PNAS RESEARCH ARTICLE | GENETICS STATISTICS OPEN ACCESS

Distinguishing direct interactions from global epistasis using rank statistics

Maryn O. Carlson[a,b,1], Bryan L. Andrews[c,d,e,1], and Yuval B. Simons[a,f,g,1]

Affiliations are included on p. 10.

Edited by Marcus Feldman, Stanford University, Stanford, CA; received April 28, 2025; accepted August 11, 2025

The phenotypic effect of a mutation may depend on the genetic background in which it occurs, a phenomenon referred to as epistasis. One source of epistasis in proteins is direct interactions between residues in close physical proximity to one another. However, epistasis may also occur in the absence of specific interactions between amino acids if the genotype-to-phenotype map is nonlinear. Disentangling the contributions of these two phenomena—specific and global epistasis—from noisy, high-throughput mutagenesis experiments is highly nontrivial: The form of the nonlinearity is generally not known and model misspecification may lead to over- or underestimation of specific epistasis. In contrast to previous approaches, we do not attempt to model the fitness measurements directly. Rather, we begin with the observation that global epistasis, under the assumption of monotonicity, imposes strong constraints on the rank statistics of a combinatorial mutagenesis experiment. Namely, the rank-order of mutant phenotypes should be preserved across genetic backgrounds. We exploit this constraint to devise a simple semiparametric method to detect specific epistasis in the presence of global epistasis and measurement noise. We apply this method to three high-throughput mutagenesis experiments, uncovering known protein contacts with similar accuracy to existing, more complicated procedures. Our method immediately generalizes beyond proteins, providing a simple, yet powerful framework for interpreting the epistasis observed in combinatorial datasets.

global epistasis | genotype-to-phenotype map | deep mutational scanning | protein | fitness landscape

Significance

Epistasis, the dependence of a mutation's effect on its genetic background, may arise from direct interactions or from choice of scale. Distinguishing these two sources of epistasis is fundamental to our understanding of fitness landscapes, but represents an outstanding statistical challenge. We propose rank statistics as a natural framework to tease apart these two sources of epistasis: Under monotonic transformations of the measurement scale, the rank order of mutational effects is preserved across different genetic backgrounds. Based on this idea, we develop a rank-based method for detecting specific interactions between mutations. Applying this method to combinatorial mutagenesis experiments reveals that it is possible to accurately detect direct interactions from noisy data without assuming or estimating the form of global epistasis.

The question of measurement scale has been central to the definition and detection of genetic interactions—referred to henceforth as epistasis—since R. A. Fisher introduced the notion of "epistacy" in the context of quantitative traits (1, 2). For example, the analysis of genetic effects which combine multiplicatively on an additive scale would result in the appearance of widespread epistasis—a phenomenon Fisher referred to as "metrical bias" (3). He supposed that metrical bias could often be removed by applying an appropriate, likely rank-preserving, transformation (3–6).

The question of the appropriate measurement scale has continued to animate studies of interactions across disciplines (e.g., refs. 7–11), including protein biophysics (e.g., refs. 12 and 13). Here, contemporary mutagenesis experiments, referred to as deep mutational scans (DMSs), assay the phenotypes of thousands to millions of protein mutants simultaneously with high-throughput sequencing-based methods (14, 15). In DMSs assaying the combined fitness effects of two or more mutations (Fig. 1 *A* and *B*), observed epistasis is often categorized into two types: specific epistasis (SE), where the effect of a mutation at one position depends on the identity of the amino acid at another via a direct interaction; and global epistasis (GE)—the analog of Fisher's metrical bias—where apparent interactions between mutations emerge from the presence of nonlinearities in the genotype-to-phenotype map (12, 13, 16, 17). The former, SE, is most often associated with amino acids in close proximity in the protein structure (Fig. 1*C*; 18, 19), while GE may be attributed to many causes, including a thermodynamic equilibrium between conformational states or detection limits imposed by an experimental assay (Fig. 1*D*; 12, 13, 16, 17, 20, 21). Another body of work suggests that GE can arise from widespread SE (22–24), a scenario which we return to in *Discussion*. The extent to which SE and GE contribute to the epistasis observed in a given experiment and, in turn, shape protein fitness landscapes is an open question (25–27), with consequences for how proteins evolve and function (28–30). In the simulated example in Fig. 1*E*, GE almost completely obscures SE.

Author contributions: M.O.C., B.L.A., and Y.B.S. designed research; M.O.C., B.L.A., and Y.B.S. performed research; M.O.C., B.L.A., and Y.B.S. contributed new reagents/analytic tools; M.O.C. analyzed data; M.O.C. and B.L.A. created the figures and revised the paper; Y.B.S. revised the paper; and M.O.C. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: mocarlson@uchicago.edu, andrewsb@uchicago.edu, or yuval.simons@uchicago.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2509444122/-/DCSupplemental.

Published September 23, 2025.

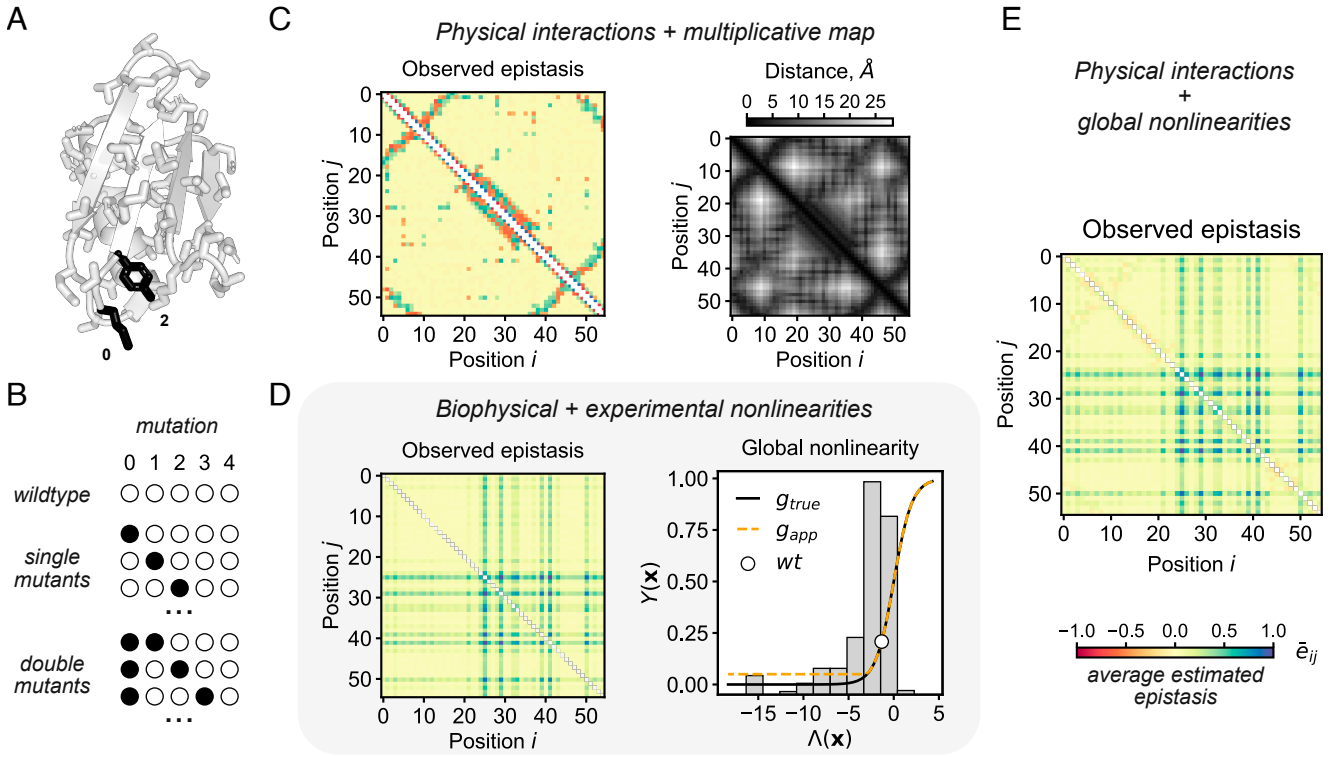

Fig. 1. Potential sources of observed epistasis in proteins. (*A*) A crystal structure of protein GB1 (PDB: 2J52). Mutated positions, here 0 and 2, are denoted in black. (*B*) In a deep mutational scan, mutations (black circles) are introduced into a wildtype background (white circles). (*C*) If the genotype-to-phenotype map is multiplicative, as is simulated here, we expect the detected epistasis, $\bar{e}_{ij}$ (*Left*), to reflect specific epistasis induced by the contact map (*Right*, derived from *A*), when epistatic effects are computed with respect to a multiplicative fitness map. (*D*) Nonlinearities in the genotype-to-phenotype (**x** to Λ to *Y*) map of biological (*Right*, g_{true}, solid black line) and/or experimental origin (g_{app}, dashed orange line), may introduce epistasis (*Left*), even in the absence of specific interactions. Here, the preponderance of deleterious mutations—illustrated in the histogram of simulated single mutant Λ values (*Right*)—coupled with saturation at the lower end of the measurement range induces widespread positive epistasis. (*E*) Observed epistasis from simulations when both physical interactions and global nonlinearities determine the genotype-to-phenotype map. GE may significantly obscure SE, as is simulated here. For simulation procedures, see *SI Appendix*, section 1H.

