Inferring the shape of global epistasis

Jakub Otwinowski[a,1], David M. McCandlish[b], and Joshua B. Plotkin[a]

[a]Biology Department, University of Pennsylvania, Philadelphia, PA 19104; and [b]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724

Edited by Richard E. Lenski, Michigan State University, East Lansing, MI, and approved June 20, 2018 (received for review March 7, 2018)

Genotype–phenotype relationships are notoriously complicated. Idiosyncratic interactions between specific combinations of mutations occur and are difficult to predict. Yet it is increasingly clear that many interactions can be understood in terms of global epistasis. That is, mutations may act additively on some underlying, unobserved trait, and this trait is then transformed via a nonlinear function to the observed phenotype as a result of subsequent biophysical and cellular processes. Here we infer the shape of such global epistasis in three proteins, based on published high-throughput mutagenesis data. To do so, we develop a maximum-likelihood inference procedure using a flexible family of monotonic nonlinear functions spanned by an I-spline basis. Our analysis uncovers dramatic nonlinearities in all three proteins; in some proteins a model with global epistasis accounts for virtually all of the measured variation, whereas in others we find substantial local epistasis as well. This method allows us to test hypotheses about the form of global epistasis and to distinguish variance components attributable to global epistasis, local epistasis, and measurement error.

deep mutational scanning | fitness landscape | genotype–phenotype map | protein | evolution

The mapping from genetic sequence to biological phenotype in large part determines the course of evolution. Nonadditive interactions between sites, called epistasis, can either accelerate or severely constrain the pace of adaptation. Due to the high dimensionality of sequence space, studies of evolution on epistatic genotype–phenotype maps, or fitness landscapes, have historically been limited to mathematical models, such as NK landscapes (1, 2), or computational models of simple biophysical phenotypes, such as RNA folding (3, 4).

Recently, high-throughput sequencing has made it possible to measure genotype–phenotype maps for proteins (5) and across genomes (6). While many thousands of genotype–phenotype pairs can now be assayed in a single experiment, this is still only a minuscule fraction of all possible sequences, and so the sampled genotypes typically take the form of a scattered cloud centered on the wild type. Statistical models are therefore required to derive biological insight from these sparsely sampled genotype–phenotype maps. One approach has been to fit models that include terms for each pairwise interaction between genetic sites (7, 8). Such models can sometimes predict protein contacts or mutational effects from protein sequence alignments (9), but they do a poor job at capturing a complete picture of the genotype–phenotype map—so that their predictions far from the observed data are often wildly inaccurate (10, 11).

Models of a genotype–phenotype map that contain terms for every possible pairwise interaction seem reasonable if we believe that epistasis arises primarily through the idiosyncratic effects of particular pairs of mutants—for instance, pairs of residues that contact each other in a folded protein. While there is abundant biochemical evidence for epistasis of this type, geneticists have long suspected that a substantial fraction of observed epistasis is due not to specific pairwise interactions, but rather to inherent nonlinearities in molecular phenotypes, cellular fitness, organismal physiology, and reproductive success. Apparent pairwise interactions may be caused by mutations that contribute additively to some underlying trait, combined with a nonlinear relationship between the underlying trait and the measured phenotype, or fitness.

A classic argument for this form of epistasis is the physiological theory of dominance (12), which held that dominance arises from the inherently nonlinear relationship between gene activity or dosage and measured phenotype [e.g., saturation phenomena in metabolic networks (13)]. Another classical example arose in the attempt to reconcile apparent contradictions between patterns of genetic polymorphism and the tolerable degree of genetic load, by positing models of nonlinear selection—for instance, truncation selection—operating on an underlying additive trait such as the number of deleterious mutations or heterozygous genes (14–18). Following these classic examples, research into the shape of nonlinear selection on an observed quantitative trait using either quadratic (19) or more general (20) forms became a major interest of evolutionary quantitative genetics (21). Contemporary studies on the fitness effects of mutations often incorporate epistasis based on thermodynamic arguments, where the probability of transcription factor binding (22) or protein folding (23–25) is expressed as a nonlinear function of a binding or folding energy that is additive across sites. Today, the idea that epistatic interactions are the result of a nonlinear mapping from an underlying additive trait has been called "unidimensional" (26, 27), "nonspecific" (28), or "global" (29) epistasis, and several heuristic techniques have been proposed to infer epistasis of this form (30–33).

Here we present a maximum-likelihood framework for inferring models of global epistasis, and we apply this framework to several large genotype–phenotype maps of proteins derived

Significance

How does an organism's genetic sequence govern its measurable characteristics? New technologies provide libraries of randomized sequences to study this relationship in unprecedented detail for proteins and other molecules. Deriving insight from these data is difficult, though, because the space of possible sequences is enormous, so even the largest experiments sample a tiny minority of sequences. Moreover, the effects of mutations may combine in unexpected ways. We present a statistical framework to analyze such mutagenesis data. The key assumption is that mutations contribute in a simple way to some unobserved trait, which is related to the observed trait by a nonlinear mapping. Analyzing three proteins, we show that this model is easily interpretable and yet fits the data remarkably well.

Author contributions: J.O., D.M.M., and J.B.P. designed research; J.O. performed research; and J.O., D.M.M., and J.B.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The code used for inference and analysis, as well as input–output data, is available on GitHub at https://github.com/jotwin/GlobalEpistasis.jl.

[1]To whom correspondence should be addressed. Email: jakubo@sas.upenn.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1804015115/-/DCSupplemental.

Published online July 23, 2018.

from deep mutational scanning (DMS) experiments. Under the assumption that the observed phenotype is a monotonic, nonlinear function of some unobserved additive trait, we find the best-fit model of global epistasis, test hypotheses about the form of this nonlinear relationship, infer the additive coefficients for the unobserved trait together with their confidence intervals, and estimate the extent of additional epistasis due to specific interactions between sites. Our approach shares much of the simplicity of additive models, but it can capture far more of the phenotypic variation using only a handful of additional parameters.

A Model of Global Epistasis

Statistical models can infer genotype–phenotype maps from data by assuming some form of underlying structure. The starting point of our analysis is a nonepistatic model. Given a sequence of amino acids a_i for sites $i = 1$ to L and an associated measured phenotype y, the nonepistatic model is

$$y = \beta_0 + \sum_i^L \beta_{i,a_i} + \varepsilon,$$

where the noise term ε represents deviations between model and data. The additive effects $\beta_{i,\alpha}$ can be visualized as a matrix of effects for each position and amino acid with $\beta_{i,a_i} = 0$ if there is no substitution (i.e., $a_i = a_i^{\mathrm{WT}}$), so that the wild-type phenotype is modeled by the constant term β_0. With many pairs of measured sequences and phenotypes the additive effects $\beta_{i,\alpha}$ can be estimated by linear regression. Nonepistatic models often explain much of the variance in genotype–phenotype maps, but there is usually a significant portion of unexplained variance that cannot be accounted for, given the known magnitudes of measurement errors (34).

