PNAS RESEARCH ARTICLE | EVOLUTION OPEN ACCESS

Higher-order epistasis and phenotypic prediction

Juannan Zhou[a,b,1], Mandy S. Wong[c,2], Wei-Chia Chen[b,3], Adrian R. Krainer[c], Justin B. Kinney[b], and David M. McCandlish[b,1]

Edited by Marcus Feldman, Stanford University, Stanford, CA; received March 9, 2022; accepted July 18, 2022

Contemporary high-throughput mutagenesis experiments are providing an increasingly detailed view of the complex patterns of genetic interaction that occur between multiple mutations within a single protein or regulatory element. By simultaneously measuring the effects of thousands of combinations of mutations, these experiments have revealed that the genotype–phenotype relationship typically reflects not only genetic interactions between pairs of sites but also higher-order interactions among larger numbers of sites. However, modeling and understanding these higher-order interactions remains challenging. Here we present a method for reconstructing sequence-to-function mappings from partially observed data that can accommodate all orders of genetic interaction. The main idea is to make predictions for unobserved genotypes that match the type and extent of epistasis found in the observed data. This information on the type and extent of epistasis can be extracted by considering how phenotypic correlations change as a function of mutational distance, which is equivalent to estimating the fraction of phenotypic variance due to each order of genetic interaction (additive, pairwise, three-way, etc.). Using these estimated variance components, we then define an empirical Bayes prior that in expectation matches the observed pattern of epistasis and reconstruct the genotype–phenotype mapping by conducting Gaussian process regression under this prior. To demonstrate the power of this approach, we present an application to the antibody-binding domain GB1 and also provide a detailed exploration of a dataset consisting of high-throughput measurements for the splicing efficiency of human pre-mRNA 5′ splice sites, for which we also validate our model predictions via additional low-throughput experiments.

genotype–phenotype map | Gaussian processes | genetic interaction | splicing | protein G

Understanding the relationship between genotype and phenotype is difficult because the effects of a mutation often depend on which other mutations are already present in the sequence, a phenomenon known as epistasis (1–3). Recent advances in high-throughput mutagenesis and phenotyping have for the first time provided a detailed view of these complex genetic interactions, by allowing phenotypic measurements for the effects of tens of thousands of combinations of mutations within individual proteins (4–18), RNAs (19–24), and regulatory or splicing elements (25–31). Importantly, it has now become clear that the data from these experiments cannot be captured by considering simple pairwise interactions, but rather higher-order genetic interactions between three, four, or even all sites within a functional element are empirically common (2, 12, 32–44) and indeed often expected based on first-principles biophysical considerations (12, 23, 32, 35, 36, 41, 45, 46). However, the enormous number of possible combinations of mutations makes these higher-order interactions both difficult to conceptualize and challenging to incorporate into predictive models.

From a very basic perspective, data from combinatorial mutagenesis experiments provide us with observations of phenotypic values for individual genotypes, the effects of specific mutations on specific genetic backgrounds, epistatic coefficients between pairs of mutations on specific backgrounds, etc. The essential problem in modeling data like this then comes down to the question of how to combine these observed quantities to make phenotypic predictions for unobserved genotypes. That is, given that we have already seen the results of a specific mutation in several different genetic backgrounds, how should we combine these observations to predict its effect in a new background?

Here we provide an answer to this question based on the intuition that when making these predictions, we should focus on the observed effects of mutations that are nearby in sequence space to the genetic background we are making a prediction for, rather than observations of mutational effects that are more distant. We do this by considering a key comprehensible aspect of higher-order epistasis, namely, the decay in the predictability of mutational effects (12, 47), epistatic coefficients of double mutants, and observed phenotypes (33, 48, 49), as one moves through sequence space.

More specifically, we use the observed pattern of decay in phenotypic correlation as a function of genetic distance to estimate the fraction of variance due to each order of

Significance

One core goal of genetics is to systematically understand the mapping between the DNA sequence of an organism (genotype) and its measurable characteristics (phenotype). Understanding this mapping is often challenging because of interactions between mutations, where the result of combining several different mutations can be very different than the sum of their individual effects. Here we provide a statistical framework for modeling complex genetic interactions of this type. The key idea is to ask how fast the effects of mutations change when introducing the same mutation in increasingly distant genetic backgrounds. We then propose a model for phenotypic prediction that takes into account this tendency for the effects of mutations to be more similar in nearby genetic backgrounds.

Author affiliations: [a]Department of Biology, University of Florida, Gainesville, FL 32611; [b]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724; and [c]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724

Author contributions: J.Z., M.S.W., W.-C.C., A.R.K., J.B.K., and D.M.M. designed research; J.Z., M.S.W., W.-C.C., and J.B.K. performed research; and J.Z., M.S.W., W.-C.C., A.R.K., J.B.K., and D.M.M. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: juannanzhou@ufl.edu and mccandlish@cshl.edu.

[2]Present address: Beam Therapeutics, Cambridge, MA 02142.

[3]Present address: Department of Physics, National Chung Cheng University, Chiayi 62102, Taiwan, R.O.C.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2204233119/-/DCSupplemental.

Published September 21, 2022.

interaction in our observed data. Based on these point estimates, we then construct a prior distribution over all possible sequence-to-function mappings where the expected decay in the predictability of mutational effects in the prior matches that observed in the data. Finally, we conduct Bayesian inference under this prior using Gaussian process regression (50) and employ Hamiltonian Monte Carlo (51) to sample from the resulting high-dimensional posterior distribution. The end result is a procedure that automatically weights the contributions of our observations to our predictions in the manner suggested by the type and extent of higher-order epistasis present in the data, while simultaneously accounting for the effects of measurement noise and quantifying the uncertainty in our predictions.

We call this method empirical variance component regression (VC regression) because it uses an empirical Bayes (52) prior defined by the variance components. To demonstrate the performance of our method, we apply it to two datasets. The first dataset is derived from a combinatorial mutagenesis experiment for protein G (37), a streptococcal antibody-binding protein that has served as a model system for studies of the genotype–phenotype map in proteins. The second dataset consists of high-throughput measurements of the splicing efficiency of human $5'$ splice sites (31), which are RNA sequence elements crucial for the assembly of the spliceosome during pre-mRNA splicing. For this latter dataset, we also present low-throughput validation measurements for our model predictions, as well as a qualitative exploration of the complex patterns of epistasis in splicing efficiency observed in this system.

Results

Experimental observations of the genotype-phenotype map typically consist of measurements of phenotypic values for a subset of possible genotypes. From these observations, we can calculate a number of more derived quantities, such as mutational effects (i.e., the difference in phenotype between two mutationally adjacent genotypes) and double-mutant epistatic coefficients (i.e., the difference between the observed phenotype of a double mutant and its expected phenotype based on the sum of the single-mutant effects). The central question of phenotypic prediction is then deciding how to combine these various mutational effects, local epistatic coefficients, and individual phenotypic values to produce accurate predictions for the phenotypic values of unmeasured genotypes.

Different prediction methods typically reflect different overall strategies for how to combine these experimentally derived quantities. For example, when we fit an additive or nonepistatic model (53), we are implementing a strategy based on the assumption that the phenotypic effects of observed mutations are the same regardless of the presence/absence of other mutations. Thus, fitting an additive model can be thought of as a generalization of the simple heuristic procedure of making predictions by 1) averaging over all the times the effect of each possible point mutation is observed and then 2) adding up these average effects to make a prediction for any given genotype. In a similar way, it is easy to show that while a pairwise interaction model (54) allows the effects of individual mutations to vary across genetic backgrounds, the epistatic interaction observed in double mutants for any specific pair of mutations is constant across backgrounds (*SI Appendix*). Thus, fitting a pairwise model is conceptually closely related to the heuristic of determining the interaction between a pair of mutations by averaging over the local epistatic coefficients for this pair of mutations that are observed in the data and then assuming that this pair of mutations has this same epistatic coefficient regardless of what genetic background these mutations occur on.

Here we introduce a prediction method corresponding to a different heuristic, one that implements the intuitions that 1) all orders of genetic interaction can be important and helpful in making predictions and 2) observations of mutational effects and epistatic coefficients in nearby genetic backgrounds should influence our predictions more than observations in distant genetic backgrounds.

Genetic Interactions and the Predictability of Mutational Effects. How consistent are mutational effects, double-mutant genetic interactions, etc., in increasingly distant genetic backgrounds? The answer to this question largely depends on the type and amount of epistasis present and has important practical implications for phenotypic prediction. For example, if the genotype–phenotype map is highly additive, long-distance extrapolation may be feasible, and relatively few measurements are required to make accurate predictions, whereas in a fully uncorrelated genotype–phenotype map (55), accurate prediction may be impossible even at short distances.

One common way to quantify the type and amount of epistasis present is to define the fraction of variance due to a particular interaction order as the increase in the R^2 of a least squares fit when one adds interaction terms of that order to a regression model that already includes all interactions of lower order (*SI Appendix*). These values are often referred to as variance components or as the normalized amplitude spectrum (33, 56). Fig. 1*A* shows these values for a simulated genotype–phenotype map that contains both a large additive component and a substantial amount of higher-order epistasis.

Now, suppose we consider the effects of point mutations and ask how predictable the effects of mutations tend to be on increasingly distant genetic backgrounds. Recent theoretical results (12, 47) have shown that given the variance components of a genotype–phenotype map, together with knowledge of the number of sites and number of alleles per site, one can calculate the curve describing how the correlation between mutational effects decays as a function of the distance between genetic backgrounds (i.e., the distance correlation function for mutational effects). Such a curve for our simulated landscape is shown in Fig. 1*B*. In *SI Appendix*, we extend these results to show that the decay in the predictability of double-mutant epistatic coefficients and local interactions of all orders can likewise be calculated given knowledge of the variance components. For instance, Fig. 1 *C* and *D* show the decay in the predictability of local pairwise and three-way interactions, respectively, for our simulated genotype–phenotype map.

The key intuition is that each variance component in Fig. 1*A* contributes a specific shape to the curves in Fig. 1 *B*–*D*, where more weight on the higher-order variance components results in a faster decay. Fig. 1 *E* and *F* give an illustration of this principle for the correlation between phenotypic values at different mutational distances, shown in Fig. 1*G*. In particular, the shape contributed by each of the variance components is given by a set of orthogonal polynomials known as the Krawtchouk polynomials (33, 48, 49, 57), which we show visually in Fig. 1 *E* and *F*. These functions naturally fall into two groups: 1) orders of interaction that contribute locally positive correlations and hence increase the similarity between adjacent phenotypes (Fig. 1*E*) and 2) orders of interaction that contribute locally negative correlations and hence decrease the similarity between adjacent phenotypes (Fig. 1*F*). (These two groups correspond to orders of interaction less than the expected Hamming distance between two random genotypes and orders of interaction greater than or equal to this quantity, respectively; *SI Appendix*).