Formally, under a model of GE, each single mutation i has an independent effect λ_i on a latent additive trait Λ, where Λ may, for example, correspond to the energy associated with protein folding or ligand binding (e.g., ref. 31). The phenotype, Y, is a potentially nonlinear, monotonic function g of Λ,

$$\Lambda(\mathbf{x}) \coloneqq \Lambda_{wt} + \sum_{i=1}^{L} \lambda_i x_i \quad \text{and} \quad Y(\mathbf{x}) = g[\Lambda(\mathbf{x})], \qquad \textbf{[1]}$$

where Λ_{wt} is the value of Λ for the wildtype sequence (Fig. 1 *D*, *Right*; e.g., ref. 13). For exposition, we let $\mathbf{x} \in \{0,1\}^L$ be a binary, L-length protein sequence. In practice, mutations often occupy a larger state space (e.g., the 20 amino acids).

We consider two factors that may result in deviations from Eq. **1**. A set of mutation pairs i and j may exhibit second-order effects, λ_{ij}, on the additive trait (Eq. **2a**). We refer to the λ_{ij} as specific epistatic effects. In addition, the observed phenotype may be a noisy measurement of Y. Incorporating these two features, the estimated phenotype associated with **x**, $\hat{Y}(\mathbf{x})$, is modeled as,

$$\Lambda(\mathbf{x}) \coloneqq \Lambda_{wt} + \sum_{i=1}^{L} \lambda_i x_i + \sum_{j<i} \lambda_{ij} x_i x_j \qquad \textbf{[2a]}$$

$$\hat{Y}(\mathbf{x}) \coloneqq g[\Lambda(\mathbf{x})] + \epsilon, \qquad \textbf{[2b]}$$

where ϵ is the measurement error with a potentially unknown distribution (e.g., ref. 13).

In the event that g is nonlinear, estimation of the λ_{ij} under the assumption of an additive model—equivalent to assuming that g is a linear function—may lead to spurious inference of many nonzero higher-order coefficients (compare Fig. 1 *C*–*E*). More broadly, misspecification of the form of g is apt to distort features of the genotype-to-phenotype map, over- or underrepresenting the importance of higher-order interactions (12, 13, 31).

Thus motivated, researchers have developed several methods to account for GE when estimating genotype-to-phenotype maps. Several of these methods rely on strong assumptions about the form of the nonlinearity: For example, assuming that g is logistic (25), follows from a thermodynamic model (31, 32), or is otherwise of a prespecified form (12, 21, 33, 34). Although other procedures impose fewer modeling assumptions, they do not provide a principled hypothesis testing framework for identifying epistasis (19), or they involve potentially cumbersome fitting procedures (13, 21, 35).

Indeed, developing a hypothesis testing framework in the presence of GE is highly nontrivial. First, fitness measurements are often derived from noisy, sequencing-based assays with systematic variation in precision across the measurement range (36, *SI Appendix*, section 1H). This heteroskedasticity arises from the fact that less fit variants are associated with relatively fewer read counts and implies that statistical power to detect SE in a well-calibrated statistical test should similarly vary across the measurement range. Second, if not properly accounted for, uncertainty in the estimation of g may lead to over- or underestimation of the prevalence of SE.

In contrast to previous approaches for detecting SE in the presence of GE, we do not attempt to explicitly estimate the form of the nonlinearity, nor to model the fitness measurements directly. We do not even assume that g is nonlinear. Rather, our work begins with the observation that GE imposes strong constraints on the rank statistics of a DMS. Namely, if the latent space is unidimensional, which we will assume going forward (though see *Model Misspecification* and *SI Appendix*, section 2D), and if the nonlinearity g is monotonic and strictly increasing (or decreasing), then g is also order preserving in the absence of measurement noise. In other words, when g is monotonic, the ordering of mutations is shared across genetic backgrounds. This observation similarly motivated (37) to introduce a rank-based loss function for phenotypic prediction in the context of GE. Our work further exploits the (assumed) monotonicity of GE to detect SE, which disrupts the ordering of mutations.

We first demonstrate that rank statistics are a natural framework for the analysis of combinatorial datasets, a notion which has only recently been developed in the literature (38, 39) and applied to protein DMSs (37). We then define a rank-based measure of SE and use it to develop a semiparametric test for deviations of mutation pairs from GE that accounts for heteroskedasticity, which may arise due to variation in the shape of g or scale of the measurement noise (or both) across the measurement range. Our procedure—referred to as Resample and Reorder or R&R—requires minimal preprocessing of the data beyond generating variant read counts, is invariant under monotonic transformations of the data, and is agnostic to the form of the nonlinearity beyond monotonicity. We apply R&R to simulated and empirical DMSs of proteins, demonstrating its ability to recover true epistatic effects and physical contacts, respectively. Finally, we explore the consequences of misspecification of the GE model on the results of our inference procedure. In particular, we consider a scenario where there are two, rather than one, latent additive traits. While we motivate R&R using DMSs of proteins, the method generalizes almost immediately to other types of combinatorial datasets.

Modeling Framework

Rank Statistics As a Natural Framework for Global Epistasis. Under the assumption of GE (Eq. **1**), the monotonicity of the nonlinearity g implies that the rank-order of mutations should be preserved regardless of the background in which the mutations occur. As a consequence, in the absence of measurement noise and SE, the Spearman's correlation between mutant phenotypes measured in distinct backgrounds i and j, $\hat{\rho}_{ij}$, is equal to one. Measurement noise, however, may result in Spearman's correlations that deviate substantially from one, even when SE is sparse or absent (Fig. 2*B*).

Consider the ordering of two mutations m and n in the background of mutation i. If the difference between their true fitness values, Y_{im} and Y_{in}, is small relative to the magnitude of the measurement noise, σ_ϵ, then each of the possible orderings of their *estimated* fitness values, $\hat{Y}_{im} > \hat{Y}_{in}$ and $\hat{Y}_{im} < \hat{Y}_{in}$, is approximately equally likely. In the absence of SE, such small differences arise when 1) differences in the mutations' effects on the latent trait, λ_m and λ_n, are small; 2) the slope of the nonlinearity g in the neighborhood of the background i is small; or, 3) σ_ϵ is large. More succinctly,

$$\mathbb{P}\{\hat{Y}_{im} > \hat{Y}_{in}\} \approx \frac{1}{2} \iff g'(\lambda_i)[\lambda_m - \lambda_n] \ll \sigma_\epsilon, \quad [3]$$

where, for the sake of exposition, we have assumed that both λ_m and λ_n are small (*SI Appendix*, section 2F). Therefore, variation in the slope of the nonlinearity and the magnitude of noise across the measurement range may induce systematic variation in $\hat{\rho}_{ij}$ as a function of single mutant fitness (Fig. 2*B*).

To illustrate how the presence of a nonlinearity and noise introduce variation in the $\hat{\rho}_{ij}$ values among mutant backgrounds, we simulate a DMS under the assumption of a two-state thermodynamic model where the effects of single mutations on binding are specified by their estimated values from Otwinowski (31) (Fig. 2*A*). In addition, we introduce SE between amino acids at nearby positions ($\leq$5 Å; *SI Appendix*, section 1H). An important feature of this model is saturation at both very small and large values of Λ: In the background of a very deleterious mutation, additional deleterious mutations will not further reduce fitness, and vice versa for very fit mutations.

At these two extremes, fitness differences among mutants will be small relative to measurement noise, and thus, the double mutant phenotypes will be approximately uniformly ordered. This implies that for a very deleterious (or beneficial) mutation i, $\hat{\rho}_{ij} \approx 0$ for all mutations $j \neq i$. In addition, row maximum will systematically increase with single mutant rank until reaching a critical rank at which it begins to decrease (Fig. 2 *B* and *C* and see *SI Appendix*, section 2F). In these GB1-based simulations, reductions in the row maximum of $\hat{\rho}_{ij}$ among the fittest mutations are small due to the oversampling of deleterious mutations (Fig. 1 *D*, *Right* and 2 *C*, *Right*). Additionally, SE is sufficiently sparse to preserve the systematic variation in $\hat{\rho}_{ij}$ induced by GE.

In other words, the estimated rank of a mutation m in the background of a very deleterious mutation i, $\hat{R}_{im}$, is uncorrelated with its estimated single mutant rank, $\hat{R}_m$ (Fig. 2 *C*, *Left*). In contrast, mutations in the background of a typical mutation will be well-ordered, with small deviations due to measurement noise and large deviations due to SE (Fig. 2*C*, *Center*). We will exploit the latter feature to detect SE.

