Empirical variance component regression for sequence-function relationships

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

[a]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724
[b]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724
[1]jzhou@cshl.edu
[2]mccandlish@cshl.edu

Abstract

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 genetic interactions not only between pairs of sites, but also higher-order interactions between 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.). Based on these estimated variance components, we then define an empirical Bayes prior that in expectation matches the observed pattern of epistasis, and reconstruct the sequence-function 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 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.

Introduction

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 [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 [4–15], RNAs [16–20], and regulatory or splicing elements [21–24]. Importantly, it has now become clear that the data from these experiments cannot be captured by considering simple pairwise interactions, but rather that higher-order genetic interactions between three, four, or even all sites within a functional element are empirically common [2, 12, 25–35] and indeed often expected based on first-principles biophysical considerations [12, 20, 25, 28, 34, 36]. 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 the effects of specific mutations on specific genetic backgrounds, epistatic coefficients between pairs of mutations on specific backgrounds, phenotypic values for individual genotypes, 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 seen the results of a specific mutation in several different genetic backgrounds already, how should we combine these observations to make a prediction for the effect of this mutation 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, epistatic coefficients of double mutants, and observed phenotypes, as one moves through sequence space. We show analytically that the shape of how precisely this predictability decays as a function of distance is completely determined by the fraction of phenotypic variance due to each order of genetic interaction (additive, pair-wise, three-way, etc.). Thus, rather than conceptualizing higher-order epistasis in terms of innumerable interaction terms between larger and larger number of sites, we suggest that: (1) we can understand a great deal about higher-order epistasis by considering simple diagrams showing how the correlations between mutational effects, epistatic coefficients, etc. decay as a function of genetic distance; and (2) these same diagrams suggest a method for making phenotypic predictions by weighting our observations in terms of the degree of information they provide for mutations on a genetic background of interest.

We implement these ideas in terms of a Gaussian process regression [37] framework with an empirical Bayes [38] prior. 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 interaction in our observed data. We then use these point estimates of the variance components to construct a prior distribution over all possible sequence-to-function mappings where the expected decay in the predictability of mutational effects matches that observed in the data. Finally, we conduct Bayesian inference under this prior, using Hamiltonian monte carlo [39] 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 overall form of higher-order epistasis present in the data, while simultaneously accounting for the effects of measurement noise and quantifying the uncertainty in our predictions.

To demonstrate the performance of this technique, we present an analysis of combinatorial mutagenesis data from protein G [30], a streptococcal antibody-binding protein that has served as a model system for studies of the genotype-phenotype map in proteins, as well as a high-throughput dataset measuring splicing efficiency of human $5'$ splice sites [40], which are RNA sequence elements crucial for the assembly of the spliceosome for pre-mRNA splicing. For this latter dataset, we also present low-throughput validation of our model predictions as well as a qualitative exploration of the complex patterns of epistasis in splicing efficiency observed in this system.

Results

The key question in phenotypic prediction is deciding how to combine the selective effects, local epistatic coefficients and individual phenotypic values observed in experiments, when assigning phenotypic values to unmeasured genotypes. For example, when we fit an additive or non-epistatic model [41], we are assuming that to the extent that the phenotypic effects of observed mutations generalize across genotypes, the effects of any specific mutation are the same no matter where it occurs. That is, in an additive model,

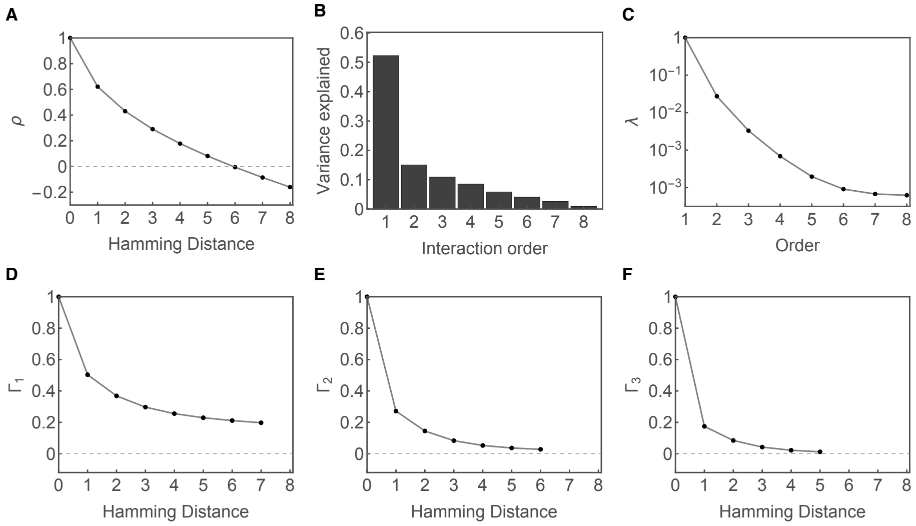

Figure 1: Summary statistics of a simulated sequence-to-function mapping on sequences of length 8 with 4 alleles per site. (A) Empirical distance correlation function (ρ), which is the correlation between the phenotypic values for all pairs of sequences separated by the specified number of mutations. (B) Empirical variance components, equal to the fraction of phenotypic variance due to each order of genetic interaction. (C) Mean square magnitude of individual genetic interaction terms (λ) as a function of interaction order. (D-F) Distance correlation of epistatic coefficients (Γ_k) of order k = 1-3. Note that Γ_1 measures the correlation of mutational effects. All panels represent the expected summary statistics of a random field model [43, 44] specified by the interaction term magnitudes shown in panel (C). Formulas for calculating these statistics can be found in *Materials and Methods*.

the effect of any given mutation is assumed to be constant across all genetic backgrounds, and 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 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 [42] allows the mutational effects of individual mutations to vary across genetic backgrounds, the epistatic interaction observed in double mutants for any specific pair of mutations is again constant across backgrounds (see *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 epistatic coefficients for this pair of mutations that are observed in the data and then assuming that this pair of mutations has the same interaction regardless of what genetic background these mutations occur on.

Putting the underlying strategies of additive and pairwise interaction models in these simple terms helps clarify the deficiencies of these models. Both models assume that only interactions between a certain number of mutations are relevant to prediction (i.e. additive effects of single mutations in non-epistatic models and interactions between two sites in pairwise interaction models). And both models make assumptions that these interactions or mutational effects are consistent over sequence space, first by pooling information across of all observed sequences to estimate these interactions or mutational effects and then making predictions that extrapolate these observations to all of sequence space —even to areas of sequence space where we have little or no data.

Here we would like to build 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.

Higher-order epistasis and phenotypic prediction

To implement a strategy of this type, it will be helpful to present some general results concerning higher-order epistasis. We first consider the case where all phenotypic values are known, before proceeding to our main problem of predicting unknown phenotypic values.

Our first task is to understand the relationship between the overall smoothness of the sequence-function relationship, the amount of higher order epistasis, and the typical magnitude of epistatic interactions of various orders. These features of the sequence-function relationship are illustrated for a simulated complete sequence-function mapping in Figure 1A-C. Figure 1A shows the distance correlation function (ref. [26, 44, 45] and *Materials and Methods*), which plots how correlations between phenotypic values drop off as one moves through sequence space. Figure 1B shows the decomposition of the sequence-function relationship into variance components (*Materials and Methods*), where the variance due to a particular interaction order is equal to the increase in the R^2 of a least squares fit when one e.g. adds pairwise terms to a model with only additive terms, three-way terms to a model with pair-wise and additive terms, etc., which in the literature is known as the (normalized) amplitude spectrum [26, 44]. Figure 1C shows how large the individual interaction terms of a given order (ref. [26, 44, 46] and *Materials and Methods*) tend to be, by plotting the mean square interaction size as a function of interaction order.

Because our goal is to understand how to combine the mutational effects, observed epistatic coefficients, etc., we can also plot how the predictability of these effects drops off as we move through sequence space [12, 47]. These are calculated for mutational effects, local pairwise epistatic coefficients, and local three-way interactions using Eq. 14 and shown in Figure 1D-F, respectively.

These pictures, particularly the plots of correlations as a function of distance in genotypic space, are quite informative for our intuitive goal of determining how to combine our observations of mutational effects, local epistatic coefficients, etc. when making predictions. We see for example from Figure 1D that, for this particular sequence-function relationship, mutational effects remain moderately correlated across all of sequence space, dropping from having a Pearson correlation coefficient of roughly 0.5 in adjacent genetic backgrounds to a correlation coefficient of roughly 0.2 in maximally distant backgrounds. However, from Figure 1E we see that the predictability of interactions in double-mutants decays much more rapidly, and so our observations are only really informative in genetic backgrounds up to two mutations away, and Figure 1F shows that three-way interactions are only substantially informative in immediately adjacent genetic backgrounds. These results suggest that when making predictions it might e.g. be sensible to extrapolate our observations of mutational effects throughout sequence space, but only allow our observations of interactions in local double mutant cycles to influence our predictions in relatively nearby genetic backgrounds.

How can we convert these intuitions based on examining the decay in the consistency of observed interactions into a rigorous method of phenotypic prediction? The key in answering this question lies in the fact that all 6 panels of Figure 1 are actually intimately related with each other and with previously proposed methods for phenotypic prediction.

In particular, it is classically known that the three pictures in Figure 1A-C in fact contain identical information, so that for any given sequence-function relationship, having any one of the panels in the top row of Figure 1 allows us to compute the other two (ref. [26, 43, 44], *Materials and Methods*). Here, we extend this result, showing that in fact having any of the pictures in Figure 1A-C allows us to draw all three panels in the bottom row of Figure 1 as well as their higher-order generalizations (i.e. how the predictability of local k-way interactions decays as we move through sequence space). Specifically, we show that the distance correlation function of k-th order epistatic coefficients depends only on the variance components of order k and higher (*Materials and Methods*).

Moreover, knowledge of any one panel in the first row of Figure 1 also defines a natural prior distribution for sequence-function relationships that can be used to derive specific predictions from partial data. Given e.g. the fraction of variance due to each order of interaction shown in Figure 1B we can draw epistatic interaction coefficients from a zero-mean normal distribution with variance given by the values in Figure 1C, which results in a sequence-function relationship that in expectation produces the patterns of correlation shown in Figure 1A and Figure 1C-D.

The above construction results in a natural family of priors for sequence-function relationships, where this prior distribution can be parameterized in terms of the fraction of variance due to each order of genetic interaction (i.e., the prior is a "random field model", [43, 44]). Importantly, various previously developed methods can be subsumed as particular (limiting) cases of inference under this class of priors. For example, the additive model and our recently proposed method of minimum epistasis interpolation [48] both arise as particular limiting cases where the fraction of variance due to additive effects goes to 1, and the pair-wise interaction model [42] arises as a limiting case where the total fraction of variance due to additive and pairwise effects goes to 1 (see Supplemental Figure 1). Thus, in a rigorous manner 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 Figure 1.

Finally, a key fact about this family of priors is that they are Gaussian, and so under the assumption that experimental errors are normally distributed, we can do inference under this prior using Gaussian process regression (see [37] for a review), which allows us to write down analytical expressions for the corresponding posterior distribution. 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 noise estimates 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

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

and covariance matrix

$$\mathbf{K} - \mathbf{K}_{\cdot B}(\mathbf{K}_{BB} + \mathbf{E})^{-1}\mathbf{K}_{B\cdot} \tag{2}$$

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.

Estimating variance components from partial data

To summarize the previous section, if we know the fraction of phenotypic variation due to each order of epistatic interaction, then we can derive a simple method of making phenotypic predictions that uses the corresponding covariance structure to appropriately generalize from observed phenotypic effects, double mutant epistatic interactions, phenotypic values, etc. While several existing methods of phenotypic prediction essentially come down to making specific assumptions about these variance components, our analysis suggests that a natural approach would be to make our predictions using variance components estimated from the data itself, i.e. an empirical Bayes approach in which we determine what prior to use by looking at the covariance structure of our observations. Conceptually, we want to make phenotypic predictions by assuming that the observed pattern of distance correlation of mutational effects, local epistatic interactions, etc. generalize to regions of the sequence space with no data. Practically, we can implement this idea by doing inference under a prior consisting of random sequence-function relationships where the effects of mutations and epistatic coefficients decay in the same way as in our data.

A naive 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 for 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. To see what the difficulty is, it is helpful to take another look at the relationship between higher-order epistasis and the distance correlation function.