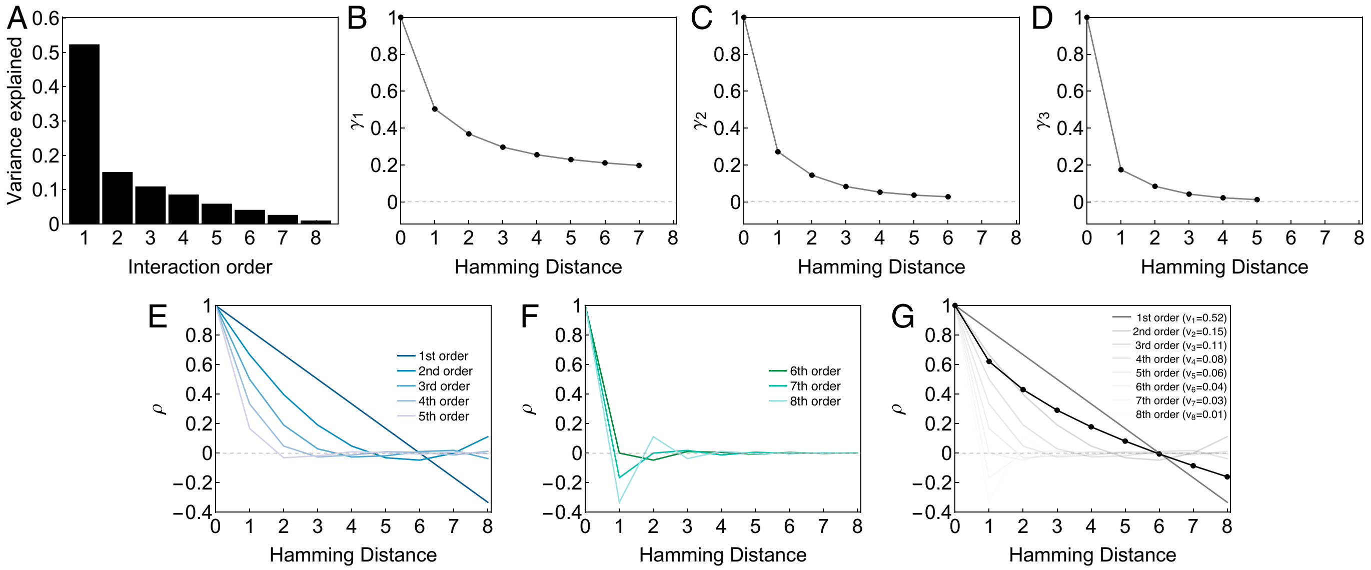

Fig. 1. (*A*–*D*) Summary statistics of a simulated genotype–phenotype map on sequences of length $\ell = 8$ with four alleles per site. (*A*) Decomposition of the genotype–phenotype map into the proportion of variance due to each order of genetic interaction. (*B*) Distance correlation function for mutational effects (γ_1), (*C*) distance correlation function for local double-mutant epistatic coefficients (γ_2), and (*D*) distance correlation function for local triple-mutant interactions (γ_3). Formulas for calculating these statistics can be found in *SI Appendix*. (*E*) Distance correlation functions for interaction orders contributing to positive local correlations. (*F*) Distance correlation functions for interaction orders contributing to negative local correlations. (*G*) The distance correlation function of the simulated genotype–phenotype map is a weighted average of curves shown in *E* and *F*, with the weights given by the variance components (v_k, $k = 1, \cdots, \ell$, shown in parentheses in the legend and graphically in *A*), and each curve in *G* is shaded proportionally to its weight.

As has long been known (33, 56, 58), the decay in correlations between observed phenotypic values shown in Fig. 1*G* can be obtained by weighting the corresponding curves from Fig. 1 *E* and *F* by the variance components in Fig. 1*A*. In *SI Appendix*, we extend these results by showing that the decay in local epistatic interaction coefficients of all orders can likewise be reconstructed as weighted sums of Krawtchouk polynomials with weights derived from the variance components (for an example, see *SI Appendix*, Fig. S1). Thus, for sequences of length ℓ, the problem of knowing how far our experimental observations generalize across increasingly divergent genetic backgrounds can in large part be characterized by these variance components ($\ell - 1$ parameters, since the variance components must sum to 1).

Bayesian Phenotypic Prediction. To incorporate our understanding of how the effects of mutations, epistatic coefficients, etc., change across increasingly divergent backgrounds, we take a Bayesian approach. Specifically, we wish to construct a prior distribution over all possible genotype–phenotype maps where the prior is concentrated on genotype–phenotype maps whose predictability decays in the desired manner.

In fact, such a family of priors is already well established in the literature in the form of random field models (33, 49, 56), which are parameterized in terms of the amount of variance due to each order of genetic interaction. Such models take the form of multivariate Gaussian distributions and can be constructed by drawing certain epistatic coefficients from zero mean normal distributions with appropriately chosen variances (*SI Appendix*).

Importantly, various previously developed methods for phenotypic prediction can be subsumed as particular (limiting) cases of Bayesian inference under this family of priors. For example, the solutions of the additive model and our recently proposed method of minimum epistasis interpolation (59) both arise as the maximum a posteriori (MAP) estimates in particular limiting cases where the prior fraction of variance due to additive effects goes to 1 (*SI Appendix*, Fig. S2). Similarly, the pairwise model (54) can be specified as the MAP estimate in a particular limit where the total fraction of variance due to additive and pairwise effects goes to 1 (*SI Appendix*, Fig. S2). Thus, we can view these previously proposed methods as encoding specific assumptions about how the predictability of mutational effects, epistatic coefficients, and phenotypic values changes as we move through sequence space, where these assumptions take the form of particular shapes for the curves in Fig. 1 *B*–*D* and *G*, which are in turn fully specified by the variance components, i.e., Fig. 1*A*.

Because these priors are multivariate Gaussian, under the assumption that experimental errors are also normally distributed, we can use Gaussian process regression (see ref. 50 for a review) to conduct inference under this prior. In particular, suppose our prior distribution is a mean zero Gaussian with covariance matrix $\mathbf{K}$, $\mathbf{y}$ is our vector of observations, and $\mathbf{E}$ is a diagonal matrix with estimates of the variance due to experimental noise for each of our observations down the main diagonal. Then the posterior distribution for our vector of predicted phenotypes $\mathbf{f}$ is normally distributed with mean

$$\widehat{\mathbf{f}} = \mathbf{K}_{\cdot B}(\mathbf{K}_{BB} + \mathbf{E})^{-1}\mathbf{y}, \qquad [1]$$

and covariance matrix

$$\mathbf{K} - \mathbf{K}_{\cdot B}(\mathbf{K}_{BB} + \mathbf{E})^{-1}\mathbf{K}_{B\cdot}, \qquad [2]$$

where $\mathbf{K}_{BB}$ is the submatrix of $\mathbf{K}$ indexed by the set of observed sequences B, and $\mathbf{K}_{B\cdot}$ and $\mathbf{K}_{\cdot B}$ are the submatrices of $\mathbf{K}$ consisting, respectively, of the rows and columns indexed by members of B (50). While at first glance, Eqs. **1** and **2** would appear difficult to evaluate at scale due to the need to invert the dense matrix $\mathbf{K}_{BB} + \mathbf{E}$, we show in *SI Appendix* that by exploiting the symmetries of sequence space the problem can be reformulated using only sparse matrix multiplication and that Hamiltonian Monte Carlo (51) can be used to efficiently draw samples from the posterior distribution. Thus, in practice, we are able to make comprehensive phenotypic predictions for genotypic spaces containing up to low millions of sequences.

Estimating Variance Components from Partial Data. The above analysis suggests that in order to make phenotypic predictions that appropriately incorporate the observed decay in the predictability of mutational effects, epistatic coefficients, etc., we should conduct Bayesian inference under a prior where these effects decay in a similar manner. One naïve implementation of this approach would be to simply use our observed distance correlation function to build the covariance matrix $\mathbf{K}$ for our prior by setting the covariance between each pair of sequences at distance d equal to the covariance between sequences at distance d in our data.

However, there is a subtle problem with this idea because the distance correlation function is a weighted sum of distance correlation curves for the various orders of interaction with the weights equal to the fraction of variance due to each order, as shown in Fig. 1*G*. The fact that these weights need to be positive and sum to 1 puts strong constraints on the shape that the correlation function can take for a function defined over all of sequence space, but these constraints need not hold for a partial sample, and unfortunately, using such a function to define a the matrix $\mathbf{K}$ would not result in a valid prior (in particular, $\mathbf{K}$ would not be positive definite; *SI Appendix*).

Thus, rather than using the observed covariance function to define our prior, we instead attempt to find the closest valid prior. We do this using weighted least squares, where the squared error for the correlation at distance d is weighted by the number of pairs of sequences at distance d (*SI Appendix*); this technique is formally equivalent to the idea of choosing a prior based on kernel alignment in the Gaussian processes literature (60). In addition, when producing this weighted least squares estimate, we take into account the magnitude of the experimental noise so as to distinguish between experimental uncertainty and the influence of any true uncorrelated component of the genotype–phenotype map, and we apply regularization to ensure that the resulting prior includes interactions of all orders (*SI Appendix*).

Validation on Simulated Data. Before using empirical variance component regression to analyze experimental data, we will first examine its characteristics on simulated data, where the ground truth is known. In particular, we will consider its behavior on a class of random genotype–phenotype maps where each possible combination of alleles at each possible combination of sites makes a separate additive contribution to the observed phenotype, and the magnitudes of these contributions are drawn from a standard normal distribution (*SI Appendix*).

Fig. 2*A* shows the distance correlation function for a simulated genotype–phenotype map of this class for the case where there are two alleles per site and $\ell = 16$ sites (65,536 possible genotypic states), and we have sampled 80% of these sequences to use as training data. For this case, the distance correlation function decays quite rapidly due to the substantial contribution of genetic interactions of orders 2 through 9 (Fig. 2*B*), and the effects of mutations become almost completely uncorrelated in genetic backgrounds that differ by five or more mutations (Fig. 2*C*). Based on this 80% sample, Figs. 2 *A*–*C* also show the posterior estimates (lines with error bars) for the correlation functions and variance components; on a 2021 Macbook laptop with 32 GB of RAM, inferring the MAP solution took less than 60 s, and we can generate upward of 10 posterior samples per minute. The corresponding MAP estimate produced an out-of-sample R^2 of 0.96 against the ground-truth data, indicating that given sufficient data, our method is able to accurately model genotype–phenotype maps with a substantial amount of higher-order epistasis.