One obvious way to incorporate epistasis into this additive model is by including explicit terms for interactions between each pair of sites and possibly higher-order interactions as well (7, 8). By contrast, we develop a model and inference procedure for a different form of epistasis, motivated by examining the deviations between empirical datasets and their best-fit nonepistatic models. Such deviations often show a smooth nonlinearity, as shown in *SI Appendix*, Fig. S1*A* for protein G. This observed nonlinearity suggests a global coupling between all sites that determines the observed phenotype. This idea can be formalized as a semiparametric model

$$y = g(\phi) + \varepsilon \quad [1]$$

$$\phi = \beta_0 + \sum_i^L \beta_{i,a_i}, \quad [2]$$

where ϕ is an inferred, but unobserved, additive trait that depends on the genetic sequence, and $g(\phi)$ is an arbitrary nonlinear function that represents the shape of global epistasis relating the additive trait to the measured data. We refer to the coefficients $\beta_{i,\alpha}$ in this model as the additive effects, even though the full model is nonlinear. The assumption underlying this model is that mutations affect the measured phenotype only via a single unobserved additive trait. To model the function $g(\phi)$ we choose a flexible parametric family of monotonic functions, I-splines (35), to avoid fitting an arbitrarily complex model and to reduce the inference to standard maximum likelihood. With sufficient data, the additive effects $\beta_{i,\alpha}$ and the shape of global epistasis $g(\phi)$ can be inferred simultaneously.

We also estimate the amount of epistasis that is not captured by the global nonlinearity in our model. When independent estimates of the measurement error for each sequence are available, as is often the case, we can infer how much of the remaining variation in the measured phenotype is due to other forms of epistasis. We refer to this last component of variation in the measured phenotype as house-of-cards (HOC) epistasis, because in our statistical framework we model the phenotypic contribution of this epistasis as an independent random draw, for each genotype, from a Gaussian distribution—akin to a fully uncorrelated house-of-cards model (36). In particular, we estimate the HOC epistasis component, σ^2_{HOC}, by setting the total variance of our per-sequence Gaussian likelihood to be $\sigma^2_{\mathrm{HOC}} + \sigma^2_m$, where σ^2_m is the independent estimate of phenotype measurement error (*Materials and Methods*).

We infer the model described above by maximum likelihood. Doing so allows us to compare models of different complexity by means of a likelihood-ratio test via parametric bootstrap (detailed in *Materials and Methods*). This approach allows us to ask whether there is statistical support for certain features, but not others, in the shape of global epistasis and to assess the uncertainty in our estimates of both global epistasis and the coefficients describing the impact of mutations on the underlying additive trait.

Global Epistasis in Protein GB1

As a first application of our framework, we characterized global epistasis in the IgG-binding domain of protein G (GB1), a model system of protein folding and stability (Fig. 1), using data from a study by Olson et al. (37). In particular, Olson et al. (37) targeted 55 sites for mutation and measured the binding to an immunoglobulin fragment (IgGFC) for all single–amino-acid substitutions (1,045) and 95% of all double substitutions (509,693) (37). We find that for GB1 our model of global epistasis outperforms a purely nonepistatic model ($P < 0.0003$, likelihood-ratio test, *SI Appendix*, Table S1; see *Materials and Methods* for statistical methodology), and it substantially reduces the extent of unmodeled epistasis ($\sigma_{\mathrm{HOC}} = 0.35$, reduced from $\sigma_{\mathrm{HOC}} = 0.56$ for the nonepistatic model) with a 10-fold cross-validated $r^2 = 93.5\%$, compared with $r^2 = 86.2\%$ for the nonepistatic model.

Looking first at the shape of global epistasis inferred by our model (Fig. 1*A*), we see that it includes both diminishing and increasing returns, depending on the value of the underlying additive trait. In particular, our inferred nonlinear function has a negative second derivative in the region around the wild type (negative in 95% of bootstraps for $-0.97 < \phi < 1.72$), indicating a pattern of diminishing-returns epistasis, but the slope remains strictly positive at high trait values ($P < 0.003$; *SI Appendix*, Table S1), suggesting that further increases in binding affinity outside the range of observed data are possible. For deleterious mutations, we observe saturation with decreasing additive trait values, as indicated by a positive second derivative (i.e., increasing returns; positive in 95% of bootstraps for $-4.27 < \phi < -1.04$). We find significant support for a nonzero slope on the left-hand side ($P < 0.003$; *SI Appendix*, Table S1), but our estimate of this slope is extremely small, with an expected change of only 0.0037 (CI 0.0034–0.0046) per mutation, in units of relative log binding. This slope is very small compared with the overall range of binding scores observed (from around -5 to $+2$) and it is consistent with control experiments by ref. 37 that established a lower bound on the binding score in their assay. More generally, our estimates of the shape of global epistasis are extremely precise, with an average 95% CI of width 0.13 (*Materials and Methods*) over the full range of the observed binding affinities (light gray in Fig. 1*A*, too small to be visible).

For the effects on the unobserved additive trait (Fig. 1*D*), we find 234 beneficial and 774 deleterious mutations with 95% CIs that exclude zero, out of a total of 1,045 mutations (*SI Appendix*, Fig. S1*D*). These additive effects are similar to what would be inferred under the purely additive model, but they are

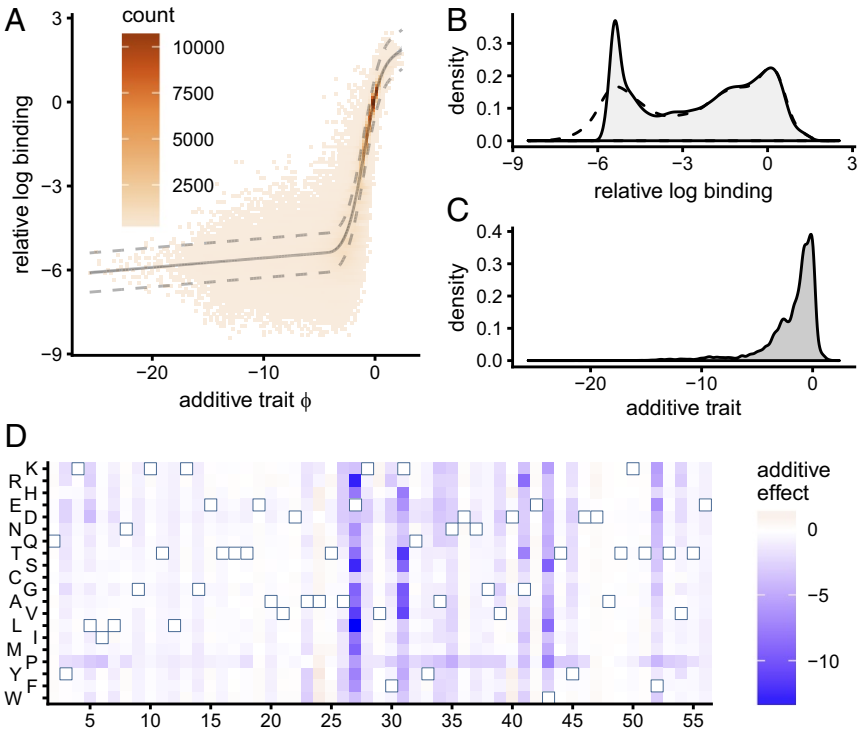

Fig. 1. (*A*) Global epistasis as a nonlinear function of an intermediate additive trait (Eq. **1**) for binding of protein GB1 to an immunoglobulin fragment (IgGFC) (37). Solid black line indicates the nonlinear function $g(\phi)$ (95% confidence interval in light gray, too small to be visible). Based on the inferred magnitude of nonglobal epistasis, we estimate that the true phenotypes for 95% of genotypes lie between the two dashed lines ($2\sigma_{\text{HOC}} = 0.7$). Colors are a histogram of observed binding. (*B* and *C*) The distribution of binding values is bimodally distributed (measured dashed line, inferred solid line) (*B*), while the distribution of inferred additive phenotypes is unimodal (*C*). (*D*) The impact of each possible amino acid substitution on the underlying additive trait. Open squares are the wild-type amino acid. Additive effects and the underlying trait are scaled so that the wild type has trait value 0 and the average mutation changes the trait by distance 1 (*Materials and Methods*).

more widely distributed for deleterious mutations (*SI Appendix*, Fig. S1*B*). Interestingly, the distribution of inferred additive traits is unimodal whereas the observed distribution of binding affinities is bimodal, with one peak around the wild-type affinity and another at low binding (Fig. 1 *B* and *C*). This apparent contradiction is resolved by the form of global epistasis, which maps all sequences with low values of the additive trait to similar values for the observed phenotype, creating a second mode.