A deep result from the literature on the mathematical theory of fitness landscapes states that the contribution of each particular order of interaction (e.g. additive, pairwise, three-way, etc.) to the distance correlation function takes a very specific shape. Technically, these shapes are given by a set of orthogonal polynomials known as the Krawtchouk polynomials [26, 43, 49, 50], but for our purposes it is suffices to look at the functions visually, as in Figure 2A and B. The orders of epistatic interactions split naturally into two groups with different qualitative interpretations, shown in panels Figure 2A and B, respectively, and which group an epistatic interaction falls into depends on whether the order of interaction is greater than or less than the expected distance between two random sequences (*Materials and Methods*).

Epistatic interactions of order less than the distance between two random sequences contribute positive local correlations, so that genotypes that are near to each other in sequence space tend to have similar phenotypes. These are shown in Figure 2A, and we can see that the main qualitative effect of increasing interaction order among this group is that these locally positive correlations decay increasingly rapidly.

Epistatic interactions of order greater than or equal to the distance between random sequences contribute negative local correlations, i.e. they make mutationally adjacent sequences tend to have anti-correlated values (if the order is equal to the expected distance between random sequences, then the correlation at distance 1 is zero, but it will be negative at distance 2). These orders of interaction are shown in Figure 2B, and qualitatively they oscillate increasingly rapidly as the order increases.

Now the distance correlation function itself is simply a weighted average of these curves, with the weights given by the mean square interaction terms of different orders (illustrated by Figure 2C). The fact that these weights need to be positive and sum to one puts strong constraints on the shape that the correlation function can take for a function defined over all of sequence space. For example, positive local correlations cannot decay any more slowly than they would for a purely additive model. However, for incompletely sampled sequence spaces, these constraints need not not hold (e.g. if the sampling consisted of several clusters of sequences with identical phenotypes separated from each other with missing sequences, one could have a perfect correlation within the smaller distance classes). 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, see *SI Appendix*). Thus, rather than using the observed covariance function to define our prior, we instead find the closest valid prior using weighted least squares, where the squared error for for the correlation at distance d is weighted by the number of pairs of sequences at distance d (*Materials and Methods*); this technique is formally equivalent to the idea of choosing a prior based on "kernel alignment" in the Gaussian processes literature, see ref. [51].

Practical implementation

One major challenge in solving Eq. 1 and 2 is that the computation involves inverting the $m \times m$ dense matrix $\mathbf{K}_{BB}$, a problem whose complexity scales cubically with m in time and quadratically with m in space. As a result, Gaussian process regression becomes computationally expensive when the training data size m is larger than several thousand [52].

To circumvent this difficulty, we provide an implementation that leverages the symmetries of sequence space to allow practical computations for sequence spaces containing up to low millions of sequences. The basic strategy is to rephrase our problem so that the solution can be found iteratively using only sparse matrix-vector multiplication.

In particular, notice that Eq. 1 can be solved by first finding a vector $\boldsymbol{\alpha}$ that satisfies $(\mathbf{K}_{BB}+\mathbf{E})\boldsymbol{\alpha} = \mathbf{y}$. Also, notice that matrix $\mathbf{K}_{BB}$ is a principle submatrix of $\mathbf{K}$, so that we can write $\mathbf{K}_{BB} = \mathbf{I}_{\cdot B}^T \mathbf{K} \mathbf{I}_{\cdot B}$ where $\mathbf{I}_{\cdot B}$ consists of the columns of the identity matrix $\mathbf{I}$ that correspond to our set of observed sequences B. Since the entries of $\mathbf{K}$ depend only on the Hamming distance between the corresponding sequences, $\mathbf{K}$ can expressed as a polynomial in the graph Laplacian (*i.e.* the matrix $\mathbf{L}$ whose i, j-th entry is -1 if i is adjacent to j, $\ell(\alpha - 1)$ if $i = j$, and 0 otherwise) that is, $\mathbf{K} = \sum_{k=0}^{\ell} b_k \mathbf{L}^k$ [53, 54] for some $b_0, \ldots, b_\ell$ that we can find analytically and thus that $\mathbf{K}\mathbf{v}$ can be found by iteratively applying the sparse matrix $\mathbf{L}$ to $\mathbf{v}$ at most ℓ times. Using these results, we can rewrite our original equation $(\mathbf{K}_{BB} + \mathbf{E})\boldsymbol{\alpha} = \mathbf{y}$ using only sparse matrices as $(\mathbf{I}_{\cdot B}^T (\sum_{k=0}^{\ell} b_k \mathbf{L}^k) \mathbf{I}_{\cdot B} + \mathbf{E})\boldsymbol{\alpha} = \mathbf{y}$, which we solve using the conjugate gradient

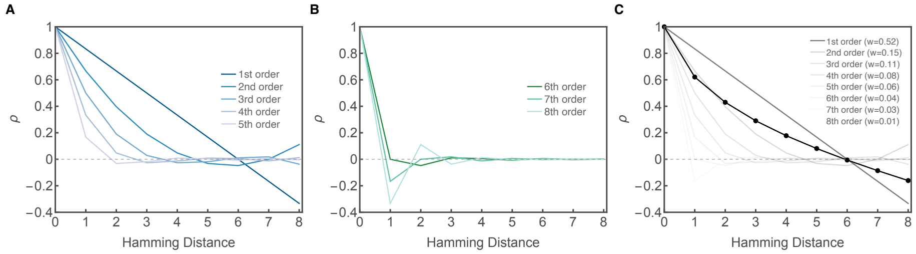

Figure 2: Superimposition of distance correlation functions for pure k-th order interactions for sequences of length 8 with 4 alleles per site. (A) Distance correlation function for locally correlated orders of genetic interaction (in this case, interaction orders k=1–5). (B) Distance correlation function for locally anticorrelated components (in this case, k=6–8). (C) The distance correlation function (solid line) is a weighted sum of elementary autocorrelation functions (gray lines) with the weights (denoted as w in the figure legend) given by the variance components. Distance correlation function and variance components are identical to those shown in Figure 1.

algorithm.

Application to protein G

We first apply our method to a dataset derived from a deep mutational scanning study of the IgG-binding domain of streptococcal protein G (GB1) [30]. This experiment attempted to assay all possible combination of mutations at four sites (V39, D40, G41, and V54; 20^4 = 160000 protein variants) that had been previously shown to exhibit high levels of pairwise epistasis [7]. The library of protein variants were 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 low coverage of the input library, the original data do not provide the binding score for 6.6% of the variants.

We began by inferring the variance components of the GB1 landscape from the empirical autocorrelation function using our least squares procedure applied to all available data (93.6% of all possible sequences), Figure 3A (see *Materials and Methods* for details). In Figure 3B, we note that the majority of the variance in the data is estimated to be explained by the additive and pairwise components (56% and 36% of total variance, respectively). The third-order component is estimated to have a small but non-negligible contribution (8% of total variance), and the estimated contribution of the 4th order component is negligible.

We can use the results from the previous section to understand the practical meaning of these estimates for our task of phenotypic prediction. For example, in Figure 3C, we plot the correlation of mutational effects as a function of Hamming distance [47] (*Materials and Methods*). 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 is characteristic of non-additivity and 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 sequence, and that therefore when making predictions we should be giving local observations of mutational effects approximately twice as strong a weight as distant observation of mutational effects.

At a broader scale, our analysis above also provides qualitative insights into the overall structure of the sequence-function relationship. For example, we stated above that the orders of epistatic interaction can be divided into the locally correlated and the locally anti-correlated groups, depending on whether the order of the interaction is greater than or less than the expected distance between two random sequences. Random protein sequences of length 4 differ at $(1 - \frac{1}{20})4 = 3.8$ sites on average, so interaction orders 1 through 3 correspond to the sequence-function relationship being locally correlated, whereas order 4 controls the strength of local anti-correlation. Thus, our estimated variance components suggest that the GB1 sequence-function relationship is dominated by locally positive correlations, with essentially no anti-correlated component.

Within our overall inference procedure, the estimated variance components discussed above are used to construct a prior probability distribution over all sequence-function relationships where in expectation mutational effects, epistatic interactions and observed phenotypes generalize across sequence space in the same manner as observed in the data. The next step is to use the observed data to update this prior distribution, which was based solely on the coarse summary statistics encapsulated in the distance correlation function, using the fine-scale information from the individual observations. An immediate question is the extent to which the statistical features of the resulting posterior distribution are similar or different from that of the prior. We drew 2000 samples from the resulting posterior distribution and plotted the results in Figure 3A-C using error bars to indicated 95% credible intervals. We find that the posterior gives very tight estimates of the variance components and correlation structure of the true sequence-function relationship, but that these estimates differ somewhat from the prior, with the 3rd order interactions being roughly 1.6 times as strong in the posterior (Figure 3B), which results in a slightly faster decay in the predictability of mutational effects as we move through sequence space (Figure 3C). Thus, we conclude that our prior distribution provided a qualitatively reasonable estimate of the overall statistical features of the data.

Obviously, another important question is the performance of the predictions made by our method. Since the GB1 landscape is relatively well sampled, we were able to assess this performance for a large range of sampling regimes, from quite sparse to extremely dense, by using our method to make predictions for randomly sampled held-out data with increasing amounts of training data (critically, the variance component estimates were re-computed for each of these random samples in order to provide a realistic test of the entire inference pipeline in the low-data regime). For comparison we also fit an additive model using ordinary least squares, regularized pairwise and 3-way regression models. Since both L_1 and L_2 regularized regression have been used to model data of sequence-function relationships [32, 42, 55], here we fit the pairwise and three-way models using elastic net regression (*Materials and Methods*) where the penalty term for model complexity is a mixture of L_1 and L_2 norms [56] with the relative weight of the two penalties chosen through crossvalidation. This allows us to compare our method against the regression models fitted using regularization most appropriate for the a particular training dataset. In addition to the linear regression models, we also fit a global epistasis model [36] where the binding score is modeled as a nonlinear transformation of a latent additive phenotype on which each possible mutation has a background-independent effect (*Materials and Methods*).

We compared the predictive accuracy of these five models by plotting out-of-sample R^2 against a wide range of training sample size, Figure 3D. We first note that the out-of-sample R^2 of the additive model and the global epistasis model stay constant regardless of training sample size, consistent with their low number of model coefficients and flexibility. The low R^2 of the global epistasis model also indicates a substantial degree of specific epistasis (i.e. interactions between specific subsets of sites, [27]). 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 3-way model performs strongly with a large amount of data, but under-performs when data are sparse. We see that our empirical variance component regression (VC regression) method performs equivalently to the pairwise model at low data density and similar to the three-way model at high data density (remaining marginally superior at very high sampling), and thus provides the strongest overall performance.

Application to human $5'$ splice site data

To provide an application of our method to a nucleic acid sequence-function relationship, we turn to an analysis of a high-throughput splicing assay that attempted to measure the activity of all possible $5'$ splice sites [40]. The $5'$ splice site ($5'$ss) is a 9-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

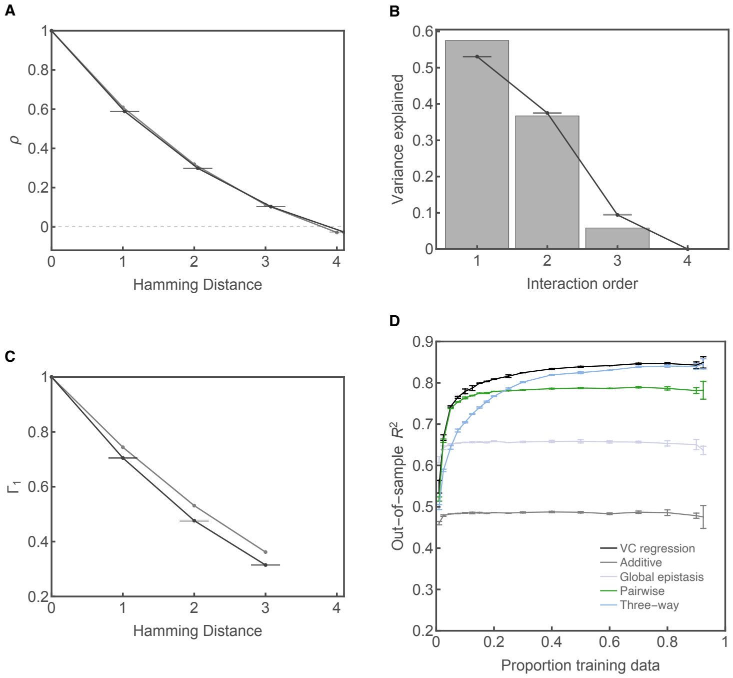

Figure 3: Analyses of the GB1 combinatorial mutagenesis dataset. (A) Distance correlation of phenotypic values. (B) Variance components. (C) Distance correlation of mutational effects. In A-C, gray represents statistics of the prior distribution inferred from the full dataset consisting of 149361 genotypes (93.6% of all possible sequences), black represents the posterior statistics estimated based on 2000 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 5 replicates. Additive models were fit using ordinary least squares. Pairwise and 3-way regression models were fit using elastic net regularization with regularization parameters chosen by 10-fold cross-validation (*Materials and Methods*). Global epistasis model assumes the binding score is a nonlinear transformation of an unobserved additive phenotype and was fitted following ref. [36]. Error bars represent one standard deviation.