To see how these results depend on the quantity of training data, we repeated this inference procedure over a range of sampling densities where we train on between 1 and 99% of the data (Fig. 2*D*). For comparison we also fit regularized pairwise and three-way regression models. Since both L_1 and L_2 regularized regression have been used in the literature to infer genotype–phenotype maps (39, 41, 54, 61, 62), here we fit the pairwise and three-way models using elastic net regression (*SI Appendix*) where the penalty term for model complexity is a mixture of L_1 and L_2 norms (63) and the regularization parameters are chosen via cross-validation. In addition to the linear regression models, we fit a global epistasis model (45) where the phenotype score is modeled as a nonlinear transformation of a latent additive phenotype on which each possible mutation has a background-independent effect (*SI Appendix*). Fig. 2*D* shows that empirical variance component regression consistently outperforms these

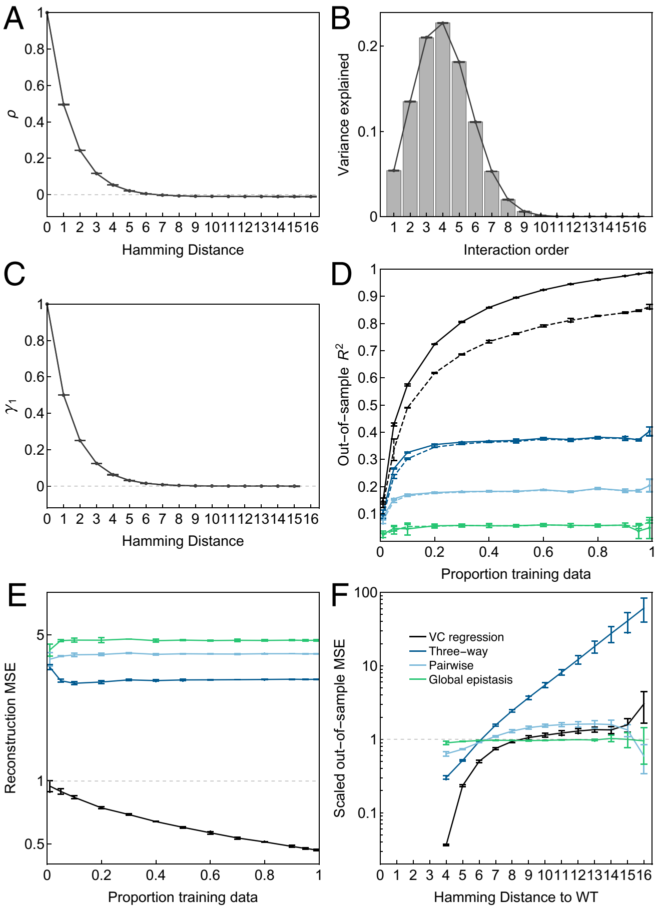

Fig. 2. Simulated biallelic genotype–phenotype map with $\ell = 16$. (*A*) Distance correlation of phenotypic values. (*B*) Variance components. (*C*) Distance correlation of the effects of single mutations. In *A*–*C*, gray represents statistics of the prior distribution inferred using 80% of the data, and black represents the posterior statistics estimated based on 2,000 Hamiltonian Monte Carlo samples from the resulting posterior (these curves are closely overlapping). Error bars indicate 95% credible intervals. (*D*–*F*) Comparison of model performance on the above simulated data, with error bars indicating 1 SD ($n = 3$ replicate simulations) and color legend given in *F*. Pairwise and three-way regression models were fit using elastic net regularization with regularization parameters chosen by 10-fold cross-validation (*SI Appendix*). The global epistasis model assumes the phenotype is a nonlinear transformation of an unobserved additive trait and was fit following ref 45. (*D*) Out-of-sample R^2 for a range of training sample sizes. Dashed lines give results for noised training data where the R^2 between the noised and true values is 0.8. (*E*) In-sample MSE as compared to the true values when trained on noised data (where we have scaled the phenotype so that the noise variance is 1). (*F*) Out-of-sample MSE as a function of Hamming distance from the wild-type (WT), for models trained on data generated using simulated mutagenesis (where we have scaled the phenotype so that the realized phenotypic variance is 1). The dashed line gives the expected mean square error for the trivial model that assigns the same phenotypic value to all genotypes.

other methods and continues to improve across the entire range of sampling frequencies.

Another important consideration is how our method is affected by measurement noise. An ideal regression procedure would not be overly affected by such noise and, indeed, could even provide a degree of in-sample denoising. To test the properties of our method in the presence of noise, we added uniform uncorrelated Gaussian noise of a magnitude chosen to reduce the R^2 between the true values and the simulated experimental values to 0.8. The dashed lines in Fig. 2*D* show that the accuracy of our predictions is reduced but that empirical variance component regression still outperforms the pairwise and three-way models. *SI Appendix*, Fig. S3*A* shows the performance of empirical variance component regression under a broader array of noise levels and sampling depths and demonstrates that for well-sampled, but noisy, datasets, our method can sometimes produce better out-of-sample predictions than would be obtained by direct experimental observation.

These observations suggest that empirical variance component regression should also have an in-sample denoising effect. To test this, we again trained on the noised data but instead considered the in-sample performance. We see that our method consistently provides a small to moderate denoising effect, reducing the noise variance by roughly a factor of 2 at high sampling levels, while the other models used for comparison are too misspecified to produce useable denoised estimates (Fig. 2*E*) (here the phenotype is scaled so that the noise has variance 1, and a mean squared error [MSE] less than 1 indicates denoising). *SI Appendix*, Fig. S3*B* shows the magnitude of this denoising effect over a broader range of experimental noise magnitudes and sampling depths.

While the above analyses were based on random sampling of genotypes, often empirical datasets are more localized in sequence space because the genotypes to be assayed are constructed via mutagenesis of a reference sequence. To assess how well our method behaves in this alternative sampling regime, we constructed a distribution over genotypic space based on random mutagenesis and then sampled from this distribution until we obtained 5% of all sequences to use as a training set. This resulted in a training set containing the reference sequence, all sequences at Hamming distances 1 to 3, 87% of sequences at distance 4, 17.5% of sequences at distance 5, 2.5% of sequences at distance 6, .26% of sequences at distance 7, .03% of sequences at distance 8, and no sequences at distance 9 or greater.

Training on this more localized dataset, we found that our out-of-sample predictions (Fig. 2*F*) were by far the most accurate at short distances (roughly, Hamming distance 4 to 7). At greater distances, our predictions were essentially the mean of the phenotypic distribution, consistent with our initial observation that for this genotype–phenotype map, phenotypic correlations decay rapidly to near zero within a few mutations. In contrast, the regularized three-way model inappropriately attempts to generalize the observations at short Hamming distances across all of sequence space, resulting in pathologically large MSEs far from the data (points above the dashed line; Fig. 2*F*). Critically, our empirical Bayes method uses the training data to determine how far generalization is possible, and thus, while it only extrapolates to short distances in highly epistatic genotypic–phenotype maps, it can also adapt to allow long-distance extrapolation for genotype–phenotype maps with larger additive and lower-order epistatic components (*SI Appendix*, Fig. S4).

In addition to the above results for the biallelic case, we also conducted simulation studies (*SI Appendix*, Fig. S5) of genotype–phenotype maps for the 4-allele case (e.g., for nucleic acid sequences, here $\ell = 8$) and the 20-allele case (for proteins, where we studied $\ell = 4$). Moreover, we repeated all these simulations using a sparse interaction variant (39) where 90% of the phenotypic contributions for combinations of alleles are set to zero (*SI Appendix*). Overall, we find that the results are very similar regardless of whether interactions are sparse or dense and that the $\ell = 8$ nucleic acid case is qualitatively similar to the biallelic case above. However, the protein sequence case with $\ell = 4$ was somewhat different in that the empirical variance component regression model, while arguably still being the best performing overall, behaved much more similarly to the three-way interaction model. This makes sense because for $\ell = 4$, a three-way interaction model contains almost all the orders of possible genetic interactions, so here the differences between the three-way and variance component models are primarily driven by how they weight additive vs. pairwise vs. three-way interactions rather than being driven primarily by the inability of lower-order models to capture higher-order epistasis as in our other examples. Finally, *SI Appendix*, Fig. S6 demonstrates the performance of our method on a simulated dataset where a model with pairwise interactions is transformed by a global epistasis nonlinearity.

Application to Protein G. As a first empirical example, we apply our method to a dataset derived from a deep mutational scanning study of the IgG-binding domain of streptococcal protein G (GB1) (37). This experiment assayed nearly all possible combinations of mutations at four sites (V39, D40, G41, and V54; $20^4 = 160,000$ protein variants), where pairs of mutations at these sites were previously shown to exhibit strong interactions (7). The library of protein variants was sequenced before and after binding to IgG-Fc beads, and the binding scores were determined as the log enrichment ratio (logarithm of ratio of counts before and after selection, normalized by subtracting the log ratio of the wild-type). Due to incomplete coverage of the input library, these data lack binding scores for 6.6% of possible variants.

We began by inferring the variance components of the GB1 landscape from the empirical autocorrelation function (Fig. 3*A*) using our least squares procedure applied to the empirical distance correlation function of all available data (93.6% of all possible sequences; see also *SI Appendix* for details). In Fig. 3*B*, we see that the vast majority of the variance in the data is estimated to be explained by the additive, pairwise, and three-way components (59, 37, and 4% of total variance, respectively), while the estimated contribution of the fourth-order component, which in this case is the only component that contributes to local anticorrelations, is negligible.

What is the practical meaning of these estimates for our task of phenotypic prediction? As an example, in Fig. 3*C*, we plot the correlation of mutational effects as a function of Hamming distance. We observe that the correlation of the effect of a random mutation is 0.72 between two genetic backgrounds that differ by one mutation and 0.32 for two maximally distinct backgrounds (Hamming distance = 3). This decay shows that while the effects of point mutations remain positively correlated across sequence space, the extent of this correlation is approximately twice as high in nearby sequences as opposed to maximally distant sequences and that therefore, when making predictions, we should give local observations of mutational effects approximately twice as strong a weight as distant observations of mutational effects.

