Inference of fitness landscapes with heterogeneous patterns of epistasis across sites

Carlos Martí-Gómez[1,*] and David M. McCandlish[1,*]

[1]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724

*Correspondence: martigo@cshl.edu, mccandlish@cshl.edu

Abstract

Fitness landscapes provide a framework for understanding how genetic variation shapes evolutionary outcomes. Although these landscapes were long treated as abstract conceptual objects, recent advances in genetic engineering and high-throughput phenotyping have enabled the empirical measurement of phenotypic values across large combinatorial sequence spaces. These developments create a need for statistical frameworks that can summarize, infer, and interpret fitness landscapes in the presence of complex genetic interactions. Here, we introduce a framework for summarizing the structure of genetic interactions across sites based on the average squared local k-way epistatic coefficients between mutations at different subsets of sites, and derive the precise manner in which the variance in these local k-way epistatic coefficients across backgrounds relates to epistasis of orders higher than k. These statistics can be computed exactly for complete combinatorial landscapes and are related to classical statistics in the fitness landscape literature. Moreover, they can be estimated from empirical correlations when data are incomplete or noisy, and used to define an empirical Bayes prior for fitness landscape inference that differentially penalizes interactions involving different subsets of sites. We apply this inference method to diverse high-throughput protein and RNA combinatorial mutagenesis datasets and find that fitness landscapes often show highly structured patterns of genetic interactions across positions. Finally, we use this model to infer a fitness landscape for a dynamic self-splicing intron comprising 65,536 genotypes, and describe in detail the main genetic interactions that shape the structure of this landscape and how they relate to the underlying molecular mechanism. Together, these results provide new tools for summarizing and modeling complex fitness landscapes, and for linking large-scale empirical data to the mathematical theory of fitness landscapes.

Keywords: fitness landscape; epistasis; Gaussian process

Introduction

The fitness landscape is a fundamental concept in evolutionary biology and genetics. First introduced by Wright (1932), it describes the mapping between genotypes and their associated fitness. The structure of a fitness landscape, including the number, distribution, and connectivity of fitness peaks, plays a crucial role in how populations evolve and diverge over time (Kauffman and Levin 1987; Kondrashov *et al.* 2002; Gavrilets 2004; Weinreich *et al.* 2006; McCandlish 2011; De Visser and Krug 2014; Fragata *et al.* 2019; Manrubia *et al.* 2021; Bank 2022; Johnson *et al.* 2023). Understanding the mapping from genotype to fitness is not only important for explaining and predicting evolution, but also has critical applications in cancer and human disease (Moore and Williams 2009; Dasari *et al.* 2021) as well as plant and animal breeding (De Los Campos *et al.* 2013; Sackton and Hartl 2016; Soyk *et al.* 2020; Dwivedi *et al.* 2024). However, despite its importance, characterizing this mapping is inherently challenging due to the high dimensionality of sequence space. Because the number of possible sequences grows exponentially with sequence length, such high-dimensional fitness landscapes are often summarized by computing low-dimensional summary statistics that characterize the ruggedness or structure of the landscape, e.g. the number of local optima, lengths of adaptive walks, and the number of alternative local optima accessible from starting genotype (Kauffman and Levin 1987; Szendro *et al.* 2013; Ferretti *et al.* 2018). Another common approach is to characterize how mutational effects change across genetic backgrounds, for instance by quantifying the average magnitude of local epistatic coefficients (Zhou and McCandlish 2020) or by measuring the correlation of mutational effects between genotypes separated by increasing numbers of mutations (Weinberger 1990; Stadler 1996; Neidhart *et al.* 2013; Bank *et al.* 2016; Ferretti *et al.* 2016).

Historically, the scarcity of comprehensive experimental data has motivated the development of theoretical and computational models of fitness landscapes, which have provided a framework for understanding how summary statistics behave across different classes of landscapes. One approach is to consider families of fitness landscapes drawn from a probability distribution, generally known as random field models (Kauffman and Levin 1987; Stadler and Happel 1999). Classical examples include the House of Cards model (Kingman 1978), the NK model (Kauffman and Levin 1987), and the Rough Mount Fuji landscape (Aita and Husimi 1998). Within these frameworks, one can compute the expected values of summary statistics and analyze how they depend on parameters that control the smoothness or ruggedness of the landscape (Kauffman and Levin 1987; Schmiegelt and Krug 2014; Neidhart *et al.* 2014; Hwang *et al.* 2018; Reddy

and Desai 2021).

