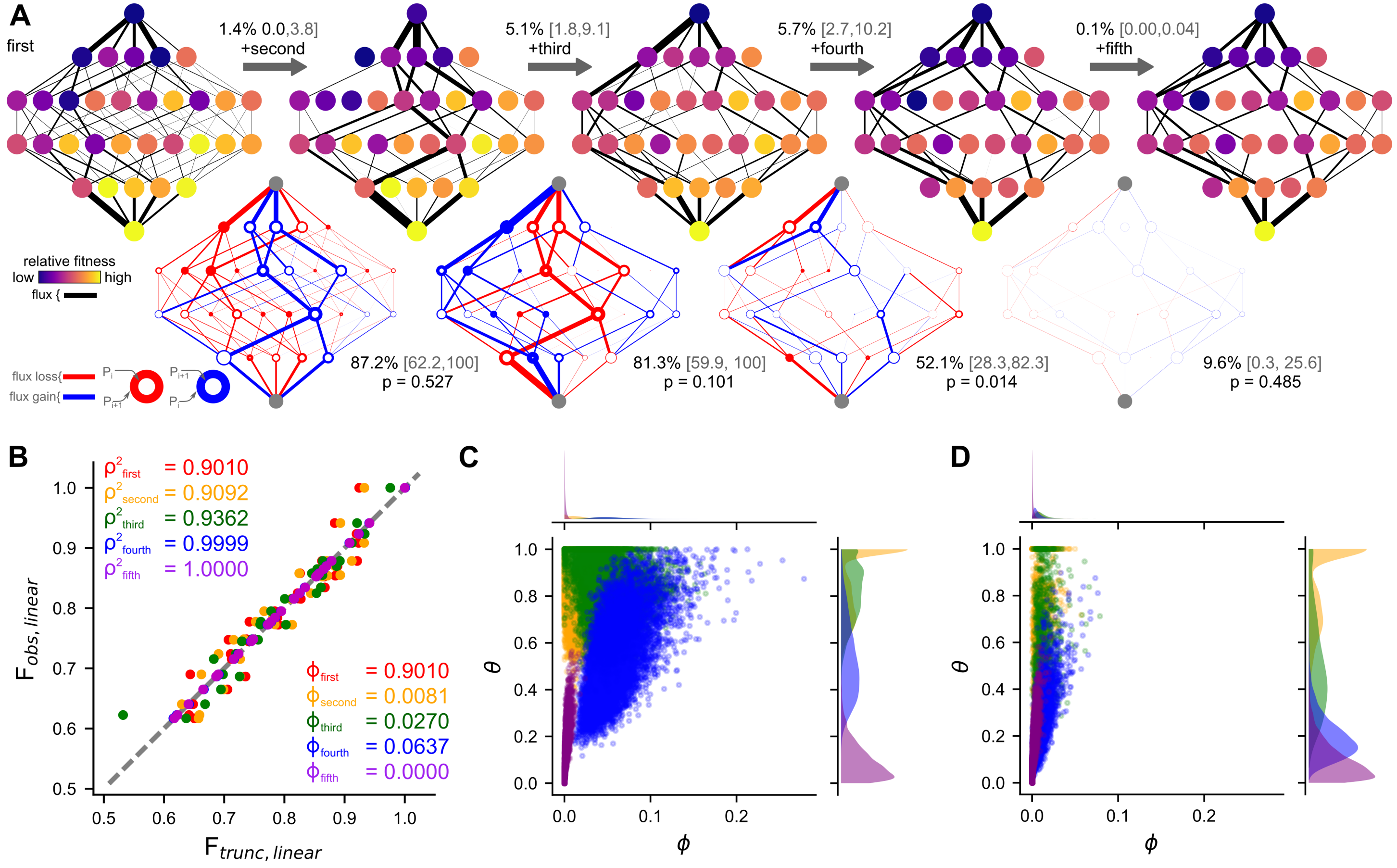


Figure 38: Epistasis alters trajectories in dataset III. A) Colors, panel layouts, and statistics are as in Fig 2. B) Colors, panel layouts, and statistics are as in Fig 1A. C-D): Colors and panel layouts are as in Fig 3.

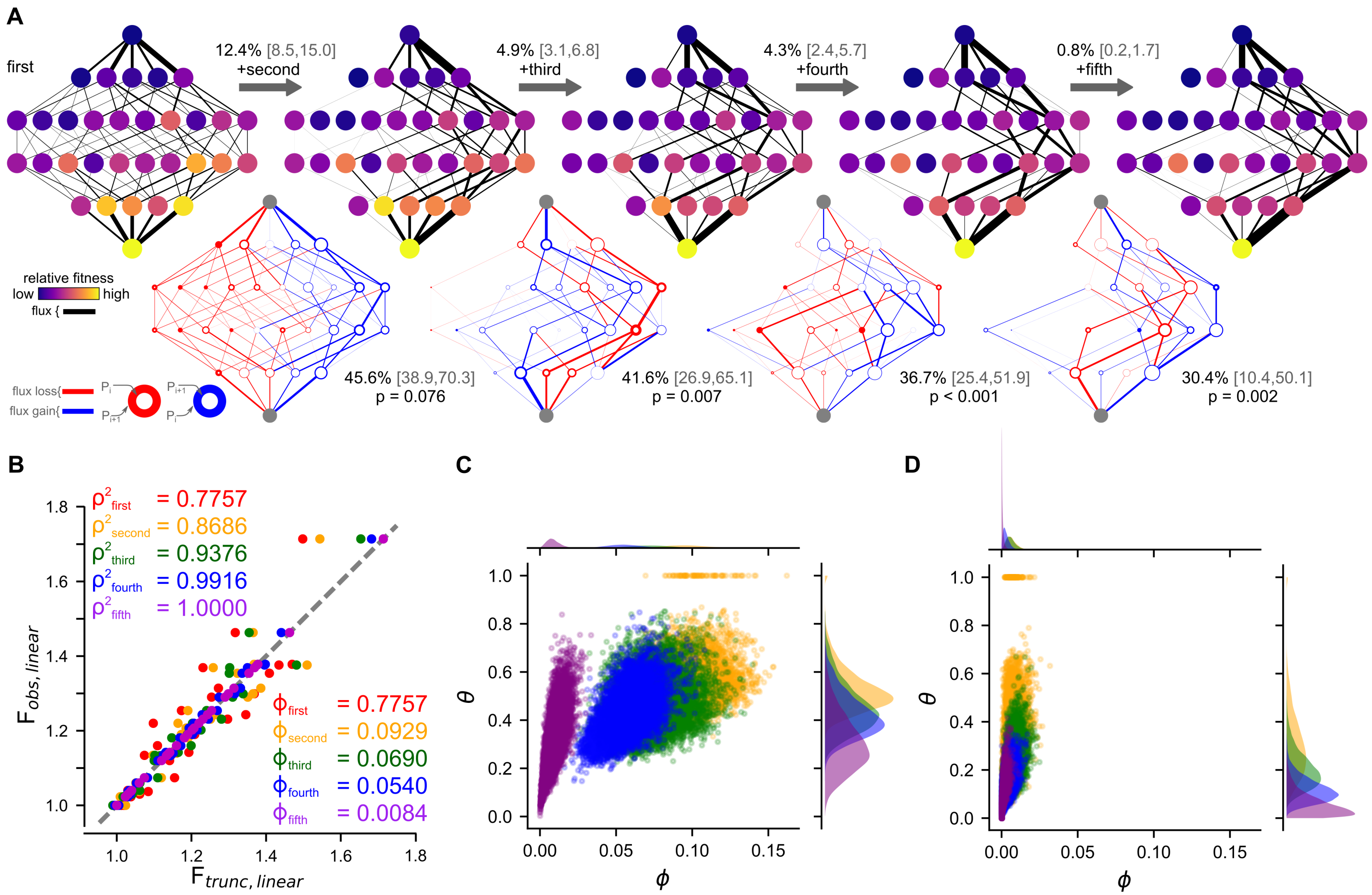


Figure 39: Epistasis alters trajectories in dataset IV. A) Colors, panel layouts, and statistics are as in Fig 2. B) Colors, panel layouts, and statistics are as in Fig 1A. C-D): Colors and panel layouts are as in Fig 3.