Taken together, our model for the underlying additive trait and for the nonlinear function ϕ provides a simple explanation for the observed binding data. The model accounts for essentially all of the variation in binding measurements beyond what would be expected under measurement noise alone. In particular, the root-mean-square error in our predictions is $\sqrt{\sum (y-\hat{y})^2/N} = 0.59$ whereas even under a perfect model we would expect a root-mean-square error of $\sqrt{\sum \sigma_m^2/N} = 0.52$ due to the Poisson variation in the counts used to calculate the binding scores. The extent of these errors is very small compared with the overall range of phenotypes observed, and it constitutes a sequence–function relationship that is dominated by global epistasis with little room for epistasis of other kinds. This can be understood visually by examining Fig. 1*A*, where we predict that the true phenotypes for 95% of genotypes fall between the two dashed lines.

Complex Epistasis in GB1

While our model captures the form of global epistasis and quantifies the extent of remaining HOC epistasis, deviations from our model can provide evidence for one remaining type of epistasis, which we call complex epistasis. This form of epistasis arises from interactions between specific combinations of sites. Unlike global epistasis, which depends in a nonlinear way on a single additive trait, or HOC epistasis, which consists of independent Gaussian effects associated with each genotype, complex epistasis can be attributed to nonlinear interactions between specific sites on the measured phenotype.

To investigate complex epistasis we compared the predictions of our model, fitted to data from Olson et al. (37), to an independent dataset from a follow-up experiment by ref. 38. Wu et al. (38) measured the binding of GB1 variants mutated to all possible combinations of amino acids at four sites specifically chosen because they were suspected to exhibit idiosyncratic epistatic interactions.

On the whole, our results confirm that the pattern of global epistasis inferred using shallow mutagenesis across the whole domain persists at a deeper level of mutagenesis for these four sites. In particular, the predictions of the global epistasis model, fitted to Olsen et al. (37) data, are close to the mean binding curve measured by Wu et al. (38), across the entire range of the inferred additive trait (*SI Appendix*, Fig. S2*A*). However, the width of the deviations from our binding predictions is much broader than the variation predicted by the error model fitted to the Olsen et al. (37) experiment (Fig. 1*A*). Some of this difference can be attributed to the much larger measurement noise in the ref. 38 dataset [$\sqrt{\sum \sigma_m^2/N} = 0.52$ (Olsen et al.) (37) vs. $\sqrt{\sum \sigma_m^2/N} = 0.77$ (Wu et al.) (38)] due to the lower read depth in Wu et al. (38). However, the deviations from the model are also skewed—so that around 13% of variants measured by Wu et al. (38) have significantly higher binding than predicted, even after the difference in measurement noise is taken into account (*SI Appendix*, Fig. S2*B*). We interpret this skew as evidence of

positive complex epistasis among the specific sites mutagenized by Wu et al. (38).

To infer which combinations of sites and mutations in GB1 exhibit complex epistasis we marginalize the distribution of deviations from our model predictions for each pair of substitutions. Pairs of mutations with large mean-square deviations (*SI Appendix*, Fig. S2*C*) indicate systematic deviation from the global epistasis model. For example, the pair of mutations 41L/54G shows the largest deviation in our analysis, and indeed it is known to be associated with structural changes in GB1 (37). By contrast, the pair 41Q/54P exhibits a similar magnitude of pairwise epistasis calculated directly from binding data and yet, in our analysis, the deviations from the model are small, indicating this pairwise epistasis is primarily caused by global epistasis. Accounting for global epistasis can therefore be instrumental in identifying the specific sites that have idiosyncratic, complex interactions for a measured phenotype.

Global Epistasis in GFP

The GB1 dataset of Olsen et al. (37) contains only single and double mutants, which covers a limited cloud of sequence space near the wild-type protein sequence. To characterize genotype–phenotype maps more broadly we applied our model of global epistasis to a DMS study that measured the fluorescence of 51,715 variants of a green fluorescent protein (avGFP), including up to 11 mutations per sequence and 3.7 mutations on average (31).

Applied to the GFP data, our model of global epistasis infers a sharp threshold that delineates fluorescing from nonfluorescing proteins, with remarkably few outliers (true positive rate 0.9966, true negative rate 0.9787, where we define positive sequences to have log relative fluorescence of −1.25 or greater; Fig. 2).

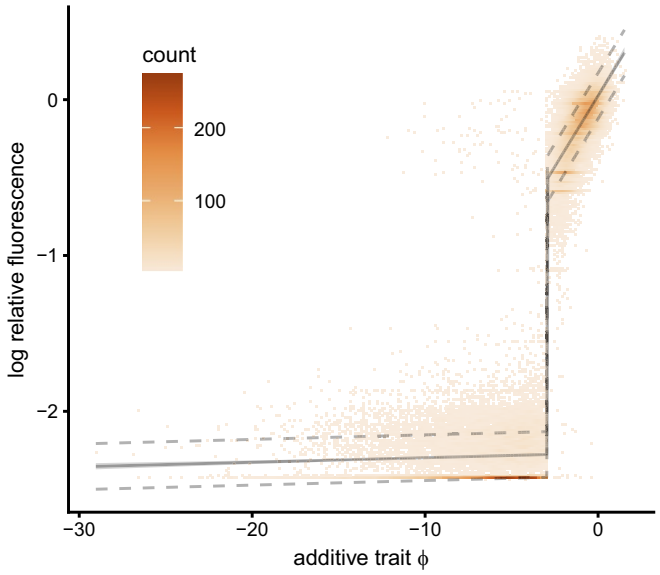

Fig. 2. The shape of global epistasis in a DMS study of GFP (31) shows a sharp threshold in the additive trait. Below the threshold there is low fluorescence and above the threshold fluorescence is high and the slope of the nonlinear mapping is positive (likelihood-ratio test, $P < 0.003$). Data consist of 51,715 protein variants, with 3.7 mutations on average. HOC epistasis has magnitude $\sigma_{\mathrm{HOC}} = 0.073$, and 10-fold cross-validated $r^2 = 0.931 \pm 0.003$. $g(\phi)$ is indicated by the black solid line with light gray (too small to be visible) indicating the 95% bootstrap CI. Our analysis suggests that the true values for 95% of genotypes will lie between the dashed lines. The underlying additive trait is scaled so that the wild type has trait value 0 and the mean absolute magnitude mutation changes the trait by distance 1 (*Materials and Methods*). Additive effects are shown in *SI Appendix*, Fig. S4.

Below the threshold, where the fluorescence is at background levels, the trait–fluorescence relation is almost flat (slope 0.00032 log fluorescence per mutation, 95% CI 0.00022–0.00041); while above the threshold the global epistasis function is increasing (likelihood-ratio test, $P < 0.003$; *SI Appendix*, Table S1; slope 0.020 log fluorescence per mutation, 95% CI 0.019–0.021). In other words, mutations observed to have no effect below the threshold may have a substantial effect in a different genetic background.