Within our overall inference procedure, after estimating the variance components and using these variance components to construct a prior, the next step is to calculate the posterior distribution using the fine-scale information from the individual observations. Specifically, we calculated the MAP solution using all available data and drew 2,000 samples from the resulting posterior distribution. Using a 2021 MacBook laptop with 32 GB RAM, calculating the MAP solution took less than 1 min, and we could also produce samples from the posterior distribution at a rate of one to two samples per minute.

One immediate question about this posterior distribution is the extent to which its distance correlation function and variance components are similar or different from that of the prior (Fig. 3 *A*–*C*). We find that the posterior gives very tight estimates of the variance components and correlation structure of the true genotype–phenotype map but that these estimates differ somewhat from the prior, with the third-order interactions being roughly 1.6 times as strong in the posterior (Fig. 3*B*), which results in a slightly faster decay in the predictability of mutational effects as we move through sequence space (Fig. 3*C*). Thus, we conclude that our prior distribution provided a qualitatively reasonable estimate of the overall statistical features of the data, but the rough estimate used to define the prior can be further refined using our full inference procedure.

We compared the predictive accuracy of empirical variance component regression against several other prediction methods by plotting out-of-sample R^2 against a wide range of training sample sizes (Fig. 3*D*). We see that the out-of-sample R^2 of the additive model and the global epistasis model stay nearly constant, regardless of training sample size, consistent with their low number of model parameters and low flexibility. The modest R^2 for the global epistasis model also indicates a substantial degree of specific epistasis [i.e., interactions between specific subsets of sites, as opposed to interactions due to a global nonlinearity (34)]. In terms of the regression models that do include these specific interactions, the pairwise model is among the top models for low training sample size but fails to improve beyond 20% training data, while the three-way model performs strongly with a large amount of data but underperforms when data are sparse. We see that our empirical variance component regression method performs equivalently to the pairwise model at low data density and similarly to the three-way model at high data density (remaining marginally superior at very high sampling) and thus provides the strongest overall performance.

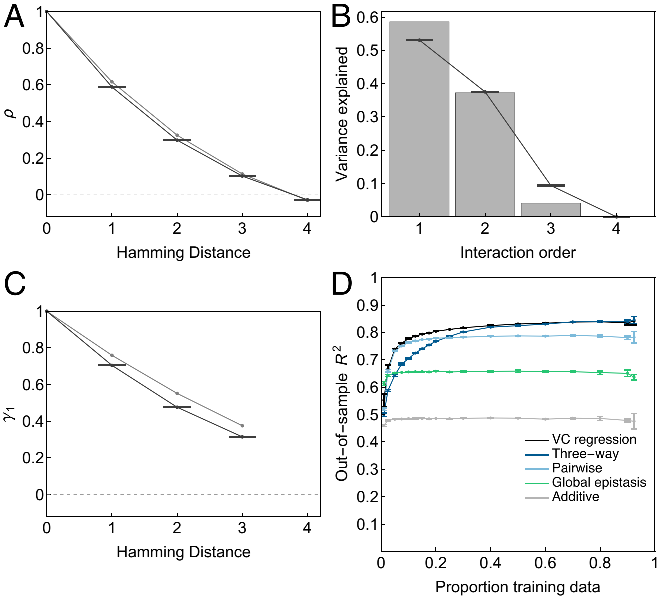

Fig. 3. Analyses of the GB1 combinatorial mutagenesis dataset (37). (*A*) Distance correlation of phenotypic values. (*B*) Variance components. (*C*) Distance correlation of the effects of single mutations. In *A*–*C*, gray represents statistics of the prior distribution inferred from the full dataset consisting of 149,361 genotypes (93.6% of all possible sequences), and black represents the posterior statistics estimated based on 2,000 Hamiltonian Monte Carlo samples. Error bars indicate 95% credible intervals. (*D*) Comparison of model performance in terms of out-of-sample R^2 for a range of training sample sizes calculated for five replicates. Additive models were fit using ordinary least squares. Pairwise and three-way regression models were fit using elastic net regularization with regularization parameters chosen by 10-fold cross-validation (*SI Appendix*). The global epistasis model assumes the binding score is a nonlinear transformation of an unobserved additive phenotype and was fit following ref. 45. Error bars represent 1 SD.

Application to Human 5′ Splice Site Data. To provide an application of our method to a nucleic acid genotype–phenotype map, we turn to an analysis of a high-throughput splicing assay that measured the activity of nearly all possible 5′ splice sites (31). The 5′ splice site (5′ss) is a nine-nucleotide sequence that spans the exon–intron junction. It comprises 3 nt at the end of the upstream exon (denoted as positions −3 to −1) and 6 nt at the beginning of the intron (coded +1 to +6). The consensus 5′ss sequence in humans is CAG/GUAAGU, with the slash denoting the exon–intron junction. At the beginning of the splicing reaction, the 5′ss is recognized by a small RNA known as the U1 small nuclear RNA (snRNA) whose 5′ sequence is complementary to the consensus 5′ss sequence (64). In ref. 31, the authors used a massively parallel splicing assay to estimate the splicing efficiency of 93.8% of the 32,768 possible 5′ss sequences of the form NNN/GYNNNN for intron 7 of the gene *SMN1*, using a barcoded minigene library transiently transfected into human cells. Splicing efficiency was measured in units of percent spliced in (PSI), which was estimated as the ratio between the exon inclusion read count and the total read count (which comprises both exon inclusion and exon skipping reads) divided by the corresponding ratio for the consensus sequence, expressed as a percentage. Computational times were somewhat faster for this dataset than for GB1, with the MAP estimate taking roughly 25 seconds on a 2021 MacBook and samples from the posterior distribution being produced at a rate of roughly 10 samples per minute.

Fig. 4*A* shows the distance correlation function of PSI for the observed sequences. These correlations drop off quite rapidly, with sequences differing at five or more positions having PSIs that are essentially uncorrelated. The associated estimated variance components are shown in Fig. 4*B*. These indicate that pairwise interactions account for the largest proportion of the sample variance (42.2%) but that there are also substantial higher-order interactions, with the variance due to five-way interactions (13.7%) being comparable to those of the additive and three-way component. The orders of genetic interaction corresponding to locally negative correlations (order ≥6 since the Hamming distance between two random sequences is equal to $\frac{3}{4} \times 8 = 6$) are estimated to play a relatively small but perhaps nonnegligible role, accounting for 6.0% of the total variance. In Fig. 4*C*, we found the correlation of mutational effects for two backgrounds that differ by one mutation is roughly 50% but decays to roughly zero for distant genetic backgrounds. Sampling from the posterior distribution, we see that the statistical characteristics of the posterior again have very small credible intervals and remain similar to those estimated using our least squares procedure, with a slightly increased contribution of pairwise and three-way interactions and a decreased contribution of five-way interactions (Fig. 4*B*). Overall, the splicing landscape appears to be dominated by interactions of order 2 through 5, resulting in positive correlations between the splicing activity of nearby genotypes but a relatively limited ability to generalize our observations to distant regions of sequence space. This qualitative behavior is consistent with our mechanistic intuition; e.g., mutations that substantially decrease U1 snRNA binding in the context of a functional splice site are likely to have no impact in an already nonfunctional sequence context.

To see how our model performs when greater or lesser amounts of data are available, we compared the predictive power of the same five models as in Fig. 3*D* by plotting their out-of-sample R^2 against a wide range of training sample sizes (Fig. 4*D*). The rank order of the models is largely consistent throughout the sampling range. More importantly, we see that the variance component model adapts to increasing data density at a much faster rate than the other models. For example, with low sampling density (training sample size $<20\%$ of all possible sequences) the three-way model has similar performance to our model, but the performance gap between the two models quickly widens as the training data become dense. The empirical variance component regression model is able to achieve a final $R^2 = 0.83$ with 93% of the sequence space assigned as training data ($n = 30,474$), compared with the three-way model which had $R^2 = 0.72$. This difference in model performance is consistent with the observation of a substantial contribution of higher-order interactions ($k > 3$), which the low-order regression model is unable to accommodate.

Another question is the qualitative nature of the genetic interactions captured by our model. We note that the global epistasis model provides a remarkably good fit to the data, considering that it has only a few more parameters than a simple additive model. In *SI Appendix*, Fig. S7, we see that the global epistasis model approximates the splicing landscape with a sigmoid-like function that maps an unobserved additive trait to the PSI scale. This is as we might expect under a simple biophysical model where each position in the splice site makes a context-independent contribution to the binding energy of the U1 snRNA to the 5′ss, and then this binding energy is mapped via a nonlinear function to PSI (3). However, the global epistasis model fails to capture some important features of the data, most notably a group of false-negative sequences that are predicted to be nonfunctional by the global epistasis model but experimentally show moderate to high measured PSI (*SI Appendix*, Fig. S8*A*). Using variance component regression, we were able to accurately predict these outlier sequences (*SI Appendix*, Fig. S8*B*). We thus conclude that while the global epistasis model provides a relatively simple first-pass understanding of the landscape of 5′ss activity, our empirical variance component regression is able to capture more of the fine-scale features of this particular genotype–phenotype map.

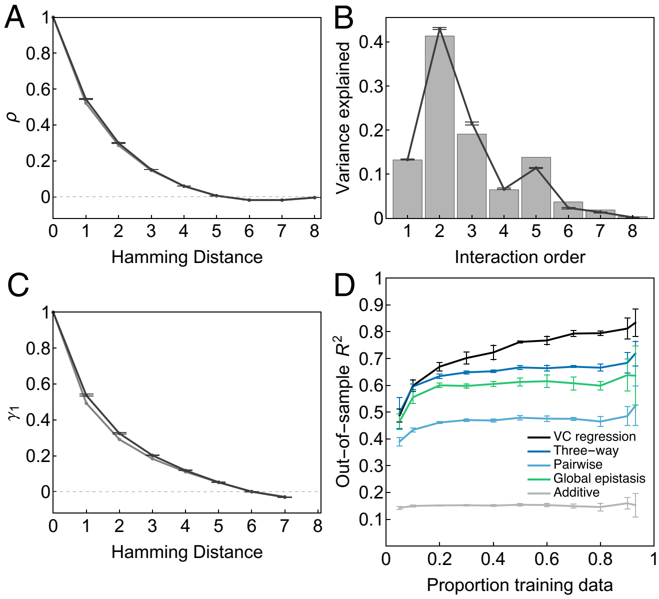

Fig. 4. Analyses of the *SMN1* 5′ss combinatorial mutagenesis dataset (31). (*A*) Distance correlation function of the splicing phenotype (PSI). (*B*) Variance components. (*C*) Distance correlation of single-mutant effects. Gray represents statistics of the prior distribution inferred from the full dataset consisting of 30,732 genotypes (93.8% of all possible splice sites), and black represents the posterior statistics estimated using 2,000 Hamiltonian Monte Carlo samples. Error bars indicate 95% credible intervals. (*D*) Out-of-sample R^2 of the five models plotted against a range of training sample sizes. Error bars represent 1 SD calculated for five replicates for each sample size.