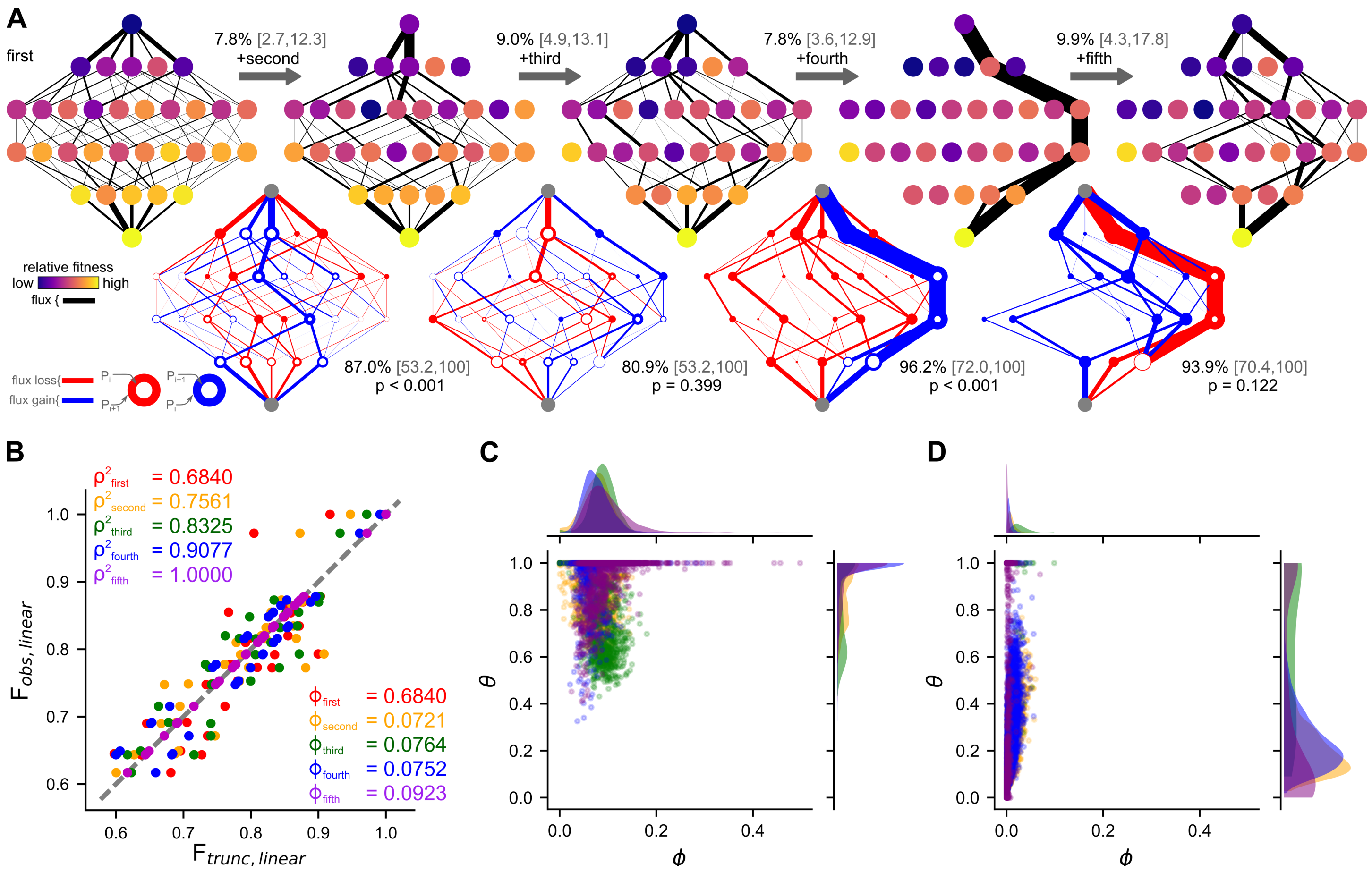


Figure 40: Epistasis alters trajectories in dataset V. A) Colors, panel layouts, and statistics are as in Fig 2. B) Colors, panel layouts, and statistics are as in Fig 1A. C-D): Colors and panel layouts are as in Fig 3.

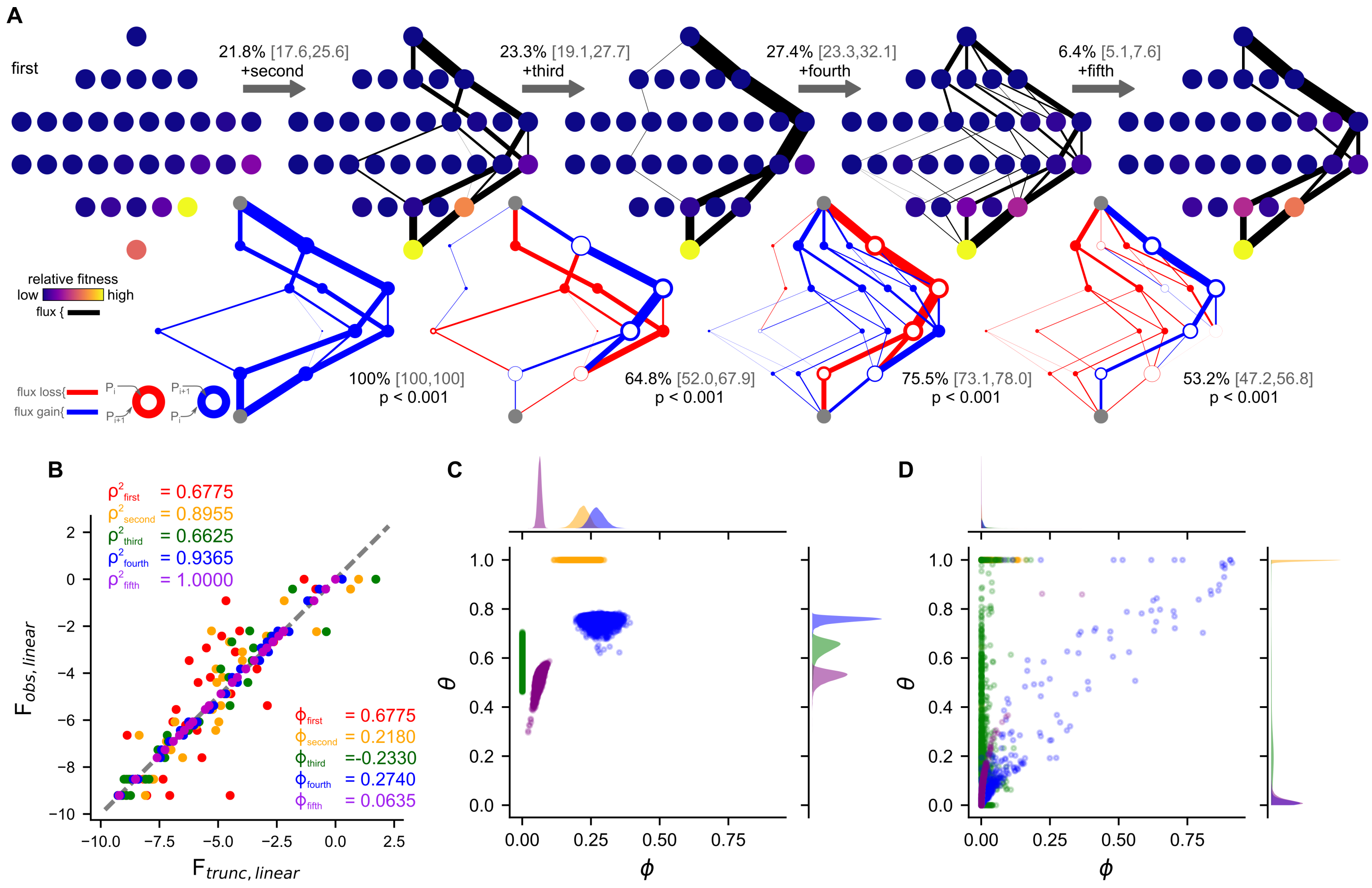


Figure 41: Epistasis alters trajectories in dataset VI. A) Colors, panel layouts, and statistics are as in Fig 2. B) Colors, panel layouts, and statistics are as in Fig 1A. C-D): Colors and panel layouts are as in Fig 3.

Effect of population size on predictability

We explored how population size changed the effect of ensemble-induced epistasis on evolutionary predictability. We repeated our evolutionary predictions using the variable population size variant of Gillespie's fixation model:

$$\pi_{x \to x+1} = \frac{1 - e^{-(w_{x+1}/w_x - 1)}}{1 - e^{-N_e \cdot (w_{x+1}/w_x - 1)}}, \tag{22}$$

where N_e is the effective population size [80]. Because the total number of accessible trajectories becomes very large without selection to winnow out low fitness trajectories, we limited our simulations to subsets of trajectories for this calculation. We selected these subsets by making only three mutational steps from each genetic background rather than all possible moves. For each genotype, we calculated $\Delta\Delta G^\circ$ for all possible mutations. We then randomly selected three of these mutations, weighted by their relative probability given our fitness function. We introduced each of these individually, creating a 3-way branching set of trajectories. We repeated this protocol for seven steps, yielding a final, discrete set of $3^7 = 2,187$ trajectories. We then attempted to predict the probabilities of trajectories within this sampled set, rather than completely open-ended evolutionary trajectories.

Fig S1A shows the overlap between the predicted and actual trajectory probabilities after seven steps for effective population sizes ranging from one to one million. At a high population size ($N_e = 1 \times 10^6$), we recapitulate our results from an infinite population model (Fig 3D). At the lowest population size ($N_e = 1$), we exactly predict the probabilities of all evolutionary trajectories. At first blush, this looks like evolution has become predictable, but this is not the case. At $N_e = 1$, drift dominates and all trajectories have equal probabilities. (This is the population genetics equivalent of infinite temperature). As a result, evolution is completely unpredictable. Our model can correctly predict that all trajectories are equally accessible because Eq. 22

reduces to one for all w: $\pi_{x\rightarrow x+1}$ no longer depends on the $\Delta\Delta G^{\circ}$ values we calculate. By $N_e = 10$, we already observe a small amount of divergence between our predicted and true trajectories.