Comparing the epistatic and nonepistatic models (*SI Appendix*, Fig. S3*A*) shows that including global epistasis improves the fit substantially ($r^2_{CV} = 0.935$ compared with $r^2_{CV} = 0.688$ for the nonepistatic model), and the inferred fluorescent variance attributable to HOC epistasis is much smaller after accounting for global epistasis ($\sigma_{\mathrm{HOC}} = 0.073$ vs. $\sigma_{\mathrm{HOC}} = 0.31$). Whereas for GB1 we found very small CIs both for $g(\phi)$ and for the additive effects of mutations on the underlying trait (e.g., *SI Appendix*, Fig. S1*D*), for GFP we find that the bootstrapped 95% CIs for $g(\phi)$ are very small (mean 0.18), but that the inferred additive coefficients (*SI Appendix*, Fig. S3*C*) show much greater uncertainty than observed for GB1. In particular, we find that only 62% of the additive coefficients have CIs excluding zero (141 positive and 1,131 negative coefficients out of 1,810 total), and the average width of these CIs was equal to 0.91 times the magnitude of an average mutation. These results show the importance of quantifying uncertainty, and they suggest that although the GFP experiment had sufficient enough power to infer the form of global epistasis, it was underpowered with respect to inferring the effects of individual mutations on the unobserved, additive trait.

Overall, our results indicate that the sequence–function relationship for GFP can be adequately understood in terms of a sharp threshold imposed on an underlying additive trait. Below this threshold, fluorescence is zero, and above this threshold fluorescence is essentially additive in the effects of mutations. More specifically, our model suggests that 95% of genotypes will fall between the two dashed lines in Fig. 2, and the error in our predictions is little more than would be expected under measurement noise alone (mean-square error in our predictions is $\sum(y - \hat{y})^2/N = 0.18$, whereas we would expect a root-mean-square error of 0.16 due to measurement error alone; *Materials and Methods*). Thus, like GB1, we find that global epistasis plays a dominant role in the GFP genotype–phenotype map, with little need to invoke other forms of epistasis to explain the observed data. These results are essentially consistent with the results of the neural network approach used by ref. 31 in their original analysis (based on a neural network architecture with a sharp bottleneck to represent the latent trait), but our approach provides a simpler, easier to understand picture of the sequence–function relationship together with better quantification of uncertainty.

Genotype-by-Environment Interactions in β-Lactamase

Deep mutational scanning studies often measure phenotypes across a library of variants in several different conditions—such as antibiotic activity at several concentrations of antibiotic. While testing in multiple conditions can in principle provide more insight into the function of a protein and should provide more robust results than measurements in a single condition, it also introduces a problem of interpretability. In particular, it becomes difficult to summarize the effects of any given mutation across the experimental conditions. Here we address this problem by extending our model to analyze genotype-by-environment interactions. We do this by assuming that the genotype and environmental condition each contribute additively to the latent underlying trait ϕ, which then determines the observed phenotype via a nonlinear function. While not universally applicable, such models make biological sense in cases where both the

genotype and the environment can act on some latent underlying trait. For instance, it is reasonable to assume that the action of an antibiotic depends on its "local" concentration in the cellular or extracellular environment, which can be modulated either through changes to the activity of an enzyme that degrades the antibiotic or by changes to the overall concentration of antibiotic in the medium.

Here we apply this approach to a study of β-lactamase, an enzyme that degrades β-lactam antibiotics, which has been a model system for DMS (30, 39, 40) and protein evolution (41–44). Stiffler et al. (45) measured the effects of 4,997 single mutants of TEM-1 β-lactamase in five different concentrations of the antibiotic ampicillin. They quantified β-lactamase activity by measuring the growth rate of each mutant in each condition. Our model takes these five growth rate measurements and synthesizes them into a single score capturing the activity of that particular genotype, together with a nonlinear function capturing the mapping from the latent trait (e.g., effective local concentration of antibiotic) to the observed growth rate.

The results of fitting this global-epistatic model are shown in Fig. 3, where the x axis displays the activity of each mutant (i.e., the impact of that mutant on the underlying trait), and the environmental conditions determine a series of curves, one for each condition. Due to our assumption that the environment and genotype interact additively to determine the underlying trait, these curves are simply translations of one another, and each curve produces a prediction of the growth rate of each mutant for one environment. Overall, we find the model produces a very good fit to the data (cross-validated $r^2 = 0.92$), far better than a purely additive model ($P < 0.0003$; *SI Appendix*, Table S1). The performance is also comparable to that of the biophysically based mechanistic model originally fitted by ref. 45, which uses approximately 5,000 more parameters.

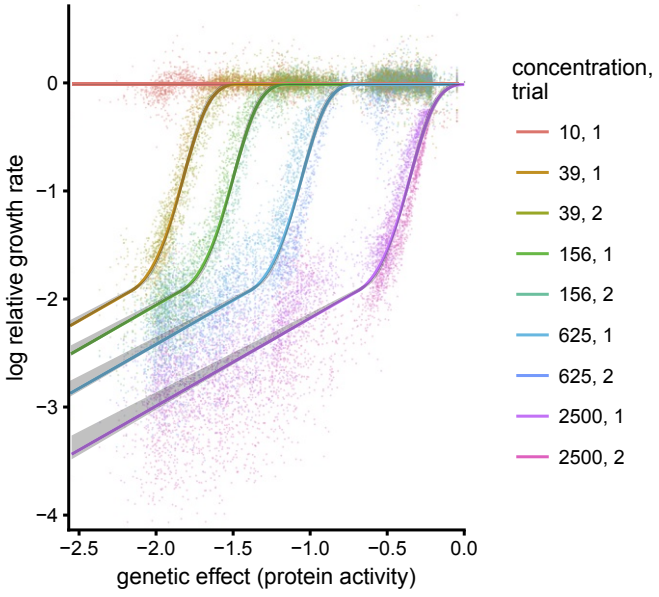

Fig. 3. Gene-by-environment interactions in a DMS study of β-lactamase (45). Data consist of 4,997 single mutants, measuring growth rate under different concentrations of antibiotic (ampicillin) and two replicates (45). Log relative growth rate is plotted against the inferred protein activity for each substitution; the mapping between protein activity for log relative growth rate for each antibiotic concentration is given by the colored curves and each such curve is surrounded by a gray region giving its 95% CI. Growth rate measurements were made using two biological replicates, shown as two slightly different hues for each antibiotic concentration. Cross-validated $r^2 = 0.923 \pm 0.002$.

The global-epistasis model provides a score for each mutation's additive effect on the underlying trait. We again normalize so that the wild type has a score equal to zero and the mean absolute effect of a mutation on the additive trait equals 1. We infer that the single mutation effects (*SI Appendix*, Fig. S5) are largely deleterious: 3,295 out of 3,312 single mutations have CI below zero, and none has CI above zero (*SI Appendix*, Fig. S6*B*), consistent with the observation that no mutant displays a growth rate consistently higher than wild type.

The inferred activity scores are mapped to growth rates via a nonlinear function that shifts, depending on the antibiotic concentration. This function shows first increasing and then decreasing returns (for instance, in the highest antibiotic concentration, the second derivative has positive 95% CI for $-0.61 < \phi < -0.38$ and negative CI for $-0.33 < \phi < -0.09$, where ϕ is the genetic effect on the underlying trait), and it ultimately saturates with zero slope at high concentrations ($P = 0.39$; *SI Appendix*, Table S1). These results indicate a threshold-like phenomenon where, for any given genotype, increasing the antibiotic concentration up to a certain point produces no growth defect, whereas further increases in antibiotic produce rapidly declining growth rates. Interestingly, while the slope of the nonlinear function becomes shallower for low-activity genotypes (slope 0.090 fitness change per mutation on left-hand side, CI 0.073–0.095), it remains significantly greater than 0 ($P < .0003$; *SI Appendix*, Table S1). This result may have clinical relevance: Although growth rates experience a sudden decline with increasing antibiotic concentration, increasing the concentration beyond the point of first inhibition is still capable of producing a further decrease in bacterial growth rate. Finally we tested the hypothesis that the environmental coefficients were a linear function of the log antibiotic concentration vs. a more general model that fitted an arbitrary coefficient for each condition. We found strong evidence against the linear model ($P < 0.003$; *SI Appendix*, Table S1), consistent with the visual pattern of increasingly severe antibiotic effects at higher concentrations (Fig. 3).