Detecting Specific Epistasis. We formally define SE as the presence of a nonzero interaction between two mutations i and m, with respect to the latent additive trait Λ, i.e., $\lambda_{im} \neq 0$ in Eq. **2a**. If the magnitude of λ_{im} is large enough, it will lead to deviations from the expected ordering under a model of GE (Eq. **1**). For exposition, consider again the ordering of the phenotypes of two mutations m and n in the background of mutation i. Suppose, without loss of generality, that $\lambda_n > \lambda_m$ and, in accordance, the estimated fitness of single mutant n exceeds that of mutant m, i.e., $\hat{Y}_n > \hat{Y}_m$. If mutations i and m exhibit SE, and $\lambda_{im} > 0$, the probability that $\hat{Y}_{im}$ is greater than $\hat{Y}_{in}$ can be approximated by,

$$\mathbb{P}\{\hat{Y}_{im} > \hat{Y}_{in}\} \approx \mathbb{P}\{g'(\lambda_i)[\lambda_m - \lambda_n + \lambda_{im}] > \epsilon_{im} - \epsilon_{in}\}, \quad [4]$$

when λ_m, λ_n, and λ_{im} are small (and $\lambda_{in} = 0$); and, ϵ_{im} and ϵ_{in} are the measurement errors of the double mutants (*SI Appendix*, section 2F). Therefore, when the epistatic effect λ_{im} exceeds a threshold set by the measurement noise, the slope of the nonlinearity in the neighborhood of λ_i, and the difference in first-order fitness effects, i.e.,

$$\lambda_{im} \gtrsim \frac{\sigma_\epsilon}{g'(\lambda_i)} + (\lambda_n - \lambda_m), \quad [5]$$

SE will likely result in a change in the ordering of the double mutants with respect to that of the single mutants: $\hat{Y}_{im} > \hat{Y}_{in}$ while $\hat{Y}_m < \hat{Y}_n$ (*SI Appendix*, section 2F). Moreover, for any n

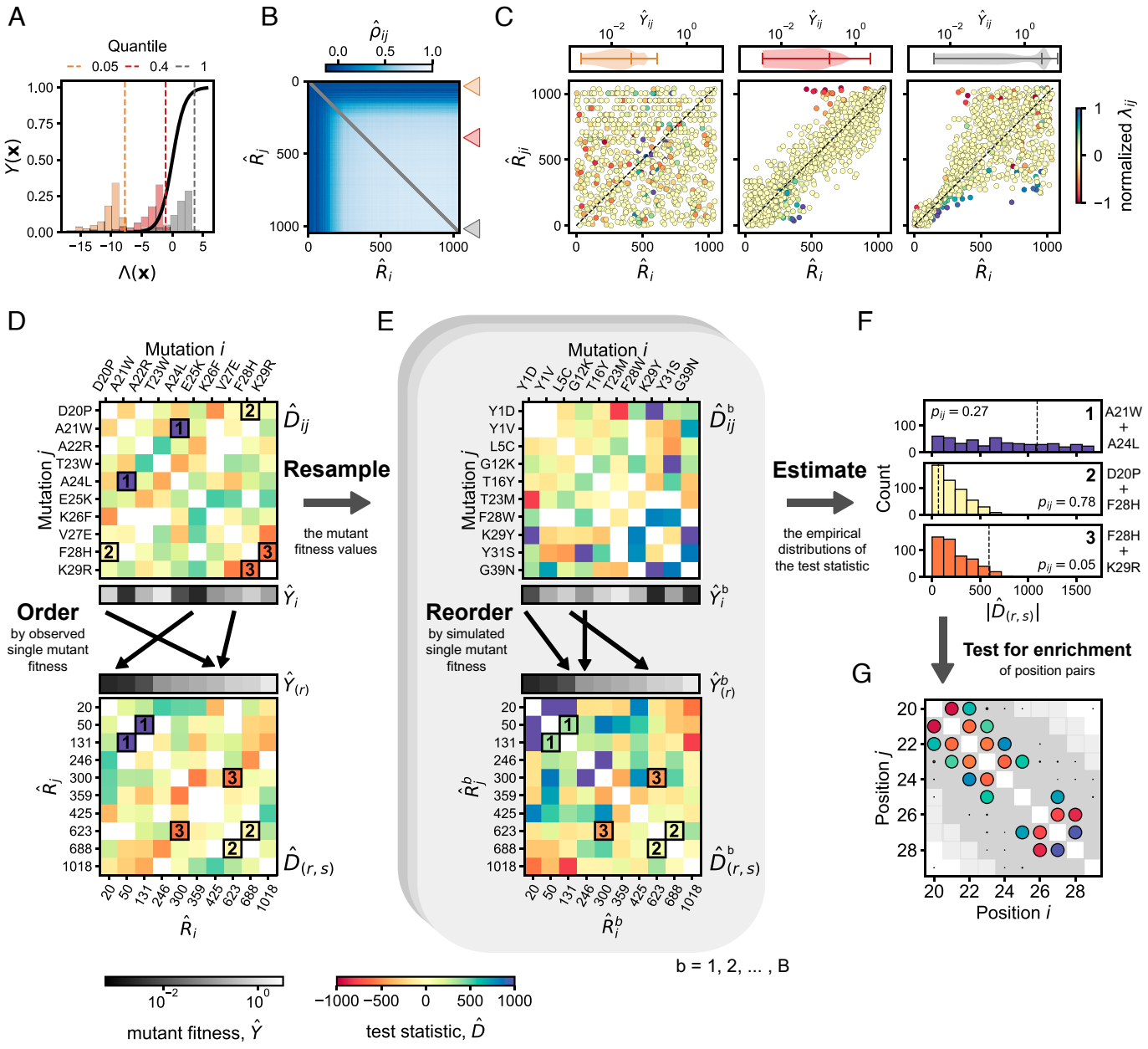

Fig. 2. Rank statistics as a natural framework for detecting specific epistasis in the presence of global epistasis. (*A*) Probability of binding, Y, as a function of the latent trait Λ (solid black line). Three focal mutations, in brown, red, and gray, with distinct Λ values (dashed vertical lines) from a deep mutational scan simulated under the assumption of a two-state thermodynamic model. Histograms correspond to the Λ values of double mutants containing the focal mutations, representing the 0.05, 0.4, and 1.0 quantiles for single mutant fitness. (*B*) Spearman's correlation matrix for all pairs of mutations i and j, computed from the estimated double mutant phenotypes $\hat{Y}_{ij}$, and ordered by single mutant ranks $\hat{R}_i$. (*C*) Distributions of $\hat{Y}_{ij}$ for each of the focal backgrounds (*Top row*). Double mutant ranks $\hat{R}_{ji}$ for each background as a function of $\hat{R}_i$ (*Bottom*). Each point is colored by its true specific epistatic effect, λ_{ij}. (*D*) Observed values of $\hat{D}_{ij}$ for a subset of mutations at adjacent positions (20 to 29) in position order (*Top*). Mutations are reordered (*Bottom*) by their single mutant fitness values $\hat{Y}_i$, denoted $\hat{Y}_{(r)}$ (gray scale heatmap). Three focal rank pairs are outlined with solid black squares. (*E*) New values of $\hat{Y}_i$ and $\hat{Y}_{ij}$ are generated under the assumption of a particular error model (*Top*), shuffling which mutation is associated with a given rank in a given bootstrapped replicate b (*Bottom*). (*F*) Empirical distributions of $|\hat{D}|$ plotted for the three focal pairs denoted in (*D*). (*G*) Position pairs enriched for SE colored by average sign and shown with respect to physical contacts (≤5 Å, dark gray) and less proximate positions (≤8 Å, light gray). The size of the point is proportional to the $-log_{10}$ P-value of enrichment.

such that Eq. **5** holds, $\hat{Y}_{im} > \hat{Y}_{in}$ is a likely outcome. Thus, a natural summary of the magnitude of SE in the context of rank statistics is the difference between the rank of the double mutant im compared to the rank of the single mutant m.

A rank-based test statistic. The aim of our inference procedure is to assess whether a given pair of mutations exhibits SE, while allowing for the presence of a global nonlinearity (Eq. **2**). Namely, the null hypothesis is given by

$$H_0 : \lambda_{ij} = 0, \qquad \text{for } i \neq j. \qquad [6]$$

Under the null hypothesis and the assumption of GE, and in the absence of measurement noise, the rank of mutation j in the background of mutation i, denoted as R_{ij}, is equal to the rank of mutation j in the wildtype background, R_j. More concisely, $R_{ij} = R_j$. Thus, a natural, rank-based estimate of SE is given by,

$$\hat{D}_{ij} := \hat{R}_{ji} - \hat{R}_i + \hat{R}_{ij} - \hat{R}_j, \qquad [7]$$

where we have summed the two rank deviations to obtain a symmetric test statistic. When calculating $\hat{D}_{ij}$ in practice, we