We can more rigorously characterize the role of drift versus population size by calculating the Shannon entropy of all trajectories through the map:

$$S = -\sum_{i\in\{T\}} p_i \mathrm{log}(p_i), \tag{23}$$

where p_i represents the probability of an individual trajectory, summed over the set of all trajectories $\{T\}$. When all trajectories have equal probability ($N_e = 1$), the entropy is high. As specific trajectories begin to increase in probability, the entropy drops. This is shown in Fig S1B. For $N_e = 1$, the entropy is 7.69. For $N_e = 1 \times 10^6$, the average entropy is 4.55.

A comparison of Fig S1A and S1B reveals an inverse correlation between the effect of the ensemble on predictability and the effect of drift. As the effect of drift decreases, predictability due to the ensemble increases. Put another way: the more selection favors specific trajectories, the more ensemble-induced epistasis makes evolution unpredictable. The less selection favors specific trajectories, more drift make evolution unpredictable. There is no sweet spot in population size at which both drift and ensemble-induced epistasis have negligible effects on predictability.

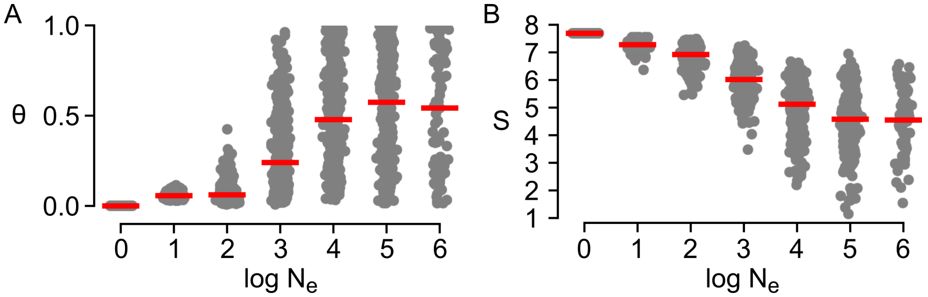


Figure 42: Unpredictability from the ensemble and neutral drift trade off with effective population size. Panel A shows a jitter plot of divergence between the true trajectories and trajectories predicted using a pairwise epistasis model using a full ensemble lattice model. Each series shows the divergence after seven steps using the effective population size shown. Panel B shows the Shannon entropy for the "true" trajectories from the simulations on the left as a function of population size. Gray points represent θ and S for individual maps. Red bars indicate the means.

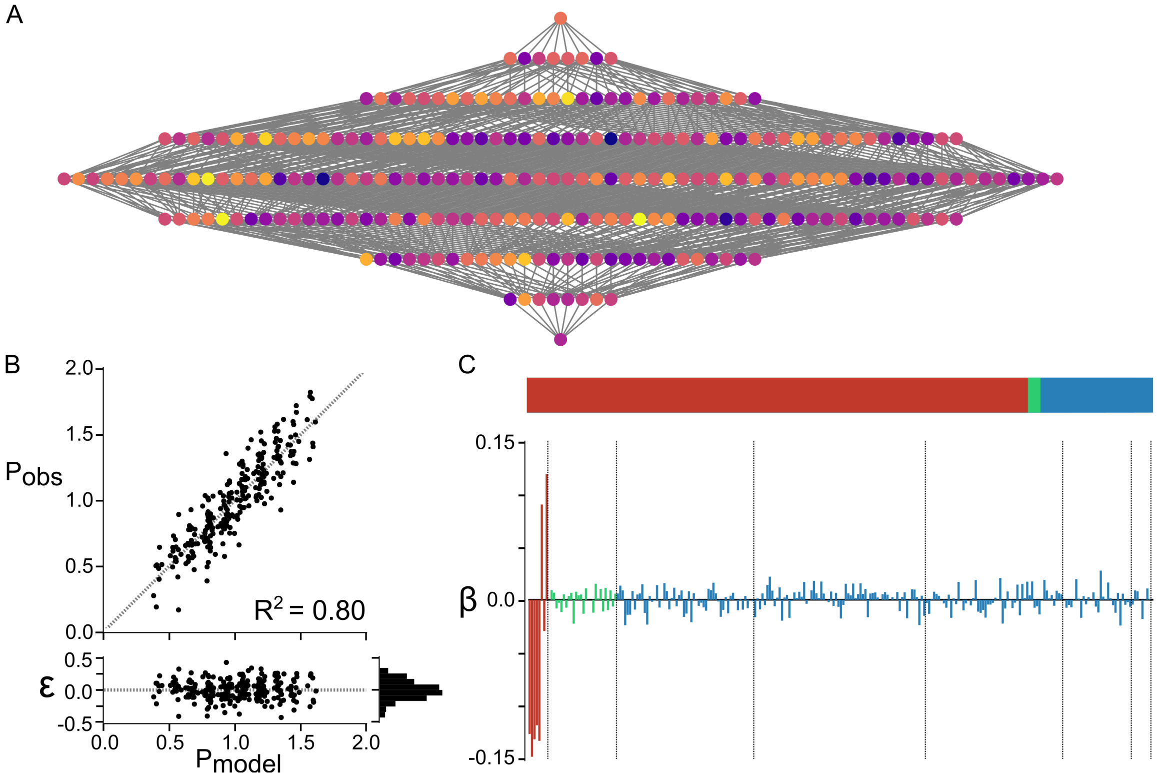


Figure 43: Epistasis in a simulated genotype-phenotype map. Panel A shows simulated genotype-phenotype map using Hadamard coefficients. Each node is a genotype; each edge is a single point mutation. Colors indicate quantitative phenotypes. Panel B shows the correlation between the observed phenotypes and the additive model, with fit residuals shown below the plot. Panel C shows the magnitude of epistasis in the map (top sub panel) and the values of all model coefficients (bottom sub panel). Colors indicate additive components (red), pairwise components (green), and high-order components (blue). In the bottom sub panel, each bar shows the value of a single model coefficient: the red bars correspond to the 8 additive coefficients, the green bars to the 28 pairwise coefficients, and the blue bars to the 220 high-order coefficients.

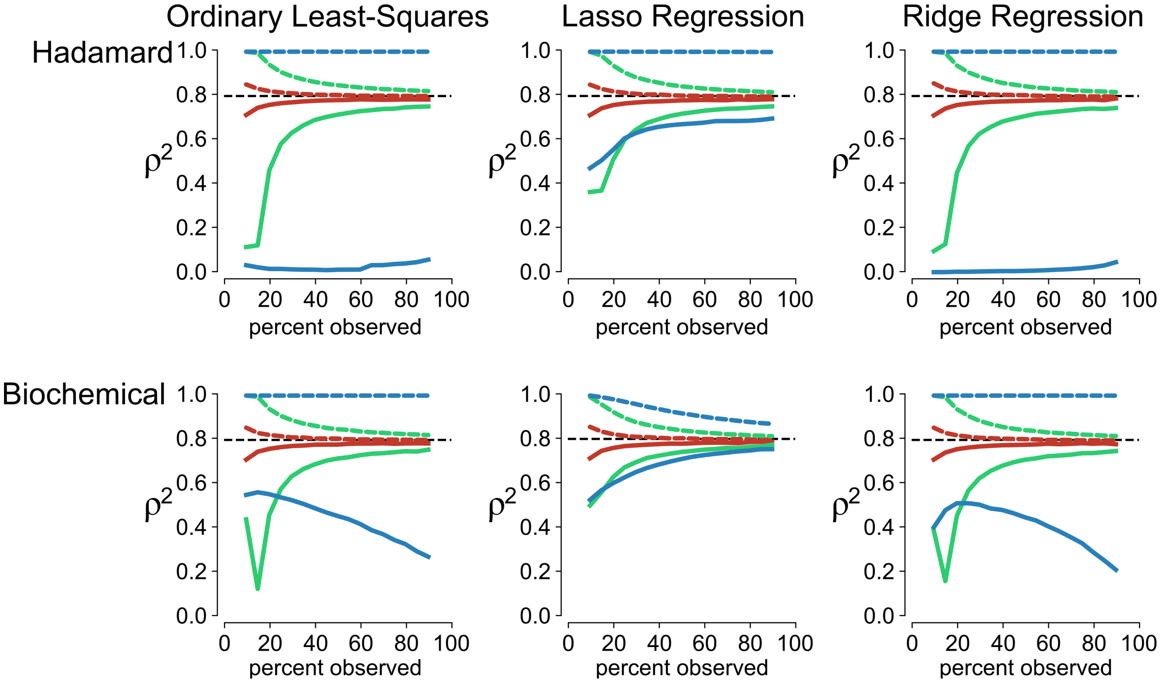


Figure 44: Linear epistatic coefficients cannot be estimated from an incomplete, simulated genotype-phenotype map. Panels show the fitting and prediction performaces of different epistasis models (Hadamard or biochemical) and regression methods (ordinary least-squares, lasso, or ridge). Each subpanel shows the fit scores versus the fraction of data used to train the model, from 10% to 90%, for different epistasis models. The dashed gray line indicates the amount of additive variation in the map (80%). Colors indicate model: additive (red), pairwise epistasis (green), and high-order epistasis (blue). Dashed lines indicate ρ^2_{train} and solid lines indicate ρ^2_{test}.

