Uninterpretable interactions: epistasis as uncertainty

Zachary R. Sailer[1,2] & Michael J. Harms[1,2]

1. Institute of Molecular Biology, University of Oregon, Eugene, OR 97403
2. Department of Chemistry & Biochemistry, University of Oregon, Eugene, OR 97403

Abstract

Epistasis is a common feature of genotype-phenotype maps. Understanding the patterns of epistasis is critical for predicting unmeasured phenotypes, explaining evolutionary trajectories, and for inferring the biological mechanisms that determine a map. One common approach is to use a linear model to decompose epistasis into specific pairwise and high-order interactions between mutations. Such interactions are then used to identify important biology or to explain how the genotype-phenotype map shapes evolution. Here we show that the coefficients extracted from such analyses are likely uninterpretable. They cannot be extracted reliably from experimental genotype-phenotype maps due to regression bias. Further, we can generate epistatic "interactions" indistinguishable from those in experimental maps using a completely random process. From this, we conclude that epistasis should be treated as a random, but quantifiable, variation in these maps. This perspective allows us to build predictive models with known error from a small number of measured phenotypes. It also suggests that we need mechanistic, nonlinear models to account for epistasis and decompose genotype-phenotype maps.

Introduction

Epistasis—that is, non-additivity between mutations—is a ubiquitous feature of genotype-phenotype maps (Fowler et al., 2010; Weinreich, 2011; Weinreich et al., 2013; Yokoyama et al., 2014; Anderson et al., 2015; Palmer et al., 2015; Podgornaia and Laub, 2015; Doud and Bloom, 2016; Boyle et al., 2017; Hopf et al., 2017; Chan et al., 2017; Sailer and Harms, 2017a; Starr et al., 2017; Domingo et al., 2018; Weinreich et al., 2018). Epistasis can provide mechanistic insight into the determinants of phenotypes (Schreiber and Fersht, 1995; Horovitz, 1996; Ritchie et al., 2001; Segrè et al., 2005); however, it also complicates predicting unmeasured phenotypes (de Visser and Krug, 2014; Miton and Tokuriki, 2016; Sailer and

Harms, 2017b; Nyerges et al., 2018), as the effect of a mutation changes depending on the presence or absence of other mutations. Despite a century of work (Fisher, 1918), epistasis remains challenging to analyze and interpret (Cordell, 2002; Phillips, 2008; Crow, 2010; Weinreich et al., 2013, 2018).

One approach is to decompose epistasis into specific pairwise and high-order interactions between mutations (Heckendorn and Whitley, 1999; Weinreich et al., 2013; Poelwijk et al., 2016; Sailer and Harms, 2017a; Poelwijk et al., 2017; Weinreich et al., 2018). This is often done by treating each coefficient as a linear and independent perturbation to the additive phenotype (Heckendorn and Whitley, 1999; Poelwijk et al., 2016). 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 (Ritchie et al., 2001; Segrè et al., 2005). 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 (Schreiber and Fersht, 1995; Horovitz, 1996).

Although linear epistasis models are very commonly used (Weinreich et al., 2013; Yokoyama et al., 2014; Anderson et al., 2015; Palmer et al., 2015; Starr et al., 2017; Domingo et al., 2018), 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 (Otwinowski and Plotkin, 2014). 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 (Sailer and Harms, 2017b). 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 (Otwinowski and Plotkin, 2014). 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 (Szendro et al., 2013; Ferretti et al., 2018). 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 (Fisher, 1918; Poelwijk et al., 2016). (The assumption of linearity contrasts with other models, such as a Potts model, in which mutations sum in a nonlinear fashion (Hopf et al., 2017)). There are two

common formulations a linear epistasis model, the Hadamard model (sometimes called a Walsh or Fourier model) and the biochemical model (Poelwijk et al., 2016). The approaches differ in their choice of coordinate origin. Each model has been described in detail elsewhere (Heckendorn and Whitley, 1999; Weinreich et al., 2013; Poelwijk et al., 2016). The two models are related by a simple set of linear transformations (Poelwijk et al., 2016). 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 (Heckendorn and Whitley, 1999; Weinreich et al., 2013; Poelwijk et al., 2016; Sailer and Harms, 2017a). 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,

$$P = \beta_{\text{origin}} + \sum_{i}^{L} \beta_i x_i \tag{1}$$

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. 1. For pairwise interactions, this has the form:

$$P = \beta_{\text{origin}} + \sum_{i}^{L} \beta_i x_i + \sum_{j<i}^{L} \beta_{ij} x_i x_j \tag{2}$$