Although predictions on held-out data provide one means of testing model performance, a stronger test is to conduct low-throughput experiments to validate the predictions of our method on sequences that were not measured in the original experiment. The *SMN1* dataset provides a suitable case study for this application since the original dataset does not report the PSI of 2,036 sequences (6.2% of all possible 5′ss) due to low read counts. To assess the predictive power of our method for these truly missing sequences, we first made predictions for all unsampled sequences using all available data. We then selected 40 unsampled sequences whose predicted values are evenly distributed on the PSI scale. The true PSIs of these sequences were then quantitatively measured using low-throughput radioactive RT-PCR (31) (Materials and Methods and Fig. 5*A*). Overall, our method achieves a reasonable qualitative agreement with the low-throughput measurements (Fig. 5*B*) but differs systematically in that the transition between nearly 0 and nearly 100 PSI is more rapid in the low-throughput measurement than in our predictions. Intuitively, we can understand the source of this discrepancy in terms of the geometry of the splicing landscape, which features a bimodal distribution of PSIs with separate modes near 0 and 100 (31) and a sharp transition between these two sets of sequences in sequence space (*SI Appendix*, Fig. S7). Because phenotypic observations generalize farther in most regions of sequence space than they do near this boundary between low and high PSI, our method tends to smooth anomalously sharp features of this type. This results in out-of-sample predictions that are more smoothly graded, rather than threshold-like, in the vicinity of this boundary.

Structure of the SMN1 splicing landscape. Besides making accurate phenotypic predictions, it is important to understand the qualitative features of a genotype–phenotype map, both with regard to how the underlying mechanisms result in observed genetic interactions and how these genetic interactions affect other processes, such as molecular evolution and disease. For simple models, such as pairwise interaction models or global epistasis models, extracting these qualitative insights can often be achieved by examining the inferred model parameters. Here we take a different approach and attempt to understand these major qualitative features by constructing visualizations based on the entire inferred activity landscape. Because we previously conducted a detailed analysis of this type for the GB1 dataset (59), we focus here on the inferred activity landscape for the 5′ss.

In particular, our visualization method (65) is based on constructing a model of molecular evolution under the assumption that natural selection is acting to preserve the molecular function measured in the assay. The resulting visualization optimally represents the expected time it takes to evolve from one sequence to another (*SI Appendix*) and naturally produces clusters of genotypes where the long-term evolutionary dynamics are similar for a population starting at any genotype in that cluster (e.g., genotypes on the slopes leading up to a fitness peak will tend to be plotted near that peak). To make such a visualization for our splicing data, we built a model of molecular evolution based on the MAP estimate obtained above (*SI Appendix*). We then used the subdominant eigenvectors of the transition matrix for this model as coordinates for the genotypes in a low-dimensional representation; these coordinates are known as diffusion axes (66) since they relate closely to how the probability distribution describing the genotypic state of a population evolving under the combined action of selection, mutation, and genetic drift is likely to diffuse through sequence space (65, 67). Note that the ability of empirical variance component regression to produce a complete estimate of the genotype–phenotype map is important for allowing this type of analysis because otherwise missing genotypes would effectively be treated as inviable.

The resulting visualization using the first three diffusion axes is shown in Fig. 6 *A* and *B*. Here genotypes are points (colored by the number of times that particular 5′ss is used in the human genome; Materials and Methods), and edges connect genotypes that differ by single point mutations. Remarkably, each of these diffusion axes turns out to have an interpretable meaning. First, diffusion axis 1 separates functional 5′ss (large positive values) from nonfunctional 5′ss (negative values), as can be seen in Fig. 6C, which plots the estimated PSI against diffusion axis 1. Second, diffusion axis 2 captures the typical physical location of consensus nucleotides within high-activity 5′ss. Specifically, as one moves up diffusion axis 2, the mean position of consensus nucleotides shifts from the exonic portion (5′-end) of the splice site to the intronic (3′) portion (Fig. 6*D*). This reflects a previously observed "seesaw" linkage pattern (68–71) between the intronic and exonic portions of the splice site, where nonconsensus nucleotides are typically clustered in one or the other of these regions but not both. Finally, we find that diffusion axis 3 encodes whether or not mutations are present at the +3 position (Fig. 6*B*, *Inset*), where 5′ss with the consensus A tend to be plotted at negative values on diffusion axis 3, and 5′ss with mutant nucleotides on +3 tend to be plotted at more positive value.

To reveal more detailed structures in the sublandscape of high-activity 5′ss, we focus on the 818 5′ss with PSI > 80. These sequences are plotted using diffusion axes 2 and 3 in Fig. 6*E*. Fig. 6*F* further groups these high-activity 5′ss by their mutant states (consensus vs. mutant) at six positions, −1, −2, +3, +4, +5, and +6, and represents each group by a dot (see also *SI Appendix*, Fig. S10). On a coarse level, we see two main groups of 5′ss separated along diffusion axis 3 that correspond to sequences with the canonical A (bottom) and mutant nucleotides (predominantly G, top) at position +3 and also a distinction into three main groups along diffusion axis 2, where the three groups correspond to mutant +5 (left), consensus for both +5 and −1 (center), and mutant −1 (right). This separation between sequences containing mutations at positions +5 and −1 (see also Fig. 6 *A*, *Inset*) arises due to a previously noted incompatibility between mutations at these two sites (31, 69–71), so that evolutionary trajectories that maintain high levels of splicing must typically wait for a reversion of a +5 mutation before fixing a mutation at the −1 position (and vice versa).

On a finer scale, Fig. 6*F* and *SI Appendix*, Fig. S10 show that each functional combination of mutant states on the three major positions (−1, +3, +5) can be thought of as defining a cube (plotted with dark edges) corresponding to the $2^3 = 8$ possible mutant states on the three minor positions (+2, +4, +6). Whereas having either a single −1 or +5 mutation is compatible with having many different combinations of mutations at the three minor positions (complete cubes on the bottom half of Fig. 6*F* and *SI Appendix*, Fig. S10), in a +3 mutant background combined with either a −1 or +5 mutation, most combinations of minor mutations result in low activity (two partial cubes on the top half of Fig. 6*F* and *SI Appendix*, Fig. S10).

While in broad strokes the above pattern is consistent with a global epistasis model in that minor mutations are mostly tolerable except in weak genetic backgrounds, the global epistasis model further requires that any specific mutation that is tolerated in a weaker background must also be tolerated in a stronger background. However, we instead observe a pattern where in weak backgrounds mutations are only tolerable if an adjacent major site is already mutated (*SI Appendix*, Fig. S11). In particular, −2 mutations are often tolerable in a −1 + 3 background but never tolerable in a +3+5 background, and +6 mutations are often tolerable in a +3+5 background but not in a −1 + 3 background. More specifically, the deleterious effect of the +6 mutation in the −1+3 mutant background is almost completely abrogated in the +3+5 mutant background (*SI Appendix*, Fig. S11), consistent with previous findings (70). Likewise, we observe that −2 mutations are often tolerated in the −1+3 mutant background (median effect = −18.6 PSI, calculated for sequences with consensus bases on all other positions and with PSI > 80) but have much larger effects when +3+5 are mutated (median effect = −93.8 PSI). The global epistasis model is even more clearly violated by the fact that the +3+5+6 mutant background can also tolerate certain +4 mutations that would not be tolerable in the absence of a +6 mutation (*SI Appendix*, Fig. S10). Specifically, these +4 mutations result in two highly functional 5′ss: CAG/GUUGUA, which binds to U1 snRNA using a noncanonical geometry known as an asymmetric loop (72), and AAG/GUGGAC, which does not seem to correspond to any known alternative binding geometry but occurs as an annotated splice site 14 times in the human genome (Materials and Methods) and has a high level of splicing activity as confirmed via low-throughput validation (*SI Appendix*, Fig. S12).

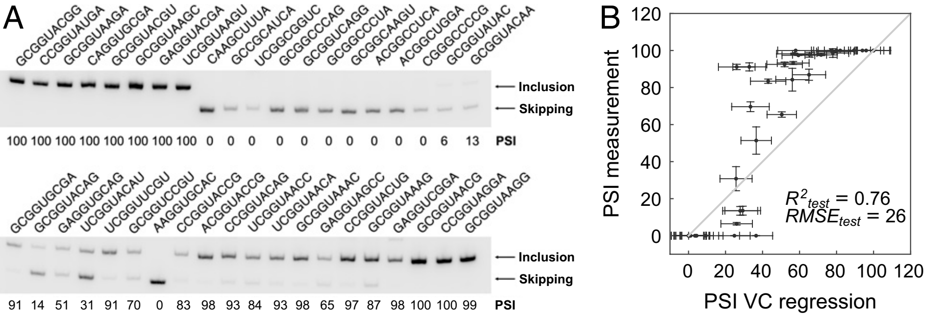

Fig. 5. Manual validation of predicted PSI for 40 unmeasured *SMN1* 5′ss. (*A*) Gel images of manually validated sequences. For each lane, the top band corresponds to mRNA product containing exon 7 (exon inclusion), while the bottom band correspond to mRNA product without exon 7 (exon skipping). PSI is indicated below each lane. Gel images are representative of triplicates. (*B*) Scatterplot showing measured PSI values versus PSI values predicted by the variance component regression. Horizontal error bars correspond to 1 SD of the posterior distribution. Vertical error bars correspond to 1 SD around the mean PSI estimated using three replicates in the manual validation. Since unlike the high-throughput measurements, the low-throughput PSIs are inherently restricted to the range 0 to 100, in this analysis we likewise cap the predicted PSIs to lie in this same range (see *SI Appendix*, Fig. S9 for the unrestricted predictions).