The simplicity of our model together with its graphical representation provides insights into the design and behavior of this experiment. For instance, in Fig. 3 we have shown the two biological replicates in each condition in slightly different colors. The plot shows that the replicates are not completely overlapping, which indicates biological variation between the replicates, and indeed we can reject our basic model in favor of a model with a different environmental coefficient for each antibiotic concentration in each replicate ($P < 0.003$; *SI Appendix*, Table S1). Another fine-scale pattern can be observed in the CIs for the additive effects (*SI Appendix*, Fig. S6*B*), which widen for genotypes that have WT-like growth rates at one antibiotic concentration and then very low growth rates in the next concentration. This suggests a finer grid of antibiotic concentrations would be optimal for future experiments.

Additive Traits and Biophysical Quantities

Since we infer an underlying additive trait when fitting a global epistasis model, it is natural to ask whether the underlying trait corresponds to some known biophysical quantity. While this question will surely depend upon the details of the assay used in a given DMS experiment—e.g., whether the assay is measuring binding of a protein to a substrate, enzymatic activity, etc.—thermodynamic stability for a protein's native conformation is almost always an important phenotype, and there is a consistent thread in the literature suggesting that the phenotypic effects of most mutations are mediated via their effects on thermodynamic stability and the nonlinear relationship between free energy of folding and probability of being folded (23–25, 46–48). Moreover, low-throughput measurements of mutational effects on thermodynamic stability ($\Delta\Delta G$) for mutations relative to

WT are typically additive (49, 50) and relatively conserved over evolutionary time (51). Thus, under the hypothesis that most mutations affect phenotype via their energetic effects of thermal stability, the inferred additive coefficients under our model should be highly correlated with the stability effects of these same mutations.

To test this hypothesis, we calculated the correlation between our additive effects and previously published estimates of the stability effects of these mutations. For protein G, we do not find a correlation between the measured stability effects and our inferred effects on the underlying additive phenotype (*SI Appendix*, Fig. S1*C*; $P = 0.14$), a result consistent with Olson et al.'s (37) observation that mutational effects on the binding phenotype of single mutants around the wild-type sequence are not correlated with their effects on protein stability (37, 52). Olson et al. (37) assayed protein G for binding to IgGFC, and presumably many mutations affect the stability of the binding interface as well as the stability of the fold in this small protein. Therefore, it is not surprising that the effects we infer, which are confounded with binding stability, do not correlate with independent measurements of fold stability. A recent biophysical model that separates binding and folding energies found a strong correlation with independent $\Delta\Delta G$ measurements (53). Similarly for β-lactamase, independent measurements of $\Delta\Delta G$ are not significantly correlated with our inferred additive effects (*SI Appendix*, Fig. S6*A*; $P = 0.08$). For GFP, on the other hand, the inferred additive effects are negatively correlated ($\rho = -0.62$, $P = 2 \times 10^{-16}$; *SI Appendix*, Fig. S3*B*) with computational predictions by ref. 31 of energetic effects for single mutations on stability, as would be predicted if the fitness effects for most mutations were mediated through their effects on folding stability. Thus, while the good fit of our model is consistent with the hypothesis that measured phenotypic effects are mostly determined by a nonlinear function of the free energy of folding (23–25, 46), the inconsistent relationship between our inferred additive effects and the estimated stability effects of mutations provides at best equivocal support for the widely held hypothesis that free energy of folding is the underlying trait.

Discussion

We have proposed a maximum-likelihood framework for inferring genetic interactions from high-throughput mutagenesis experiments. Our key assumption is that genetic interactions arise from a global nonlinear mapping between an unobserved, additive trait and the measured phenotype. Modeling this nonlinear mapping by a flexible class of monotonically increasing functions, we simultaneously infer the form of the nonlinear relationship and the contributions of each possible mutation to the underlying additive trait. Analyzing three well-studied proteins, we have shown that this form of global epistasis can provide a remarkably good fit while providing an easy-to-understand summary of the underlying sequence–function relationship. Our approach also provides a principled statistical framework for testing hypotheses and quantifying uncertainty.

Understanding genotype–phenotype relationships provides a substantial challenge due to the enormous size and large dimensionality of the space of possible genotypes. An ideal model of this relationship would exhibit comprehensible behavior, make accurate predictions, and give mechanistic insight into the underlying biology. In practice, however, models must make trade-offs between these competing demands. Our strategy is based on the gambit that, despite abundant evidence for complex genetic interactions between specific sites, a simple-to-understand model of global epistasis can account for most phenotypic variation. The resulting models can be summarized visually by a pair of graphics: a heat map of coefficients showing the effects of each possible single–amino-acid substitution on an underlying additive trait and a curve showing the shape of the nonlinear relationship between the additive trait and the observed phenotype.

Besides being easy to display, the behavior of the model over sequence space is easy to understand. Because we assume that the nonlinear relationship is monotonically increasing with increasing values of the additive underlying trait, the resulting model is single peaked, and it shows no sign epistasis (54). In particular, the sign of the effect of any particular mutation is constant across backgrounds, and the magnitude of the effect depends only on the current trait value rather than on the details of the genetic sequence (29). At a more technical level, models of this form can be analyzed using the now-classical information-theoretic treatment introduced by Berg and von Hippel (55). In particular, given a global epistasis model one can readily approximate important descriptive statistics such as the fraction of random sequences that achieve at least a given level of functionality [a quantity known in the literature as the "functional information" (56, 57)], as well as more detailed predictions such as the site-specific amino acid use among highly functional sequences (55). This simplicity and analytical tractability contrast with the highly complex, rugged landscapes typically inferred by fitting a model with all possible pairwise interactions between sites, where assessing even qualitative features of the landscape from the fitted coefficients can be extremely challenging. For instance, finding the global maximum in a pairwise interaction fit is a nondeterministic polynomial time (NP)-complete problem (58), whereas the optimal genotype can be read off directly from the heat map of additive coefficients in our model of global epistasis.

Our results show that sometimes a simple model of global epistasis can provide accurate predictions that account for virtually all of the observed experimental variation. Because of the high dimensionality of protein sequence space, even a completely additive model will typically have several thousand parameters. By augmenting an additive model with just a handful of additional parameters, which control the nonlinear function, we can produce a dramatically better fit and capture the majority of observed variation with several hundred times fewer parameters than in pairwise models (e.g., there are $\sim$500,000 pairwise interactions for GB1). The large dimensionality of genotypic space also puts a premium on quantifying uncertainty. Perhaps surprisingly, we find that the inferred CIs for the form of global epistasis are much tighter than the CIs for the additive effects of individual mutations, e.g., in our reanalysis of GFP (31), suggesting that robust measurements of global epistasis will be possible even from limited mutagenesis data.

The global epistasis model also makes reasonable predictions for genotypes far from those sampled in the mutagenesis assay, unlike the qualitatively incorrect out-of-sample predictions exhibited by pairwise models (10, 11). Extrapolation can be easily understood under the global epistasis framework, because it is based on the biologically plausible assumption that the physiological factors responsible for saturation or potentiation operate consistently across genetic backgrounds. Furthermore, our global epistasis model is additive outside the observed range of phenotypes and so it provides highly controlled, conservative behavior when extrapolating to extreme phenotypes.