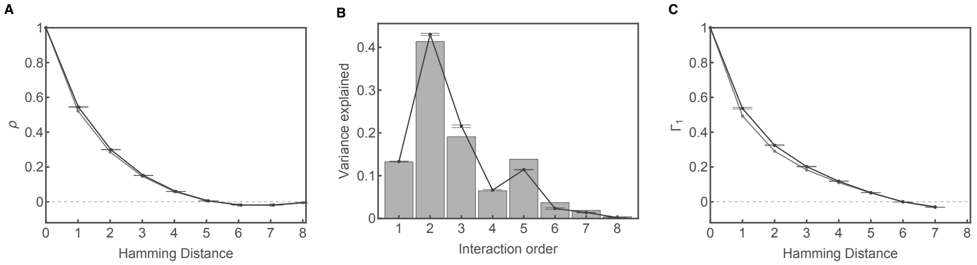

Figure 4: Analyses of the *SMN1* $5'$ss combinatorial mutagenesis dataset. (A) Distance correlation function of the splicing phenotype (PSI). (B) Variance components. (C) Distance correlation of mutational effects. Gray represents statistics of the prior distribution inferred from the full dataset consisting of 30732 genotypes (93.8% of all possible splice sites), black represents the posterior statistics estimated using 2000 Hamiltonian Monte Carlo samples. Error bars indicate 95% credible intervals.

slash denoting the exon-intron junction. At the beginning of the splicing reaction, the $5'$ss is recognized by the U1 snRNP of the spliceosome through direct base pairing between $5'$ss and the U1 snRNA [57], whose $5'$ sequence is complementary to the consensus $5'$ss sequence. In ref. [40], the authors used a massively parallel splicing assay to estimate the splicing efficiency of 94.8% of the 32768 possible $5'$ss sequences of the form NNN/GYNNNN for intron 7 of the gene *SMN1* using a minigene library in human cells. Splicing efficiency was measured in units of relative percent spliced in (PSI), defined as the ratio of read counts corresponding to exon inclusion to total read counts (including both exon inclusion and exon skipping) divided by the ratio for the consensus sequence and then expressed as a percentage.

In Figure 4A, we first show the distance correlation function of PSI for the observed sequences. These correlations appear to drop off quite rapidly, with sequences differing at 5 or more positions having PSIs that are essentially uncorrelated. The associated estimated variance components are shown in Figure 4B. These indicate that pairwise interaction accounts for the largest proportion of the sample variance (42.2%), but there are also substantial higher-order interactions with the variance due to 5-way interactions (13.7%) being comparable to those of the additive and three-way component. The orders of genetic interactions 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 non-neglible role, accounting for 2.2% of the total variance. In Figure 4C, 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 splicing landscapes again have very small credible intervals and remain similar to those estimated using our least squares procedure, with a slightly increased contribution of pair-wise and third order interactions and a decreased contribution of the five-way interactions. 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, consistent with the mechanistic intuition that mutations that e.g. substantially decrease U1 snRNA binding in the context of a functional splice site are likely to have no impact in an already non-functional sequence context.

We next compare the predictive power of our method against the four models used earlier on the GB1 dataset, namely the additive model, the global epistasis model, and the pairwise and three-way interaction models fit using elastic net regularization. We first compare the predictive power of the five models by randomly assigning a subset of our data as training examples corresponding to 80% of all possible sequences (i.e. we assigned 26,214 of the observed sequences as training data). Figure 5A-E shows the scatter plots of the true PSI vs. out-of-sample predictions for the five models in the order of increasing R^2. First, we see that the additive model performs poorly with an out-of-sample $R^2 = 0.15$. The inclusion of pairwise interaction terms substantially improves the performance with an out-of-sample $R^2 = 0.48$. Unlike the GB1 dataset, the global epistasis model exceeds the pairwise model in performance by a large margin with $R^2 = 0.60$. This is followed by the three-way interactions model ($R^2 = 0.67$).

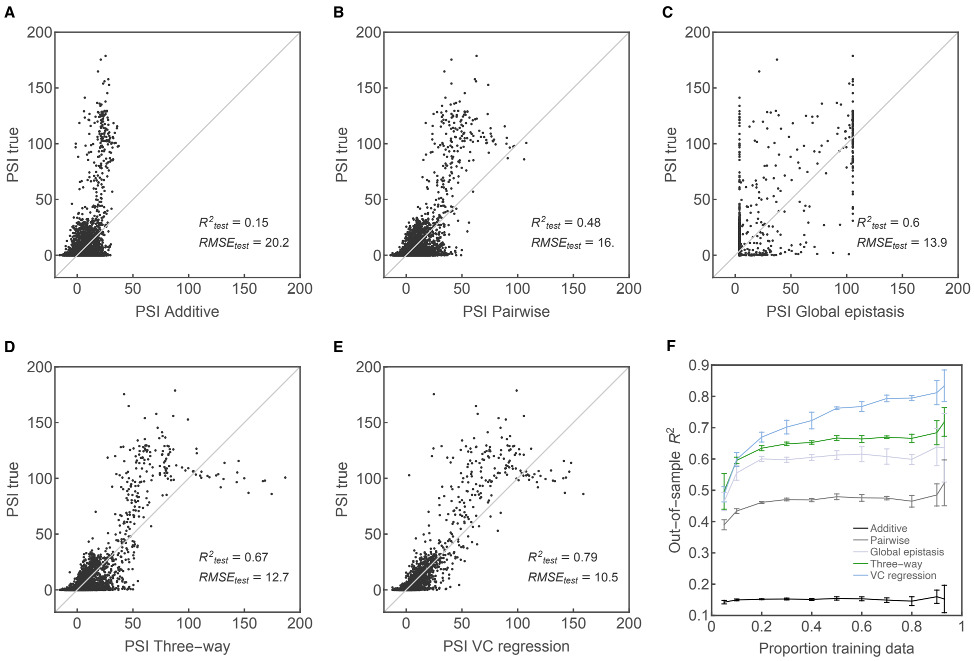

Figure 5: Model performance for the *SMN1* 5′ss combinatorial mutagenesis dataset. Additive models were fit using ordinary least squares. Pairwise and 3-way regression models were fit using elastic net regularization with regularization parameters chosen by 10-fold cross-validation (*Materials and Methods*). Global epistasis model models PSI as a nonlinear function of an unobserved additive phenotype and was fitted following ref. [36]. was fit following ref. [36]. (A-E) Scatter plots of out-of-sample predictions for the additive model, pairwise regression, global epistasis model, three-way regression, and variance component regression using one training dataset consisted of 80% of all 5′ss ($n = 26215$) assigned as training data. (F) Out-of-sample R^2 of the five models plotted against a range of training sample sizes. Error bars represent one standard deviation calculated for 5 replicates for each sample size.

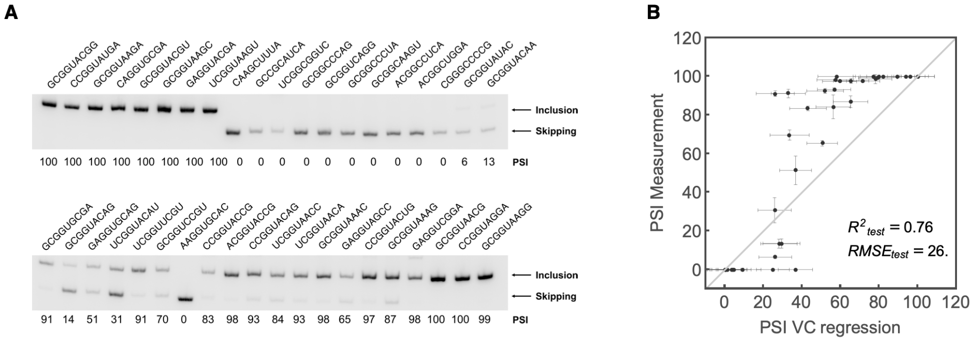

Figure 6: Manual validation of predicted PSI for 40 unmeasured *SMN1* 5′ss. (A) Gel images of manually validated sequences. For each lane, the top band corresponds to mRNA product containing exon 7 (exon inclusion), while the bottom band correspond to mRNA product without exon 7 (exon skipping). Percent spliced in (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 one standard deviation of the posterior distribution. Vertical error bars correspond to one standard deviation around the mean PSI estimated using three replicates in the manual validation. Since the low-throughput PSIs are inherently restricted to the range 0–100, in this analysis we capped the raw predicted PSIs to lie in this same range.

Finally, the variance component regression substantially outperforms the other models with $R^2 = 0.79$.

To see how these various models perform when greater or lesser amounts of data are available, we compared the predictive power of the five models by plotting their out-of-sample R^2 against a wide range of training sample sizes, Figure 5F. 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, at the sampling density (training sample size $< 20\%$ of all possible sequences), the three-way model has similar performance as our model. However, the performance gap between the two models quickly widens as the training data become dense. The variance component model is able to achieve a final $R^2 = 0.83$ with 93% of the sequence space assigned as training data ($n = 30474$), compared with the three-way model $R^2 = 0.72$. This difference in model performance is consistent with the observation of 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 Supplemental Figure 2, 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 with the 5′ss, and then this binding energy is mapped via a nonlinear function to PSI [3]. However, we also note that this simple model fails to capture some important features of the data, most notably a group of false-negative sequences that are predicted to be non-functional by the global epistasis model but experimentally show moderate to high measured PSI (Supplemental Figure 3A). Using the variance component regression, we were able to accurately predict these outlier sequences (Supplemental Figure 3B). We thus conclude that while the global epistasis model provides a good intuitive first-pass understanding of the splicing landscape, our empirical variance component regression is able to capture more of the fine-scale features of the sequence-function relationship measured here.

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 2036 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 measured using a low-throughput experiment [40], see *Materials and Methods*, Figure 6A. Overall, our method achieves a reasonable qualitative agreement with the low throughput measurements PSI (Figure 6B), 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 [40] and a sharp transition between these two sets of sequences in sequence space (Supplemental Figure 2). 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, resulting 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 sequence-function relationship, both with regard to how the underlying mechanisms result in the 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 have previously conducted a detailed analysis of this type for the GB1 dataset [see 48] we will focus on the inferred activity landscape for $5'$ss.

In particular, our visualization method [58] is based on constructing a model of molecular evolution under the assumption that natural selection is acting to preserve the molecular functionality measured in the assay. The resulting visualization optimally represents the expected time it takes to evolve from one sequence to another (*Materials and Methods*), 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 first inferred the full *SMN1* splicing landscape using Empirical Variance Component Regression and built a model of molecular evolution based on the MAP estimate (*Materials and Methods*). Then we 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 [59] 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 [58, 60].

The resulting visualization is shown in Figure 7A and Supplemental Figure 4, where 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 connected by single point mutations. It turns out that each of the first three diffusion axes has a simply interpretable meaning. Figure 7B plots the estimated PSI against Diffusion Axis 1, showing that Diffusion Axis 1 separates functional splice sites (large positive values) from non-functional splice sites (negative values). Diffusion Axes 2 (Supplemental Figure 4) and 3 then separate different groups of functional splice sites from each other. Figure 7A shows two major branches of functional splice sites that are separated along Diffusion Axis 3. Examination of sequence composition within each branch reveals that the major distinction between the two clusters lies at position $+3$, where sequences in the bottom cluster retain the consensus base A, while the top sequences possess $+3$ mutations that are predominantly G. To see the meaning of Diffusion Axis 3, we cut away to show only the most highly functional sequences (818 sequences with PSI $> 80\%$) and plot these sequences using diffusion Axes 2 and 3, Figure 7C. This figure shows a hierarchy of clusters of functional sequences.