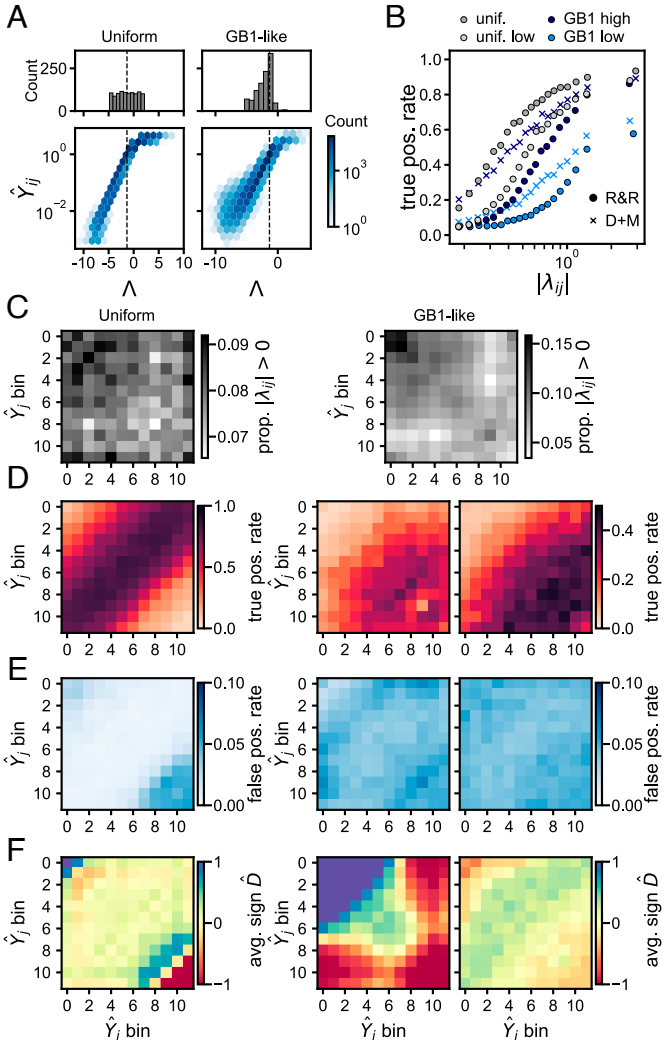

Fig. 3. Resample and reorder accurately identifies specific epistasis in simulations. A deep mutational scan was simulated under the assumption of a two-state thermodynamic model with specific epistasis (SE). (*A*) Histograms of the single mutant latent trait values Λ_i (*Top*), and double mutant phenotypes as a function of Λ_{ij} (*Bottom*) in the Uniform (*Left*) and GB1-low simulations (*Right*). The latent trait value for the wildtype is denoted by a black dashed line. (*B*) True positive rate as a function of the magnitude of the midpoint of each λ_{ij} bin, defined by evenly spaced quantiles, for the Uniform (gray), GB1-high (dark blue), and GB1-low (light blue) simulations analyzed with R&R (circles) and D+M (x's). (*C*) The proportion of nonzero interactions, $\lambda_{ij} \neq 0$, among positions within evenly spaced single mutant fitness, $\hat{Y}_i$, bins for the Uniform (*Left*) and GB1-low (*Right*) simulations. (*D*) True positive rate for each pair of bins for R&R analysis of the Uniform (*Left*) and GB1-low (*Middle*) simulations, and D+M analysis of the latter (*Right*). (*E*) False positive rates, as in (*D*). (*F*) Average sign of test statistics with *P*-values below α. For R&R, $\alpha = 0.1$. For D+M, $\alpha = 0.06$.

account for missing data by transforming the ranks onto the same scale (*Materials and Methods* and *SI Appendix*, section 1C).

A deviation of $\hat{D}_{ij}$ from zero would naïvely provide evidence for SE between mutations i and j. However, in the presence of measurement noise and GE, the variances of the constituent quantities may vary systematically across the measurement range (see Fig. 2*C* for $\hat{R}_{ji}$ as a function of $\hat{R}_i$ in three different backgrounds j). In addition, even in the absence of SE, the expected value of $\hat{D}_{ij}$ is not necessarily equal to zero. For example, the value of the test statistic for the two most deleterious mutations will be nonnegative, as ranks are bounded below by zero. Thus, correct calibration of the proposed hypothesis test requires estimation of the distribution of $\hat{D}_{ij}$ for each pair of ranks, $\hat{R}_i$ and $\hat{R}_j$, in the presence of measurement noise.

Resample and reorder, a bootstrapping approach. To estimate the null distributions of $\hat{D}_{ij}$ for each pair of mutations $i \neq j$, we take a simple, yet non-standard, bootstrapping approach. The bootstrapping procedure requires specification of an error model—the parametric component of R&R. The choice of error model, however, is not conceptually fundamental to our approach and necessarily depends on the application. When fitness estimates are defined by the ratio of sequencing reads after and before selection—as in the DMSs considered in this study—a Poisson error model is a natural choice (13, 36, 40, *Materials and Methods*; *SI Appendix*, section 1H).

In each bootstrap replicate $b = 1, \ldots, B$, we generate a synthetic dataset by sampling new fitness estimates for all observed single and double mutants from the specified error model (Fig. 2 *E*, *Top*). The mutants are then ranked according to their fitness estimates, generating a new set of single mutant ranks, $\hat{R}_i^b$, for all mutants $i = 1, \ldots, M$, and a new matrix of double mutant ranks, $\hat{R}_{ij}^b$ for $j \neq i$. The test statistic is then computed for every pair of mutations in each replicate, yielding $\hat{D}_{ij}^b$ values for all possible pairs of mutations i and j (Fig. 2 *E*, *Bottom*).

The key feature of our procedure is that the mutants with given ranks r and s will vary across simulations (Fig. 2*E*). This shuffling allows us to estimate the distribution of $\hat{D}$ as a function of the ranks of the constituent mutations, rather than their identities, ultimately yielding empirical distributions of $\hat{D}_{(r,s)}$, for each pair of mutant ranks r and s (Fig. 2*F*). The *P*-value for a given pair of mutations with observed ranks $\hat{R}_i = r$ and $\hat{R}_j = s$ is then computed with reference to the empirical distribution of the absolute value of test statistic, $|\hat{D}_{(r,s)}|$ (Fig. 2*F*, *Materials and Methods*). While bootstrapping has been used previously to estimate uncertainty in rankings (e.g., ref. 41), R&R additionally capitalizes on the randomness in the identities of mutants with a given pair of ranks to estimate the distributions of the rank-indexed test statistics.

To identify SE at the level of position pairs, and to overcome the multiple testing burden, each pair of positions is tested for enrichment of *P*-values below a fixed threshold α among all observed amino acid combinations at the two positions (Fig. 2*G* and *SI Appendix*, section 1G).

Evaluating R&R on Simulated and Empirical Datasets

R&R Identifies True Epistatic Effects in Simulations. To evaluate R&R, we simulate a DMS inspired by Olson et al. (42) and Otwinowski (31) in which the global nonlinearity is specified by a two-state thermodynamic model, and fitness has been estimated for all single and most double mutants (*SI Appendix*, section 1H).

We first consider an ideal scenario in which 1) all single and all double mutants are represented at equal frequencies in the initial library, respectively, and 2) the effects λ_i on the latent trait Λ are uniformly distributed (Fig. 3 *A*, *Top Left*). To introduce SE, we randomly sample λ_{ij} for nearby mutation pairs (within 5 Å), and otherwise set λ_{ij} to zero. Pre- and post selection read counts for each variant are sampled from Poisson distributions parameterized by initial cell count and true fitness to generate fitness estimates (Fig. 3 *A*, *Bottom* row; and see *SI Appendix*, section 1H). To apply R&R, we generate $B = 1{,}000$ bootstrap

samples under the assumption of a Poisson error model (*Materials and Methods*).

Not surprisingly, true positive rate increases as a function of the magnitude of λ_{ij} (Fig. 3*B*, gray dots). However, statistical power is lowest for combinations of poorly or highly ranked single mutations despite the uniform distribution of true epistatic effects over the measurement range (Fig. 3 *C* and *D*, *Left* column). Two factors likely contribute to the systematic reduction in power for these extreme pairs. First, in the two-state model, the slope of the nonlinearity approaches zero for both deleterious and beneficial mutations. As a consequence, the variance of the test statistic is larger for extreme pairs. Second, even in the absence of saturation, power to detect beneficial or deleterious interactions is asymmetric. A deleterious interaction ($\lambda_{ij} < 0$) will only minimally reduce the fitness of a double mutant composed of two very deleterious mutations (and likewise for $\lambda_{ij} > 0$ and two beneficial mutations), resulting in systematic biases in the inferred sign of epistasis (Fig. 3 *F*, *Left*). The extreme pairs also contribute disproportionately to the false positive rate, presumably due to greater variance of the test statistic (Fig. 3 *E*, *Left*).

Next, we simulate more realistic DMSs with the effects of single mutations on energy λ_i specified (with modification) by their estimates from a DMS of GB1 (31, 42), nonspecific binding, and variation in measurement precision (Fig. 3 *A*, *Right* and *SI Appendix*, section 1H). Unlike in our prior simulations, SE is concentrated among pairs of deleterious mutations, as mutations at positions engaged in many physical contacts tend to reduce fitness (Fig. 3 *C*, *Right*). In *SI Appendix*, we consider the effects of several of these amendments in isolation.

When R&R is applied to this GB1-like DMS, we observe a wholesale reduction in power relative to the uniform simulations (Fig. 3 *B* and *D*). Reductions in power are exacerbated when we reduce the initial cell counts by a factor of ten, further decreasing measurement precision—a scenario referred to as GB1-low in Fig. 3*B*. Compared to the uniform simulations, we observe more pronounced asymmetry in R&Rs ability to detect positive and negative epistasis in different areas of the measurement range. In particular, R&R almost exclusively detects negative epistasis among pairs of deleterious and beneficial mutations (Fig. 3*F* and *SI Appendix*, Fig. S2). This asymmetry likely arises due to saturation at the low end of the measurement range, and could potentially be mitigated by a two-sided hypothesis test (*SI Appendix*, section 1J and Fig. S10).

To further benchmark R&R, we compare its performance to that of an existing procedure which combines the fitness estimates of DiMSum (36) with a neural network-based GE inference framework, MoCHI (21), referred to henceforth as D+M. D+M tests for SE by examining the residuals of the double mutant fitness estimates with respect to a fitted GE model, here a sum of sigmoid functions.