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

$$P = \beta_{\text{origin}} + \sum_{i}^{L} \beta_i x_i + \sum_{j<i}^{L} \beta_{ij} x_i x_j + \sum_{k<j<i}^{L} \beta_{ijk} x_i x_j x_k + ... \tag{3}$$

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 (Chou et al., 2011; Tokuriki et al., 2012; Schenk et al., 2013; Sailer and Harms, 2017a; Otwinowski et al., 2018). This violates the assumption of linearity inherent in the epistasis models (Fisher, 1918; Cordell, 2002; Sailer and Harms, 2017a). 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) (Sailer and Harms, 2017a; Otwinowski et al., 2018). 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 (Box and Cox, 1964; Szendro et al., 2013; Sailer and Harms, 2017a; Otwinowski et al., 2018):

$$\vec{P}_{\text{obs}} = T(\vec{P}_{\text{add}}) + \vec{\varepsilon}. \tag{4}$$

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 (Otwinowski et al., 2018).

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:

$$\vec{P}_{\text{obs}} = \mathbf{X}\vec{\beta} + \vec{\varepsilon} \tag{5}$$

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. 3, 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 (Otwinowski and Plotkin, 2014; Poelwijk et al., 2017). 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 1 and 3. 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* (van der Walt et al., 2011; McKinney, 2010). We used the Python package *scikit-learn* to perform ordinary-, lasso-, and ridge- regression (Pedregosa et al., 2011). We used the Python

package *lmfit* to perform nonlinear-least squares regression (Newville et al., 2018). Plots were generated using *matplotlib* and *Jupyter* notebooks (Hunter, 2007; Perez and Granger, 2007). Our full software packages are available in the *gpmap* (https://harmslab.github.com/gpmap) and *epistasis* (https://harmslab.github.com/epistasis) packages on Github.

Data availability statement

All software is available for download from:

- https://github.com/harmslab/gpmap
- https://github.com/harmslab/epistasis.

Data sets are available from:

- https://github.com/harmslab/genotype-phenotype-maps.

Supplemental figures S1-S4 are available via GSA Figshare.

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.

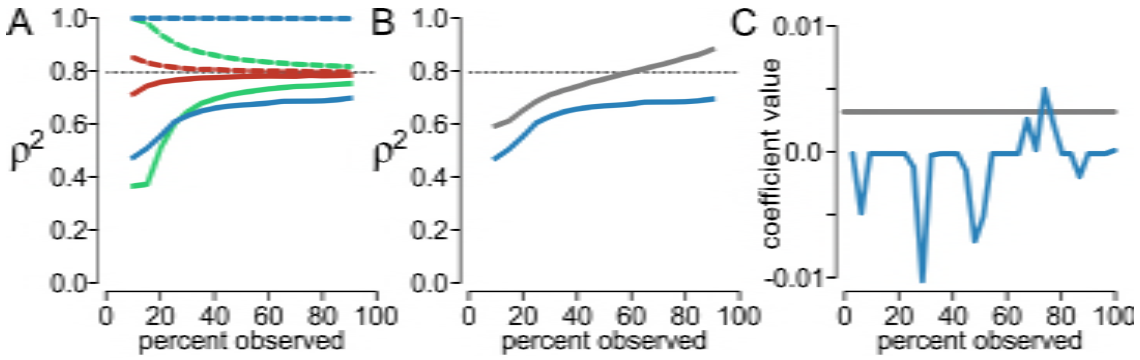

Figure 1: **Linear epistatic coefficients cannot be estimated from an incomplete, simulated genotype-phenotype map.** A) Fit scores versus the percent of the genotypes in the map used to train the model, from 10% to 90%. The dashed gray line indicates the amount of additive variation in the map (80%). 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 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. 1). 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 1A). 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.