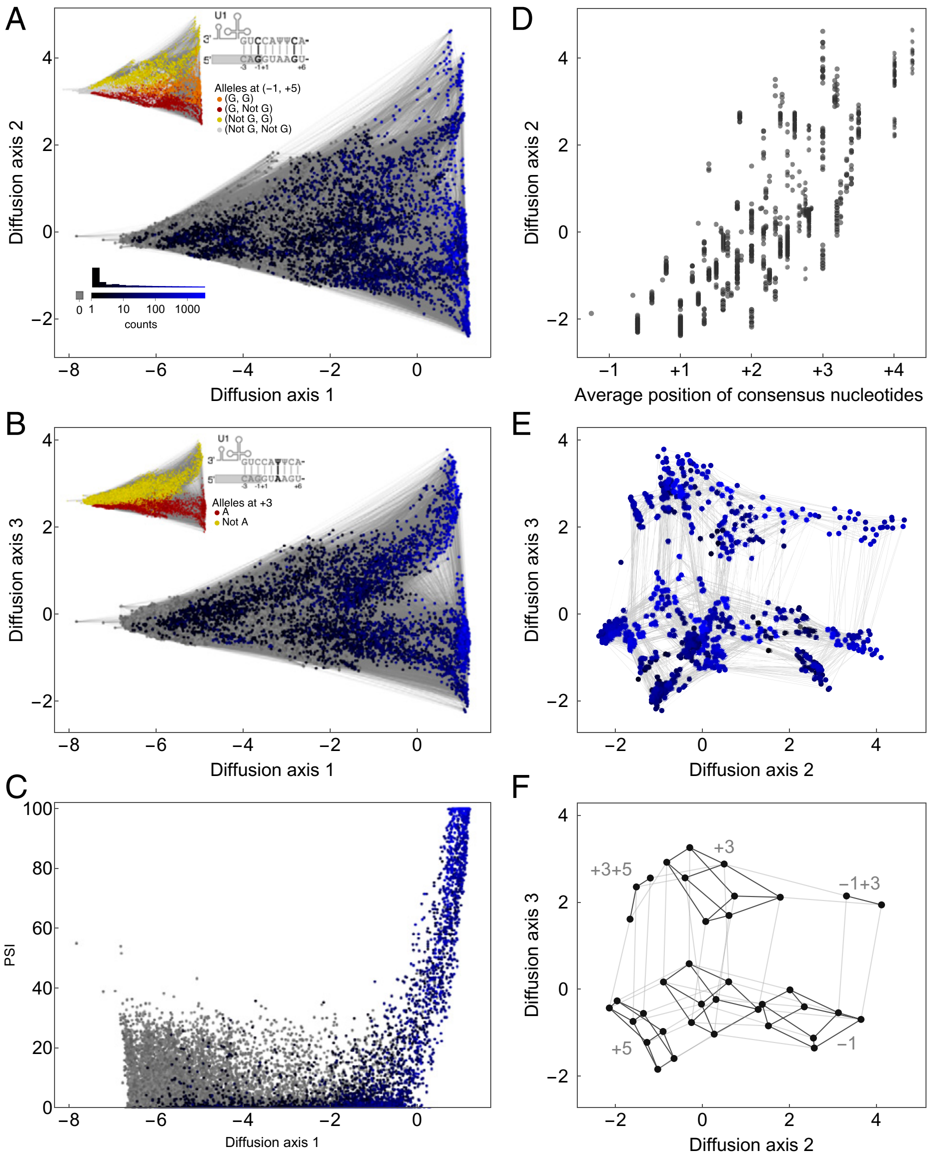

Fig. 6. Visualization of the *SMN1* splicing landscape reconstructed using empirical variance component regression. Genotypes are plotted using the dimensionality reduction technique from ref. 65 (*SI Appendix*). (*A*) Visualization of all 32,768 splice sites using diffusion axes 1 and 2. Two splice sites are connected by an edge if they differ by a point mutation. (*B*) Visualization of all 32,768 splice sites using diffusion axes 1 and 3. (*C*) PSI versus diffusion axis 1. We see that diffusion axis 1 separates high PSI versus low PSI sequences. Genotypes are colored in *A*–*C* according to the number of times they are observed as annotated splice sites in the human reference genome (hg38); see *Inset* in *A* for the color scale and a histogram of the numbers of counts. Gray dots represent sequences not present as annotated splice sites (65.9% of all sequences of the form NNN/GYNNNN). (*D*) Diffusion axis 2 versus the average physical position of the consensus nucleotides of the 818 splice sites with predicted PSI $> 80\%$. (*E*) Visualization of the 818 splice sites with predicted PSI $> 80\%$ using diffusion axes 2 and 3. Genotypes colored as in *A*–*C*. (*F*) Abstracted version of *E*. Splice sites are grouped by mutational states (consensus vs. mutated) at positions −1, −2, +3, +4, +5, and +6. Each dot corresponds to a group of sequences with a prescribed pattern of consensus or mutated states on the six sites. Two groups are connected by an edge if they differ in mutational state at exactly one site. Gray lines represent differences at positions −1, +3, and +5. Black lines represent differences at positions −2, +4, and +6. Only groups containing splice sites with >80% PSI are shown, resulting in six (in)complete cubes with black edges, each representing a combination of mutational states on the three major sites −1, +3, and +5. The incompleteness of a cube indicates the absence of a combination of mutational states at positions −2, +4, and +6. Note that no cubes contain both −1 and +5 mutant states, indicating a major incompatibility between mutations at these two sites.

In summary, we conclude that the 5′ss activity landscape contains many qualitatively different types of genetic interactions. At a broad scale, the splicing landscape can be understood in light of the global epistasis model, where PSI is modeled as a nonlinear function of an underlying additive phenotype, and interactions between major mutations arise due to the sharp threshold of the global nonlinearity. However, at a finer scale we discover that the effect of a mutation can be strongly modulated by other mutations in ways that are incompatible with the global epistasis model, both due to specific pairwise interactions, such as the interaction between the +5 and +6 positions, but also due to more complex interactions associated with changes in the physical geometry of U1 snRNA binding (72).

Discussion

In this paper, we address the problem of how to model the complex genetic interactions observed in high-throughput mutagenesis experiments. Our method is based on the simple idea that the type and extent of epistasis that we predict outside our observed data should be similar to the type and extent of epistasis observed in the data itself. This information about the type and extent of epistasis can be extracted from how correlations between phenotypic values decay as one moves through sequence space, and the decay of these correlations is determined by ℓ variance components, where ℓ is the sequence length. By estimating these variance components from the data, we can construct a prior distribution over all possible genotype–phenotype maps that is concentrated on the subset of genotype–phenotype maps where the effects of mutations generalize in the same manner as occurs in our observations. Conducting Bayesian inference under this prior results in estimates that reflect the character of epistasis in the training data, so that data that appear largely additive result in largely additive predictions, even far from the data, whereas more epistatic training data result in a rugged prior such that the model only attempts to make informative predictions near the data cloud. The method can be used to predict phenotypic values for specific unmeasured genotypes, to reduce the impact of experimental noise, and to construct combinatorially complete estimates of the genotype–phenotype map that are amenable to downstream analysis and visualization.

One way to understand our contribution here is to see it as an integration between practical Gaussian process-based methods for analyzing genotype–phenotype maps (73–75) and the classical spectral theory of fitness landscapes (49, 56, 58), which provides the most sophisticated mathematical theory of genetic interactions currently available. Within this theoretical literature, so-called random field models identical to the family of priors we propose have been extensively studied (33, 49, 56), and we have leveraged this existing knowledge to craft priors that encode comprehensible beliefs about the structure of high-dimensional genotype–phenotype maps.

Our results here also provide some significant additions to the spectral theory of fitness landscapes that help to provide a more intuitive view of this complex area of mathematical theory. First, we suggest that higher-order epistatic interactions can be qualitatively classified into two types, corresponding to interactions that result in locally positive correlations or locally negative correlations. The idea of an anticorrelated component to a genotype–phenotype map has been discussed previously in the literature in terms of the "eggbox" component (12, 47) which is perfectly anticorrelated between adjacent genotypes (i.e., whether the phenotypic value is high or low flips with each step one takes through sequence space, similar to the alternating peaks and valleys of an egg carton). Our analysis shows that there is actually a whole set of orders of genetic interaction with a similar character, corresponding to all orders of genetic interaction higher than the average number of differences between two random sequences. However, our main interest is in the components that produce locally positive correlations (which appear more likely to arise under most conceivable physical mechanisms), with the balance between these higher-order locally correlated components controlling how precisely phenotypic correlations decay with increasing Hamming distance.

Second, we defined a summary statistic $\gamma_k(d)$ which, beyond simple phenotypic correlations, measures how mutational effects ($k = 1$) or epistatic coefficients ($k > 1$) decay as the distance d between genetic backgrounds increases. The correlation of mutational effects as a function of distance between genetic backgrounds has been previously termed $\gamma(d)$, which is used to measure the ruggedness of the landscape (12, 47). Here we generalize this measure to epistatic coefficients of any order and show that the distance correlation of epistatic coefficients of order k is in fact determined solely by the components of the landscape of order larger than k (see *SI Appendix*, where we provide a simple formula showing the relationship between different orders). This result can also help us understand why our method outperforms pairwise and three-way epistatic models. Specifically, we show that models that include only up to kth-order epistatic interactions in fact make the very strong assumption that any observed kth-order interactions generalize across all genetic backgrounds. Incorporating higher-order interactions is then equivalent to relaxing this strong assumption and allowing these lower-order interactions to change as one moves through sequence space.

Third, in *SI Appendix* we provide some additional results to better understand the possible geometries produced by any given order of genetic interaction. In particular, we consider the mean phenotype as a function of Hamming distance from some focal genotype, which is a classical coarse descriptor of genotype–phenotype map and fitness landscape structure (e.g., refs. 76–78). We show that for a pure kth-order interaction this mean function is in fact equal to its distance correlation function up to a multiplicative constant. As a consequence of this result, the distance mean function for a model containing up to only kth-order terms must be a kth-order polynomial, so that, e.g., in a pairwise interaction model the mean fitness at a given distance from a focal genotype is always a quadratic function of distance. However, from a biological perspective, we might often expect mean fitness to take more complex shapes, such as a sigmoid (45, 79) (which obviously cannot be well-approximated by a quadratic), providing an explanation for the need to incorporate higher-order interactions in order to provide qualitatively reasonable fits.