More recent efforts have increasingly turned toward the experimental characterization of empirical fitness landscapes by measuring growth rates or other measures of biological functionality for many different combinations of mutations (De Visser and Krug 2014; Fragata *et al.* 2019). The size of the first empirically reconstructed fitness landscapes was limited by the difficulty of engineering large numbers of genotypes and measuring their fitness experimentally (Khan *et al.* 2011; Chou *et al.* 2011; Flynn *et al.* 2013; Szendro *et al.* 2013; Ogbunugafor *et al.* 2016; Weinreich *et al.* 2018; Gao *et al.* 2022; Aguirre *et al.* 2023; Zebell *et al.* 2025). However, recent advances in high-throughput assays (Kinney *et al.* 2010; Fowler and Fields 2014; Kinney and McCandlish 2019) have substantially expanded this scope, enabling the parallel measurement of thousands to millions of genotypes. These techniques have been used to characterize fitness landscapes across a range of biological systems, including regulatory sequences (Noderer *et al.* 2014; Rosenberg *et al.* 2015; Bonde *et al.* 2016; Evfratov *et al.* 2017; Rabani *et al.* 2017; Wong *et al.* 2018; Baeza-Centurion *et al.* 2019; Kuo *et al.* 2020; Komarova *et al.* 2020; de Boer *et al.* 2020; Vaishnav *et al.* 2022; Liao *et al.* 2023; Westmann *et al.* 2024b,a; Kuo *et al.* 2025; Chattopadhyay *et al.* 2025; Agarwal *et al.* 2025), RNAs (Domingo *et al.* 2018; Bendixsen *et al.* 2019; Soo *et al.* 2021; Rotrattanadumrong and Yokobayashi 2022), proteins (O'Maille *et al.* 2008; Bank *et al.* 2016; Wu *et al.* 2016; Starr *et al.* 2017; Poelwijk *et al.* 2019; Lite *et al.* 2020; Jalal *et al.* 2020; Bryant *et al.* 2021; Somermeyer *et al.* 2022; Moulana *et al.* 2023; Papkou *et al.* 2023; Sundar *et al.* 2024; Zarin and Lehner 2024; Johnston *et al.* 2024; Faure *et al.* 2024b; Escobedo *et al.* 2025; Herrera-Álvarez *et al.* 2025), and genome-wide gene interactions (Bakerlee *et al.* 2022; Nguyen Ba *et al.* 2022; Matsui *et al.* 2022; N'Guessan *et al.* 2025). Despite these advances, contemporary datasets are often noisy and typically do not cover the full sequence space, so that a key challenge is to develop flexible statistical methods for inferring full fitness landscapes from empirical data without distorting the rich fitness landscape geometry revealed by these high-throughput measurements.

One powerful approach to address this problem is to combine theoretical models of fitness landscapes with empirical measurements by recasting these theoretical models as Bayesian priors for reconstructing complete landscapes from incomplete and noisy data (Zhou and McCandlish 2020; Chen *et al.* 2021; Zhou *et al.* 2022, 2025; Petti *et al.* 2025; Martí-Gómez *et al.* 2026b). Such an approach can leverage our mathematical understanding of these models to define prior distributions that confer the overall inference procedure with desirable properties. For example, Minimum Epistasis Interpolation defines a prior that depends on the average squared epistatic coefficients between all pairs of mutations (Zhou and McCandlish 2020), favoring reconstructions that are locally approximately additive, which results in reconstructions that can capture genetic interactions of all orders where data is abundant but extrapolates additively far from the data. As another example, Empirical Variance Component regression (Zhou *et al.* 2022) constructs a prior parametrized by the variance explained by genetic interactions of each possible order, resulting in reconstructions that accurately reflect how quickly the predictability of mutational effects decay in increasingly distant genetic backgrounds.