We next tried to improve our predictive power by adding coefficients describing either pairwise interactions between mutations (Eq. 2) or all interactions (up to eighth-order) (Eq. 3). Because Eq. 3 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 1A). 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 1A, dashed blue line), even while test set correlation languished around 0.6 (Fig 1A, solid blue line).

This result arises because regression yields biased estimates of the epistatic coefficients (Otwinowski and Plotkin, 2014). 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 1B. The blue line reproduces ρ^2_{test} for the high-order model from Fig 1A. 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 1C 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).

ID	genotype	phenotype	L	reference
I	genomic mutations	*E. coli* fitness	5	(Khan et al., 2011)
II	point mutants	bacterial fitness	5	(Weinreich et al., 2006)
III	chromosomes	*A. niger* fitness	5	(de Visser et al., 2009)
IV	point mutants	binding affinity	5	(Anderson et al., 2015)
V	alleles in network	*S. cerevisiae* growth rate	6	(Hall et al., 2010)
VI	alleles in network	*S. cerevisiae* growth rate	6	(Hall et al., 2010)
VII	genomic mutations	*E. coli* fitness	5	(Flynn et al., 2013)
VIII	genomic mutations	*E. coli* fitness	5	(Flynn et al., 2013)
IX	chromosomes	*A. niger* fitness	5	(de Visser et al., 2009)
X	alleles in network	*S. cerevisiae* sporulation	6	(Hall et al., 2010)
XI	alleles in network	*S. cerevisiae* mating	6	(Hall et al., 2010)
XII	genomic mutations	*E. coli* fitness	6	(Palmer et al., 2015)

Table 1: **Published experimental genotype-phenotype maps.**

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 1). 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.

We started by linearizing the experimental genotype-phenotype maps (see methods) (Sailer and Harms, 2017a; Otwinowski et al., 2018). 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 (Heckendorn and Whitley, 1999; Poelwijk et al., 2016), avoiding complications arising from regression. We found that all of these maps exhibited statistically significant pairwise and high-order