Our analyses of experimental data, particularly our analysis of $5'$ss splicing efficiency, provide insight into the context dependence of mutational effects. For example, we see that U1 snRNA can bind to the $5'$ss in different physical configurations, such as binding largely to the exonic versus intronic portion of the splice site or binding in an alternative conformation, and that these differences result in different regions of sequence space where specific mutations are functionally tolerable. Because the spatial scale of this heterogeneity in mutational effects determines the sampling density needed to make accurate predictions, our results also suggest that small pilot studies sufficient to provide estimates of the variance components may be useful in the design of this type of experiment. More generally, as a Gaussian process-based method, empirical variance component regression is well suited to approaches that iterate model fitting and additional experimental data acquisition, with the goal of either further refining the model or generating highly optimized sequences (44, 73–75, 80).

The method we propose here has some commonalities with minimum epistasis interpolation (59), another method we recently proposed for phenotypic prediction that includes genetic interactions of all orders, but the two methods differ in their aims. Minimum epistasis interpolation aims to produce a highly conservative reconstruction of the genotype–phenotype map by making the effects of mutations as consistent as possible between adjacent genetic backgrounds. In contrast, empirical variance component regression aims to produce a more realistic reconstruction, where the extent and type of epistasis present in the reconstruction should be similar to the extent and type of epistasis present in the data itself. Depending on the needs of the user, both methods can be conducted either in a Bayesian manner or as a form of L_2-regularized regression (81) (where our MAP estimate is equivalent to the L_2 regularized solution; *SI Appendix*). From a regularization perspective, the main difference between these methods is that they penalize different orders of genetic interaction differently, either with a penalty that increases quadratically with the order of interaction, in the case of minimum epistasis interpolation (see also ref. 82), or a penalty determined by the empirically estimated variance components, in the case of empirical variance component regression. In *SI Appendix*, Fig. S13, we provide a comparison of model performance between these two methods in the GB1 and *SMN1* dataset. We observe that empirical variance component regression consistently outperforms minimum epistasis interpolation in both the GB1 and *SMN1* datasets. This difference in model performance is likely due to the misalignment between the quadratically increasing penalty imposed by minimum epistasis interpolation and the actual variance components in the data. Overall, empirical variance component regression is likely the superior method if high predictive performance is desirable. On the other hand, minimum epistasis interpolation is a more conservative approach and has many simple theoretical properties (59). Thus, it should be preferred when theoretical guarantees for model behavior are more important.

Our method also has commonalities with several other techniques for the analysis of genotype-phenotype maps. In their original paper suggesting that random field models could be used as theoretical approximations for observed genotype–phenotype maps, Happel and Stadler (49) also attempted to estimate variance components for computational models of the genotype–phenotype map using distance correlation functions but found that they could not do so reliably. Importantly, their method imposed a sparse reconstruction wherein only a few variance components had nonzero contributions, which in our context would correspond to inappropriately strong priors precluding the inclusion of most orders of genetic interaction. Hordijk and Stadler (83) were subsequently able to produce somewhat better estimates containing all orders of interaction for computational models of the genotype–phenotype map based on random walk correlation functions but still found that these estimates contained high uncertainty. Our kernel alignment estimates are likely performing better primarily because we are working with shorter length sequences, but more generally, in our procedure, kernel alignment is only being used to produce a reasonable choice of prior, and so high-precision estimates are less important here. Our use of Gaussian process regression with a distance-based covariance function to predict missing phenotypes is also similar to that employed more recently by Agarwala and Fisher (84), who use the history of fitnesses and mean selection coefficients encountered along an adaptive walk to predict the distribution of fitness effects for the next step in the walk in the limit of long sequence length. Their focus on theoretical results in the limit of long sequence length is complementary to the methods for empirical data analysis proposed here.

One potential limitation of our approach is our choice to select the hyperparameters based on the point estimates supplied by our training data, i.e., by kernel alignment (60). It may well be possible to produce more accurate predictions by choosing hyperparameters by maximizing the evidence (50) or via a hierarchical Bayesian model where we integrate over our uncertainty in the values of these hyperparameters, at the cost of a much greater computational burden. However, an advantage of the empirical Bayes approach is that it provides a clear conceptual separation between the first step of estimating the type and extent of epistasis present in a set of phenotypic observations, and the second step of using these estimates to make additional phenotypic predictions.

Another limitation concerning empirical variance component regression is that it is unable to explicitly model any overall nonlinearity that may be present in the genotype–phenotype map; i.e., it does not explicitly model nonspecific or global epistasis (3, 10, 34, 35, 45, 85, 86). Rather, empirical variance component regression must learn any such global structure based on consistent patterns in the observations themselves. For instance, whereas the global epistasis model is able to easily handle the saturation of PSI at 0 and 100%, empirical variance component regression must learn these flatter regions based on the consistently small effects of mutations in particular regions of sequence space, rather than via an overall nonlinearity that is assumed by the structure of the model. Incorporating the possibility of such global nonlinearities would be an important extension to the methods presented here.

A final limitation concerns the applicability of the method we propose to very large datasets. In our implementation, we take advantage of the isotropic property of the prior distribution (i.e., that covariance depends only on Hamming distance) and the highly symmetric graph structure of sequence space. This allows us to express the covariance matrix and its inverse as polynomials in the highly sparse matrix known as the graph Laplacian, which makes inference possible on sequence spaces containing up to low millions of sequences. However, due to the exponential growth of biological sequence space as a function of sequence length, practically, this still limits us to nucleic acid sequences of length 11 or less and amino acid sequences of length 5 or less. Alternatively, working directly with the dense covariance matrices in Eqs. **1** and **2**, it is possible to analyze sequence of any length, but this is only computationally feasible for datasets containing up to a few tens of thousands of observed sequences. Although here we have successfully analyzed datasets that contain up to hundreds of thousands of sequences, more work is needed to scale these methods to even larger datasets and sequence spaces.

Materials and Methods

Low-Throughput Validation of Unsampled *SMN1* 5′ss. To assess the predictive accuracy of our method for the activity of truly unsampled splice sites, we selected 40 5′ss absent in the *SMN1* dataset that are evenly distributed on the predicted PSI scale. We quantified the splicing activities of the selected 5′ss in the context of an *SMN1* minigene that spans exons 6 to 8 with the variable 5′ss residing in intron 7. The minigene construct is the same as the one used to generate the high-throughput data (31) (minigene sequence is available at https://github.com/jbkinney/15_splicing). Specifically, minigenes containing variable 5′ss were inserted in to the pcDNA5/FRT expression vector (Invitrogen). Then, 1 μg of minigene plasmid was transiently transfected into HeLa cells, which were collected after 48 h. RNA was isolated from the minigene-expressing HeLa cells using TRIzol (Life Technologies) and treated with RQ1 RNase-free DNase (Promega). cDNA was made using Improm-II Reverse Transcription System (Promega), following the manufacturer's instructions. The splicing isoforms were then amplified with minigene-specific primers (F: CTGGCTAACTAGAGAACC CACTGC; R: GGCAACTAGAAGGCACAGTCG) and ^{32}P-labeled dCTP using Q5 High-Fidelity DNA Polymerase (New England Biolabs) following the manufacturer's instructions. PCR products were separated on a 5.5% nondenaturing polyacrylamide gel and were detected using a Typhoon FLA7000 phosphorimager. Finally, we used ImageJ (NIH) to quantify isoform abundance, in the process accounting for cytosine content. All 5′ss were assessed in triplicate.

Acquisition of Human 5′ Splice Sites. Human 5′ss were extracted from GENCODE Release 34 (GRCh38.p13) (available at https://www.gencodegenes.org/human/).

Data, Materials, and Software Availability. A Python command-line interface, vcregression, that implements the empirical variance component regression method described here has been deposited in GitHub (https://github.com/davidmccandlish/vcregression) (87), together with the other scripts necessary to replicate the results presented here. Previously published data were used for this work (31, 37).

ACKNOWLEDGMENTS. This work was funded by NIH grant R35GM133613 and an Alfred P. Sloan Research Fellowship (awarded to D.M.M.), as well as funding from the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory. J.B.K. acknowledges support from NIH grant R35GM133777. M.S.W. and A.R.K. acknowledge support from NIH grant R37GM42699.