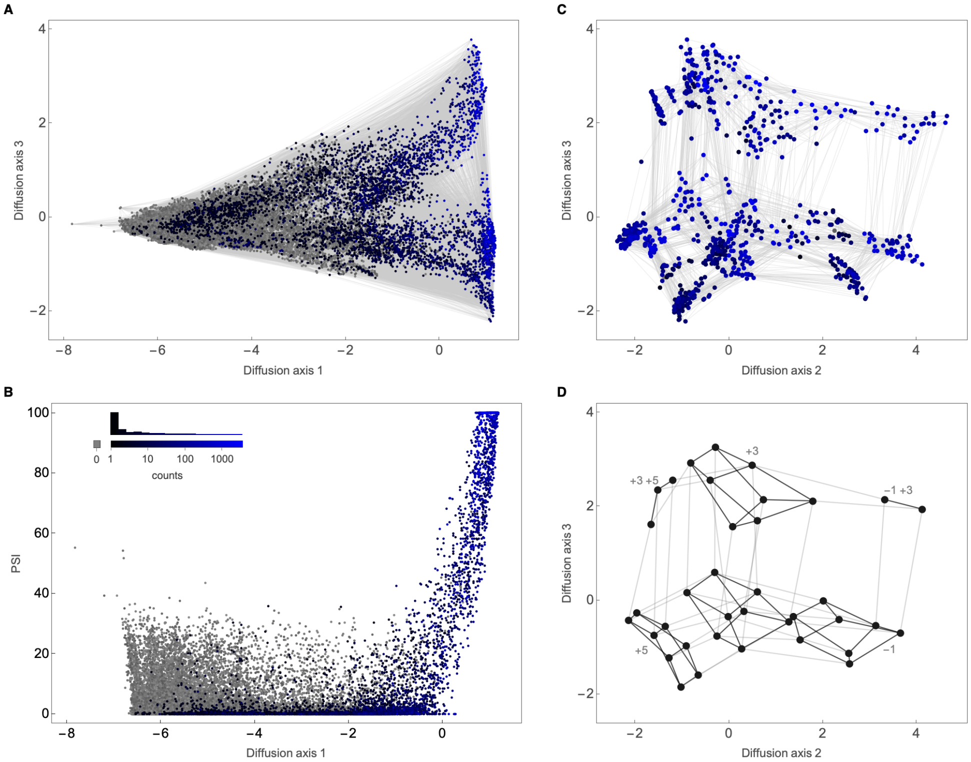

Figure 7: Visualization of the *SMN1* splicing landscape reconstructed using Empirical Variance Component regression. Genotypes are plotted using the dimensionality reduction technique from [58] (see "Methods"). (A) Visualization of all 32768 splice sites using Diffusion Axes 1 and 3. Two splice sites are connected by an edge if they differ by a point mutation. (B) Diffusion Axis 1 largely corresponds to the separation of low and high PSI splice sites. (C) Visualization of all 818 splice sites with predicted PSI $>$ 80% using Diffusion Axes 2 and 3. In A-C, splice sites are colored according to the number of times that sequence is used as a splice site in the hg38 reference genome. Gray dots represent splice sites not present as functional splice sites (65.9% of all possible splice sites). (D) Abstracted version of panel C. 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 position -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 position -2, +4, and +6. Note that no cubes contain both the -1 and +5 mutation, indicating a major incompatibility between mutations at these two sites.

Examining the sequences within each of these small clusters revealed that the small clusters correspond largely to whether a consensus or mutant nucleotide was present at each of positions -2, -1, +3, +4, +5, and +6, (Figure 7D). We see then that Diffusion Axis 2 encodes whether or not mutations are present at the -1 and +5 positions, where functional genotypes with mutant nucleotides at the +5 position are plotted at negative values on Diffusion Axis 2, functional genotypes with mutant nucleotides at the -1 position are plotted at positive values on Diffusion Axis 2, and functional genotypes with mutations at neither position are plotted in between.

The above analysis reveals a complex pattern of genetic incompatibilities between mutations at positions -2, -1, +3, +4, +5, and +6, as some but not all combinations of mutations at these positions are compatible with high splicing activity (see also Supplemental Figure 5). The overall structure is dominated by a major incompatibility between mutations at the -1 and +5 positions, since no sequences with mutations on both -1 and +5 have strong splicing activity ($>$ 80% PSI). This is consistent with previous findings of a negative interaction between -1 and +5 based on genomic comparisons [61–63], maximum entropy model fitting [64], and high-throughput splicing assays [40]. As a result of this interaction, a population constrained by natural selection to maintaining splicing function with a non-consensus nucleotide at the -1 position must typically evolve a consensus nucleotide at -1 before it can evolve to a sequence with a mutation at +5, resulting in long waiting times to evolve from a sequence with a -1 mutation to sequences with a +5 mutation.

The next most prominent structure revealed in Figure 7D corresponds to having a G mutation at the +3 position (upper portion of y-axis in Figure 7D), so together we consider positions -1, +3 and +5 as being the major mutations. Whereas having either a single -1 or +5 mutation is compatible with having many different combination of minor mutations at positions -2, +4, and +6 (complete cubes on the bottom half of Figure 7D), in a +3 mutant background combined with either a -1 or +5 mutation, we observe complex interactions between these minor mutations. In particular, in the -1+3 mutant background, the only additional minor mutation compatible with maintaining functionality is -2. However, in the presence of +3+5 mutations we see a different pattern where in this background only a +6 mutation can be tolerated, but in the presence of this additional +6 mutation, a mutation at the +4 position also changes from being intolerable to sometimes being tolerable.

How can we explain this complex pattern of genetic interactions? The overall structure of the splicing landscape with a flat, nonfunctional region where mutations have little effect and then a functional region where they have greater effects is typical of non-specific epistasis and especially compatible with a global epistasis model. Moreover, the pattern of interactions between the major -1, +3 and +5 mutations can also be accounted for by a global epistasis model with a sharp threshold-like nonlinearity where -1 and +5 mutations have large effects and +3 mutations have moderate effects such that the combination of a +3 mutation with a -1 mutation brings the sequence near the threshold, but a -1 together with a +5 mutation brings the sequence over the threshold, resulting in a loss of functionality (consistent with our inferred global epistasis fit, Supplemental Figure 2).

However, the more complex interactions involving the minor mutations are qualitatively incompatible with the global epistasis model. This is because under global epistasis any mutation that is tolerated in a weaker background must also be tolerated in a stronger background. So if +3+5 is a stronger background than -1+3, and the -2 mutant is tolerated in the -1+3 background, then it should also be tolerated in the +3+5 background. However, we observe that the effect of a mutation at position -2 when -1 and +3 are mutated is often tolerated (median effect of -18.6 PSI, calculated for sequences with consensus bases on all other positions), but when +3 +5 are mutated the -2 mutation typically has a much larger effect (median effect is -93.8 PSI for sequences with consensus bases on all other positions), which always results in a PSI $<$ 80. However, if -1+3 is the stronger background, then it should also tolerate a mutation at the +6 position, which it does not. Rather, the tolerability of the +6 mutation in the +3 +5 background appears to be due to a specific interaction between the +5 and +6 mutations, where +6 mutations have little or no effect in any background where a +5 mutation is present (Supplemental Figure 6). In particular, we find that the deleterious effect of mutations at +6 over all functional backgrounds with the consensus +5G (median $= -43.0$, calculated in backgrounds with PSI > 80) is almost completely abrogated in functional backgrounds where +5 is mutated (median $= -2.0$, calculated in backgrounds with PSI > 80). This observation would be consistent with the biophysical hypothesis that the major mutation at +5 results in the dissociation of all distal nucleotides from the 3' end of the RNA-RNA duplex, and hence any further mutation at +6 has little deleterious effect, since +6 is no longer involved in direct base-paring with the U1 snRNA [c.f., 65].

Finally, the functionality of +4 mutations in the +3+5+6 mutant background but not in the apparently stronger +3+5 background is also highly incompatible with the global epistasis hypothesis. We found two specific highly functional 5′ss sequences with this combination of mutations, CAG/GUUGUA and AAG/GUGGAC. The first sequence has been found to bind to U1 snRNA through a noncanonical binding geometry known as an asymmetric loop [66] where an uneven number of unpaired nucleotides are found in an internal loop, allowing the 3′ GUA of the splice site to form 3 additional basepairs with the U1 snRNA. The second sequence (AAG/GUGGAC) does not seem to correspond to any known alternative binding geometry. However, it does naturally occur 14 times in the human genome as a putatively functional splice site (*Materials and Methods*). Furthermore, we have verified its functionality via a low-throughput method (mean PSI $\pm$ 1SD $= 96.9 \pm 5.33$, $n = 3$, Supplemental Figure 7), suggesting that it operates via some unknown mechanism.

In summary, we conclude that the 5′ss activity landscape contains many qualitatively different types of genetic interactions. At a coarse level, the splicing landscape can be understood in light of the global epistasis model, where interactions between major mutations arise due to a threshold effect. At a finer level, however, 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, where PSI is modeled as a nonlinear function of an underlying additive phenotype, both in the form of specific pairwise interactions such as the interaction between the +5 and +6 positions, but also highly complex interactions associated with substantial changes in the physical geometry of U1 snRNA binding [66].

Discussion

In this paper, we address the problem of how to model the complex genetic interactions observed in high-throughput mutagenesis experiments in order to predict phenotypic values for unmeasured genotypes. 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. We show that 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 that: (1) this same distance correlation function also determines the degree to which our observations of mutational effects, double mutant epistatic coefficients, and observed interactions between three or more mutants generalize across increasingly distant genetic backgrounds; and (2) the distance correlation function can be parameterized in terms of the fraction of phenotypic variance due to each order of genetic interaction (i.e. the 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 sequence-function relationships that is concentrated on the subset of sequence-function relationships where the effects of mutations generalize in the same manner as occurs in our observed data. Conducting Bayesian inference under this prior then produces phenotypic estimates that reflect the belief that the extent and types of epistasis in unobserved regions of sequence space are similar to the extent and type of epistasis in regions of sequence space that we have already observed.

One way to understand our contribution here is to see it as an integration between practical Gaussian process-based methods for analyzing sequence-function relationships [67] and the classical spectral theory of fitness landscapes [43, 44, 46], which provides the most sophisticated mathematical theory of genetic interactions currently available. Within this theoretical literature, the so-called "random field models" identical to the family of priors we propose have been extensively studied [26, 43, 44], and we have leveraged this existing knowledge to craft priors that encode comprehensible beliefs about the structure of sequence-function relationships that overcome the inherent difficulty of understanding these high-dimensional objects.

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 anti-correlated component to a sequence-function relationship has been discussed previously in the literature in terms of the "eggbox" component [12, 47] which is perfectly anti-correlated 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 Γ_k which, beyond simple phenotypic correlations, measures how mutational effects ($k = 1$) or epistatic coefficients ($k > 1$) decay as one moves through sequence space. The correlation of mutational effects as a function of distance between genetic backgrounds has been previously termed γ, 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 k-th order epistatic interactions in fact make the very strong assumption that any observed k-th 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.

The method we propose here also has some commonalities with minimum epistasis interpolation [48], another method we recently proposed for phenotypic prediction that includes genetic interactions of all orders. The most important difference is based on the criterion for parsimony being employed in each instance. Minimum epistasis interpolation attempts to find a reconstruction that minimizes the expected squared epistatic coefficient between a random pair of mutants introduced on a random genetic background. Thus, minimum epistasis interpolation is based on imposing an a priori assumption that the sequence-function relationship should be simple in the sense of being locally smooth (i.e. locally non-epistatic). In contrast, empirical variance component regression takes the view that a reconstruction is parsimonious if the extent and type of epistasis present in the reconstruction are similar to the extent and type of epistasis present in the data itself. Depending on the needs of the user, both minimum epistasis interpolation and empirical variance component regression can be conducted either in a Bayesian manner or as a form of L_2-regularized regression [68] (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 the different orders of genetic interaction differently, either with a quadratically increasing penalty in the case of minimum epistasis interpolation, or a penalty determined by the empirically estimated variance components in the case of minimum epistasis interpolation

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 [51]. It may well be possible to produce more accurate predictions by choosing hyperparameters by maximizing the evidence [37] or via a hierarchical Bayesian model where we integrate over our uncertainty in the values of these hyperparameters. However, here we prefer a simpler empirical Bayes procedure, because it corresponds better to the underlying philosophy of the method, in that we estimate the extent and type of epistasis present in our observations and then directly incorporate these estimates into our prior.