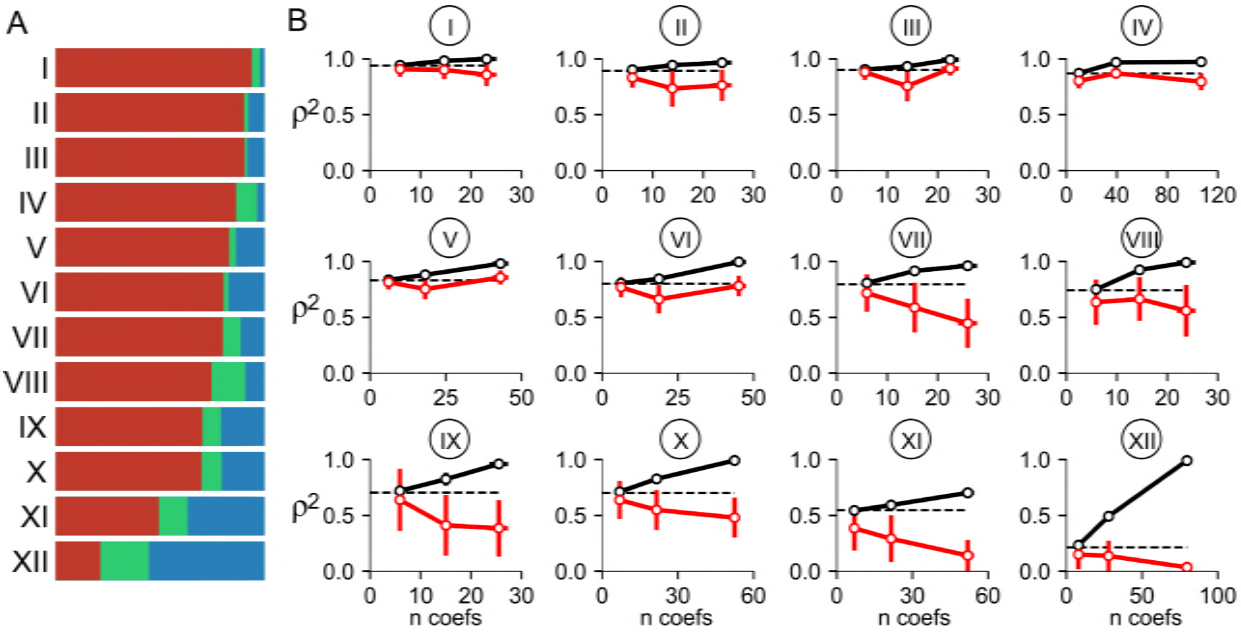

Figure 2: **Predictive epistatic coefficients cannot be resolved from experimental genotype-phenotype maps.** A) Bars show the fraction of variation in phenotype explained by additive effects (red), pairwise epistasis (green), or any order of high-order epistasis (blue). Each bar is for one of the twelve 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.

epistatic interactions. Epistasis contributed from 6% to 79% of the variation in these maps (Fig 2A).

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 pseudo-replicate training and test sets on each map.

As with our simulations, we found we could not reliably extract predictive epistatic coefficients (Fig 2B). 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 (Cordell, 2002).

We decomposed the epistasis in Map VIII using all 32 measured phenotypes and compared the resulting epistasis to our null model. Fig 3A-C shows the epistasis extracted from the experimental map. In this map 26% of the variation in phenotype is due to epistasis (Fig 3B). The residuals between the additive model and the observed phenotypes are normally distributed (Fig 3B). 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.

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 3E. 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 3F).

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 3C 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