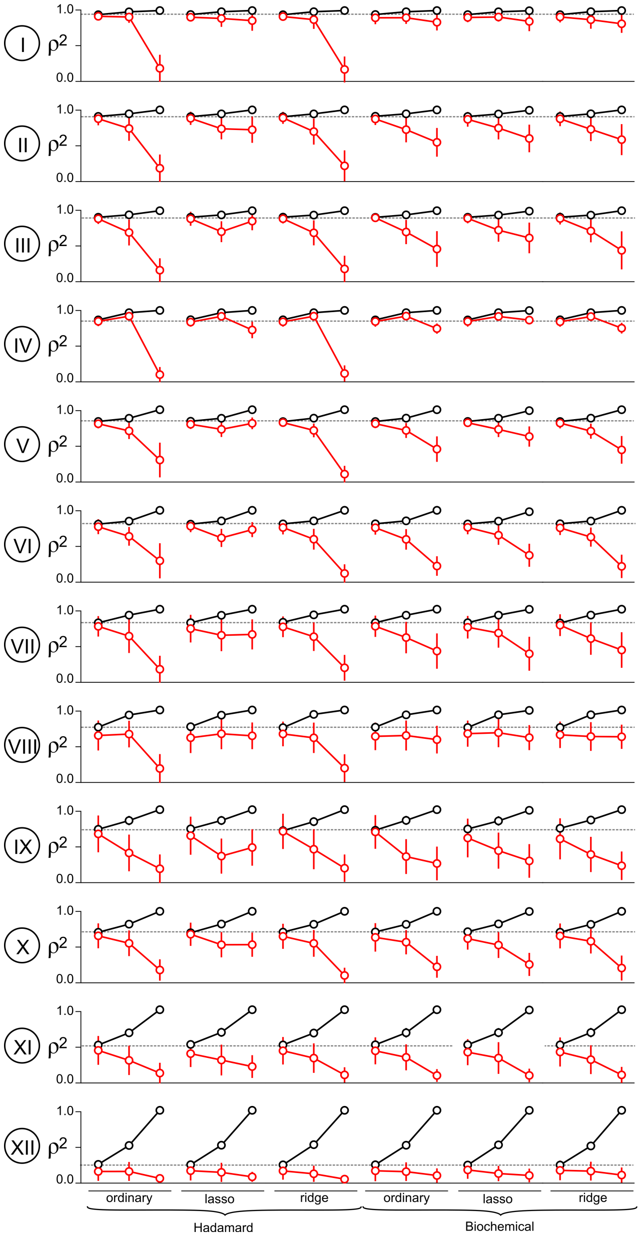


Figure 45: Predictive epistatic coefficients cannot be extracted from experimental genotype-phenotype maps. Each row shows ρ^2_{train} (black) and ρ^2_{test} (red) for a different experimental map (indicated to the left of each row) as epistatic orders are added to the model. The x-axis shows the results for various epistasis models (Hadamard or biochemical) and regression methods (ordinary least-squares, lasso, or ridge). Points are, from left to right: additive, pairwise, and high-order epistasis. Points and lines indicate the mean of 1,000 pseudoreplicate samples. Error bars are standard deviation of pseudoreplicate results. The dashed lines indicate the fraction of the variation in the map explained by the additive model.

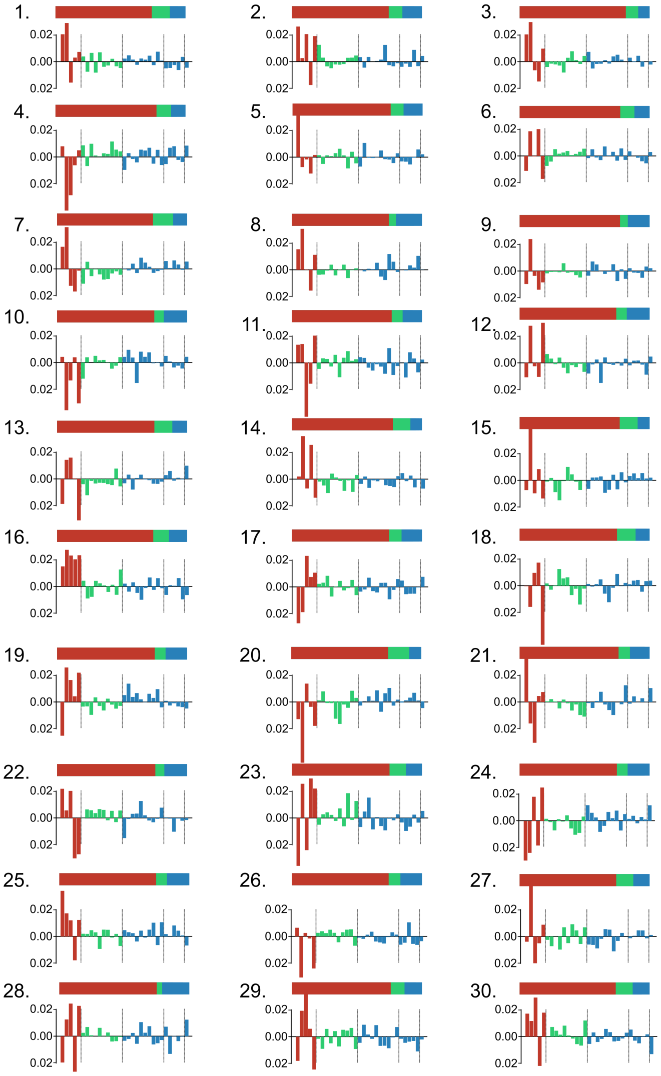


Figure 46: Epistasis in experimental maps looks like random epistasis. Panels show the epistatic coefficients extracted from 29 decoy, simulated genotype-phenotype maps and 1 experimental genotype-phenotype map (dataset VIII). All maps have a fraction of epistasis equal to 26%. Epistasis was added to the 29 decoy maps by drawing a random epistasis parameter from a normal distribution for each phenotype. Panel 20 shows the results for the experimental map. In each panel, the top subpanel indicates the magnitude of epistasis in each map and the bottom subpanel indicates the values of all model coefficients. Colors indicate additive components (red), pairwise components (green), and high-order components (blue). In the bottom sub panel, each bar shows the value of a single model coefficient: the red bars correspond to the 5 additive coefficients, the green bars to the 10 pairwise coefficients, and the blue bars to the 17 high-order coefficients.

REFERENCES CITED

[1] Douglas M. Fowler, Carlos L. Araya, Sarel J. Fleishman, Elizabeth H. Kellogg, Jason J. Stephany, David Baker, and Stanley Fields. High-Resolution Mapping of Protein Sequence-Function Relationships. *Nature Methods*, 7(9):741–746, September 2010.

[2] Mark A. DePristo, Daniel M. Weinreich, and Daniel L. Hartl. Missense Meanderings in Sequence Space: A Biophysical View of Protein Evolution. *Nat Rev Genet*, 6(9):678–687, September 2005.

[3] Erich Bornberg-Bauer and Hue Sun Chan. Modeling Evolutionary Landscapes: Mutational Stability, Topology, and Superfunnels in Sequence Space. *PNAS*, 96(19):10689–10694, September 1999.

[4] John Maynard Smith. Natural Selection and the Concept of a Protein Space. *Nature*, 225(5232):563–564, February 1970.

[5] C. Brandon Ogbunugafor and Daniel L. Hartl. A New Take on John Maynard Smith's Concept of Protein Space for Understanding Molecular Evolution. *PLOS Computational Biology*, 12(10):e1005046, October 2016.

[6] J. Arjan G. M. de Visser, Su-Chan Park, and Joachim Krug. Exploring the Effect of Sex on Empirical Fitness Landscapes. *The American Naturalist*, 174(s1):S15–S30, July 2009.

[7] Ivan G. Szendro, Martijn F. Schenk, Jasper Franke, Joachim Krug, and J. Arjan G. M. de Visser. Quantitative analyses of empirical fitness landscapes. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(01):P01005, 2013.

[8] Frank J. Poelwijk, Daniel J. Kiviet, Daniel M. Weinreich, and Sander J. Tans. Empirical Fitness Landscapes Reveal Accessible Evolutionary Paths. *Nature*, 445(7126):383–386, January 2007.

[9] J. Arjan G. M. de Visser and Joachim Krug. Empirical Fitness Landscapes and the Predictability of Evolution. *Nat Rev Genet*, 15(7):480–490, July 2014.

[10] Lizhi Ian Gong, Marc A. Suchard, and Jesse D. Bloom. Stability-Mediated Epistasis Constrains the Evolution of an Influenza Protein. *eLife*, 2:e00631, May 2013.

[11] Robert L. Summers, Anurag Dave, Tegan J. Dolstra, Sebastiano Bellanca, Rosa V. Marchetti, Megan N. Nash, Sashika N. Richards, Valerie Goh, Robyn L. Schenk, Wilfred D. Stein, Kiaran Kirk, Cecilia P. Sanchez, Michael Lanzer, and Rowena E. Martin. Diverse mutational pathways converge on saturable chloroquine transport via the malaria parasite's chloroquine resistance transporter. *Proceedings of the National Academy of Sciences*, 111(17):E1759–E1767, April 2014.