While the global epistasis model is readily comprehensible, and it is sometimes sufficient to capture the major features of the genotype–phenotype relationship, it is unclear whether the model provides an unambiguous mechanistic interpretation in terms of molecular biology and biophysics. If most epistasis in proteins is indeed due to essentially additive effects on the free energy of folding that are then converted by a nonlinear transformation between folding energy and the probability of folding (23–25, 46–48), then our model should fit well and the theory of stability-mediated epistasis would provide a mechanistic basis

for the observed additive trait. On the other hand, the observation that our model fits well is not alone sufficient to infer the existence of a mechanistically meaningful underlying additive trait. For example, if the phenotype is a saturating function of several underlying traits (e.g., refs. 37, 52, and 59), the additive trait inferred under our model may correspond to an amalgamation of multiple mechanistic traits.

The interpretation of the global epistasis model as the simplest possible epistatic extension of an additive model, together with its capacity to capture a large proportion of the observed epistasis, suggests that it should be used as a standard first analysis of mutagenesis assay data, instead of an additive fit. More expressive models, capable of capturing complex epistasis, can then be used to describe the idiosyncratic interactions between the small set of sites whose interactions deviate substantially from the large-scale patterns identified by the global-epistasis fit. Moreover, differences in experimental protocol can be viewed as changes in the nonlinear mapping from the additive trait to the observed phenotype, and so we expect the inferred additive coefficients to be more consistent and reproducible across laboratories and experiments than the raw measurements. Indeed for TEM-1, we find that our inferred coefficients are more highly correlated with previous measurements of TEM-1 activity (39) than the measured growth rates in any single antibiotic concentration ($r^2 = 0.96$ for coefficients, 0.06, 0.74, 0.90, 0.91, 0.79 in individual conditions).

A subtle, but important, technical note concerns our assumption that the nonlinear global epistasis function $g(\phi)$ is a monotonic function spanned by an I-spline basis. Some restriction on the form of $g(\phi)$ is essential because a model allowing $g(\phi)$ could exactly match the measured phenotypes and would therefore provide no useful simplification. In our case, we restrict the function $g(\phi)$ to be monotonic for comprehensibility—even a quadratic mapping can produce extremely complex patterns of epistasis (60). Although other families of monotonically increasing functions are certainly possible (ref. 61; see also ref 62, which used very similar methods to show there is no global nonlinearity in antibody-binding energies), we have found that a four-element I-spline basis is sufficient for the datasets explored here and adds only a handful of additional parameters relative to an additive model.

In contrast, recent heuristic approaches to model global epistasis essentially suffer from imposing too strong or too weak constraints on the nonlinear function $g(\phi)$. For instance, refs. 31 and 32 use a neural network approach that lacks easy comprehensibility, whereas ref. 33 assumes that the nonlinear function is a power transformation, which is comprehensible but too constrained to capture physiologically realistic patterns of saturation typical of metabolic and developmental systems (13). A subtler distinction between our method and the method proposed by Sailer and Harms (33) is that their fitting procedure assumes that the form of the nonlinearity can be accurately inferred from the observed residuals when fitting a nonepistatic model. While approximately true for the datasets analyzed by Sailer and Harms (33), this is not the case for the more complex nonlinearities analyzed here because the additive coefficients inferred based on our nonlinear mapping sometimes differ substantially from the additive coefficients inferred under a standard nonepistatic fit (e.g., *SI Appendix*, Fig. S1*B*).

The intuition that complicated behavior in a high-dimensional space can often be summarized by learning a nonlinear function of a latent additive trait has a long history in biology—as detailed in the Introduction. And techniques for inferring such relationships have been rediscovered multiple times across a breadth of scientific disciplines [e.g., single-index models in econometrics (63), projection pursuit regression in statistics (64), and linear–nonlinear models in neuroscience (65)]. Here we have shown that these ideas can be productively applied to modeling genotype–phenotype relationships in high-throughput mutagenesis data. While the form of epistasis is very simple and easy to understand, our model nonetheless captures the major features of genetic and genotype-by-environment interaction for GB1, GFP, and TEM-1 β-lactamase. We suggest that it provides a natural first model to apply to any high-throughput mutagenesis dataset.

Materials and Methods

Preprocessing of Genotype–Phenotype Data. As input our procedure requires pairs of genotype–phenotype measurements and optional estimates of measurement error and environmental condition.

For the GB1 datasets (37, 38), we defined a functional score and variance from the sequence counts before and after selection based on a Poisson approximation

$$y = \log \frac{c_1}{c_0} - \log \frac{c_1^{\text{WT}}}{c_0^{\text{WT}}}$$

$$\sigma_m^2 = \frac{1}{c_0} + \frac{1}{c_1} + \frac{1}{c_0^{\text{WT}}} + \frac{1}{c_1^{\text{WT}}},$$

where c_0 and c_1 are the counts before and after selection, respectively, and both include a pseudocount of $\frac{1}{2}$ (66). The score is defined relative to the WT sequence, using counts c_0^{WT} and c_1^{WT}.

For the GFP dataset (31), we used the fluorescence measurement (minus the WT fluorescence) and variance as provided.

For the β-lactamase dataset (45), trials 1 and 2 were combined, with given functional measurements relative to WT (and no estimates of measurement error). The WT sequences were not given, and therefore we added them to our data with value equal to zero for each trial and antibiotic concentration. Mutations were excluded that showed small effects relative to the distribution of neutral variants (absolute difference from WT less than 0.25 under all concentrations of antibiotic) since the effect of mutations that show no fitness defect across all conditions is not well defined under our best-fit model (below).

Inference of Nonepistatic Model. The additive trait $\phi(a)$, defined in Eq. **2**, is the prediction of a nonepistatic model given a sequence a and parameters β_0 and β_{i,a_i}. In practice, sequences are defined as a set of zeros and ones for each site, representing each possible amino acid substitution from the WT (dummy coding), and coefficients are defined consistently such that ϕ can be calculated by matrix multiplication. We assume a Gaussian likelihood $y_a \sim \mathcal{N}(\phi(a), \sigma_{m_a}^2 + \sigma_{\text{HOC}}^2)$ for each observation (representing binding, fluorescence, etc.) with mean equal to the additive trait and variance equal to the given measurement error for each sequence plus the HOC epistasis. We find the $19L + 1$ coefficients for the trait and σ_{HOC}^2 which maximize the total log-likelihood $\sum_a \log \mathcal{N}(\phi(a), \sigma_{m_a}^2 + \sigma_{\text{HOC}}^2)$, where the sum is over all observed sequences. We use a gradient-based algorithm, L-BFGS (NLopt library), with parameters initialized from the coefficients of a nonepistatic model fitted to the data using ordinary least-squares regression.

In the GFP (31) dataset, estimated errors are based on the fluorescence variance of sequences with the same barcode. Many of the sequences appeared only once, with no associated variance estimates. In this case, we extended our likelihood such that if there was measurement error associated with a sequence, we set the total variance in the likelihood to be $\sigma_{\text{HOC}}^2 + \sigma_{m'}^2$. This single new parameter $\sigma_{m'}^2$ is estimated simultaneously with the other parameters. It assumes the same amount of measurement noise for each sequence with only one barcode, which is a reasonable approximation for the GFP data, but may not be reasonable for other experiments.

In the β-lactamase dataset, no measurement noise estimates were used, and the log-likelihood reduces to a mean-square error.