Another limitation concerning variance component regression is that it is unable to explicitly model any overall nonlinearity of the measurement scale, i.e. it does not explicitly model nonspecific or global epistasis [3, 27, 28, 36, 69–71]. 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 consistent minimal effects of mutations in a particular region 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, particularly when the underlying latent trait being modeled is the true object of scientific interest, rather than the observed phenotype (e.g. in the case of nonlinearity due to the measurement process, or where the latent trait has a specific biophysical meaning such as a binding energy). However, empirical variance component regression as presented here may still be preferred if the primary interest is in the specific phenotype being measured, since the effects of a physiological biophysical nonlinearity on the generalizability of mutational effects and epistatic interactions is itself an issue of considerable scientific interest.

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 the sequence space, which 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, this still limits us to nucleic acid sequences of length 11 or less, and amino acid sequences of length 5 or less. Using the kernel trick [72], it is possible to work with much longer sequences, but at the cost of only being able to accommodate up to low tens of thousands of observed sequences, due to the resulting dense kernel matrix. Although we provide analyses of datasets in the current manuscript that contain tens to hundreds of thousands of sequences, more work is needed to scale the methods proposed here to even larger datasets and sequence spaces.

Methods

Summary statistics

This section defines various summary statistics used in the paper, including the quantities plotted in Figure 1. We also show how different quantities can be transformed from one to another. Here we simply list the main results without proof. Detailed derivations can be found in *SI Appendix*.

Given an alphabet A of size α, we use A^ℓ to denote the sequence space which is the set of all tuples of A, equipped with a metric that is the Hamming distance D, such that $D(x, x')$ measures the number of mutations that separate the two sequences $x, x' \in A^\ell$. Given a sequence space of size $G = \alpha^\ell$, a landscape f is a function that maps every sequence $x \in A^\ell$ to its phenotypic value. Throughout this paper, we use the boldface $\mathbf{f} \in \mathbb{R}^G$ to denote the G-dimensional column vector indexed by sequences in A^ℓ and f to denote the function which allows us to evaluate the phenotype of a sequence $f(x)$ such that $f(x) = \mathbf{f}_x$. We define the autocovariance function of f as [43, 44, 49]:

$$C(d) = \frac{1}{N_d} \sum_{x,x':D(x,x')=d} (f(x) - \overline{f})(f(x') - \overline{f}), \tag{3}$$

where $N_d = \alpha^\ell \binom{\ell}{d}(\alpha - 1)^d$ is the number of ordered pairs of sequences at Hamming distance d and $\overline{f}$ is the mean phenotypic value. We can also define the autocorrelation function by normalizing $C(d)$ with the empirical variance:

$$\rho(d) = \frac{C(d)}{C(0)}. \tag{4}$$

Now suppose we only have noisy observations $\mathbf{y} = \mathbf{f}_B + \mathbf{e}$ on a subset of sequences $B \subset A^\ell$. Here $\mathbf{e}$ is the noise vector which we assume is drawn from a normal distribution: $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{E})$, with $\mathbf{E}$ being a diagonal matrix. We can still extract the empirical covariance function by averaging over pairs of sequences in B for different distance classes. Specifically, let $\mathcal{D}(B) = (\mathcal{D}_0, \mathcal{D}_1, \ldots, \mathcal{D}_\ell)$ be the distance distribution of the set B, where $\mathcal{D}_i$ is the number of pairs of ordered sequences that are at Hamming distance i. Define the empirical autocovariance function

$$c(d) = \begin{cases} \frac{1}{|B|} \sum_{x \in B} (y(x) - \overline{y})^2 - \overline{\sigma^2} & d = 0 \\ \frac{1}{\mathcal{D}_d} \sum_{\{x,x' \in B: D(x,x')=d\}} (y(x) - \overline{y})(y(x') - \overline{y}) & d = 1, \cdots, \ell. \end{cases} \tag{5}$$

For $d = 0$, we substract the mean variance of the noise components $\overline{\sigma^2}$ from the raw empirical variance $\frac{1}{|B|} \sum_{x \in B} (y(x) - \overline{y})^2$ so that $c(0)$ is not inflated by the observation noise. The noise component is not accounted for when $d > 0$ since we assume the noise distribution is independent with mean zero, making the contribution from noise to $c(d)$ for $d > 0$ negligible. Note that $C(d)$ and $c(d)$ coincide if we have data for all sequences and the data is noise-free.

The space of all possible sequence-function relationships form a G-dimensional vector space isomorphic to $\mathbb{R}^G$ and can be naturally decomposed as $\ell + 1$ orthogonal subspaces,

$$\mathbb{R}^G = V_0 \oplus V_1 \oplus \cdots \oplus V_\ell. \tag{6}$$

Here V_k corresponds to the space of functions of pure k-th order interactions and has dimension $m_k = \binom{\ell}{k}(\alpha - 1)^k$ (*SI Appendix*); in particular, V_k is the eigenspace of $\mathbf{L}$ associated with the eigenvalue αk. Next let $\mathbf{f}_k$ be the projection of $\mathbf{f}$ onto V_k so that $\mathbf{f} = \sum_{k=0}^{\ell} \mathbf{f}_k$. Since the different components $\mathbf{f}_k$ are orthogonal, we find $\|\mathbf{f}\|^2 = \sum_{k=0}^{\ell} \|\mathbf{f}_k\|^2$. We can now define a quantity that measures the contribution of $\mathbf{f}_k$ to the total variance in $\mathbf{f}$

$$\Omega_k = \frac{\|\mathbf{f}_k\|^2}{\sum_{i=1}^{\ell} \|\mathbf{f}_i\|^2} = \frac{\|\mathbf{f}_k\|^2}{\|\mathbf{f} - \overline{\mathbf{f}}\|^2}. \tag{7}$$

Ω_k measures the amount of variance in $\mathbf{f}$ that is due to k-th order interactions alone, and therefore is known as the empirical variance component or amplitude spectrum of order k of the landscape $\mathbf{f}$ [26, 44].

Recall that V_k is a m_k-dimensional subspace. Now let $\mathbf{Q}_k \in \mathbb{R}^{G\times m_k}$ be a matrix whose columns form an orthonormal basis for V_k. Since $\mathbf{f}_k \in \text{col}(\mathbf{Q}_k)$, we can express it as $\mathbf{f}_k = \mathbf{Q}_k\mathbf{a}_k$, where $\mathbf{a}_k = \left[a_{k,i}\right]_{1\leq i\leq m_k}$ is a vector containing $m_k = \binom{\ell}{k}(\alpha-1)^k$ entries know as the Walsh coefficients of order k. Therefore, the quantity

$$\lambda_k \equiv \frac{\|\mathbf{f}_k\|^2}{m_k} = \frac{\mathbf{a}_k^T\mathbf{Q}_k^T\mathbf{Q}_k\mathbf{a}_k}{m_k} = \frac{\|\mathbf{a}_k\|^2}{m_k} = \frac{\sum_{i=1}^{m_k} a_{k,i}^2}{m_k}. \tag{8}$$

is equal to the mean squared k-th order Walsh coefficient of $\mathbf{f}$. In the special case when $k = 0$, $\lambda_0 = \frac{\|\mathbf{f}_0\|^2}{m_0} = \|\mathbf{f}_0\|^2$, since m_0 is equal to 1. Here $\mathbf{f}_0$ is the projection onto V_0, the constant subspace, which is spanned by the unit vector $\mathbf{u} = \alpha^{-\frac{\ell}{2}}\mathbf{1}$, where $\mathbf{1}$ is the vector of all ones. Therefore, we find

$$\lambda_0 = \|\mathbf{f}_0\|^2 = \|(\mathbf{f}^T\mathbf{u})\mathbf{u}\|^2 = (\mathbf{f}^T\mathbf{u})^2 = (\alpha^{-\frac{\ell}{2}}\sum_x f(x))^2 = \alpha^\ell\overline{f}^2. \tag{9}$$

Next, it can be shown that any function $\boldsymbol{\phi}$ with unit norm drawn from the subspace V_k has the same autocovariance function [43]:

$$C_{\boldsymbol{\phi}}(d) = \frac{1}{m_k}w_k^\ell(d), \tag{10}$$

where $w_k^\ell(d)$ is the known as the Krawtchouk polynomial [43, 44, 73] and is given by

$$w_k^\ell(d) = \frac{1}{\alpha^\ell}\sum_{q=0}^{\ell}(-1)^q(\alpha-1)^{k-q}\binom{d}{q}\binom{\ell-d}{k-q}. \tag{11}$$

Since $\mathbf{f}_k \in V_k$ and has norm $\|\mathbf{f}_k\|^2$, its autocovariance function is

$$C_{\mathbf{f}_k}(d) = \frac{\|\mathbf{f}_k\|^2}{m_k}w_k^\ell(d) = \lambda_k w_k^\ell(d), \tag{12}$$

The landscape $\mathbf{f}$ is a linear combination of orthogonal components $\mathbf{f}_k$. It turns out its autocovariance function is simply the sum of autocovariance functions of the components $\mathbf{f}_k$ [43]:

$$C_{\mathbf{f}}(d) = \sum_{k=1}^{\ell} C_{\mathbf{f}_k}(d) = \sum_{k=1}^{\ell}\lambda_k w_k^\ell(d). \tag{13}$$

Therefore, knowing the λ_k, or equivalently the Ω_k together with the variance $C(0)$ of the full landscape, allows us to write down the autocovariance function $C_{\mathbf{f}}(d)$. Conversely, Eq. 13 also allows us to solve for the λ_k for $k > 0$ if we are given the autocovariance function.

Given a pair of alleles on a site, we can calculate the effect of mutation from one allele to the other in all genetic backgrounds. Therefore, we can naturally measure the covariance of mutational effects as a function of distance between background sequences [47] similar to how we measure phenotypic correlation using $C(d)$. Here we generalize this notion of distance covariance of mutational effects to epistatic coefficients of any order $< l$, which is a generalization of the classical epistatic coefficient to $k \geq 2$ sites (*SI Appendix*). Specifically, we define $\Gamma_k(d)$ as the distance covariance of k-th order epistatic coefficients averaged over the whole landscape (*SI Appendix*) and show that $\Gamma_k(d)$ can be expressed in terms of the autocovariance function $C(d)$ or, alternatively, the list of λ_i truncated so as to begin with λ_k:

$$\Gamma_k(d) = 2^k\sum_{q=0}^{k}(-1)^q\binom{k}{q}C(d+q) = 2^k\sum_{k'=k}^{\ell}\lambda_{k'}w_{k'-k}^{\ell-k}(d). \tag{14}$$

To summarize, we have defined a number of summary statistics and shown how to transform between them for any complete landscape $\mathbf{f}$. In situations where we only have the noised incomplete observation $\mathbf{y}$, we cannot directly calculate the underlying λ_k and Ω_k. However, we can still calculate the empirical autocovariance function $c(d)$. We can then estimate the λ_k using a least squares technique that we outline below. The λ_k's then allow us to calculate all summary statistics listed above.

Gaussian process regression

Given noisy observations on a subset of all possible sequences, our aim is to reconstruct the full landscape $\mathbf{f}$ so that the reconstructed landscape reflects the statistical features of the observed data. Since the underlying landscape $\mathbf{f}$ is unknown, we use a Bayesian strategy by treating it as a random function that is drawn from a Gaussian prior, that is

$$\mathbf{f} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K}), \tag{15}$$

where $\boldsymbol{\mu} \in \mathbb{R}^{G\times 1}$ and $\mathbf{K} \in \mathbb{R}^{G\times G}$ are the mean vector and covariance matrix, respectively. Throughout this paper, we assume the prior distribution has mean zero, i.e. $\mu = \mathbf{0}$.