While Minimum Epistasis Interpolation and Empirical Variance Component regression can incorporate epistatic interactions of all orders, the corresponding priors are still only weakly informative in the sense that they are "isotropic" (Stadler 1996, 2002), i.e. the prior treats all sites and all mutations equally. However, in reality some sites and alleles are more influential and more likely to be involved in epistatic interactions than others (Weinreich *et al.* 2005; Kvitek and Sherlock 2011; Ferretti *et al.* 2016; Bank *et al.* 2016; Pokusaeva *et al.* 2019; Reddy and Desai 2021). Reddy and Desai (2021) recently proposed a new family of theoretical random field models whose parameters control the site-specific probability that mutations participate in genetic interactions, and these models were then extended by Zhou *et al.* (2025) to include allele-specific and mutation-specific parameters. By treating these site-, allele-, or mutation-specific parameters as hyperparameters of an informative Bayesian prior, the resulting model learns which mutations most strongly influence the predictability of other mutations, and incorporates this information when inferring the fitness landscape from data, achieving state-of-the-art predictive performance (Zhou *et al.* 2025). Nonetheless, these models still implicitly assume that the the propensity for a set of sites to interact is determined solely by these site-specific parameters, while empirical observations from both pairwise interaction models (Marks *et al.* 2012; Haldane *et al.* 2016, 2018) and the posterior distributions of models containing interactions of all orders (Chen *et al.* 2021; Martí-Gómez *et al.* 2026b) suggest that patterns of epistatic interaction are often sparse (Poelwijk *et al.* 2019) or modular (Rojas Echenique *et al.* 2019; Hwang *et al.* 2018).

Here, we present a method for fitness landscape inference incorporating a prior that can encode this type of highly structured tendency for specific sets of sites to interact with each other. We begin by proposing a simple approach to summarize the structure of genetic interactions of all orders by computing, for each pair of sites, the average squared local epistatic coefficient between mutation at those sites, which enables the identification of sets of sites involved in epistatic interactions of arbitrary order in complete combinatorial fitness landscapes. Based on these summary statistics, we define a new family of prior distributions that differentially penalize interactions involving different pairs of sites. The parameters of these priors can be estimated from empirical correlations in fitness values between sequences that differ at specific subsets of sites, and we then use the resulting priors to infer complete fitness landscapes from several empirical datasets. Finally, we apply our method to the fitness landscape of a dynamic self-splicing intron (Soo *et al.* 2021), and show how the higher-order interactions in this system have an interpretable structure wherein many aspects of the genetic architecture are systematically rewired depending on the nucleotide identities at one specific pair of sites.

Results

Epistatic coefficients for subsets of sites

In this section, our aim is to quantify the overall amount of epistasis between a specific pair of sites in an arbitrary fitness landscape f. We start by considering a simple two-locus biallelic fitness landscape and assuming that we know the fitness values of the four possible genotypes f_{AB}, f_{Ab}, f_{aB} and f_{ab}. We can measure how much the effect of a mutation $A \to a$ in one locus changes in the presence of an additional mutation $B \to b$ at the other locus via the traditional double-mutant epistatic coefficient:

$$\epsilon = (f_{Ab} - f_{ab}) - (f_{AB} - f_{aB}), \tag{1}$$

and we can use the squared epistatic coefficient ϵ^2 to quantify the magnitude of epistasis in a manner that is independent of which

physical allele is encoded as e.g."A" or "a". As we add a third locus, the epistatic coefficient between mutations $A \to a$ and $B \to b$ can be calculated in the presence of alleles C and c at this new locus, where these two different genetic backgrounds can be viewed as defining parallel faces in sequence space (Figure 1). Thus, one way of quantifying the extent of epistasis between sites 1 and 2 is by taking the average of the squared local epistatic coefficients defined on each of these two parallel faces. Similarly, we can quantify epistasis between sites 1 and 3 by averaging the local squared epistatic coefficients found on the corresponding pair of parallel faces, and we can quantify epistasis between sites 2 and 3 by averaging the squared local epistatic coefficients from the final remaining pair of faces (Figure 1).