Inference of Global Epistasis. Our model of global epistasis is an extension of the nonepistatic model, replacing ϕ with $g(\phi)$ in the likelihood. The nonlinearity g is implemented by a flexible family of third-order monotonic I-splines (35). I-splines are a basis set of piecewise polynomials, defined by a set of control points, or knots, that are linearly combined to form a monotonic nonlinear function. We found that four evenly spaced knots were sufficient for each analysis and had sufficient flexibility without overfitting, resulting in five basis functions. The nonlinear function is a linear combination of these basis functions plus a constant, $g(\phi) = c_\alpha + \sum_m \alpha_m I_m(\phi)$, with nonnegative α_m coefficients. I-splines are defined only within the knot boundaries, so we extend g to linearly extrapolate beyond the knot

boundaries. We estimate c_α, five α_m coefficients, the additive effects, and the variance parameters simultaneously by maximum likelihood, as described above with $g(\phi)$ replacing ϕ.

While it is possible for the optimization to get stuck in local optima, resulting in a poor fit, the following choice of initial parameters produced very good solutions in all datasets we tested. First, a nonepistatic model was fitted, resulting in $\beta_{i,\alpha}$ and σ^2_{HOC}. The additive effects were then rescaled such that $0 \leq \phi \leq 1$, and these were used as initial parameters for the global epistasis fit. For the initial parameters of the I-spline we chose four evenly spaced knots between zero and one, and we chose the spline coefficients, α_m, that produce a linear function g, by means of a separate optimization holding other parameters fixed. We used these spline coefficients as the initial parameters in the global epistasis fit.

One subtlety in the inference procedure is that the latent additive trait is defined only up to an affine transformation. This occurs because, for any invertible affine transformation A of the underlying phenotype, a corresponding change in the global epistasis can give identical predictions: $g(\phi) = g'(A(\phi))$, where g' is g composed with A^{-1}. During the inference, therefore, we fix the affine transformation by choosing the knot positions and keeping these fixed throughout the maximization. We then postprocess the inferred parameters to have a natural biological interpretation by choosing the affine transformation that sets the WT phenotype to 0 and makes the average absolute value of the additive coefficients equal to 1.

Separating Genetic and Environmental Effects. We separate genetic and environmental effects with a model where the measured phenotype is a nonlinear function of an intermediate trait plus the environmental effect, $g(\phi + \phi_e)$. For a categorical environmental effect where z is the environmental state, $\phi_e = \sum_\gamma \zeta_\gamma \delta_{\gamma,z}$, where ζ_γ is the effect of being in environment γ (relative to the reference environment) and $\delta_{\gamma,z} = 1$ when γ is equal to z and zero otherwise. For a continuous variable $\phi_e = \zeta z$.

As noted in the main text, our best-fit nonlinear function g for the β-lactamase data had slope equal to zero, $\frac{\partial g}{\partial \phi} = 0$, to the right of the highest knot position. Strictly speaking, the associated coefficients $\beta_{i,\alpha}$ for any genotype that is in the flat portion of the curve for all environmental conditions become unidentifiable, as a change in the coefficient does not change the prediction. In these instances, the optimization is allowed to finish, and then the $\beta_{i,\alpha}$ that are affected are decreased to be equal to the rightmost knot position.

When the data consist of single mutants in multiple environments, the cross-validation of the model requires some modification to avoid assigning all data on a given mutation to the test set, since we have no way of making predictions for mutations that have never been observed. To avoid this problem, we generate the training and testing subsets based on the environmental conditions. For each mutation that is observed under several conditions/replicates, we assign a random permutation of integers between one and the number of replicates and conditions (i.e., the nine categories in Fig. 3). These integers define the subsets used for training and testing. For each fold of cross-validation, all observations assigned to a particular integer are used as the test set and the remaining observations are used as training.

Bootstrapped Confidence Intervals. We estimated confidence intervals by a parametric bootstrap based on the maximum-likelihood predictions. The bootstrapped data were generated as $y' = g(\phi) + \sqrt{\sigma^2_{\text{HOC}} + \sigma^2_m}\eta$, where η is an instance of a standard normal random variable. Because the values for the underlying trait are defined only up to an affine transformation, for each bootstrap b, we find parameters that linearly transform the original trait estimates into the bootstrap trait scale, $\phi_b = m_b \phi_{\text{ML}} + c_b$ via least-squares regression, where the ϕ_b and ϕ_{ML} are vectors describing the intermediate traits in the bootstrapped and maximum-likelihood models, and m_b and c_b are the inferred parameters. For calculating the CI of g_{ML}, we make a distribution based on each bootstrapped model $g_b(m_b\chi + c_b)$ for each value of χ. We choose 101 linearly spaced values for χ in the nonlinear region (between zero and one) and 101 linearly spaced values for the each of two linearly extrapolated regions (from the knot boundaries to the maximum/minimum ϕ). For the CI of maximum-likelihood $\beta_{i,\alpha}$ we make a distribution based on $\beta_{i,\alpha,b}/m_b$.

Model Comparison. We compared nested maximum-likelihood models by a bootstrapped likelihood-ratio test. First the difference in log-likelihood was computed between the simpler (null) model and the more complex (hypothesis) model. The likelihood optimization of the more complex model had initial parameters determined by the simpler (null) model. P values were calculated by comparing to a bootstrapped log-likelihood difference distribution, with y' generated by a parametric bootstrap from the null model as described above.

We validated this procedure by bootstrapping the P values for the first test in *SI Appendix*, Table S1. That is, for each bootstrap in the test, we calculate a P value by means of a second (or recursive) likelihood-ratio test based on the bootstrapped data and models (with 100 bootstraps). The resulting P-value distribution (*SI Appendix*, Fig. S7) should be uniform in the ideal case; however, the resulting distribution has excess probability in the center, indicating that the test is somewhat conservative.

ACKNOWLEDGMENTS. We thank Jesse Bloom for comments on the bioRxiv preprint and the anonymous referees for helpful comments. J.O. was supported by NIH Grant T32AI055400, and J.B.P. acknowledges support from the David and Lucile Packard Foundation and the US Army Research Office (W911NF-12-R-0012-04).