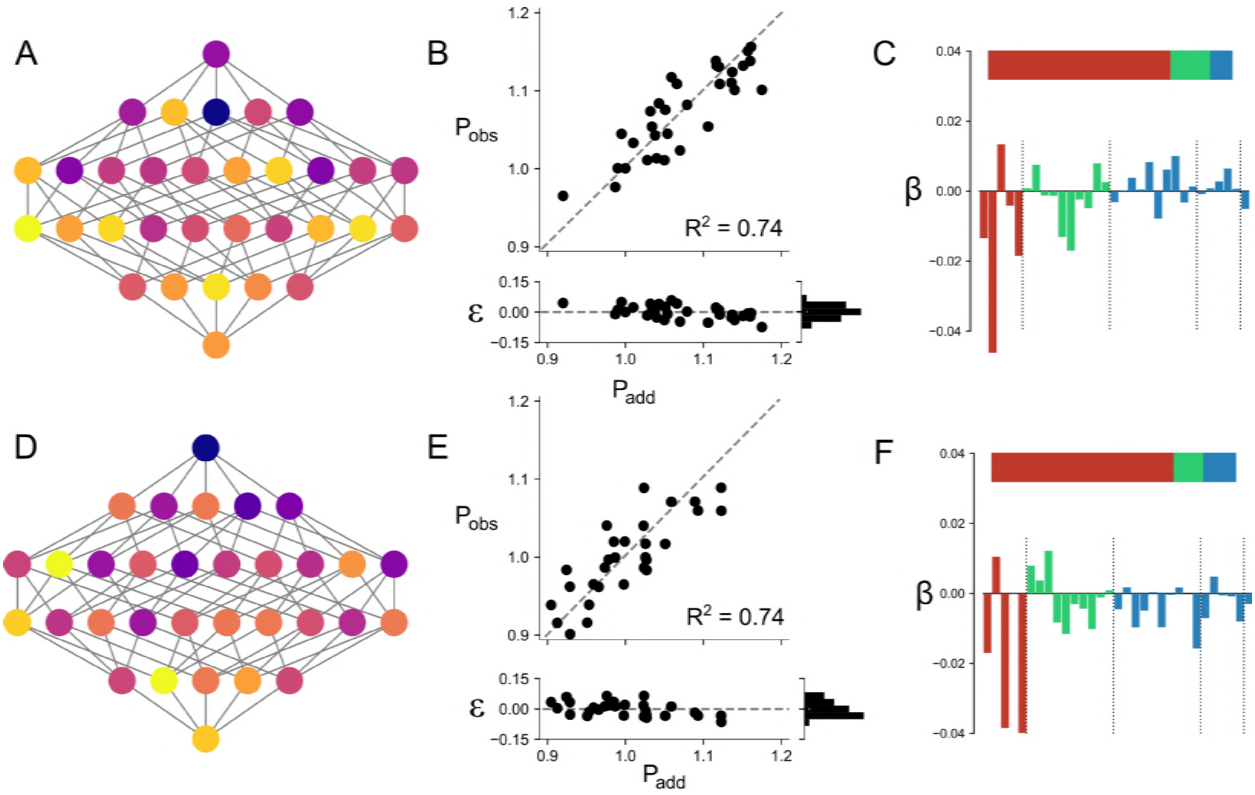

Figure 3: **Experimental maps resemble random maps.** Panels A and D show genotype-phenotype maps. Each node is a genotype; each edge is a single point mutation. Colors indicate quantitative phenotype. Panels B and E show the correlation between the observed phenotypes and the additive model, with fit residuals shown below the plot. Panels C and F indicate the magnitude of epistasis in each map as in Fig 2A (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.

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 4A). 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 (Szendro et al., 2013; Sailer and Harms, 2017a; Otwinowski et al., 2018), the residuals between $\vec{P}_{\text{obs}}$ and $\vec{P}_{\text{add}}$ will be normally distributed (Fig 3B). As a result, the phenotype of a genotype g is given by:

$$P_{\text{obs,g}} = P_{\text{add,g}} \pm \xi \tag{6}$$

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 (Fisher, 1918), 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.

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

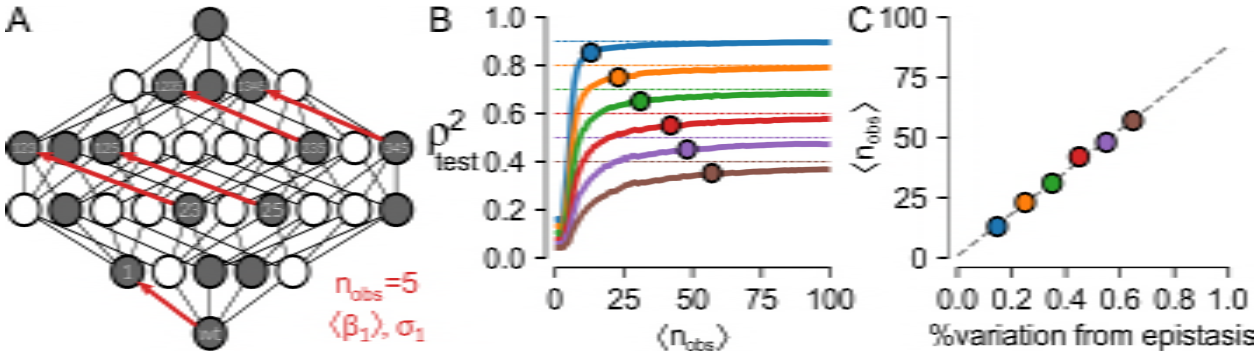

Figure 4: **Epistasis as uncertainty.** A) A partially characterized map. Circles represent genotypes, some of which have been measured (filled), some of which have not (unfilled). Lines represent single point mutations. Given these observations, we measure the effect of mutation 1 in five different 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.

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 4B). 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 4B. 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 4C).