[12] Rhys Hayward, Kevin J. Saliba, and Kiaran Kirk. The pH of the digestive vacuole of Plasmodium falciparum is not associated with chloroquine resistance. *Journal of Cell Science*, 119(6):1016–1025, March 2006.

[13] David A. Fidock, Takashi Nomura, Angela K. Talley, Roland A. Cooper, Sergey M. Dzekunov, Michael T. Ferdig, Lyann M. B. Ursos, Amar bir Singh Sidhu, Bronwen Naudé, Kirk W. Deitsch, Xin-zhuan Su, John C. Wootton, Paul D. Roepe, and Thomas E. Wellems. Mutations in the P. falciparum Digestive Vacuole Transmembrane Protein PfCRT and Evidence for Their Role in Chloroquine Resistance. *Molecular Cell*, 6(4):861–871, October 2000.

[14] R.A. Fisher. The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Philosophical Transactions of the Royal Society of Edinburgh*, (52):399–433, 1918.

[15] Heather J. Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468, October 2002.

[16] Patrick C. Phillips. Epistasis — the Essential Role of Gene Interactions in the Structure and Evolution of Genetic Systems. *Nat Rev Genet*, 9(11):855–867, November 2008.

[17] Hsin-Hung Chou, Hsuan-Chao Chiu, Nigel F. Delaney, Daniel Segrè, and Christopher J. Marx. Diminishing Returns Epistasis Among Beneficial Mutations Decelerates Adaptation. *Science*, 332(6034):1190–1192, March 2011.

[18] Aisha I. Khan, Duy M. Dinh, Dominique Schneider, Richard E. Lenski, and Tim F. Cooper. Negative Epistasis Between Beneficial Mutations in an Evolving Bacterial Population. *Science*, 332(6034):1193–1196, March 2011.

[19] Sergey Kryazhimskiy, Daniel P. Rice, Elizabeth R. Jerison, and Michael M. Desai. Global Epistasis Makes Adaptation Predictable despite Sequence-Level Stochasticity. *Science*, 344(6191):1519–1522, June 2014.

[20] Karen S. Sarkisyan, Dmitry A. Bolotin, Margarita V. Meer, Dinara R. Usmanova, Alexander S. Mishin, George V. Sharonov, Dmitry N. Ivankov, Nina G. Bozhanova, Mikhail S. Baranov, Onuralp Soylemez, Natalya S. Bogatyreva, Peter K. Vlasov, Evgeny S. Egorov, Maria D. Logacheva, Alexey S. Kondrashov, Dmitry M. Chudakov, Ekaterina V. Putintseva, Ilgar Z. Mamedov, Dan S. Tawfik, Konstantin A. Lukyanov, and Fyodor A. Kondrashov. Local Fitness Landscape of the Green Fluorescent Protein. *Nature*, 533(7603):397–401, May 2016.

[21] Jakub Otwinowski, David M. McCandlish, and Joshua B. Plotkin. Inferring the shape of global epistasis. *Proceedings of the National Academy of Sciences*, page 201804015, July 2018.

[22] Zachary R. Sailer and Michael J. Harms. Detecting High-Order Epistasis in Nonlinear Genotype-Phenotype Maps. *Genetics*, 205(3):1079–1088, March 2017.

[23] Zachary R. Sailer and Michael J. Harms. Molecular ensembles make evolution unpredictable. *Proceedings of the National Academy of Sciences*, 114(45):11938–11943, November 2017.

[24] Daniel M. Weinreich, Richard A. Watson, and Lin Chao. Perspective: Sign Epistasis and Genetic Costraint on Evolutionary Trajectories. *Evolution*, 59(6):1165–1174, June 2005.

[25] Daniel M. Weinreich, Nigel F. Delaney, Mark A. DePristo, and Daniel L. Hartl. Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science*, 312(5770):111–114, July 2006.

[26] Shimon Bershtein, Michal Segal, Roy Bekerman, Nobuhiko Tokuriki, and Dan S. Tawfik. Robustness–epistasis Link Shapes the Fitness Landscape of a Randomly Drifting Protein. *Nature*, 444(7121):929–932, December 2006.

[27] Jack da Silva, Mia Coetzer, Rebecca Nedellec, Cristina Pastore, and Donald E. Mosier. Fitness Epistasis and Constraints on Adaptation in a Human Immunodeficiency Virus Type 1 Protein Region. *Genetics*, 185(1):293–303, May 2010.

[28] Sergey Kryazhimskiy, Jonathan Dushoff, Georgii A. Bazykin, and Joshua B. Plotkin. Prevalence of Epistasis in the Evolution of Influenza A Surface Proteins. *PLoS Genet*, 7(2):e1001301, February 2011.

[29] Merijn L. M. Salverda, Eynat Dellus, Florien A. Gorter, Alfons J. M. Debets, John van der Oost, Rolf F. Hoekstra, Dan S. Tawfik, and J. Arjan G. M. de Visser. Initial Mutations Direct Alternative Pathways of Protein Evolution. *PLOS Genet*, 7(3):e1001321, March 2011.

[30] Michael S. Breen, Carsten Kemena, Peter K. Vlasov, Cedric Notredame, and Fyodor A. Kondrashov. Epistasis as the Primary Factor in Molecular Evolution. *Nature*, 490(7421):535–538, October 2012.

[31] Danielle M. Tufts, Chandrasekhar Natarajan, Inge G. Revsbech, Joana Projecto-Garcia, Federico G. Hoffmann, Roy E. Weber, Angela Fago, Hideaki Moriyama, and Jay F. Storz. Epistasis Constrains Mutational Pathways of Hemoglobin Adaptation in High-Altitude Pikas. *Mol Biol Evol*, page msu311, November 2014.

[32] Dave W. Anderson, Alesia N. McKeown, and Joseph W. Thornton. Intermolecular Epistasis Shaped the Function and Evolution of an Ancient Transcription Factor and Its DNA Binding Sites. *eLife Sciences*, page e07864, June 2015.

[33] Premal Shah, David M. McCandlish, and Joshua B. Plotkin. Contingency and Entrenchment in Protein Evolution under Purifying Selection. *PNAS*, page 201412933, June 2015.

[34] Charlotte M. Miton and Nobuhiko Tokuriki. How Mutational Epistasis Impairs Predictability in Protein Evolution and Design. *Protein Science*, 25(7):1260–1272, July 2016.

[35] Zachary R. Sailer and Michael J. Harms. High-order epistasis shapes evolutionary trajectories. *PLOS Computational Biology*, 13(5):e1005541, May 2017.

[36] Michael J. Harms and Joseph W. Thornton. Historical Contingency and Its Biophysical Basis in Glucocorticoid Receptor Evolution. *Nature*, 512(7513):203–207, August 2014.

[37] Nicholas C. Wu, Lei Dai, C. Anders Olson, James O. Lloyd-Smith, and Ren Sun. Adaptation in Protein Fitness Landscapes Is Facilitated by Indirect Paths. *eLife*, 5:e16965, July 2016.

[38] Jamie T. Bridgham, Eric A. Ortlund, and Joseph W. Thornton. An Epistatic Ratchet Constrains the Direction of Glucocorticoid Receptor Evolution. *Nature*, 461(7263):515–519, September 2009.

[39] Michael J. Harms and Joseph W. Thornton. Evolutionary Biochemistry: Revealing the Historical and Physical Causes of Protein Properties. *Nature reviews. Genetics*, 14(8):559–571, August 2013.

[40] Marylyn D. Ritchie, Lance W. Hahn, Nady Roodi, L. Renee Bailey, William D. Dupont, Fritz F. Parl, and Jason H. Moore. Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer. *The American Journal of Human Genetics*, 69(1):138–147, July 2001.

[41] Shozo Yokoyama, Ahmet Altun, Huiyong Jia, Hui Yang, Takashi Koyama, Davide Faggionato, Yang Liu, and William T. Starmer. Adaptive Evolutionary Paths from UV Reception to Sensing Violet Light by Epistatic Interactions. *Science Advances*, 1(8):e1500162, September 2015.

[42] Jiya Sun, Fuhai Song, Jiajia Wang, Guangchun Han, Zhouxian Bai, Bin Xie, Xuemei Feng, Jianping Jia, Yong Duan, and Hongxing Lei. Hidden Risk Genes with High-Order Intragenic Epistasis in Alzheimer's Disease. *J. Alzheimers Dis.*, 41(4):1039–1056, 2014.

[43] Ting Hu, Yuanzhu Chen, Jeff W. Kiralis, Ryan L. Collins, Christian Wejse, Giorgio Sirugo, Scott M. Williams, and Jason H. Moore. An Information-Gain Approach to Detecting Three-Way Epistatic Interactions in Genetic Association Studies. *J Am Med Inform Assoc*, pages amiajnl–2012–001525, February 2013.

[44] Daniel M Weinreich, Yinghong Lan, C Scott Wylie, and Robert B. Heckendorn. Should Evolutionary Geneticists Worry about Higher-Order Epistasis? *Current Opinion in Genetics & Development*, 23(6):700–707, December 2013.