The *P*-values of R&R and D+M are highly correlated though they exist on vastly different scales (Spearman's $\rho \approx 0.75$). And, like R&R, D+M is relatively less powered to detect SE among deleterious mutations (Fig. 3 *D*, *Center* and *Right* and *SI Appendix*, Fig. S2). To provide a meaningful comparison between the methods, we fix the false positive rate across the two methods in each simulation scenario (*SI Appendix*, section 1I). Under this parameterization, D+M is better powered to detect weak SE relative to R&R, with less appreciable differences in power for the GB1-low simulations (Fig. 3*B*). R&Rs reduced power relative to D+M is not surprising given overdispersion in the bootstrap sample (*SI Appendix*, section 2B) and the fact that D+M explicitly estimates the form of g. Rather, we emphasize that R&R can achieve comparable results—particularly when fitness measurements are less precise—without estimating the nonlinearity and at a fraction of the computational cost.

Resample and Reorder Identifies Protein Contacts in Empirical Datasets. We apply R&R to two DMSs, for which previous studies support models of single-trait GE and a strong association between inferred SE and the physical proximity of amino acids in the crystal structure (19, 32, 43).

In the first, Diss and Lehner (32) conducted a DMS of two alpha-helical proteins, Fos and Jun, which form a heterodimeric transcription factor in vivo (Fig. 4*A*). The authors estimated the interaction strength of all single mutants and the majority of double trans-mutants—pairwise combinations of single mutations in each protein—in a high-throughput, sequencing-based assay. In the second, Zarin and Lehner (43) estimated the binding affinities of almost all single and the majority of double trans-mutants of a large portion of the third PDZ domain of PSD-95, PDZ3, for its 8-residue cognate ligand CRIPT (Fig. 4*F*). Due to batch effects and variable sequencing coverage, we separately analyze two 43 amino acid segments of PDZ3, referred to henceforth as blocks 1 and 2 (*SI Appendix*, section 2E).

We implement R&R with $B = 1{,}000$ bootstrap replicates under the Poisson model (*Materials and Methods*). In *SI Appendix*, section 2, we summarize the results of R&R at the level of position–amino acid pairs. Here, we follow previous work (18, 19, 43) and test whether specific position pairs are enriched for SE, defined as mutated amino acid pairs with *P*-values below a threshold α (*Materials and Methods*).

Fos–Jun. Consistent with prior studies (19), enriched position pairs after correcting for multiple testing (44) are sparse and predominantly constrained to protein contacts, defined as amino acids within 5 Å in the crystal structure (Fig. 4 *B* and *D*), with R&R achieving comparable or higher contact prediction accuracy relative to previous analyses (18, 19, 32) and D+M (*SI Appendix*, Fig. S12).

Specific epistasis among enriched protein contacts—position pairs which are both significantly enriched after correcting for multiple testing (44, $\alpha_{BH} = 0.1$) and contacts—is disproportionately positive, consistent with ref. 32 (Fig. 4*C*). In Fig. 4*E*, we highlight an extreme example—positions L4 and L4 in Fos and Jun, respectively—for which all of the ≈30 amino acid interactions with *P*-values below $\alpha = 0.05$ are positive (*SI Appendix*, Fig. S5). The predominance of positive interactions is likely, in part, explained by two considerations: 1) the majority of mutations at positions engaged in protein–protein contacts are deleterious in isolation (*SI Appendix*, Fig. S3) and 2) R&R is relatively underpowered to detect negative interactions among the most deleterious mutations (Fig. 3). The latter limitation is not unique to R&R (see *SI Appendix*, Figs. S2, S10, and S11 for a comparison with D+M).

Positions T8 Fos and V8 Jun present a notable exception to the widespread positive epistasis among protein contacts, as ≈95% of interactions between T8 and V8 with *P*-values below α are negative in each replicate, a result that cannot readily be explained by methodological biases nor technical artifacts (Fig. 4*E* and *SI Appendix*, Fig. S5).

PDZ3-CRIPT. Sequencing coverage in the PDZ3-CRIPT dataset was lower than for that of Fos-Jun. As a consequence, overestimation of single mutant fitness values due to low initial read counts may have resulted in spurious detection of negative SE (*SI Appendix*, section 2E and Fig. S19).

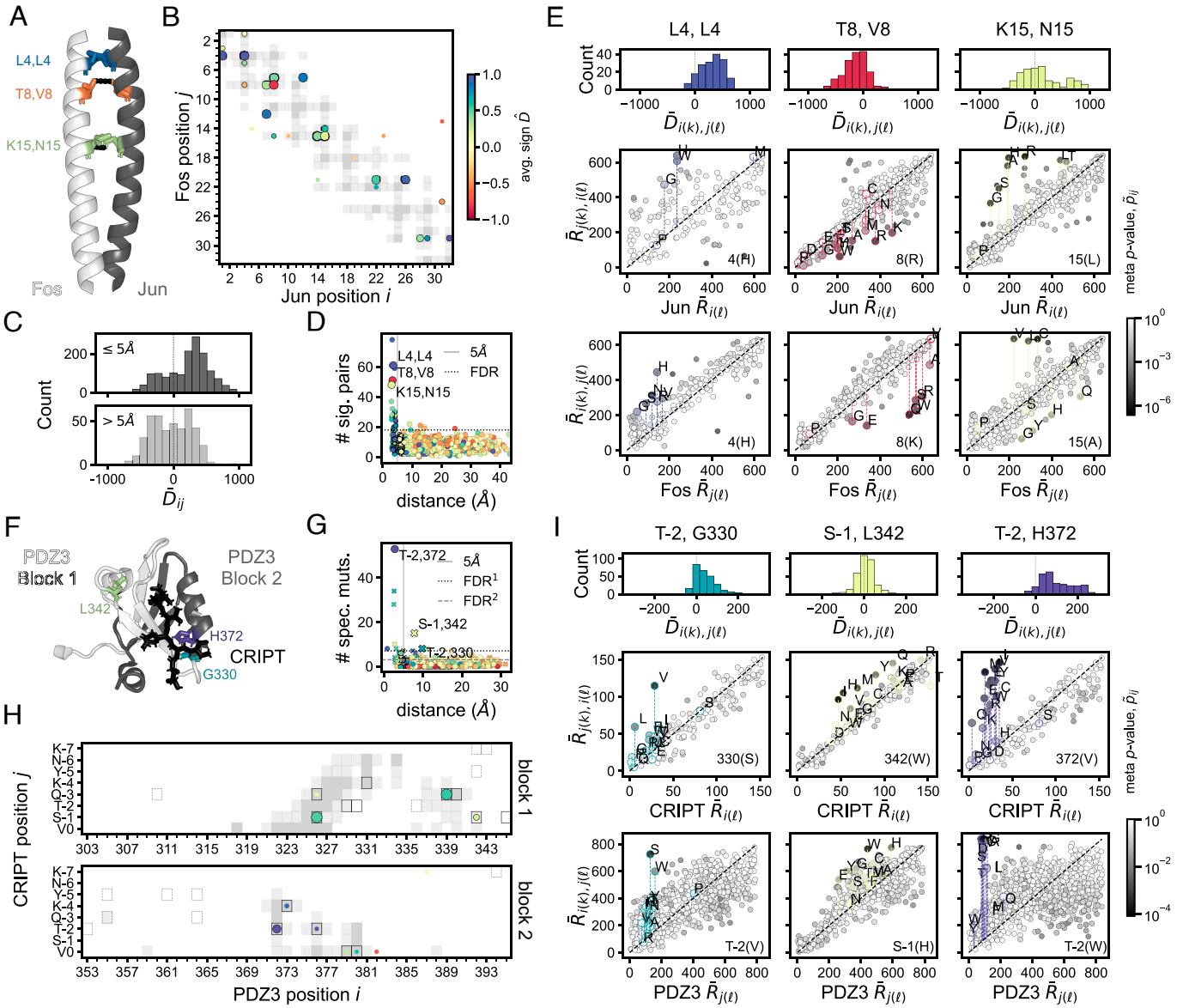

Fig. 4. Resample and reorder identifies protein contacts in empirical datasets. (*A*) Crystal structure of the mutated region of the Fos–Jun complex (PDB: 1FOS). Three position pairs enriched for specific epistasis (SE; $\alpha_{BH} \leq 0.01$) are highlighted and colored according to the average sign of the test statistic, $\hat{D}_{ij}$. (*B*) All position pairs enriched for SE ($\alpha_{BH} = 0.1$) are indicated by points, colored by the average sign of amino acid pairs with *P*-values below α ($\alpha = 0.05$); the size of the point is proportional to the $-log_{10}$ enrichment *P*-value. Pairs significant at $\alpha_{BH} = 0.01$ are outlined in black. (*C*) Histograms of the average value of the test statistic, $\bar{D}_{ij}$, across replicates for amino acid pairs with *P*-values below α among enriched protein contacts (≤5 Å, $\alpha_{BH} = 0.1$; *Top*, dark gray) and noncontacts (>5 Å, *Bottom*, light gray). (*D*) The number of *P*-values below α associated with a given position pair as a function of physical distance (Å), where the dotted black line denotes the significance threshold ($\alpha_{BH} = 0.1$). (*E*) Each column corresponds to a different enriched position pair. Histograms of the corresponding $\bar{D}_{ij}$ values, averaged over replicates (*Top*), and average double mutant rank, $\bar{R}_{i(k),j(\ell)}$, as a function of the single Jun (*Middle*) and Fos (*Bottom*) mutant ranks. The shading of each point denotes the meta *P*-value, $\tilde{p}_{ij}$. Points corresponding to the focal pairs are outlined in color and labeled by amino acid. (*F*) Crystal structure of the mutated region of the PDZ3-CRIPT complex (PDB: 5HEB). Three significant position pairs are highlighted, as in (*A*). (*G*) The same plot as in (*D*), except the number of positive interactions with *P*-values below α is shown for each PDZ3-CRIPT pair ($\alpha_1 = 0.034$, $\alpha_2 = 0.021$). Pairs involving positions in blocks 1 and 2 are represented by x's and points, respectively. (*H*) All position pairs enriched for SE, as in (*B*). In addition, position pairs enriched for positive SE at $\alpha_{BH} = 0.1, 0.01$ are denoted by dotted and solid black boxes, respectively. (*I*) The same plots as in (*E*) for select PDZ3-CRIPT position pairs.