To derive the covariance matrix $\mathbf{K}$, we start out by defining simple distributions for the Walsh coefficients of different orders k. Specifically, we assume the Walsh coefficients are independent and Gaussian with mean 0 and identical variance for each order k. Let $\mathbf{a}_k \in \mathbb{R}^{m_k}$ be the random vector containing all k-th order Walsh coefficients, then

$$\mathbf{a}_k \sim \mathcal{N}(\mathbf{0}, \lambda_k \mathbf{I}_{m_k}). \tag{16}$$

It is easy to check that the random vector $\mathbf{f} = \sum_{k=0}^{\ell} \mathbf{Q}\mathbf{a}_k$ is also Gaussian and has mean 0. Furthermore, its covariance matrix is

$$\mathbf{K} = \mathbb{E}_{\mathbf{f}}\left[(\mathbf{f}-\boldsymbol{\mu})(\mathbf{f}-\boldsymbol{\mu})^T\right] = \mathbb{E}_{\mathbf{a}_0,\mathbf{a}_1,\cdots,\mathbf{a}_\ell}\left[\sum_{j=0}^{\ell}\mathbf{Q}_j\mathbf{a}_j(\sum_{k=0}^{\ell}\mathbf{Q}_k\mathbf{a}_k)^T\right] = \sum_{j,k}\mathbf{Q}_j\mathbb{E}_{\mathbf{a}_j,\mathbf{a}_k}\left[\mathbf{a}_j\mathbf{a}_k^T\right]\mathbf{Q}_k^T \tag{17}$$

$$= \sum_{k=0}^{\ell}\lambda_k\mathbf{Q}_k\mathbf{Q}_k^T = \sum_{k=0}^{\ell}\lambda_k\mathbf{W}_k. \tag{18}$$

Here $\mathbf{W}_k = \mathbf{Q}_k\mathbf{Q}_k^T$ is a $G \times G$ matrix whose entries are given by the Krawtchouk polynomial [44, 73] and only depends on the Hamming distance between sequences:

$$\mathbf{W}_k(x, x') = w_k^{\ell}(d(x, x')). \tag{19}$$

Therefore, we have defined a family of Gaussian prior distributions for $\mathbf{f}$ with covariance matrix $\mathbf{K} = \sum_{k=0}^{\ell}\lambda_k\mathbf{W}_k$, where $\lambda_k > 0$ serve as hyperparameters of the prior distribution, which can be specified *a priori* or inferred from the data. Furthermore, since the columns of $\mathbf{Q}_k$ are orthonormal, $\mathbf{W}_k = \mathbf{Q}_k\mathbf{Q}_k^T$ is the projection matrix to the space of k-th order interactions. As a result, the matrix $\mathbf{K}$ defined above is guaranteed to be positive-definite if $\lambda_k > 0$ for all $\ell \geq k \geq 0$, therefore is a proper covariance matrix.

Because the $\mathbf{W}_k(x, x')$ only depend on the Hamming distance between pairs of sequences, the covariance of this prior distribution likewise is a function of the Hamming distance between sequences. In other words, we have defined a Gaussian isotropic random field [26, 43, 44]. This allows us to summarize the covariance structure of our prior distribution by the following kernel function

$$K(d) = \sum_{k=0}^{\ell}\lambda_k w_k^{\ell}(d). \tag{20}$$

Note that Eq.20 is very similar to Eq.13. The main difference is that here $\lambda_k > 0$ are hyperparameters for the prior that specify the variance of Walsh coefficients of order k, whereas in Eq. 13, λ_k is the mean square Walsh coefficient of a specific landscape. Note that we also include the 0 order term λ_0 in Eq.20 because we assume that the prior distribution has zero mean and the mean of a sample from the distribution is normally distributed with variance λ_0. In fact, the expected empirical autocovariance function differs from the kernel function by a constant:

$$\mathbb{E}_{\mathbf{f}}\left[C_{\mathbf{f}}(d)\right] = K(d) - \alpha^{-\ell}\lambda_0. \tag{21}$$

We write $I = A^\ell \setminus B$ as the set of all missing sequences. Throughout this paper, we also use B and I to denote columns and rows of matrices that are indexed by A^ℓ. For example, $\mathbf{K}_{BB}$ is the $m \times m$ submatrix of $\mathbf{K}$ generated by selecting rows and columns corresponding to B, while $\mathbf{K}_{\cdot B}$ denotes the $G \times m$ matrix whose columns correspond to sequences in B.

Recall that $\mathbf{y} = \mathbf{f}_B + \mathbf{e} \in \mathbb{R}^m$ is the vector of observations for the subset B, where $\mathbf{e}$ is a vector of observation noise so that $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{E})$ with $\mathbf{E}$ being a diagonal matrix. The distribution of $\mathbf{y}$ is

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{BB} + \mathbf{E}). \tag{22}$$

Without loss of generality, we will order our sequences so that the m sequences in B whose phenotypes are known come first. The joint distribution of the full landscape $\mathbf{f}$ and $\mathbf{y}$ is then

$$\begin{bmatrix}\mathbf{f}\\ \mathbf{y}\end{bmatrix} \sim \mathcal{N}(\begin{bmatrix}\mathbf{0}_G\\ \mathbf{0}_m\end{bmatrix}, \begin{bmatrix}\mathbf{K} & \mathbf{K}_{\cdot B}\\ \mathbf{K}_{\cdot B} & \mathbf{K}_{BB} + \mathbf{E}\end{bmatrix}). \tag{23}$$

The posterior distribution for $\mathbf{f}$ is also Gaussian and is given by well-known formula for Gaussian process regression [37]

$$\mathbf{f}|\mathbf{y} \sim \mathcal{N}(\mathbf{K}_{\cdot B}(\mathbf{K}_{BB} + \mathbf{E})^{-1}\mathbf{y}, \mathbf{K} - \mathbf{K}_{\cdot B}(\mathbf{K}_{BB} + \mathbf{E})^{-1}\mathbf{K}_{B\cdot}), \tag{24}$$

where $\mathbf{K}_{\cdot B}(\mathbf{K}_{BB} + \mathbf{E})^{-1}\mathbf{y} = \widehat{\mathbf{f}}$ is know as the maximum a posterior (MAP) estimate. The posterior variance for a single sequence x can be calculated as

$$\sigma_x^2 = \mathbf{K}_{xx} - \mathbf{K}_{xB}(\mathbf{K}_{BB} + \mathbf{E})^{-1}\mathbf{K}_{Bx}\ , \tag{25}$$

where $\mathbf{K}_{Bx} = \mathbf{K}_{xB}^T \in \mathbb{R}^m$ is the column vector containing the covariance between the genotype x and every genotype in the training data B.

Inference of hyperparameters for the prior distribution

To use Gaussian process regression, we must choose the covariance matrix $\mathbf{K}$, or equivalently a kernel function $K(d)$ for $d = 0, 1, \cdots, \ell$. According to Eq. 17, the kernel function $K(d)$ of our prior distribution must take the form $K(d) = \sum_{k=0}^{\ell} \lambda_k w_k^\ell(d)$, for $d = 0, 1, \cdots, \ell$. In this paper, we take an Empirical Bayes procedure to infer the hyperparameters λ_k directly from the data. First, recall that $c(d)$ is the empirical autocovariance function extracted from the data $\mathbf{y}$ (Eq. 5). So our overall goal is to find the hyperparameters λ_k so that the kernel function $K(d)$ aligns as well as possible with $c(d)$. Here, we provide a naive method as well as a regularized least square method for estimating λ_k to accommodate various possible scenarios in the inference procedure. In the naive method, we directly solve the linear equation

$$c(d) + \overline{y}^2 = \sum_{k=0}^{\ell} \lambda_k w_k^\ell(d), \quad d = 0, 1, \cdots, \ell \tag{26}$$

for the hyperparameters λ_k's, where $\overline{y}^2$ is added to allow us to infer the 0-th order hyperparameter λ_0, since $c(d)$ does not contain the mean of the data. This procedure is equivalent to using $c(d) + \overline{y}^2$ as the kernel function $K(d)$ for our prior distribution. However, an important constraint for λ_k is that they must be nonnegative, since they are the eigenvalues of the covariance matrix $\mathbf{K} = \sum_{k=0}^{\ell} \lambda_k \mathbf{W}_k$. While it has been shown that when B is the whole sequence space, the λ_k's solved using the equation above must be nonnegative [43], no such guarantee exists when B is a proper subset of A^ℓ.

Another possible scenario where solving Eq. 26 is impossible is when the data does not contain all possible distance classes, i.e., $\mathcal{D}_i = 0$ for some i. Therefore, we introduce a second method using regularized least squares to estimate the λ_k. This method is similar to a machine learning technique called kernel alignment [51]. Briefly, our strategy is to match the empirical second moment matrix $\mathbf{y}\mathbf{y}^T$ using a nonnegative linear combination of the basis matrices $\mathbf{W}_k$'s and the noise variance matrix $\mathbf{E}$. Mathematically, we achieve this by minimizing the squared Frobenius norm ($\|\cdot\|_F$) of the difference between the target matrix $\mathbf{y}\mathbf{y}^T$ and the submatrix $\mathbf{K}_{BB} = \sum_{k=0}^{\ell} \lambda_k \mathbf{W}_{k_{BB}}$

$$\|\mathbf{y}\mathbf{y}^T - (\sum_{k=0}^{\ell} \lambda_k \mathbf{W}_{k_{BB}} + \mathbf{E})\|_F^2 = \boldsymbol{\lambda}^T \mathbf{M} \boldsymbol{\lambda} - 2\boldsymbol{\lambda}^T \mathbf{a} + \text{constant}, \tag{27}$$

where $M_{i,j} = \langle \mathbf{W}_{i_{BB}}, \mathbf{W}_{j_{BB}} \rangle_F$ and $a_i = \langle \mathbf{W}_{i_{BB}}, \mathbf{y}\mathbf{y}^T - E \rangle_F$.

To ensure the nonnegativity of λ_k, we parametrize it as $\lambda_k = \exp(\eta_k)$. Furthermore, we introduce a regularization term $\sum_{k=2}^{\ell-1} \|2\eta_k - \eta_{k-1} - \eta_{k+1}\|^2$ equal to the sum of squared second order finite differences in $\boldsymbol{\eta}$. This term is added to the cost function in Eq.(27) to penalize the deviation of $\lambda_1, \cdots, \lambda_l$ from a linear function on the log scale. We then find the optimal $\widehat{\boldsymbol{\lambda}} = \exp(\widehat{\boldsymbol{\eta}})$ by solving the following minimization problem

$$\widehat{\boldsymbol{\eta}} = \text{argmin}_{\boldsymbol{\eta} \in \mathbb{R}^{\ell+1}} (e^{\boldsymbol{\eta}})^T \mathbf{M} e^{\boldsymbol{\eta}} - 2\mathbf{a}^T e^{\boldsymbol{\eta}} + \beta \sum_{k=2}^{\ell-1} \|2\eta_k - \eta_{k-1} - \eta_{k+1}\|. \tag{28}$$

Here $\beta > 0$ is the regularization parameter. In practice, we can choose the optimal β using 10-fold crossvalidation. Since the vector $\boldsymbol{\eta}$ has only $\ell + 1$ entries, the solution is readily found using generic minimization algorithms.

Posterior sampling using Hamiltonian Monte Carlo

Eq. 25 allows us to calculate the posterior variance for individual sequences. However, since the evaluation of this function is as costly as the MAP estimate, in practice we can only acquire the posterior variance for a subset of sequences of high interest. In this section we outline an alternative method for estimating the posterior covariance matrix by directly sampling from the posterior distribution in Eq. 24. Specifically, suppose $\mathbf{f}^{(l)}$, $(l = 1, \cdots, n)$ is a set of n samples drawn from the posterior distribution using a Markov chain whose stationary distribution is our posterior distribution, then we can approximate the posterior covariance matrix with the finite sum

$$\mathbf{K} - \mathbf{K}_{\cdot B}(\mathbf{K}_{BB} + \mathbf{E})^{-1}\mathbf{K}_{B\cdot} = \mathrm{E}\left[(\mathbf{f} - \widehat{\mathbf{f}})(\mathbf{f} - \widehat{\mathbf{f}})\right] \approx \frac{1}{n} \sum_{l=1}^{n} (\mathbf{f}^{(l)} - \overline{\mathbf{f}})(\mathbf{f}^{(l)} - \overline{\mathbf{f}})^T, \tag{29}$$

where $\overline{\mathbf{f}}$ is the mean vector taken over all samples $\mathbf{f}^{(l)}$.

A major challenge for sampling from the posterior distribution is posed by the typical high dimensionality of the sequence space. Specifically, as the dimension of the sample space increases, the region of high probability of the posterior distribution (the typical set) becomes increasingly singular and concentrated in space [39, 74]. As a consequence, the diffusive behavior of popular naive random walk algorithms such as MCMC either leads to high rejection rates or highly autocorrelated samples, both making the exploration of the probability distribution extremely slow.