1. Kauffman S, Levin S (1987) Towards a general theory of adaptive walks on rugged landscapes. *J Theor Biol* 128:11–45.
2. Kauffman SA (1993) *The Origins of Order: Self Organization and Selection in Evolution* (Oxford Univ Press, New York).
3. Huynen MA, Stadler PF, Fontana W (1996) Smoothness within ruggedness: The role of neutrality in adaptation. *Proc Natl Acad Sci USA* 93:397–401.
4. Fontana W (2002) Modelling 'evo-devo' with RNA. *Bioessays* 24:1164–1177.
5. Fowler DM, Fields S (2014) Deep mutational scanning: A new style of protein science. *Nat Methods* 11:801–807.
6. Jerison ER, Desai MM (2015) Genomic investigations of evolutionary dynamics and epistasis in microbial evolution experiments. *Curr Opin Genet Dev* 35:33–39.
7. Hinkley T, et al. (2011) A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nat Genet* 43:487–489.
8. Otwinowski J, Nemenman I (2013) Genotype to phenotype mapping and the fitness landscape of the E. coli lac promoter. *PLoS One* 8:e61570.
9. Levy RM, Haldane A, Flynn WF (2017) Potts Hamiltonian models of protein covariation, free energy landscapes, and evolutionary fitness. *Curr Opin Struct Biol* 43:55–62.
10. Otwinowski J, Plotkin JB (2014) Inferring fitness landscapes by regression produces biased estimates of epistasis. *Proc Natl Acad Sci USA* 111:E2301–E2309.
11. du Plessis L, Leventhal GE, Bonhoeffer S (2016) How good are statistical models at approximating complex fitness landscapes. *Mol Biol Evol* 33:2454–2468.
12. Wright S (1934) Physiological and evolutionary theories of dominance. *Am Nat* 68:24–53.
13. Kacser H, Burns JA (1981) The molecular basis of dominance. *Genetics* 97:639–666.
14. Sved JA, Reed TE, Bodmer WF (1967) The number of balanced polymorphisms that can be maintained in a natural population. *Genetics* 55:469–481.
15. King JL (1967) Continuously distributed factors affecting fitness. *Genetics* 55:483–492.
16. Milkman RD (1967) Heterosis as a major cause of heterozygosity in nature. *Genetics* 55:493–495.
17. Kimura M, Crow JF (1978) Effect of overall phenotypic selection on genetic change at individual loci. *Proc Natl Acad Sci USA* 75:6168–6171.
18. Kondrashov AS (1995) Contamination of the genome by very slightly deleterious mutations: Why have we not died 100 times over? *J Theor Biol* 175:583–594.
19. Lande R, Arnold SJ (1983) The measurement of selection on correlated characters. *Evolution* 37:1210–1226.
20. Schluter D (1988) Estimating the form of natural selection on a quantitative trait. *Evolution* 42:849–861.
21. Kingsolver JG, et al. (2001) The strength of phenotypic selection in natural populations. *Am Nat* 157:245–261.
22. Berg J, Willmann S, Lässig M (2004) Adaptive evolution of transcription factor binding sites. *BMC Evol Biol* 4:42.
23. Bloom JD, et al. (2005) Thermodynamic prediction of protein neutrality. *Proc Natl Acad Sci USA* 102:606–611.
24. DePristo MA, Weinreich DM, Hartl DL (2005) Missense meanderings in sequence space: A biophysical view of protein evolution. *Nat Rev Genet* 6:678–687.
25. Wylie CS, Shakhnovich EI (2011) A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proc Natl Acad Sci USA* 108:9916–9921.
26. Kondrashov FA, Kondrashov AS (2001) Multidimensional epistasis and the disadvantage of sex. *Proc Natl Acad Sci USA* 98:12089–12092.
27. Kondrashov DA, Kondrashov FA (2015) Topological features of rugged fitness landscapes in sequence space. *Trends Genet* 31:24–33.
28. Starr TN, Thornton JW (2016) Epistasis in protein evolution. *Protein Sci* 25:1204–1218.
29. Kryazhimskiy S, Rice DP, Jerison ER, Desai MM (2014) Global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science* 344:1519–1522.
30. Jacquier H, et al. (2013) Capturing the mutational landscape of the beta-lactamase TEM-1. *Proc Natl Acad Sci USA* 110:13067–13072.
31. Sarkisyan KS, et al. (2016) Local fitness landscape of the green fluorescent protein. *Nature* 533:397–401.
32. Pokusaeva V, et al. (2017) Experimental assay of a fitness landscape on a macroevolutionary scale. bioRxiv:222778.
33. Sailer ZR, Harms MJ (2017) Detecting high-order epistasis in nonlinear genotype-phenotype maps. *Genetics* 205:1079–1088.

34. Szendro IG, Schenk MF, Franke J, Krug J, de Visser JAGM (2013) Quantitative analyses of empirical fitness landscapes. *J Stat Mech Theor Exp* 2013:P01005.
35. Ramsay JO (1988) Monotone regression splines in action. *Stat Sci* 3:425–441.
36. Kingman JF (1978) A simple model for the balance between selection and mutation. *J Appl Probab* 15:1–12.
37. Olson CA, Wu NC, Sun R (2014) A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr Biol* 24:2643–2651.
38. Wu NC, Dai L, Olson CA, Lloyd-Smith JO, Sun R (2016) Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife* 5:e16965.
39. Firnberg E, Labonte JW, Gray JJ, Ostermeier M (2014) A comprehensive, high-resolution map of a gene's fitness landscape. *Mol Biol Evol* 31:1581–1592.
40. Klesmith JR, Bacik JP, Wrenbeck EE, Michalczyk R, Whitehead TA (2017) Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proc Natl Acad Sci USA* 114:2265–2270.
41. Weinreich DM, Delaney NF, Depristo MA, Hartl DL (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* 312:111–114.
42. Novais A, et al. (2010) Evolutionary trajectories of beta-lactamase CTX-m-1 cluster enzymes: Predicting antibiotic resistance. *PLoS Pathog* 6:e1000735.
43. Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M (2016) Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol Biol Evol* 33:268–280.
44. Bloom JD (2014) An experimentally informed evolutionary model improves phylogenetic fit to divergent lactamase homologs. *Mol Biol Evol* 31:1–17.
45. Stiffler M, Hekstra D, Ranganathan R (2015) Evolvability as a function of purifying selection in TEM-1-lactamase. *Cell* 160:882–892.
46. Bershtein S, Segal M, Bekerman R, Tokuriki N, Tawfik DS (2006) Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* 444:929–932.
47. Gong LI, Suchard MA, Bloom JD (2013) Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife* 2:e00631.
48. Dasmeh P, Serohijos AW, Kepp KP, Shakhnovich EI (2014) The influence of selection for protein stability on dN/dS estimations. *Genome Biol Evol* 6:2956–2967.
49. Wells JA (1990) Additivity of mutational effects in proteins. *Biochemistry* 29: 8509–8517.
50. Sandberg WS, Terwilliger TC (1993) Engineering multiple properties of a protein by combinatorial mutagenesis. *Proc Natl Acad Sci USA* 90:8367–8371.
51. Risso VA, et al. (2014) Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history. *Mol Biol Evol* 32:440–455.
52. Wu NC, Olson CA, Sun R (2016) High-throughput identification of protein mutant stability computed from a double mutant fitness landscape. *Protein Sci* 25: 530–539.
53. Otwinowski J (2018) Biophysical inference of epistasis and the effects of mutations on protein stability and function. arXiv:1802.08744.
54. Weinreich DM, Chao L (2005) Rapid evolutionary escape by large populations from local fitness peaks is likely in nature. *Evolution* 59:1175–1182.
55. Berg OG, von Hippel PH (1987) Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters. *J Mol Biol* 193:723–743.
56. Carothers JM, Oestreich SC, Davis JH, Szostak JW (2004) Informational complexity and functional activity of RNA structures. *J Am Chem Soc* 126:5130–5137.
57. Hazen RM, Griffin PL, Carothers JM, Szostak JW (2007) Functional information and the emergence of biocomplexity. *Proc Natl Acad Sci USA* 104:8574–8581.
58. Barahona F (1982) On the computational complexity of Ising spin glass models. *J Phys A Math Gen* 15:3241–3253.
59. Manhart M, Morozov AV (2015) Protein folding and binding can emerge as evolutionary spandrels through structural coupling. *Proc Natl Acad Sci USA* 112:1797–1802.
60. Hwang S, Park SC, Krug J (2017) Genotypic complexity of Fisher's geometric model. *Genetics* 206:1049–1079.
61. Ramsay JO (1998) Estimating smooth monotone functions. *J R Stat Soc Ser B Stat Methodol* 60:365–375.
62. Adams RM, Kinney JB, Walczak AM, Mora T (2017) Physical epistatic landscape of antibody binding affinity. arXiv:1712.04000 [q-bio].
63. Li Q, Racine JS (2007) *Nonparametric Econometrics: Theory and Practice* (Princeton Univ Press, Princeton).
64. Friedman JH, Stuetzle W (1981) Projection pursuit regression. *J Am Stat Assoc* 76: 817–823.
65. Atencio CA, Sharpee TO, Schreiner CE (2008) Cooperative nonlinearities in auditory cortical neurons. *Neuron* 58:956–966.
66. Plackett RL (1981) *The Analysis of Categorical Data* (MacMillan, New York), 2nd Ed.