[45] Yinhua Wang, Carolina Díaz Arenas, Daniel M. Stoebel, and Tim F. Cooper. Genetic Background Affects Epistatic Interactions between Two Beneficial Mutations. *Biology Letters*, page rsbl20120328, August 2012.

[46] Mats Pettersson, Francois Besnier, Paul B. Siegel, and Örjan Carlborg. Replication and Explorations of High-Order Epistasis Using a Large Advanced Intercross Line Pedigree. *PLoS Genet*, 7(7):e1002180, July 2011.

[47] Tomoaki Matsuura, Yasuaki Kazuta, Takuyo Aita, Jiro Adachi, and Tetsuya Yomo. Quantifying Epistatic Interactions among the Components Constituting the Protein Translation System. *Molecular Systems Biology*, 5(1):297, January 2009.

[48] Marcin Imielinski and Calin Belta. Exploiting the Pathway Structure of Metabolism to Reveal High-Order Epistasis. *BMC Systems Biology*, 2:40, 2008.

[49] Chia-Ti Tsai, Juey-Jen Hwang, Marylyn D. Ritchie, Jason H. Moore, Fu-Tien Chiang, Ling-Ping Lai, Kuan-Lih Hsu, Chuen-Den Tseng, Jiunn-Lee Lin, and Yung-Zu Tseng. Renin–angiotensin System Gene Polymorphisms and Coronary Artery Disease in a Large Angiographic Cohort: Detection of High Order Gene–gene Interaction. *Atherosclerosis*, 195(1):172–180, November 2007.

[50] Jianfeng Xu, James Lowey, Fredrik Wiklund, Jielin Sun, Fredrik Lindmark, Fang-Chi Hsu, Latchezar Dimitrov, Baoli Chang, Aubrey R. Turner, Wennan Liu, Hans-Olov Adami, Edward Suh, Jason H. Moore, S. Lilly Zheng, William B. Isaacs, Jeffrey M. Trent, and Henrik Grönberg. The Interaction of Four Genes in the Inflammation Pathway Significantly Predicts Prostate Cancer Risk. *Cancer Epidemiol Biomarkers Prev*, 14(11):2563–2568, January 2005.

[51] Daniel Segrè, Alexander DeLuna, George M. Church, and Roy Kishony. Modular Epistasis in Yeast Metabolism. *Nat Genet*, 37(1):77–83, January 2005.

[52] Frank J. Poelwijk, Vinod Krishna, and Rama Ranganathan. The Context-Dependence of Mutations: A Linkage of Formalisms. *PLOS Computational Biology*, 12(6):e1004771, June 2016.

[53] Kenneth J. Rothman, Sander Greenland, and Alexander M. Walker. Concepts of Interaction. *Am. J. Epidemiol.*, 112(4):467–470, January 1980.

[54] Wayne N. Frankel and Nicholas J. Schork. Who's Afraid of Epistasis? *Nat Genet*, 14(4):371–373, December 1996.

[55] Darin R. Rokyta, Paul Joyce, S. Brian Caudle, Craig Miller, Craig J. Beisel, and Holly A. Wichman. Epistasis between Beneficial Mutations and the Phenotype-to-Fitness Map for a ssDNA Virus. *PLOS Genet*, 7(6):e1002075, June 2011.

[56] Martijn F. Schenk, Ivan G. Szendro, Merijn L. M. Salverda, Joachim Krug, and J. Arjan G. M. de Visser. Patterns of Epistasis between Beneficial Mutations in an Antibiotic Resistance Gene. *Mol Biol Evol*, 30(8):1779–1787, January 2013.

[57] François Blanquart. Properties of Selected Mutations and Genotypic Landscapes under Fisher's Geometric Model. *Evolution*, 68(12):3537–54, 2014.

[58] Robert B. Heckendorn and Darrell Whitley. Predicting Epistasis from Mathematical Models. *Evolutionary Computation*, 7(1):69–101, March 1999.

[59] David W. Hall, Matthew Agan, and Sara C. Pope. Fitness Epistasis among 6 Biosynthetic Loci in the Budding Yeast Saccharomyces Cerevisiae. *J Hered*, 101(suppl 1):S75–S84, January 2010.

[60] G. E. P. Box and D. R. Cox. An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 1964.

[61] R. J. Carroll and David Ruppert. On Prediction and the Power Transformation Family. *Biometrika*, 68(3):609–615, January 1981.

[62] Johannes Neidhart, Ivan G. Szendro, and Joachim Krug. Exact Results for Amplitude Spectra of Fitness Landscapes. *Journal of Theoretical Biology*, 332:218–227, September 2013.

[63] Chun-Ting Zhang and Ren Zhang. Analysis of Distribution of Bases in the Coding Sequences by a Digrammatic Technique. *Nucl. Acids Res.*, 19(22):6313–6317, November 1991.

[64] Herve Abdi. The Bonferonni and Sidak Corrections for Multiple Comparisons. *Encyclopedia of measurement and statistics*, 3:103–107, 2007.

[65] S. van der Walt, S. C. Colbert, and G. Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering*, 13(2):22–30, March 2011.

[66] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[67] J. D. Hunter. Matplotlib: A 2d Graphics Environment. *Computing in Science Engineering*, 9(3):90–95, May 2007.

[68] F. Perez and B. E. Granger. IPython: A System for Interactive Scientific Computing. *Computing in Science Engineering*, 9(3):21–29, May 2007.

[69] R. C. MacLean, G. G. Perron, and A. Gardner. Diminishing Returns From Beneficial Mutations and Pervasive Epistasis Shape the Fitness Landscape for Rifampicin Resistance in Pseudomonas Aeruginosa. *Genetics*, 186(4):1345–1354, December 2010.

[70] Nobuhiko Tokuriki, Colin J. Jackson, Livnat Afriat-Jurnou, Kirsten T. Wyganowski, Renmei Tang, and Dan S. Tawfik. Diminishing Returns and Tradeoffs Constrain the Laboratory Optimization of an Enzyme. *Nat Commun*, 3:1257, December 2012.

[71] Sarah Perin Otto and Marcus W. Feldman. Deleterious Mutations, Variable Epistatic Interactions, and the Evolution of Recombination. *Theoretical Population Biology*, 51(2):134–147, April 1997.

[72] Alesia N. McKeown, Jamie T. Bridgham, Dave W. Anderson, Michael N. Murphy, Eric A. Ortlund, and Joseph W. Thornton. Evolution of DNA Specificity in a Transcription Factor Family Produced a New Gene Regulatory Module. *Cell*, 159(1):58–68, September 2014.

[73] Jakub Otwinowski and Joshua B. Plotkin. Inferring Fitness Landscapes by Regression Produces Biased Estimates of Epistasis. *PNAS*, 111(22):E2301–E2309, March 2014.

[74] Joseph Lehár, Andrew Krueger, Grant Zimmermann, and Alexis Borisy. High-order Combination Effects and Biological Robustness. *Molecular Systems Biology*, 4(1):215, January 2008.

[75] Ting Hu, Nicholas A. Sinnott-Armstrong, Jeff W. Kiralis, Angeline S. Andrew, Margaret R. Karagas, and Jason H. Moore. Characterizing Genetic Interactions in Human Disease Association Studies Using Statistical Epistasis Networks. *BMC Bioinformatics*, 12:364, 2011.

[76] Matthew B. Taylor and Ian M. Ehrenreich. Higher-Order Genetic Interactions and their Contribution to Complex Traits. *Trends in Genetics*, 31(1):34–40, January 2015.

[77] Mark A Bedau and Norman H Packard. Evolution of Evolvability via Adaptation of Mutation Rates. *Biosystems*, 69(2–3):143–162, May 2003.

[78] Michael M. Desai. Reverse Evolution and Evolutionary Memory. *Nat Genet*, 41(2):142–143, February 2009.

[79] Adam C. Palmer, Erdal Toprak, Michael Baym, Seungsoo Kim, Adrian Veres, Shimon Bershtein, and Roy Kishony. Delayed Commitment to Evolutionary Fate in Antibiotic Resistance Fitness Landscapes. *Nature Communications*, 6:7385, June 2015.

[80] John H. Gillespie. Molecular Evolution Over the Mutational Landscape. *Evolution*, 38(5):1116–1129, 1984.

[81] H. Allen Orr. The Population Genetics of Adaptation: The Adaptation of Dna Sequences. *Evolution*, 56(7):1317–1330, July 2002.

[82] Guy Sella and Aaron E. Hirsh. The Application of Statistical Physics to Evolutionary Biology. *PNAS*, 102(27):9541–9546, May 2005.

[83] Hirotogu Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In Emanuel Parzen, Kunio Tanabe, and Genshiro Kitagawa, editors, *Selected Papers of Hirotugu Akaike*, Springer Series in Statistics, pages 199–213. Springer New York, 1998.

[84] John H. Gillespie. *Population Genetics: A Concise Guide*. JHU Press, December 2010.

[85] Kenneth M. Flynn, Tim F. Cooper, Francisco B.-G. Moore, and Vaughn S. Cooper. The Environment Affects Epistatic Interactions to Alter the Topology of an Empirical Fitness Landscape. *PLOS Genet*, 9(4):e1003426, April 2013.