In this paper, we employ the Hamiltonian Monte Carlo (HMC) sampling method [39, 74]. HMC is a gradient-based algorithm that is able to take advantage of the local geometry of the typical set, making it more suitable for sampling from high dimensional probability distributions. HMC first introduces an auxiliary momentum parameter to complement each dimension of our target probability space. The total energy (the Hamiltonian) of the system is then defined as the sum of the potential energy given by the log probability of the posterior and the kinetic energy, which is equal to the squared norm of the momentum vector in our case. The algorithm proceeds using a Markov chain consisting of alternate random updates to the momentum vector and deterministic integration of Hamiltonian dynamics that leaves the total energy unchanged. In practice, this integration is discretized and performed using the so-called leapfrog method [39]. Since the numerical errors accumulated during the leapfrog steps lead to changes in total energy at the end of the integration, a Metropolis step at the end of the numerical integration is used to keep the Markov chain reversible. Together, this sampling scheme allows the HMC algorithm to make large jumps in probability space while keeping the rejection rate small.

The HMC algorithm relies on the gradient of the log posterior probability to perform the Hamiltonian dynamics integration. To derive the gradient, first define the precision matrix for the posterior distribution

$$\widetilde{\mathbf{K}} = (\mathbf{K} - \mathbf{K}_{\cdot B}(\mathbf{K}_{BB} + \mathbf{E})^{-1}\mathbf{K}_{B\cdot}))^{-1}. \tag{30}$$

Since the log probability of a sample $\mathbf{f}$ is $\log(\mathbf{f}|\mathbf{y}) = -\frac{1}{2}(\mathbf{f} - \widehat{\mathbf{f}})^T\widetilde{\mathbf{K}}(\mathbf{f} - \widehat{\mathbf{f}}) + \text{costant}$, we find

$$\frac{1}{2}\nabla \log p(\mathbf{f}|\mathbf{y}) = -\widetilde{\mathbf{K}}\mathbf{f} + \widetilde{\mathbf{K}}\widehat{\mathbf{f}}. \tag{31}$$

Next, we can simplify the expression for $\widetilde{\mathbf{K}}$ by expanding the inverse using the Woodbury identity. This gives

$$\widetilde{\mathbf{K}} = \mathbf{K}^{-1} + \begin{bmatrix} \mathbf{E}^{-1} & 0 \\ 0 & 0 \end{bmatrix}. \tag{32}$$

Since Eq. 17 is also the eigendecomposition of $\mathbf{K}$ (*SI Appendix*) and all λ_k's are constrained to be positive, the inverse of $\mathbf{K}$ exists and is equal to $\mathbf{K}^{-1} = \sum_{k=0}^{\ell} \frac{1}{\lambda_k}\mathbf{W}_k$. Therefore, the evaluation of Eq. 31 involves multiplying the sample $\mathbf{f}$ by a diagonal matrix and the matrix $\mathbf{K}^{-1}$, which can be greatly sped up using a representation of $\mathbf{K}^{-1}$ as a polynomial in the sparse graph Laplacian $\mathbf{L}$.

Finally, we employ the dual averaging algorithm [75] to find the optimal step size for the leapfrog integrator during an initial tuning phase, so that the average rejection rate of the Metropolis steps is near the optimal value of 0.65 [75].

Regularized regression

We use the following linear model to fit additive, pairwise and 3-way interaction models:

$$\widehat{f}(x) = \sum_j \beta_j \phi_j(x), \tag{33}$$

where the $\phi_j(x)$ are an indicator variables encoding the presence or absence of particular alleles at particular sites in x. For the additive model, each $\phi_j(x)$ encode the presence or absence of a given allele on a given site. For the pairwise and three-way models, $\phi_j(x)$ encode the presence or absence of combinations of allelic states for each possible pair of sites or triple or sites, respectively. We can express Eq.(33) in matrix notation

$$\widehat{\mathbf{f}} = \mathbf{X}\boldsymbol{\beta}. \tag{34}$$

Given m observations, the dimension of $\mathbf{X}$ is $m \times \ell\alpha$ for the additive model, and $m \times \sum_{k=0}^{2}\binom{\ell}{k}\alpha^k$ and $m \times \sum_{k=0}^{3}\binom{\ell}{k}\alpha^k$ for the pairwise and three-way model, respectively.

We fit the additive model using ordinary least squares. The pairwise and three-way regression models were fitted using elastic net regularization, where the penalty of model complexity is a mixture of L_1 and L_2 norms. Specifically, we find our solution by minimizing

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2 + \lambda\left((1-\alpha)\|\boldsymbol{\beta}\|_2^2/2+\alpha\|\boldsymbol{\beta}\|_1\right), \tag{35}$$

where the penalty for model complexity is controlled by α, which represents a compromise between lasso ($\alpha = 1$) and ridge ($\alpha = 0$) regressions. The parameter λ controls the overall strength of the penalty. Both α and λ were chosen by 10-fold cross validation for each training sample. Elastic net regressions were fit using the R package glmnet [76].

Processing of the *SMN1* dataset

The *SMN1* raw dataset consists of enrichment ratio (number of output reads/number of input reads) across three libraries, each containing three replicates. Previous analysis discarded two replicates due to low sample quality. Since no library effect was detected [40], we consider the enrichment ratios across 7 replicates as independent samples. Depending on its presence or absence in each input sample, a splice site can have zero to 7 measured enrichment ratios. Out of the 32768 possible splice sites, 2036 are not represented in any replicates, and therefore are considered missing data.

We assume the enrichment ratios across replicates for a given genotype are log-normally distributed. First, for sequences with all positive ratios across n replicates ($1 < n \leq 7$), we use the bias corrected geometric mean [77] as the estimate of the median enrichment ratio using the formula

$$\mu = \exp(\overline{y} - \widehat{\sigma}^2/2n), \tag{36}$$

where $\overline{y}$ and $\widehat{\sigma}^2$ are the arithmetic mean and sample variance of the log-transformed enrichment ratios, respectively. For sequences containing zero enrichment ratios where the above equation is inapplicable, we simply calculate the median of the enrichment ratios across replicates.

We then estimate the variance for the log-normal distribution using the standard formula

$$\sigma^2 = (\exp(\widehat{\sigma}^2) - 1)\exp(2\widehat{\mu} + \widehat{\sigma}^2) \tag{37}$$

For sequences with zero ratios and/or with only 1 replicate, we use the modified formula

$$\sigma^2 = (\exp(\overline{\widehat{\sigma}^2}) - 1)\exp(2\mu' + \overline{\widehat{\sigma}^2}), \tag{38}$$

where μ' is the log of the median of the enrichment ratios and $\overline{\widehat{\sigma}^2}$ is the mean $\widehat{\sigma}^2$ for all sequences with only positive ratios and at least two replicates.

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 a *SMN1* minigene that spans exon 6-8 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 [40] (minigene sequence is available at `https://github.com/jbkinney/15_splicing`). The minigenes containing variable 5′ss were inserted in to the pcDNA5/FRT expression vector (Invitrogen). 1 µg of minigene plasmid was then transiently transfected into HeLa cells, which were collected after 48 hr. 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: CTGGCTAACTAGAGAACCCACTGC;

R: GGCAACTAGAAGGCACAGTCG) and P32-labelled dCTP using Q5 High-Fidelity DNA Polymerase (New England Biolabs) following the manufacturer's instructions. PCR products were separated on a 5.5% non-denaturing polyacrylamide gel and were detected using a Typhon FLA7000 phosphorimager. Finally, we used ImageJ (NIH) to quantify isoform abundance. All 5′ss were assessed in triplicates.

Visualization of the *SMN1* splicing landscape

To derive a low dimensional representation of the splicing landscape, we consider a population evolving in continuous time under weak mutation [79–81] with natural selection acting to maintain splicing activity. We first used our method to reconstruct the full landscape consisting of 65536 sequences corresponding to all combinations of alleles at the eight variable positions of the 9-nt splice site. Note that the reconstructed landscape also includes sequences with A or G at the +2 position, which do not constitute valid splice sites. Since these sequences are nonetheless accessible through mutation, we include them but set the PSI of all such sequences to be zero. Next, exon-exon junction sequencing in the original study revealed that a secondary GU at the -2 and -1 positions can be preferentially used over the GU or GC at position +1 and +2 [40], leading to a frameshift in the mature mRNA. Therefore, we set the PSI of all such sequences to be zero. Last, to ensure an appropriate degree of realism for the evolutionary Markov chain, we truncate all predicted PSI values to be between 0 and 100. We model evolution as a continuous-time Markov chain where the population moves between sequences at each fixation event based on fitness values given by the modeled PSI. The rate matrix $\mathbf{Q}$ of the Markov chain is

$$\mathbf{Q}_{x,x'} = \begin{cases} \frac{1}{\alpha-1}\frac{c(f(x')-f(x))}{1-e^{-c(f(x')-f(x))}} & d(x,x')=1 \\ -\sum_{x''\neq x}\mathbf{Q}_{x,x''} & x=x' \\ 0 & \text{otherwise,} \end{cases} \tag{39}$$

where c is the conversion factor that transforms PSI to scaled fitness (Malthusian fitness $\times\ N_e$). We choose c so that the expected PSI at stationarity is equal to 80. Time is scaled so that the total mutation rate per site is equal to 1. We use the right eigenvectors of $\mathbf{Q}$ associated with the 3 greatest nonzero eigenvalues as coordinates to embed the splicing landscape in three dimensions, where each eigenvector is scaled so that the weighted mean of its squared entries is equal to the relaxation time of the associated eigenmode where the weights are given by the frequency of each genotype at stationarity (see [58] for details). This allows our low-dimensional representation of the landscape to optimally capture the expected time for a population to evolve between sequences [58]. Note that although we included sequences with A or G at position +2 when calculating the embedding coordinates, for simplicity we omitted these sequences when plotting the final visualization.

Acquisition of human 5′ splice sites

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

Code availability

We developed vcregression, a python command-line interface that implements the Empirical Variance Component method described here (available at `https://github.com/davidmccandlish/vcregression`).

Acknowledgements

This work was funded by NIH grant R35GM133613, 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. M.S.W. and A.R.K. acknowledge support from NIH grant R37GM42699.

References

[1] Patrick C Phillips. "Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems". In: *Nat. Rev. Genet.* 9.11 (2008), pp. 855–867.

[2] Dmitry A Kondrashov and Fyodor A Kondrashov. "Topological features of rugged fitness landscapes in sequence space". In: *Trends Genet.* 31.1 (2015), pp. 24–33.

[3] Júlia Domingo, Pablo Baeza-Centurion, and Ben Lehner. "The Causes and Consequences of Genetic Interactions (Epistasis)". In: *Annu. Rev. Genomics Hum. Genet.* 20 (2019).

[4] Douglas M Fowler et al. "High-resolution mapping of protein sequence-function relationships". In: *Nat. Methods* 7.9 (2010), pp. 741–746.

[5] Lea M Starita et al. "Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis". In: *Proc. Natl. Acad. Sci. U.S.A.* 110.14 (2013), E1263–E1272.

[6] Daniel Melamed et al. "Deep mutational scanning of an RRM domain of the *Saccharomyces cerevisiae* poly (A)-binding protein". In: *RNA* 19.11 (2013), pp. 1537–1551.

[7] C Anders Olson, Nicholas C Wu, and Ren Sun. "A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain". In: *Curr. Biol.* 24.22 (2014), pp. 2643–2651.

[8] Michael B Doud, Orr Ashenberg, and Jesse D Bloom. "Site-specific amino acid preferences are mostly conserved in two closely related protein homologs". In: *Mol. Biol. Evol.* 32.11 (2015), pp. 2944–2960.

[9] Anna I Podgornaia and Michael T Laub. "Pervasive degeneracy and epistasis in a protein-protein interface". In: *Science* 347.6222 (2015), pp. 673–677.

[10] Karen S Sarkisyan et al. "Local fitness landscape of the green fluorescent protein". In: *Nature* 533.7603 (2016), p. 397.

[11] Barrett Steinberg and Marc Ostermeier. "Shifting fitness and epistatic landscapes reflect trade-offs along an evolutionary pathway". In: *J Mol Biol.* 428.13 (2016), pp. 2730–2743.