To be conservative in the detection of enriched position pairs, we chose the *P*-value threshold α in each block to maximize contact precision and recall (*SI Appendix*, Fig. S15, $\alpha_1 = 0.034$, $\alpha_2 = 0.021$). As in ref. 43, we also identify position pairs enriched for positive interactions using the same values of α, respectively. While the pairs most enriched for positive interactions are within 5 Å—partly by design—several position pairs in block 1 overtly contravene this trend. Specifically, S-1 CRIPT and L342 and G345 in PDZ3, and T-2 CRIPT and G330 PDZ3 at distances of approximately 8, 11, and 10 Å, respectively, are significant after correction for multiple testing ($\alpha_{BH} = 0.01$; Fig. 4 *G* and *H*). These position pairs were previously identified as "specificity-changing mutations" (43) and/or implicated in allosteric regulation in PDZ3 binding (45). In isolation, both T-2 CRIPT and G330 PDZ3 harbor many of the most deleterious mutations (Fig. 4*I*). However, a subset of amino acid pairs at these positions results in appreciable fitness gains (Fig. 4*I*), though with absolute fitness still below that of wildtype (*SI Appendix*, Fig. S16). In contrast, S-1 CRIPT and L342 and G345 PDZ3 single mutations have more modest effects on fitness, with some

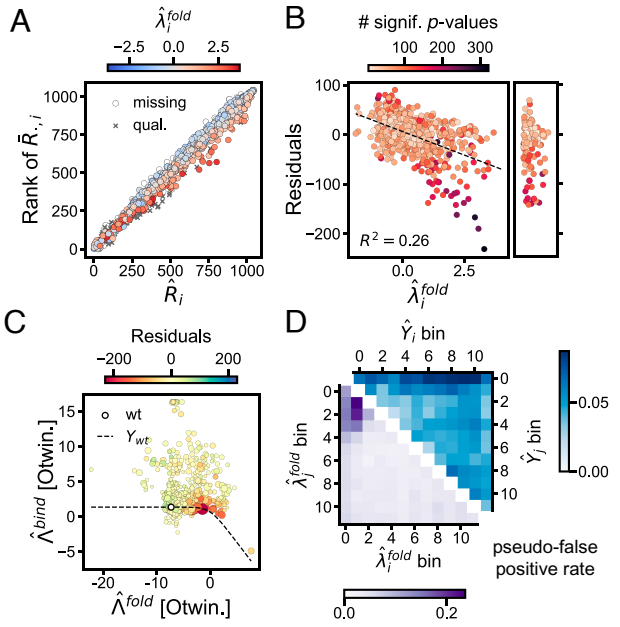

Fig. 5. Model misspecification results in spurious detection of specific epistasis. (*A*) The rank of the average rank of each mutation in GB1 (42) across all backgrounds, $\bar{R}_{\cdot,i}$, as a function of single mutant rank, $\hat{R}_i$. Where possible, each mutation is colored by its independently estimated folding energy , $\hat{\lambda}_i^{\text{fold}}$ (46). Mutations associated with "qualitative" and missing measurements are in dark gray and white, respectively (see ref. 46). (*B*) Residuals, defined as the difference between the *y* and *x*-axis values of (*C*), as a function of $\hat{\lambda}_i^{\text{fold}}$. The dotted line denotes a linear fit (to the quantitative measurements). Residuals for the qualitative measurements are shown to the *Right*. Each single mutation is colored by the number of associated *P*-values below α ($\alpha = 0.1$). (*C*) Figure 1b of ref. 31 reproduced, with each mutation additionally colored by its residual [*y*-axis in (*D*)]. The size of the point corresponds to the number of pseudo-false positives involving each mutation, defined as a *P*-value below α, where the mutations are at distances greater than 8 Å. (*D*) Pseudo-false positive rate computed for mutation pairs binned by fitness, $\hat{Y}_i$ (*Top*) and $\hat{\lambda}_i^{\text{fold}}$ (*Bottom*).

amino acid pairs exhibiting negative SE (Fig. 4*I* and *SI Appendix*, Fig. S16).

H372 PDZ3 and T-2 CRIPT exhibit the largest number of amino acid pairs with *P*-values below α, with the test statistic approaching its maximum value for several pairs (Fig. 4*I* and *SI Appendix*, Fig. S16). Though, as with the G330 PDZ3/T-2 CRIPT pairs, SE does not fully recover wildtype fitness (*SI Appendix*, Fig. S16). As the ranks of single mutations at H372 and T-2 are similarly deleterious, and many amino acid pairs exhibit large, positive rank deviations, R&R may underestimate the proportion of nonzero interactions (Fig. 4*I*). For example, the *P*-values associated with H372V and T-2R are above $\alpha = 0.05$ in each replicate. However, the observed rank deviations are unlikely to be explained by experimental noise, as suggested by the corresponding meta *P*-value ($\tilde{P} = 0.013$).

Model Misspecification Leads to Spurious Detection of Specific Epistasis. Finally, we analyze a comprehensive DMS of 55 of the 56 amino acids in the immunoglobin fragment G (IgG) binding domain of protein G, GB1 (42). Our primary interest in the DMS of GB1 is that single-trait GE is a priori thought to be insufficient to explain GB1-IgG binding. Instead, prior work supports a three-state equilibrium model, where the protein's folding and binding energies govern transitions among unfolded, folded, and ligand-bound states (31, 46). In this case, fitness—a proxy for binding affinity—is a bivariate nonlinear function of two latent additive traits, Λ^{fold} and Λ^{bind} (*SI Appendix*, Eq. **S15**). The presence of two latent traits implies that, even in the absence of measurement noise, the ranks of mutations will not necessarily be preserved across backgrounds, violating the underlying assumption of R&R. Indeed, the contact accuracy of R&R at the level of position-pairs is substantially lower than for Fos-Jun, though R&R performs comparably to or slightly below prior procedures (*SI Appendix*, Figs. S13 and S14). When D+M is used to fit a three-state model, contact prediction accuracy improves appreciably (*SI Appendix*, Fig. S13). As such, the DMS of GB1 presents an opportunity to evaluate the consequences of model misspecification, in the form of an additional latent trait, on R&Rs detection of SE in an empirical dataset.

A key feature of the three-state model is that a mutation i with an adverse effect on binding is more deleterious in the background of an unstable mutation j. In this scenario, the rank of mutation i in the background of j will likely be lower than expected given its single mutant rank, i.e., $\mathbb{E}[\hat{R}_{ji}] < \mathbb{E}[\hat{R}_i]$, resulting in a preponderance of negative $\hat{D}_{ij}$ values and spurious negative SE among all such pairs. The converse holds when a stabilizing mutation i occurs in the background of a mutation j with adverse effects on folding: $\mathbb{E}[\hat{R}_{ji}] > \mathbb{E}[\hat{R}_i]$ yielding $\hat{D}_{ij} > 0$ and spurious positive SE.

We therefore suspected that a mutation's residual, defined as the difference between its rank based on all possible backgrounds and single mutant rank (Fig. 5*A*), would correlate with its effects on folding energy, λ_i^{fold}. Indeed, variants with adverse effects on $\hat{\lambda}^{\text{fold}}$—independently measured by Nisthal et al. (46)—exhibit more negative residuals relative to more stable variants (Fig. 5*B*, linear fit, $R^2 = 0.26$), larger numbers of interactions with *P*-values below α (Fig. 5*B*), and higher pseudo-false positive rates (Fig. 5 *D*, *Lower* triangle). Indeed, $\hat{\lambda}_i^{\text{fold}}$ is a much better predictor of pseudo-false positive rate than single mutant phenotype (Fig. 5 *D*, *Upper* triangle). To better visualize this phenomenon, we reproduce Figure 1b of ref. 31, in which the inferred folding and binding energies of each GB1 mutation are plotted with respect to the wildtype sequence (Fig. 5*C*). Here, we observe that variants with estimated adverse effects on folding ($\hat{\Lambda}_i^{\text{fold}} > 0$) and fitness approximating the wildtype fitness exhibit the highest pseudo-false positive rates; in Fig. 5*C*, the size of each point is proportional to the number of false positives. Simulations under the assumption of a three-state model without SE qualitatively reproduce these results (*SI Appendix*, Fig. S14).

Discussion

Quantifying epistasis usually requires specification of the measurement scale, e.g., additive or multiplicative. Existing semi-parametric methods relax this constraint by defining (specific) epistasis as a deviation from a model of GE fitted under minimal assumptions about the form of the nonlinearity, g. For example, modeling g as a sum of monotonic functions, such as spline functions or sigmoids (13, 21). Nonparametric procedures, such as refs. 47–49, impose even fewer constraints on scale, but may overestimate the prevalence of SE in not explicitly accounting for any global nonlinearities.

By redefining epistasis in the context of rank statistics, R&R entirely circumvents the choice of scale. Our work follows from the observation that, as long as g is monotonic—but not necessarily nonlinear—SE can be detected as a deviation from a given reference order.