[86] Zachary D. Blount, Christina Z. Borland, and Richard E. Lenski. Historical Contingency and the Evolution of a Key Innovation in an Experimental Population of Escherichia Coli. *PNAS*, 105(23):7899–7906, October 2008.

[87] James F. Crow. On Epistasis: Why It Is Unimportant in Polygenic Directional Selection. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365(1544):1241–1244, April 2010.

[88] Mark Lunzer, G. Brian Golding, and Antony M. Dean. Pervasive Cryptic Epistasis in Molecular Evolution. *PLoS Genet*, 6(10):e1001162, October 2010.

[89] Jacques Monod. On Chance and Necessity. In Francisco Jose Ayala and Theodosius Dobzhansky, editors, *Studies in the Philosophy of Biology*, pages 357–375. Macmillan Education UK, 1974.

[90] Stephen Jay Gould. *Wonderful Life: The Burgess Shale and the Nature of Life*. New York: Norton, 1989.

[91] Simon Conway Morris. Evolution: Like Any Other Science It Is Predictable. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1537):133–145, January 2010.

[92] Michael Lässig, Ville Mustonen, and Aleksandra M. Walczak. Predicting Evolution. *Nature Ecology & Evolution*, 1:0077, February 2017.

[93] Ryan T. Hietpas, Jeffrey D. Jensen, and Daniel N. A. Bolon. Experimental Illumination of a Fitness Landscape. *Proceedings of the National Academy of Sciences*, 108(19):7896–7901, October 2011.

[94] M. Michael Gromiha, Motohisa Oobatake, and Akinori Sarai. Important Amino Acid Properties for Enhanced Thermostability from Mesophilic to Thermophilic Proteins. *Biophysical Chemistry*, 82(1):51–67, November 1999.

[95] Darin M. Taverna and Richard A. Goldstein. Why Are Proteins Marginally Stable? *Proteins: Structure, Function, and Bioinformatics*, 46(1):105–109, January 2002.

[96] Rafael Couñago, Stephen Chen, and Yousif Shamoo. In Vivo Molecular Evolution Reveals Biophysical Origins of Organismal Fitness. *Molecular Cell*, 22(4):441–449, May 2006.

[97] Tobias Sikosek and Hue Sun Chan. Biophysics of Protein Evolution and Evolutionary Protein Biophysics. *Journal of The Royal Society Interface*, 11(100):20140419, November 2014.

[98] Jesse D. Bloom, Claus O. Wilke, Frances H. Arnold, and Christoph Adami. Stability and the Evolvability of Function in a Model Protein. *Biophysical Journal*, 86(5):2758–2764, May 2004.

[99] Sanzo Miyazawa and Robert L. Jernigan. Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation. *Macromolecules*, 18(3):534–552, March 1985.

[100] Kit Fun Lau and Ken A. Dill. A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins. *Macromolecules*, 22(10):3986–3997, 1989.

[101] Leonid A. Mirny, Victor I. Abkevich, and Eugene I. Shakhnovich. How Evolution Makes Proteins Fold Quickly. *Proceedings of the National Academy of Sciences*, 95(9):4976–4981, April 1998.

[102] Gregorio Weber. Energetics of Ligand Binding to Proteins. In John T. Edsall C.B. Anfinsen and Frederic M. Richards, editors, *Advances in Protein Chemistry*, volume 29, pages 1–83. Academic Press, 1975.

[103] Elan Z. Eisenmesser, Oscar Millet, Wladimir Labeikovsky, Dmitry M. Korzhnev, Magnus Wolf-Watz, Daryl A. Bosco, Jack J. Skalicky, Lewis E. Kay, and Dorothee Kern. Intrinsic Dynamics of an Enzyme Underlies Catalysis. *Nature*, 438(7064):117–121, November 2005.

[104] Hesam N. Motlagh, James O. Wrabl, Jing Li, and Vincent J. Hilser. The Ensemble Nature of Allostery. *Nature*, 508(7496):331–339, April 2014.

[105] K. Gunasekaran, Buyong Ma, and Ruth Nussinov. Is Allostery an Intrinsic Property of All Dynamic Proteins? *Proteins*, 57(3):433–443, November 2004.

[106] Joseph A. Marsh and Sarah A. Teichmann. Protein Flexibility Facilitates Quaternary Structure Assembly and Evolution. *PLOS Biology*, 12(5):e1001870, May 2014.

[107] Frederic Rousseau and Joost Schymkowitz. A Systems Biology Perspective on Protein Structural Dynamics and Signal Transduction. *Current Opinion in Structural Biology*, 15(1):23–30, February 2005.

[108] Patrick A. Alexander, Yanan He, Yihong Chen, John Orban, and Philip N. Bryan. A Minimal Sequence Code for Switching Protein Structure and Function. *Proceedings of the National Academy of Sciences*, 106(50):21149–21154, December 2009.

[109] Erik D. Nelson and Nick V. Grishin. Long-Range Epistasis Mediated by Structural Change in a Model of Ligand Binding Proteins. *PLOS ONE*, 11(11):e0166739, November 2016.

[110] Mingyang Lu, Mohit Kumar Jolly, Ryan Gomoto, Bin Huang, José Onuchic, and Eshel Ben-Jacob. Tristability in Cancer-Associated microRNA-TF Chimera Toggle Switch. *J Phys Chem B*, 117(42):13164–13174, October 2013.

[111] Sylvain Bessonnard, Laurane De Mot, Didier Gonze, Manon Barriol, Cynthia Dennis, Albert Goldbeter, Geneviève Dupont, and Claire Chazaud. Gata6, Nanog and Erk Signaling Control Cell Fate in the Inner Cell Mass through a Tristable Regulatory Network. *Development*, 141(19):3637–3648, October 2014.

[112] Sergey Kryazhimskiy, Gašper Tkačik, and Joshua B. Plotkin. The Dynamics of Adaptation on Correlated Fitness Landscapes. *Proceedings of the National Academy of Sciences*, 106(44):18638–18643, March 2009.

[113] Daniel M. Weinreich. High-Throughput Identification of Genetic Interactions in HIV-1. *Nat Genet*, 43(5):398–400, May 2011.

[114] Shozo Yokoyama, Jinyi Xing, Yang Liu, Davide Faggionato, Ahmet Altun, and William T. Starmer. Epistatic Adaptive Evolution of Human Color Vision. *PLOS Genet*, 10(12):e1004884, December 2014.

[115] Anna I. Podgornaia and Michael T. Laub. Pervasive degeneracy and epistasis in a protein-protein interface. *Science*, 347(6222):673–677, February 2015.

[116] Michael B. Doud and Jesse D. Bloom. Accurate Measurement of the Effects of All Amino-Acid Mutations on Influenza Hemagglutinin. *Viruses*, 8(6):155, June 2016.

[117] Evan A. Boyle, Johan O. L. Andreasson, Lauren M. Chircus, Samuel H. Sternberg, Michelle J. Wu, Chantal K. Guegler, Jennifer A. Doudna, and William J. Greenleaf. High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proceedings of the National Academy of Sciences*, page 201700557, May 2017.

[118] Thomas A. Hopf, John B. Ingraham, Frank J. Poelwijk, Charlotta P. I. Schärfe, Michael Springer, Chris Sander, and Debora S. Marks. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2):128–135, February 2017.

[119] Yvonne H. Chan, Sergey V. Venev, Konstantin B. Zeldovich, and C. Robert Matthews. Correlation of fitness landscapes from three orthologous TIM barrels originates from sequence and structure constraints. *Nature Communications*, 8:14614, March 2017.

[120] Tyler N. Starr, Lora K. Picton, and Joseph W. Thornton. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature*, 549(7672):409–413, September 2017.

[121] Júlia Domingo, Guillaume Diss, and Ben Lehner. Pairwise and higher-order genetic interactions during the evolution of a tRNA. *Nature*, 558(7708):117–121, June 2018.

[122] Daniel M. Weinreich, Yinghong Lan, Jacob Jaffe, and Robert B. Heckendorn. The Influence of Higher-Order Epistasis on Biological Fitness Landscape Topography. *Journal of Statistical Physics*, 172(1):208–225, July 2018.

[123] Gideon Schreiber and Alan R. Fersht. Energetics of protein-protein interactions: Analysis ofthe Barnase-Barstar interface by single mutations and double mutant cycles. *Journal of Molecular Biology*, 248(2):478–486, January 1995.

[124] Amnon Horovitz. Double-Mutant Cycles: A Powerful Tool for Analyzing Protein Structure and Function. *Folding and Design*, 1(6):R121–R126, December 1996.

[125] Ákos Nyerges, Bálint Csörgő, Gábor Draskovits, Bálint Kintses, Petra Szili, Györgyi Ferenc, Tamás Révész, Eszter Ari, István Nagy, Balázs Bálint, Bálint Márk Vásárhelyi, Péter Bihari, Mónika Számel, Dávid Balogh, Henrietta Papp, Dorottya Kalapis, Balázs Papp, and Csaba Pál. Directed evolution of multiple genomic loci allows the prediction of antibiotic resistance. *Proceedings of the National Academy of Sciences*, page 201801646, June 2018.