[12] Claudia Bank et al. "On the (un)predictability of a large intragenic fitness landscape". In: *Proc. Natl. Acad. Sci. U.S.A.* 113.49 (2016), pp. 14085–14090.

[13] Tyler N Starr, Lora K Picton, and Joseph W Thornton. "Alternative evolutionary histories in the sequence space of an ancient protein." In: *Nature* 549.7672 (Sept. 2017), pp. 409–413.

[14] Victoria O Pokusaeva et al. "An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape". In: *PLos Genet.* 15.4 (2019), e1008079.

[15] Calin Plesa et al. "Multiplexed gene synthesis in emulsions for exploring protein functional landscapes". In: *Science* 359.6373 (2018), pp. 343–347.

[16] J N Pitt and A R Ferré-D'Amaré. "Rapid Construction of Empirical RNA Fitness Landscapes". In: *Science* 330.6002 (2010), pp. 376–379. DOI: 10.1126/science.1192001. URL: http://www.sciencemag.org/content/330/6002/376.abstract.

[17] José I Jiménez et al. "Comprehensive experimental fitness landscape and evolutionary network for small RNA". In: *Proc. Natl. Acad. Sci. U.S.A.* 110.37 (2013), pp. 14984–14989.

[18] Olga Puchta et al. “Network of epistatic interactions within a yeast snoRNA”. In: *Science* 352.6287 (2016), pp. 840–844.

[19] Chuan Li et al. “The fitness landscape of a tRNA gene”. In: *Science* 352.6287 (2016), pp. 837–840.

[20] Júlia Domingo, Guillaume Diss, and Ben Lehner. “Pairwise and higher-order genetic interactions during the evolution of a tRNA”. In: *Nature* 558.7708 (2018), p. 117.

[21] Justin B Kinney et al. “Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence”. In: *Proc. Natl. Acad. Sci. U.S.A.* 107.20 (2010), pp. 9158–9163.

[22] Alexander B Rosenberg et al. “Learning the sequence determinants of alternative splicing from millions of random sequences”. In: *Cell* 163.3 (2015), pp. 698–711.

[23] Philippe Julien et al. “The complete local genotype–phenotype landscape for the alternative splicing of a human exon”. In: *Nat. Commun.* 7 (2016), p. 11558.

[24] Shengdong Ke et al. “Saturation mutagenesis reveals manifold determinants of exon definition”. In: *Genome Res.* 28.1 (2018), pp. 11–24.

[25] Daniel M Weinreich et al. “Should evolutionary geneticists worry about higher-order epistasis?” In: *Curr. Opin. Genet. Dev.* 23.6 (2013), pp. 700–707.

[26] Johannes Neidhart, Ivan G Szendro, and Joachim Krug. “Exact results for amplitude spectra of fitness landscapes”. In: *J. Theor. Biol.* 332 (2013), pp. 218–227.

[27] Tyler N Starr and Joseph W Thornton. “Epistasis in protein evolution”. In: *Protein Sci.* 25.7 (2016), pp. 1204–1218.

[28] Zachary R Sailer and Michael J Harms. “Detecting High-Order Epistasis in Nonlinear Genotype-Phenotype Maps”. In: *Genetics* 205.3 (2017), pp. 1079–1088.

[29] Zachary R Sailer and Michael J Harms. “High-order epistasis shapes evolutionary trajectories”. In: *PLoS Comput. Biol.* 13.5 (2017), e1005541.

[30] Nicholas Wu et al. “Adaptation in protein fitness landscapes is facilitated by indirect paths”. In: *eLife* 5 (2016), e16965.

[31] Julian Echave and Claus O Wilke. “Biophysical models of protein evolution: understanding the patterns of evolutionary sequence divergence”. In: *Annu. Rev. Biophys.* 46 (2017), pp. 85–103.

[32] Frank J Poelwijk, Michael Socolich, and Rama Ranganathan. “Learning the pattern of epistasis linking genotype and phenotype in a protein”. In: *Nat. Commun.* 10.1 (2019), pp. 1–11.

[33] Aneth S Canale et al. “Evolutionary mechanisms studied through protein fitness landscapes”. In: *Curr. Opin. Struct. Biol.* 48 (2018), pp. 141–148.

[34] Daniel M. Weinreich et al. “The Influence of Higher-Order Epistasis on Biological Fitness Landscape Topography”. In: *J. Stat. Phys.* 172.1 (2018), pp. 208–225. ISSN: 00224715. DOI: 10.1007/s10955-018-1975-3. URL: https://doi.org/10.1007/s10955-018-1975-3.

[35] Jay F Storz. “Compensatory mutations and epistasis for protein function”. In: *Curr. Opin. Struct. Biol.* 50 (2018), pp. 18–25.

[36] Jakub Otwinowski, David Martin McCandlish, and Joshua B Plotkin. “Inferring the shape of global epistasis”. In: *Proceedings of the National Academy of Sciences* 115.32 (2018), E7550–E7558.

[37] Carl Edward Rasmussen and Christopher K I Williams. *Gaussian processes for machine learning.* Vol. 1. MIT press Cambridge, 2006.

[38] Bradley P Carlin and Thomas A Louis. *Bayes and empirical Bayes methods for data analysis.* Vol. 88. Chapman & Hall/CRC Boca Raton, 2000.

[39] Radford M Neal et al. "MCMC using Hamiltonian dynamics". In: *Handbook of markov chain monte carlo* 2.11 (2011), p. 2.

[40] Mandy S Wong, Justin B Kinney, and Adrian R Krainer. "Quantitative Activity Profile and Context Dependence of All Human 5' Splice Sites". In: *Mol. Cell* (2018).

[41] R A Fisher. "The Correlation Between Relatives on the Supposition of Mendelian Inheritance". In: *Trans R Soc Edinburgh* 52.02 (1918), pp. 399–433.

[42] Trevor Hinkley et al. "A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase". In: *Nature Genetics* 43.5 (2011), pp. 487–489.

[43] Robert Happel and Peter F Stadler. "Canonical approximation of fitness landscapes". In: *Complexity* 2.1 (1996), pp. 53–58.

[44] Peter F Stadler and Robert Happel. "Random field models for fitness landscapes". In: *J. Math. Biol.* 38.5 (1999), pp. 435–478.

[45] Manfred Eigen, John McCaskill, and Peter Schuster. "The molecular quasi-species". In: *Adv. Chem. Phys* 75 (1989), pp. 149–263.

[46] Edward D Weinberger. "Fourier and Taylor series on fitness landscapes". In: *Biol Cybern* 65.5 (1991), pp. 321–330.

[47] Luca Ferretti et al. "Measuring epistasis in fitness landscapes: The correlation of fitness effects of mutations". In: *Journal of Theoretical Biology* 396 (2016), pp. 132–143. ISSN: 10958541. DOI: 10.1016/j.jtbi.2016.01.037.

[48] Juannan Zhou and David M McCandlish. "Minimum epistasis interpolation for sequence-function relationships". In: *Nature communications* 11.1 (2020), pp. 1–14.

[49] Peter F Stadler. "Landscapes and their correlation functions". In: *J. Math. Chem.* 20.1 (1996), pp. 1–45.

[50] Peter F Stadler. "Fitness landscapes". In: *Biological Evolution and Statistical Physics*. Springer, 2002, pp. 183–204.

[51] Tinghua Wang, Dongyan Zhao, and Shengfeng Tian. "An overview of kernel alignment and its applications". In: *Artificial Intelligence Review* 43.2 (2015), pp. 179–192.

[52] John P Cunningham, Krishna V Shenoy, and Maneesh Sahani. "Fast Gaussian process methods for point process intensity estimation". In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 192–199.

[53] Peter F Stadler. "Random walks and orthogonal functions associated with highly symmetric graphs". In: *Discrete mathematics* 145.1-3 (1995), pp. 229–237.

[54] DG Higman. "Intersection matrices for finite permutation groups". In: *Journal of Algebra* 6 (1967), pp. 22–42.

[55] Daniel M Weinreich et al. "The influence of higher-order epistasis on biological fitness landscape topography". In: *Journal of Statistical Physics* 172.1 (2018), pp. 208–225.

[56] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the royal statistical society: series B (statistical methodology)* 67.2 (2005), pp. 301–320.

[57] Yasushi Kondo et al. "Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5 splice site recognition". In: *Elife* 4 (2015), e04986.

[58] David M McCandlish. "Visualizing fitness landscapes". In: *Evolution* 65.6 (2011), pp. 1544–1558.

[59] Ronald R Coifman and Stéphane Lafon. "Diffusion maps". In: *Applied and computational harmonic analysis* 21.1 (2006), pp. 5–30.

[60] David M McCandlish. “Long-term evolution on complex fitness landscapes when mutation is weak”. In: *Heredity* 121.5 (2018), pp. 449–465.

[61] Xavier Roca et al. “Features of 5-splice-site efficiency derived from disease-causing mutations and comparative genomics”. In: *Genome research* 18.1 (2008), pp. 77–87.

[62] Ido Carmel et al. “Comparative analysis detects dependencies among the 5 splice-site positions”. In: *RNA* 10.5 (2004), pp. 828–840.

[63] Chris Burge and Samuel Karlin. “Prediction of complete gene structures in human genomic DNA”. In: *Journal of Molecular Biology* 268.1 (1997), pp. 78–94.

[64] Gene Yeo and Christopher B Burge. “Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals”. In: *Journal of computational biology* 11.2-3 (2004), pp. 377–394.

[65] Syue-Ting Kuo et al. “Global fitness landscapes of the Shine-Dalgarno sequence”. In: *Genome Research* 30.5 (2020), pp. 711–723.

[66] Jiazi Tan et al. “Noncanonical registers and base pairs in human 5 splice-site selection”. In: *Nucleic acids research* 44.8 (2016), pp. 3908–3921.

[67] Philip A Romero, Andreas Krause, and Frances H Arnold. “Navigating the protein fitness landscape with Gaussian processes”. In: *Proc. Natl. Acad. Sci. U.S.A.* 110.3 (2013), E193–E201.

[68] Alexander J Smola and Risi Kondor. “Kernels and regularization on graphs”. In: *COLT*. Vol. 2777. Springer. 2003, pp. 144–158.

[69] Justin B Kinney and David M McCandlish. “Massively Parallel Assays and Quantitative Sequence–Function Relationships”. In: *Annu Rev Genomics Hum Genet* 20 (2019).

[70] Karen S Sarkisyan et al. “Local fitness landscape of the green fluorescent protein”. In: *Nature* 533.7603 (2016), pp. 397–401.

[71] Ammar Tareen et al. “MAVE-NN: Quantitative Modeling of Genotype-Phenotype Maps as Information Bottlenecks”. In: *BioRxiv* (2020).

[72] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. “A generalized representer theorem”. In: *International conference on computational learning theory*. Springer. 2001, pp. 416–426.

[73] Vladimir I Levenshtein. “Krawtchouk polynomials and universal bounds for codes and designs in Hamming spaces”. In: *IEEE Transactions on Information Theory* 41.5 (1995), pp. 1303–1321.

[74] Michael Betancourt. “A conceptual introduction to Hamiltonian Monte Carlo”. In: *arXiv preprint arXiv: 1701.02434* (2017).

[75] Matthew D Hoffman and Andrew Gelman. “The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.” In: *J. Mach. Learn. Res.* 15.1 (2014), pp. 1593–1623.

[76] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. “Regularization paths for generalized linear models via coordinate descent”. In: *Journal of statistical software* 33.1 (2010), p. 1.

[77] TB Parkin and JA Robinson. “Statistical evaluation of median estimators for lognormally distributed variables”. In: *Soil Science Society of America Journal* 57.2 (1993), pp. 317–323.

[78] James O Ramsay et al. “Monotone regression splines in action”. In: *Statistical Science* 3.4 (1988), pp. 425–441.

[79] Yoh Iwasa. “Free fitness that always increases in evolution”. In: *J. Theor. Biol.* 135.3 (1988), pp. 265–281.

[80] Guy Sella and Aaron E Hirsh. “The application of statistical physics to evolutionary biology”. In: *Proc. Natl. Acad. Sci. U.S.A.* 102.27 (2005), pp. 9541–9546.

[81] David M McCandlish, Premal Shah, and Joshua B Plotkin. “Epistasis and the dynamics of reversion in molecular evolution”. In: *Genetics* 203.3 (2016), pp. 1335–1351.