We contribute a principled hypothesis testing procedure, R&R, that accounts for heteroskedasticity in the test statistic due to variation in 1) the slope of the nonlinearity, 2) the density of single effects on the latent trait Λ, and 3) variance of the fitness estimates conditional on Λ. To "learn" the distribution of $\hat{D}$ as a function of the ranks of the constituent mutations, R&R generates an ensemble of synthetic datasets under the assumption of an error model—the only context-specific and parametric element of R&R. In the analyses of protein DMSs presented here, we employed a Poisson error model, allowing R&R to operate on minimally processed variant count data. Additional postprocessing and alternative error models, e.g., normally distributed errors, would allow R&R to combine information across replicates, account for batch effects, as in ref. 43, and explicitly model overdispersion (e.g., ref. 36), potentially improving statistical power and reducing false positives.

Despite its simplicity, R&R accurately identified true SE and protein contacts in a diverse set of simulated and empirical datasets—with substantial variation in measurement precision—demonstrating its robustness, while also revealing several limitations:

The test statistics for a given mutation are correlated across variants as they are all computed with reference to the single mutant rank. When the fitness values of single mutants are estimated with high precision, errors in the single mutant ranks will make minimal contributions to the false positive rate. However, when single mutant ranks are estimated with nonnegligible error—as in the PDZ3-CRIPT dataset—we expect to, and likely do, detect spurious SE. Alternative definitions of the test statistic could potentially mitigate susceptibility to errors in the single fitness estimate.

R&R implicitly assumes that the estimated fitness values are independent and identically distributed conditional on their latent trait values. Variation in measurement precision will increase the volatility of the bootstrap sample, reducing power to detect SE among more precise fitness measurements. To mitigate this bias, one could potentially integrate over uncertainty in the double mutant rank when computing the test statistic or devise a more sophisticated bootstrapping scheme that accounts for variation in measurement precision.

R&Rs power to detect SE depends on the density of SE in a mutation's single effect neighborhood. When SE is not sparse, R&R will detect the largest interactions, as observed for mutations of H372 PDZ3 and T-2 CRIPT, potentially underestimating the complexity of the genotype-to-phenotype map.

More profoundly, analyzing DMSs from a rank-based perspective exposes systematic biases in the detection of SE—with R&R or other procedures—across the measurement range. For example, both R&R and D+M are underpowered to detect SE, and particularly negative SE, among deleterious mutations. Our ability to derive general principles about proteins from distinct DMSs requires an understanding of how experimental and statistical biases influence the distribution of SE detected in a given experiment. For example, to conclude that SE is on average positive or negative, necessitates a proper accounting of statistical power to detect SE of one sign versus the other. To the extent possible, experimentalists can maximize power to detect SE by measuring single mutant fitness with high accuracy, increasing the measurement range, and achieving equal representation of each mutant in the library as well as uniform sequencing coverage across mutants.

In the present work, we have arguably treated the nonlinearity g as a "nuisance," which conceals the "true" epistatic landscape. When g is induced entirely by experimental design, for example, by a lower detection threshold, this treatment is uncontroversial. However, when g emerges from the physics of the protein, one can not necessarily neatly distinguish between a global nonlinearity and direct physical interactions (see ref. 50). For one, recent work demonstrates that GE can emerge from numerous microscopic interactions among mutations at many orders (23, 24). This finding reveals a fundamental ambiguity in the specification of genotype-to-phenotype maps: Is a GE model—where the latent additive trait includes sparse higher-order terms—to be privileged over a dense model, with numerous epistatic interactions at many orders, that does not explicitly account for global nonlinearities (e.g., see refs. 26 and 48)?

Our work presents immediate areas for future research. R&R relies on sufficient numbers of mutations at a given order to estimate the distributions of the rank-indexed test statistic—hence our focus on DMSs with large numbers of single and double mutants. Many DMSs, however, more sparsely sample the sequence space and include higher-order mutants. For example, "pathway" DMSs assay a combinatorially complete set of mutants that interpolate between two sequence endpoints (e.g., refs. 51–54). Extending R&R, and applying rank-based procedures for detecting interactions more generally (see refs. 38 and 39) to these more common, sparser datasets may provide additional insights into epistasis across proteins.

While monotonicity is common in biology, it is by no means a rule (16). Despite violating a fundamental assumption of R&R, nonmonotonic GE may still systematically constrain the rank statistics of combinatorial DMSs, for example, if g is unimodal. Such constraints could potentially be exploited to detect interactions in the presence of nonmonotonic GE. In addition, as we demonstrated in *Model Misspecification*, the presence of an additional latent trait results in spurious inference of SE. Extending rank-based detection of SE to higher dimensional models of GE—where mutations combine additively in a multidimensional latent space—demands advances in our mathematical understanding of how multidimensional (monotonic) GE constrains the rank statistics of combinatorial DMSs, representing a rich area for future work.

Finally, in our applications of R&R, we have assumed that single-trait, monotonic GE is the appropriate model. Non- or semiparametric tests that formally assess this hypothesis would provide a powerful tool in the analysis and interpretation of combinatorial DMSs.

Materials and Methods

Processing the Deep Mutational Scan Data. For the GB1 DMS (42), we removed variants with fewer than 21 reads in the input pool.

For the Fos–Jun DMS (32), we processed the raw sequencing reads to produce pre- and post selection read counts for each variant in each of the three replicates. We then removed observations with fewer than 11 initial reads.

For the PDZ3-CRIPT DMS (43), we removed all positions for which the average single mutant read coverage was below 20 across the three replicates, observations with fewer than 11 initial reads, and any variants with more than 95% missing data across all double mutants. Due to batch effects, we analyzed the two halves of PDZ3 separately, with blocks 1 and 2 spanning positions 303-345aa and 353-395aa, respectively.

Computing the Test Statistic. Given an L-length protein and M possible states, one can observe $L_d := (L-1) \times (M-1)$ double mutants in a given background. Thus, to compute the test statistic $\hat{D}_{ij}$ Eq. **7**, we adjust the single rank of i by excluding all mutations at position j (and vice versa), where we

have abused notation in using i and j to refer to both position and amino acid mutant.

As fewer mutations may be observed in some backgrounds due to insufficient coverage in the initial sequencing pool, all of the ranks are transformed onto the same scale such that the ranks in each background range from 0 to $L_d - 1$ (*SI Appendix*, section 1C). Tied values are resolved using mid-ranks (*SI Appendix*, section 1B). In the PDZ3-CRIPT dataset, where the number of PDZ3 mutations far outnumbered that of CRIPT, we employ a reweighted test statistic (*SI Appendix*, section 1E).

Poisson Error Model. The pre- and postselection read counts in the b-th bootstrap replicate $N_i^{0,b}$ and $N_i^{1,b}$ for mutant i, respectively, are assumed to be Poisson distributed with means specified by their observed value, $\hat{N}_i^0$ and $\hat{N}_i^1$, $N_i^{0,b} \sim \text{Poisson}(\hat{N}_i^0 + \delta)$ and $N_i^{1,b} \sim \text{Poisson}(\hat{N}_i^1 + \delta)$, where $\delta = 1$ is a pseudocount. The fitness estimate for mutant i in the b-th replicate is then given by,

$$\hat{\gamma}_i^b := \begin{cases} (N_i^{1,b} + \eta)/(N_i^{0,b} + \eta) & \text{for } N_i^{0,b} \geq N^* \\ \text{missing} & \text{else.} \end{cases} \quad [8]$$

where $\hat{N}_i^0$ is the observed initial read count and N^* is a minimum initial read count threshold, and likewise for double mutants.

Estimating the *P*-Value of Interactions Between Mutations. The bootstrapping procedure described in the main text is used to generate an ensemble of test statistics for each pair of double mutant ranks, r and s. The empirical distribution of $\hat{D}_{(r,s)}$ is then given by, $\hat{\mathcal{F}}_{(r,s)}(d) := \frac{1}{B+1}\sum_{b=1}^{B} \mathbb{1}\{|\hat{D}_{(r,s)}^b| < d\}$, where $\hat{D}_{(r,s)}^b$ is computed from Eq. **7** for the mutants ranked r-th and s-th in the b-th simulation. As certain combinations of mutant ranks are not observed in a given bootstrap replicate, the matrix $\hat{\mathbf{D}}^{(b)}$ is imputed using a nearest neighbor approach (*SI Appendix*, section 1D). The P-value for a pair of mutations i and j with estimated ranks r and s, is then given by, $p_{ij} := 1 - \hat{\mathcal{F}}_{(r,s)}\left(|\hat{D}_{ij}|\right)$, where $\hat{D}_{ij}$ is the observed test statistic.

Method Comparison. DiMSum (36) was applied to simulated and empirical datasets after filtering on initial read counts, to estimate the mean (log) relative fitness values and their SEs under a Poisson error model. The estimates from DiMSum were used as input to MoCHI (21) to fit a GE model, where the nonlinearity was assumed to take the form of 1) a sum of arbitrary sigmoid functions (simulations and Fos–Jun), 2) a two-state thermodynamic model (GB1), and 3) a three-state thermodynamic model (GB1). See *SI Appendix*, section 1I for more details.

Data, Materials, and Software Availability. All experimental data were previously made available to the public. All code required to reproduce the analyses is available at: github.com/marync/resample_and_reorder. Previously published data were used for this work (32, 42, 43).