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 5A). 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 5A). The map exhibited a highly non-linear relationship between $\vec{P}_{\text{obs}}$ and $\vec{P}_{\text{add}}$ (Fig 5B), so we linearized the map with a

5th-order spline (Eq. 4), yielding normal residuals between $\vec{P}_{\text{obs,linearized}}$ and $\vec{P}_{\text{add}}$ (Fig 5C). 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 5D). 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 $\pm 38\%$ 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.

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 1B), unstable to the addition of new data (Fig 1C), and not useful for predicting unmeasured phenotypes (Fig 2A). 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 2B). 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 3). This argues for viewing epistatic coefficients as uninterpretable decompositions of random variation, unless shown

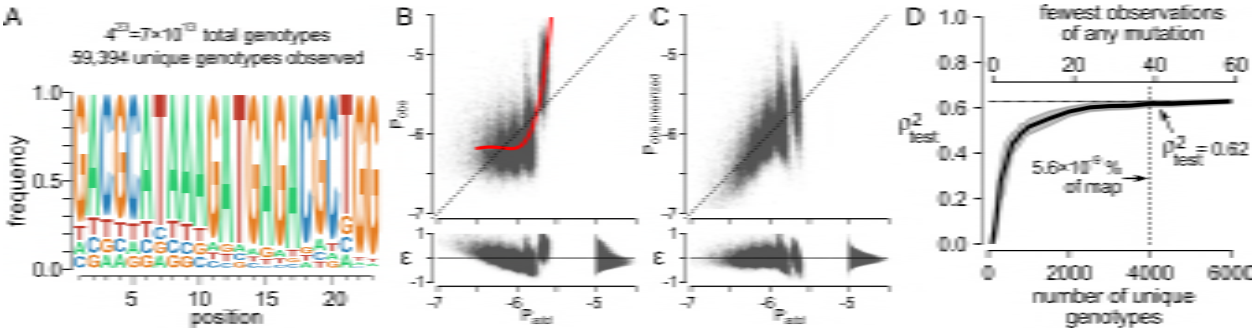

Figure 5: **A predictive, additive model can be trained on a large genotype-phenotype map.** A) Summary of the genotype-phenotype map reported in (Boyle et al., 2017). Map consists of 23 sites, each with four bases with the frequency at each site shown in the sequence logo. The total map has 7×10^{13} 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 $\approx 4,000$ 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.

otherwise.

Viewed mechanistically, this is unsurprising. The epistatic models under investigation assume linearity, but biology is nonlinear (Bershtein et al., 2006; Lehner, 2011; Chou et al., 2011; Tokuriki et al., 2012; Schenk et al., 2013; Sailer and Harms, 2017a; Otwinowski et al., 2018). There is no reason to believe that a linear model will capture a complicated non-linear 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 (Sailer and Harms, 2017b). 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" (Cordell, 2002).

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 1C, 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. 6). 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 4A). 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 4C).

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 (Sailer and Harms, 2017a; Otwinowski et al., 2018). Multiple methods exist for linearizing genotype-phenotype maps, including taking the log of phenotypes (Cordell, 2002), power transforms (Sailer and Harms, 2017a), splines (Otwinowski et al., 2018), and even mechanistic models (Schenk et al., 2013; Otwinowski, 2018). 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 5C). 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 (Szendro et al., 2013; Ferretti et al., 2018; Crona et al.). 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 (Figliuzzi et al., 2016; Hopf et al., 2017), variational auto encoders (Riesselman et al., 2017; Sinai et al., 2017), and neural networks (Wang et al., 2017; Ma et al., 2018). 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 (Tokuriki et al., 2012; Schenk et al., 2013; Otwinowski, 2018; Dutta et al., 2018). 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 (Schenk et al., 2013). 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 (Fisher, 1918). 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.

Acknowledgements

We would like to thank members of the Harms lab for helpful discussions and comments. Work was supported by start up funds from the University of Oregon (ZRS). MJH is a Pew Scholar in the Biomedical Sciences, supported by The Pew Charitable Trusts.