[126] Frank J. Poelwijk, Michael Socolich, and Rama Ranganathan. Learning the pattern of epistasis linking genotype and phenotype in a protein. *bioRxiv*, page 213835, November 2017.

[127] Luca Ferretti, Daniel Weinreich, Fumio Tajima, and Guillaume Achaz. Evolutionary constraints in fitness landscapes. *Heredity*, page 1, July 2018.

[128] Wes McKinney. Data Structures for Statistical Computing in Python. pages 51–56, 2010.

[129] Matt Newville, Renee Otten, Andrew Nelson, Antonino Ingargiola, Till Stensitzki, Dan Allan, Austin Fox, Michał, Glenn, Yoav Ram, MerlinSmiles, Li Li, Christoph Deil, Stuermer, Alexandre Beelen, Oliver Frost, gpasquev, Allan L. R. Hansen, Alexander Stark, Tim Spillane, Shane Caldwell, Anthony Polloreno, andrewhannum, colgan, Robbie Clarken, Kostis Anagnostopoulos, Jose Borreguero, deep 42-thought, Ben Gamari, and Anthony Almarza. lmfit/lmfit-py 0.9.11, June 2018.

[130] Ben Lehner. Molecular mechanisms of epistasis within and between genes. *Trends in Genetics*, 27(8):323–331, August 2011.

[131] Jakub Otwinowski. Biophysical inference of epistasis and the effects of mutations on protein stability and function. *arXiv:1802.08744 [q-bio]*, February 2018. arXiv: 1802.08744.

[132] Kristina Crona, Alex Gavryushkin, Devin Greene, and Niko Beerenwinkel. Inferring genetic interactions from comparative fitness data. *eLife*, 6.

[133] Matteo Figliuzzi, Hervé Jacquier, Alexander Schug, Oliver Tenaillon, and Martin Weigt. Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. *Molecular Biology and Evolution*, 33(1):268–280, January 2016.

[134] Adam J. Riesselman, John B. Ingraham, and Debora Susan Marks. Deep generative models of genetic variation capture mutation effects. *bioRxiv*, page 235655, December 2017.

[135] Sam Sinai, Eric Kelsic, George M. Church, and Martin A. Nowak. Variational auto-encoding of protein sequences. *arXiv:1712.03346 [cs, q-bio]*, December 2017. arXiv: 1712.03346.

[136] Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLOS Computational Biology*, 13(1):e1005324, January 2017.

[137] Jianzhu Ma, Michael Ku Yu, Samson Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, and Trey Ideker. Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*, 15(4):290–298, April 2018.

[138] Sandipan Dutta, Jean-Pierre Eckmann, Albert Libchaber, and Tsvi Tlusty. Green function of correlated genes in a minimal mechanical model of protein evolution. *Proceedings of the National Academy of Sciences*, 115(20):E4559–E4568, May 2018.

[139] R. C. Lewontin and Ken-ichi Kojima. The Evolutionary Dynamics of Complex Polymorphisms. *Evolution*, 14(4):458–472, 1960.

[140] Inna S. Povolotskaya and Fyodor A. Kondrashov. Sequence space and the ongoing expansion of the protein universe. *Nature*, 465(7300):922–926, June 2010.

[141] N. H. Barton and B. Charlesworth. Why Sex and Recombination? *Science*, 281(5385):1986–1990, September 1998.

[142] Thomas MacCarthy and Aviv Bergman. Coevolution of robustness, epistasis, and recombination favors asexual reproduction. *Proceedings of the National Academy of Sciences*, 104(31):12801–12806, July 2007.

[143] Paul E O'Maille, Arthur Malone, Nikki Dellas, B Andes Hess, Lidia Smentek, Iseult Sheehan, Bryan T Greenhagen, Joe Chappell, Gerard Manning, and Joseph P Noel. Quantitative Exploration of the Catalytic Landscape Separating Divergent Plant Sesquiterpene Synthases. *Nature Chemical Biology*, 4(10):617–623, October 2008.

[144] Elena R. Lozovsky, Thanat Chookajorn, Kyle M. Brown, Mallika Imwong, Philip J. Shaw, Sumalee Kamchonwongpaisan, Daniel E. Neafsey, Daniel M. Weinreich, and Daniel L. Hartl. Stepwise Acquisition of Pyrimethamine Resistance in the Malaria Parasite. *PNAS*, 106(29):12025–12030, July 2009.

[145] Marna S. Costanzo, Kyle M. Brown, and Daniel L. Hartl. Fitness Trade-Offs in the Evolution of Dihydrofolate Reductase and Drug Resistance in Plasmodium Falciparum. *PLOS ONE*, 6(5):e19636, May 2011.

[146] Leah E. Cowen and Susan Lindquist. Hsp90 Potentiates the Rapid Evolution of New Traits: Drug Resistance in Diverse Fungi. *Science*, 309(5744):2185–2189, September 2005.

[147] Matthew T. G. Holden, Edward J. Feil, Jodi A. Lindsay, Sharon J. Peacock, Nicholas P. J. Day, Mark C. Enright, Tim J. Foster, Catrin E. Moore, Laurence Hurst, Rebecca Atkin, Andrew Barron, Nathalie Bason, Stephen D. Bentley, Carol Chillingworth, Tracey Chillingworth, Carol Churcher, Louise Clark, Craig Corton, Ann Cronin, Jon Doggett, Linda Dowd, Theresa Feltwell, Zahra Hance, Barbara Harris, Heidi Hauser, Simon Holroyd, Kay Jagels, Keith D. James, Nicola Lennard, Alexandra Line, Rebecca Mayes, Sharon Moule, Karen Mungall, Douglas Ormond, Michael A. Quail, Ester Rabbinowitsch, Kim Rutherford, Mandy Sanders, Sarah Sharp, Mark Simmonds, Kim Stevens, Sally Whitehead, Bart G. Barrell, Brian G. Spratt, and Julian Parkhill. Complete genomes of two clinical Staphylococcus aureus strains: Evidence for the rapid evolution of virulence and drug resistance. *Proceedings of the National Academy of Sciences*, 101(26):9786–9791, June 2004.

[148] Douglas D. Richman, Terri Wrin, Susan J. Little, and Christos J. Petropoulos. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. *Proceedings of the National Academy of Sciences*, 100(7):4144–4149, April 2003.

[149] Jesse D. Bloom, Lizhi Ian Gong, and David Baltimore. Permissive Secondary Mutations Enable the Evolution of Influenza Oseltamivir Resistance. *Science*, 328(5983):1272–1275, June 2010.

[150] Robert L. Summers, Megan N. Nash, and Rowena E. Martin. Know your enemy: understanding the role of PfCRT in drug resistance could lead to new antimalarial tactics. *Cellular and Molecular Life Sciences*, 69(12):1967–1995, June 2012.

[151] D. J. Krogstad, I. Y. Gluzman, D. E. Kyle, A. M. Oduola, S. K. Martin, W. K. Milhous, and P. H. Schlesinger. Efflux of chloroquine from Plasmodium falciparum: mechanism of chloroquine resistance. *Science*, 238(4831):1283–1285, November 1987.

[152] C. A. Homewood, D. C. Warhurst, W. Peters, and V. C. Baggaley. Lysosomes, pH and the Anti-malarial Action of Chloroquine. *Nature*, 235(5332):50–52, January 1972.

[153] Andrew F. G. Slater. Chloroquine: Mechanism of drug action and resistance in plasmodium falciparum. *Pharmacology & Therapeutics*, 57(2):203–235, January 1993.

[154] Rowena E. Martin, Rosa V. Marchetti, Anna I. Cowan, Susan M. Howitt, Stefan Bröer, and Kiaran Kirk. Chloroquine Transport via the Malaria Parasite's Chloroquine Resistance Transporter. *Science*, 325(5948):1680–1682, September 2009.

[155] Heather J. Cordell. Detecting Gene–gene Interactions that Underlie Human Diseases. *Nat Rev Genet*, 10(6):392–404, June 2009.

[156] Jakub Otwinowski and Ilya Nemenman. Genotype to Phenotype Mapping and the Fitness Landscape of the E. Coli Lac Promoter. *PLoS ONE*, 8(5):e61570, May 2013.

[157] Zachary R. Sailer and Michael J. Harms. Uninterpretable interactions: epistasis as uncertainty. *bioRxiv*, page 378489, July 2018.

[158] Christopher K Williams and Carl Edward Rasmussen. Gaussian processes for machine learning. *the MIT Press*, 2(3):4, 2006.

[159] Collin B. Erickson, Bruce E. Ankenman, and Susan M. Sanchez. Comparison of Gaussian process modeling software. *European Journal of Operational Research*, 266(1):179–192, April 2018.

[160] Ian Jolliffe. Principal Component Analysis. In *International Encyclopedia of Statistical Science*, pages 1094–1096. Springer, Berlin, Heidelberg, 2011.

[161] D. R. Cox. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242, 1958.

[162] Frank E. Harrell. Ordinal Logistic Regression. In *Regression Modeling Strategies*, Springer Series in Statistics, pages 311–325. Springer, Cham, 2015.