ACKNOWLEDGMENTS. We thank Federica Ferretti, Alisdair Hastewell, Nikos Ignatiadis, Rama Ranganathan, Abigail Skwara, Rebecca Willett, and attendees of the NITMB research-in-progress for helpful conversations. We thank Taraneh Zarin for providing guidance with respect to the PDZ3-CRIPT data. M.O.C. was supported by The National Institute of General Medical Sciences of the NIH (R35GM151211) and an a National Institute for Theory and Mathematics in Biology (NITMB) fellowship supported by grants from the NSF (DMS-2235451) and Simons Foundation (MP-TMPS-00005320). Analyses were performed on the Midway cluster, supported by the Research Computing Center at the University of Chicago.

Author affiliations: [a]National Institute for Theory and Mathematics in Biology, Chicago, IL 60611; [b]The James Franck Institute, University of Chicago, Chicago, IL 60637; [c]Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, IL 60637; [d]Center for the Physics of Evolving Systems, University of Chicago, Chicago, IL 60637; [e]Center for Living Systems, University of Chicago, Chicago, IL 60637; [f]Section of Genetic Medicine, University of Chicago, Chicago, IL 60637; and [g]Department of Human Genetics, University of Chicago, Chicago, IL 60637

1. R. A. Fisher, Xv.-the correlation between relatives on the supposition of mendelian inheritance. *Earth Environ. Sci. Trans. R. Soc. Edinb.* **52**, 399–433 (1919).
2. H. J. Cordell, Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.* **11**, 2463–2468 (2002).
3. R. Fisher, F. Immer, O. Tedin, The genetical interpretation of statistics of the third degree in the study of quantitative inheritance. *Genetics* **17**, 107 (1932).
4. K. Mather, Polygenic inheritance and natural selection. *Biol. Rev.* **18**, 32–64 (1943).
5. J. F. Crow, How important is detecting interaction? *Behav. Brain Sci.* **13**, 126–127 (1990).
6. P. C. Phillips, Epistasis-the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* **9**, 855–867 (2008).
7. G. R. Loftus, On interpretation of interactions. *Mem. Cogn.* **6**, 312–319 (1978).
8. D. Wahlsten, Insensitivity of the analysis of variance to heredity-environment interaction. *Behav. Brain Sci.* **13**, 109–120 (1990).
9. A. Berrington de González, D. R. Cox, Interpretation of interaction: A review. *Ann. Appl. Stat.* **1**, 371–385 (2007).
10. E. J. Wagenmakers, A. M. Krypotos, A. H. Criss, G. Iverson, On the interpretation of removable interactions: A survey of the field 33 years after loftus. *Mem. Cogn.* **40**, 145–160 (2012).
11. J. Diaz-Colunga *et al.*, Global epistasis on fitness landscapes. *Philos. Trans. R. Soc. B* **378**, 20220053 (2023).
12. Z. R. Sailer, M. J. Harms, Detecting high-order epistasis in nonlinear genotype-phenotype maps. *Genetics* **205**, 1079–1088 (2017).
13. J. Otwinowski, D. M. McCandlish, J. B. Plotkin, Inferring the shape of global epistasis. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E7550–E7558 (2018).
14. D. M. Fowler, S. Fields, Deep mutational scanning: A new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
15. J. B. Kinney, D. M. McCandlish, Massively parallel assays and quantitative sequence-function relationships. *Annu. Rev. Genom. Human Genet.* **20**, 99–127 (2019).
16. B. Lehner, Molecular mechanisms of epistasis within and between genes. *Trends Genet.* **27**, 323–331 (2011).
17. T. N. Starr, J. W. Thornton, Epistasis in protein evolution. *Protein Sci.* **25**, 1204–1218 (2016).
18. N. J. Rollins *et al.*, Inferring protein 3D structure from deep mutation scans. *Nat. Genet.* **51**, 1170–1176 (2019).
19. J. M. Schmiedel, B. Lehner, Determining protein structures using deep mutagenesis. *Nat. Genet.* **51**, 1177–1186 (2019).
20. M. S. Johnson, G. Reddy, M. M. Desai, Epistasis and evolution: recent advances and an outlook for prediction. *BMC Biol.* **21**, 120 (2023).
21. A. J. Faure, B. Lehner, Mochi: neural networks to fit interpretable models and quantify energies, energetic couplings, epistasis, and allostery from deep mutational scanning data. *Genome Biol.* **25**, 303 (2024).
22. S. Kryazhimskiy, D. P. Rice, E. R. Jerison, M. M. Desai, Global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science* **344**, 1519–1522 (2014).
23. D. M. Lyons, Z. Zou, H. Xu, J. Zhang, Idiosyncratic epistasis creates universals in mutational effects and evolutionary trajectories. *Nat. Ecol. Evol.* **4**, 1685–1693 (2020).
24. G. Reddy, M. M. Desai, Global epistasis emerges from a generic model of a complex trait. *eLife* **10**, e64740 (2021).
25. Y. Park, B. P. Metzger, J. W. Thornton, The simplicity of protein sequence-function relationships. *Nat. Commun.* **15**, 7953 (2024).
26. T. Dupic, A. M. Phillips, M. M. Desai, Protein sequence landscapes are not so simple: On reference-free versus reference-based inference. bioRxiv [Preprint] (2024). https://doi.org/10.1101/2024.09.17.613512 (Accessed 1 April 2025).
27. A. J. Faure *et al.*, The genetic architecture of protein stability. *Nature* **634**, 995–1003 (2024).
28. Z. R. Sailer, M. J. Harms, High-order epistasis shapes evolutionary trajectories. *PLoS Comput. Biol.* **13**, e1005541 (2017).
29. T. U. Sato, K. Kaneko, Evolutionary dimension reduction in phenotypic space. *Phys. Rev. Res.* **2**, 013197 (2020).
30. K. Husain, A. Murugan, Physical constraints on epistasis. *Mol. Biol. Evol.* **37**, 2865–2874 (2020).
31. J. Otwinowski, Biophysical inference of epistasis and the effects of mutations on protein stability and function. *Mol. Biol. Evol.* **35**, 2345–2354 (2018).
32. G. Diss, B. Lehner, The genetic landscape of a physical interaction. *eLife* **7**, e32472 (2018).
33. J. B. Kinney, A. Murugan, C. G. Callan Jr., E. C. Cox, Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 9158–9163 (2010).
34. V. O. Pokusaeva *et al.*, An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. *PLoS Genet.* **15**, e1008079 (2019).
35. P. D. Tonner, A. Pressman, D. Ross, Interpretable modeling of genotype-phenotype landscapes with state-of-the-art predictive power. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2114021119 (2022).
36. A. J. Faure, J. M. Schmiedel, P. Baeza-Centurion, B. Lehner, Dimsum: An error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. *Genome Biol.* **e21**, 207 (2020).
37. D. Brookes, J. Otwinowski, S. Sinai, Contrastive losses as generalized models of global epistasis. *Adv. Neural Inf. Process. Syst.* **37**, 93374–93405 (2024).
38. K. Crona, A. Gavryushkin, D. Greene, N. Beerenwinkel, Inferring genetic interactions from comparative fitness data. *eLife* **6**, e28629 (2017).
39. C. Lienkaemper, L. Lamberti, J. Drain, N. Beerenwinkel, A. Gavryushkin, The geometry of partial fitness orders and an efficient method for detecting genetic interactions. *J. Math. Biol.* **77**, 951–970 (2018).
40. A. Sarkar, M. Stephens, Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nat. Genet.* **53**, 770–777 (2021).

41. P. Hall, H. Miller, Using the bootstrap to quantify the authority of an empirical ranking. *Ann. Stat.* **37**, 3929–3959 (2009).
42. C. A. Olson, N. C. Wu, R. Sun, A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr. Biol.* **24**, 2643–2651 (2014).
43. T. Zarin, B. Lehner, A complete map of specificity encoding for a partially fuzzy protein interaction. bioRxiv [Preprint] (2024). 10.1101/2024.04.25.591103 (Accessed 1 April 2025).
44. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
45. A. S. Raman, K. I. White, R. Ranganathan, Origins of allostery and evolvability in proteins: A case study. *Cell* **166**, 468–480 (2016).
46. A. Nisthal, C. Y. Wang, M. L. Ary, S. L. Mayo, Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 16367–16377 (2019).
47. J. Zhou, D. M. McCandlish, Minimum epistasis interpolation for sequence-function relationships. *Nat. Commun.* **11**, 1782 (2020).
48. J. Zhou *et al*., Higher-order epistasis and phenotypic prediction. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2204233119 (2022).
49. A. Aghazadeh *et al*., Epistatic net allows the sparse spectral regularization of deep neural networks for inferring fitness functions. *Nat. Commun.* **12**, 5225 (2021).
50. S. Dutta, J. P. Eckmann, A. Libchaber, T. Tlusty, Green function of correlated genes in a minimal mechanical model of protein evolution. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E4559–E4568 (2018).
51. D. M. Weinreich, N. F. Delaney, M. A. DePristo, D. L. Hartl, Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114 (2006).
52. N. C. Wu, L. Dai, C. A. Olson, J. O. Lloyd-Smith, R. Sun, Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife* **5**, e16965 (2016).
53. F. J. Poelwijk, M. Socolich, R. Ranganathan, Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nat. Commun.* **10**, 4213 (2019).
54. A. Moulana *et al*., The landscape of antibody binding affinity in SARS-CoV-2 omicron BA. 1 evolution. *eLife* **12**, e83442 (2023).