References

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. ISSN 2050-084X. doi: 10.7554/eLife.07864.

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. ISSN 0028-0836. doi: 10.1038/nature05385.

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. ISSN 0035-9246.

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. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1700557114. URL `http://www.pnas.org/content/early/2017/05/09/1700557114`.

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. ISSN 2041-1723. doi: 10.1038/ncomms14614. URL `https://www.nature.com/articles/ncomms14614`.

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. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1203799.

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. ISSN 0964-6906. doi: 10.1093/hmg/11.20.2463. URL https://academic.oup.com/hmg/article/11/20/2463/616080.

Kristina Crona, Alex Gavryushkin, Devin Greene, and Niko Beerenwinkel. Inferring genetic interactions from comparative fitness data. *eLife*, 6. ISSN 2050-084X. doi: 10.7554/eLife.28629. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5737811/.

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. ISSN 0962-8436, 1471-2970. doi: 10.1098/rstb.2009.0275.

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. ISSN 1471-0056. doi: 10.1038/nrg3744.

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. ISSN 0003-0147, 1537-5323. doi: 10.1086/599081.

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. ISSN 1476-4687. doi: 10.1038/s41586-018-0170-7. URL https://www.nature.com/articles/s41586-018-0170-7.

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. doi: 10.3390/v8060155. URL http://www.mdpi.com/1999-4915/8/6/155.

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. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1716215115. URL `http://www.pnas.org/content/115/20/E4559`.

Luca Ferretti, Daniel Weinreich, Fumio Tajima, and Guillaume Achaz. Evolutionary constraints in fitness landscapes. *Heredity*, page 1, July 2018. ISSN 1365-2540. doi: 10.1038/s41437-018-0110-1. URL `https://www.nature.com/articles/s41437-018-0110-1`.

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. ISSN 0737-4038. doi: 10.1093/molbev/msv211. URL `https://academic.oup.com/mbe/article/33/1/268/2579532`.

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.

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. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003426.

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. ISSN 1548-7091. doi: 10.1038/nmeth.1492.

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. ISSN 0022-1503, 1465-7333. doi: 10.1093/jhered/esq007.

Robert B. Heckendorn and Darrell Whitley. Predicting Epistasis from Mathematical Models. *Evolutionary Computation*, 7(1):69–101, March 1999. ISSN 1063-6560. doi: 10.1162/evco.1999.7.1.69.

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. ISSN 1546-1696. doi: 10.1038/nbt.3769. URL `https://www.nature.com/articles/nbt.3769`.

Amnon Horovitz. Double-Mutant Cycles: A Powerful Tool for Analyzing Protein Structure and Function. *Folding and Design*, 1(6):R121–R126, December 1996. ISSN 1359-0278. doi: 10.1016/S1359-0278(96)00056-9.

J. D. Hunter. Matplotlib: A 2d Graphics Environment. *Computing in Science Engineering*, 9(3):90–95, May 2007. ISSN 1521-9615. doi: 10.1109/MCSE.2007.55.

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. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1203801.

Ben Lehner. Molecular mechanisms of epistasis within and between genes. *Trends in Genetics*, 27(8):323–331, August 2011. ISSN 0168-9525. doi: 10.1016/j.tig.2011.05.007. URL `http://www.sciencedirect.com/science/article/pii/S0168952511000771`.

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. ISSN 1548-7105. doi:

10.1038/nmeth.4627. URL https://www.nature.com/articles/nmeth.4627/.

Wes McKinney. Data Structures for Statistical Computing in Python. pages 51–56, 2010. URL http://conference.scipy.org/proceedings/scipy2010/mckinney.html.

Charlotte M. Miton and Nobuhiko Tokuriki. How Mutational Epistasis Impairs Predictability in Protein Evolution and Design. *Protein Science*, 25(7):1260–1272, July 2016. ISSN 1469-896X. doi: 10.1002/pro.2876.

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. URL https://zenodo.org/record/1301254.