1. 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).
2. D. A. Kondrashov, F. A. Kondrashov, Topological features of rugged fitness landscapes in sequence space. *Trends Genet.* **31**, 24–33 (2015).
3. J. Domingo, P. Baeza-Centurion, B. Lehner, The causes and consequences of genetic interactions (epistasis). *Annu. Rev. Genomics Hum. Genet.* **20**, 433–460 (2019).
4. D. M. Fowler *et al.*, High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7**, 741–746 (2010).
5. L. M. Starita *et al.*, Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E1263–E1272 (2013).
6. D. Melamed, D. L. Young, C. E. Gamble, C. R. Miller, S. Fields, Deep mutational scanning of an RRM domain of the *Saccharomyces cerevisiae* poly(A)-binding protein. *RNA* **19**, 1537–1551 (2013).
7. 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).
8. M. B. Doud, O. Ashenberg, J. D. Bloom, Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. *Mol. Biol. Evol.* **32**, 2944–2960 (2015).
9. A. I. Podgornaia, M. T. Laub, Pervasive degeneracy and epistasis in a protein-protein interface. *Science* **347**, 673–677 (2015).
10. K. S. Sarkisyan *et al.*, Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).
11. B. Steinberg, M. Ostermeier, Shifting fitness and epistatic landscapes reflect trade-offs along an evolutionary pathway. *J. Mol. Biol.* **428**, 2730–2743 (2016).
12. C. Bank, S. Matuszewski, R. T. Hietpas, J. D. Jensen, On the (un)predictability of a large intragenic fitness landscape. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 14085–14090 (2016).
13. T. N. Starr, L. K. Picton, J. W. Thornton, Alternative evolutionary histories in the sequence space of an ancient protein. *Nature* **549**, 409–413 (2017).
14. 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).
15. C. Plesa, A. M. Sidore, N. B. Lubock, D. Zhang, S. Kosuri, Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. *Science* **359**, 343–347 (2018).
16. D. S. Tack *et al.*, The genotype-phenotype landscape of an allosteric protein. *Mol. Syst. Biol.* **17**, e10179 (2021).
17. T. N. Starr *et al.*, Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* **182**, 1295–1310.e20 (2020).
18. L. Gonzalez Somermeyer *et al.*, Heterogeneity of the GFP fitness landscape and data-driven protein design. *eLife* **11**, e75842 (2022).
19. J. N. Pitt, A. R. Ferré-D'Amaré, Rapid construction of empirical RNA fitness landscapes. *Science* **330**, 376–379 (2010).
20. J. I. Jiménez, R. Xulvi-Brunet, G. W. Campbell, R. Turk-MacLeod, I. A. Chen, Comprehensive experimental fitness landscape and evolutionary network for small RNA. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 14984–14989 (2013).
21. O. Puchta *et al.*, Network of epistatic interactions within a yeast snoRNA. *Science* **352**, 840–844 (2016).
22. C. Li, W. Qian, C. J. Maclean, J. Zhang, The fitness landscape of a tRNA gene. *Science* **352**, 837–840 (2016).
23. J. Domingo, G. Diss, B. Lehner, Pairwise and higher-order genetic interactions during the evolution of a tRNA. *Nature* **558**, 117–121 (2018).
24. V. W. C. Soo, J. B. Swadling, A. J. Faure, T. Warnecke, Fitness landscape of a dynamic RNA structure. *PLoS Genet.* **17**, e1009353 (2021).
25. R. P. Patwardhan *et al.*, High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).
26. 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).
27. S. Ke *et al.*, Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res.* **21**, 1360–1374 (2011).
28. A. B. Rosenberg, R. P. Patwardhan, J. Shendure, G. Seelig, Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* **163**, 698–711 (2015).
29. P. Julien, B. Miñana, P. Baeza-Centurion, J. Valcárcel, B. Lehner, The complete local genotype-phenotype landscape for the alternative splicing of a human exon. *Nat. Commun.* **7**, 11558 (2016).
30. S. Ke *et al.*, Saturation mutagenesis reveals manifold determinants of exon definition. *Genome Res.* **28**, 11–24 (2018).
31. M. S. Wong, J. B. Kinney, A. R. Krainer, Quantitative activity profile and context dependence of all human 5′ splice sites. *Mol. Cell* **71**, 1012–1026.e3 (2018).
32. D. M. Weinreich, Y. Lan, C. S. Wylie, R. B. Heckendorn, Should evolutionary geneticists worry about higher-order epistasis? *Curr. Opin. Genet. Dev.* **23**, 700–707 (2013).
33. J. Neidhart, I. G. Szendro, J. Krug, Exact results for amplitude spectra of fitness landscapes. *J. Theor. Biol.* **332**, 218–227 (2013).
34. T. N. Starr, J. W. Thornton, Epistasis in protein evolution. *Protein Sci.* **25**, 1204–1218 (2016).
35. Z. R. Sailer, M. J. Harms, Detecting high-order epistasis in nonlinear genotype-phenotype maps. *Genetics* **205**, 1079–1088 (2017).
36. Z. R. Sailer, M. J. Harms, High-order epistasis shapes evolutionary trajectories. *PLOS Comput. Biol.* **13**, e1005541 (2017).
37. 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).
38. J. Echave, C. O. Wilke, Biophysical models of protein evolution: Understanding the patterns of evolutionary sequence divergence. *Annu. Rev. Biophys.* **46**, 85–103 (2017).
39. F. J. Poelwijk, M. Socolich, R. Ranganathan, Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nat. Commun.* **10**, 4213 (2019).
40. A. S. Canale, P. A. Cote-Hammarlof, J. M. Flynn, D. N. A. Bolon, Evolutionary mechanisms studied through protein fitness landscapes. *Curr. Opin. Struct. Biol.* **48**, 141–148 (2018).
41. D. M. Weinreich, Y. Lan, J. Jaffe, R. B. Heckendorn, The influence of higher-order epistasis on biological fitness landscape topography. *J. Stat. Phys.* **172**, 208–225 (2018).
42. J. F. Storz, Compensatory mutations and epistasis for protein function. *Curr. Opin. Struct. Biol.* **50**, 18–25 (2018).
43. C. M. Miton, K. Buda, N. Tokuriki, Epistasis and intramolecular networks in protein evolution. *Curr. Opin. Struct. Biol.* **69**, 160–168 (2021).
44. G. Yang *et al.*, Higher-order epistasis shapes the fitness landscape of a xenobiotic-degrading enzyme. *Nat. Chem. Biol.* **15**, 1120–1128 (2019).
45. 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).
46. G. Reddy, M. M. Desai, Global epistasis emerges from a generic model of a complex trait. *eLife* **10**, e64740 (2021).
47. L. Ferretti *et al.*, Measuring epistasis in fitness landscapes: The correlation of fitness effects of mutations. *J. Theor. Biol.* **396**, 132–143 (2016).
48. P. F. Stadler, Landscapes and their correlation functions. *J. Math. Chem.* **20**, 1–45 (1996).
49. R. Happel, P. F. Stadler, Canonical approximation of fitness landscapes. *Complexity* **2**, 53–58 (1996).
50. C. E. Rasmussen, C. K. I. Williams, *Gaussian Processes for Machine Learning* (MIT Press, 2006).
51. R. M. Neal, "MCMC using Hamiltonian dynamics" in *Handbook of Markov Chain Monte Carlo*, S. Brooks, A. Gelman, G. Jones, X. L. Meng, Eds. (CRC Press, 2011), pp. 113–162.
52. B. P. Carlin, T. A. Louis, *Bayes and Empirical Bayes Methods for Data Analysis* (Chapman & Hall/CRC, Boca Raton, 2000), vol. 88.
53. R. A. Fisher, The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* **52**, 399–433 (1918).
54. T. Hinkley *et al.*, A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nat. Genet.* **43**, 487–489 (2011).
55. E. Weinberger, Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biol. Cybern.* **63**, 325–336 (1990).
56. P. F. Stadler, R. Happel, Random field models for fitness landscapes. *J. Math. Biol.* **38**, 435–478 (1999).
57. P. F. Stadler, *Fitness Landscapes in Biological Evolution and Statistical Physics* (Springer, 2002), pp. 183–204.
58. E. D. Weinberger, Fourier and Taylor series on fitness landscapes. *Biol. Cybern.* **65**, 321–330 (1991).
59. J. Zhou, D. M. McCandlish, Minimum epistasis interpolation for sequence-function relationships. *Nat. Commun.* **11**, 1782 (2020).
60. T. Wang, D. Zhao, S. Tian, An overview of kernel alignment and its applications. *Artif. Intell. Rev.* **43**, 179–192 (2015).
61. A. Aghazadeh *et al.*, Epistatic Net allows the sparse spectral regularization of deep neural networks for inferring fitness functions. *Nat. Commun.* **12**, 5225 (2021).
62. D. H. Brookes, A. Aghazadeh, J. Listgarten, On the sparsity of fitness functions and implications for learning. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2109649118 (2022).
63. H. Zou, T. Hastie, Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* **67**, 301–320 (2005).
64. Y. Kondo, C. Oubridge, A. M. M. van Roon, K. Nagai, Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5′ splice site recognition. *eLife* **4**, e04986 (2015).
65. D. M. McCandlish, Visualizing fitness landscapes. *Evolution* **65**, 1544–1558 (2011).
66. R. R. Coifman, S. Lafon, Diffusion maps. *Appl. Comput. Harmon. Anal.* **21**, 5–30 (2006).
67. D. M. McCandlish, Long-term evolution on complex fitness landscapes when mutation is weak. *Heredity* **121**, 449–465 (2018).
68. X. Roca *et al.*, Features of 5′-splice-site efficiency derived from disease-causing mutations and comparative genomics. *Genome Res.* **18**, 77–87 (2008).
69. I. Carmel, S. Tal, I. Vig, G. Ast, Comparative analysis detects dependencies among the 5′ splice-site positions. *RNA* **10**, 828–840 (2004).
70. C. Burge, S. Karlin, Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
71. G. Yeo, C. B. Burge, Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11**, 377–394 (2004).
72. J. Tan *et al.*, Noncanonical registers and base pairs in human 5′ splice-site selection. *Nucleic Acids Res.* **44**, 3908–3921 (2016).
73. P. A. Romero, A. Krause, F. H. Arnold, Navigating the protein fitness landscape with Gaussian processes. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E193–E201 (2013).
74. C. N. Bedbrook, K. K. Yang, A. J. Rice, V. Gradinaru, F. H. Arnold, Machine learning to design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization. *PLOS Comput. Biol.* **13**, e1005786 (2017).
75. J. C. Greenhalgh, S. A. Fahlberg, B. F. Pfleger, P. A. Romero, Machine learning-guided acyl-ACP reductase engineering for improved in vivo fatty alcohol production. *Nat. Commun.* **12**, 5825 (2021).
76. A. S. Kondrashov, Selection against harmful mutations in large sexual and asexual populations. *Genet. Res.* **40**, 325–332 (1982).
77. S. F. Elena, R. E. Lenski, Test of synergistic interactions among deleterious mutations in bacteria. *Nature* **390**, 395–398 (1997).
78. S. Bonhoeffer, C. Chappey, N. T. Parkin, J. M. Whitcomb, C. J. Petropoulos, Evidence for positive epistasis in HIV-1. *Science* **306**, 1547–1550 (2004).
79. S. Bershtein, M. Segal, R. Bekerman, N. Tokuriki, D. S. Tawfik, Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444**, 929–932 (2006).
80. B. L. Hie, K. K. Yang, Adaptive machine learning for protein engineering. *Curr. Opin. Struct. Biol.* **72**, 145–152 (2022).
81. A. J. Smola, R. Kondor, "Kernels and regularization on graphs" in *Learning Theory and Kernel Machines* (Springer, 2003), vol. 2777, pp. 144–158.
82. W. C. Chen, J. Zhou, J. M. Sheltzer, J. B. Kinney, D. M. McCandlish, Field-theoretic density estimation for biological sequence space with applications to 5′ splice site diversity and aneuploidy in cancer. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2025782118 (2021).
83. W. Hordijk, P. F. Stadler, Amplitude spectra of fitness landscapes. *Adv. Complex Syst.* **1**, 39–66 (1998).
84. A. Agarwala, D. S. Fisher, Adaptive walks on high-dimensional fitness landscapes and seascapes with distance-dependent statistics. *Theor. Popul. Biol.* **130**, 13–49 (2019).
85. J. B. Kinney, D. M. McCandlish, Massively parallel assays and quantitative sequence–function relationships. *Annu. Rev. Genomics Hum. Genet.* **20**, 99–127 (2019).
86. A. Tareen *et al.*, MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect. *Genome Biol.* **23**, 98 (2022).
87. D. M. McCandlish, vcregression : variance component regression for sequence-function relationships. GitHub. https://github.com/davidmccandlish/vcregression. Deposited 10 August 2020.