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. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1801646115. URL http://www.pnas.org/content/early/2018/05/30/1801646115.

Jakub Otwinowski. Biophysical inference of epistasis and the effects of mutations on protein stability and function. *arXiv:1802.08744 [q-bio]*, February 2018. URL http://arxiv.org/abs/1802.08744. arXiv: 1802.08744.

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

0027-8424, 1091-6490. doi: 10.1073/pnas.1400849111.

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. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1804015115. URL `http://www.pnas.org/content/early/2018/07/19/1804015115`.

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. ISSN 2041-1723. doi: 10.1038/ncomms8385.

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.

F. Perez and B. E. Granger. IPython: A System for Interactive Scientific Computing. *Computing in Science Engineering*, 9(3):21–29, May 2007. ISSN 1521-9615. doi: 10.1109/MCSE.2007.53.

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. ISSN 1471-0056. doi: 10.1038/nrg2452.

Anna I. Podgornaia and Michael T. Laub. Pervasive degeneracy and epistasis in a protein-protein interface. *Science*, 347(6222):673–677, February 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1257360. URL `http://science.sciencemag.org/content/347/6222/673`.

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. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004771. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004771.

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. doi: 10.1101/213835. URL https://www.biorxiv.org/content/early/2017/11/03/213835.

Adam J. Riesselman, John B. Ingraham, and Debora Susan Marks. Deep generative models of genetic variation capture mutation effects. *bioRxiv*, page 235655, December 2017. doi: 10.1101/235655. URL https://www.biorxiv.org/content/early/2017/12/18/235655.1.

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. ISSN 0002-9297. doi: 10.1086/321276.

Zachary R. Sailer and Michael J. Harms. Detecting High-Order Epistasis in Nonlinear Genotype-Phenotype Maps. *Genetics*, 205(3):1079–1088, March 2017a. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.116.195214. URL http://www.genetics.org/content/205/3/1079.

Zachary R. Sailer and Michael J. Harms. Molecular ensembles make evolution unpredictable. *Proceedings of the National Academy of Sciences*, 114(45):11938–11943, November 2017b. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1711927114. URL http://www.pnas.org/content/114/45/11938.

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. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/mst096.

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. ISSN 0022-2836. doi: 10.1016/S0022-2836(95)80064-6. URL http://www.sciencedirect.com/science/article/pii/S0022283695800646.

Daniel Segrè, Alexander DeLuna, George M. Church, and Roy Kishony. Modular Epistasis in Yeast Metabolism. *Nat Genet*, 37(1):77–83, January 2005. ISSN 1061-4036. doi: 10.1038/ng1489.

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. URL http://arxiv.org/abs/1712.03346. arXiv: 1712.03346.

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. ISSN 1476-4687. doi: 10.1038/nature23902. URL https://www.nature.com/articles/nature23902.

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. ISSN 1742-5468. doi: 10.1088/1742-5468/2013/01/P01005. URL http://stacks.iop.org/1742-5468/2013/i=01/a=P01005.

Nobuhiko Tokuriki, Colin J. Jackson, Livnat Afriat-Jurnou, Kirsten T. Wyganowski, Ren-

mei Tang, and Dan S. Tawfik. Diminishing Returns and Tradeoffs Constrain the Laboratory Optimization of an Enzyme. *Nat Commun*, 3:1257, December 2012. doi: 10.1038/ncomms2246.

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. ISSN 1521-9615. doi: 10.1109/MCSE.2011.37.

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. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005324. URL `http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005324`.

Daniel M. Weinreich. High-Throughput Identification of Genetic Interactions in HIV-1. *Nat Genet*, 43(5):398–400, May 2011. ISSN 1061-4036. doi: 10.1038/ng.820.

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. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1123539.

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. ISSN 0959-437X. doi: 10.1016/j.gde.2013.10.007.

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. ISSN 0022-4715, 1572-9613. doi: 10.1007/s10955-018-1975-3. URL `https://link.springer.com/article/10.1007/s10955-018-1975-3`.

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. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004884.