We can generalize these statistics for larger fitness landscapes composed of ℓ sites with α alleles per site represented by a vector f of length α^ℓ. In particular, epistasis between any pair of sites i, j can be quantified by computing the average squared local epistatic coefficient $\overline{\epsilon^2_{ij}}(f)$, where the average is taken over every possible pair of mutations at those sites and across every possible genetic background in which they can be introduced. It is easy to see that the average of these quantities $\overline{\epsilon^2_{ij}}(f)$ across all possible pairs of sites corresponds to the previously proposed average squared epistatic coefficient $\overline{\epsilon^2}(f)$ describing the overall local smoothness of a fitness landscape (Zhou and McCandlish 2020) (see Supplement). However, separately averaging these squared epistatic coefficients for each pair of sites provides a more granular description of epistasis, showing not only how much epistasis there is, but whether it is equally distributed across different pairs of sites or concentrated within specific subsets of sites.

For the special case of bi-allelic fitness landscapes, Ferretti *et al.* (2016) previously discussed the $\overline{\epsilon^2_{ij}}$ as part of their proposal of the statistic $\gamma_{i\to j}$, which measures the correlation in the effects of mutations at site j before and after mutating site i. Here, we generalize the $\overline{\epsilon^2_{ij}}$ statistics to multi-allelic landscapes and to higher-order epistatic coefficients between any subset of sites U (see Supplement). Moreover, we show the relationship between these statistics and the total variance explained by interactions involving the sites in U of order equal or higher than $|U|$. Specifically, let $\mathrm{Var}^{(U)}[f]$ be the total variance explained by $|U|$-way interactions among the sites in U as well as interactions of order greater than $|U|$ that also involve all the sites in U (Reddy and Desai 2021; Martí-Gómez *et al.* 2026b). In the supplement, we show that this variance $\mathrm{Var}^{(U)}[f]$ is proportional to the average squared local $|U|$-way epistatic coefficient $\overline{\epsilon^2_U}$:

$$\mathrm{Var}^{(U)}[f] = \left(\frac{\alpha^\ell}{2^{|U|}}\left(\frac{\alpha-1}{\alpha}\right)^{|U|}\right)\overline{\epsilon^2_U}. \tag{2}$$

In addition, we show that the portion of this variance explained by genetic interactions of order strictly higher than $|U|$ is proportional to the variance across backgrounds in $|U|$-way local epistatic coefficients between mutations at sites U:

$$\mathrm{Var}^{(U)}_{k>|U|}[f] = \left(\frac{\alpha^\ell}{2^{|U|}}\left(\frac{\alpha-1}{\alpha}\right)^{|U|}\right)\mathrm{Var}[\epsilon_U]. \tag{3}$$

These results provide a quantitative link between the size and variance of local $|U|$-way epistatic coefficients involving a set of sites U and the amount of $|U|$-way and higher epistatic variance explained by that set of sites.

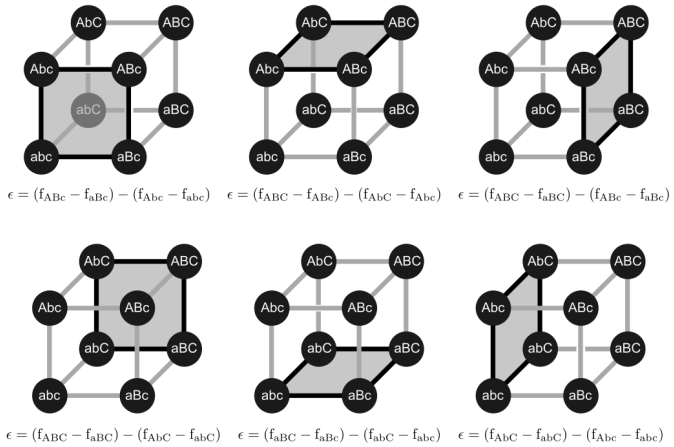

Figure 1 Local epistatic coefficients for mutations at different pairs of sites in a three-site bi-allelic fitness landscape. Local epistatic coefficients can be grouped into sets corresponding to the interaction of the same pair of mutations on different genetic backgrounds. Geometrically this corresponds to grouping epistatic coefficients on parallel sets of "faces"; here we show each set of parallel faces in its own column, with those corresponding to interactions between sites 1 and 2 in the first column, those corresponding to interactions between sites 2 and 3 in the middle column, and those corresponding to interactions between sites 1 and 3 in the last column.

Local epistasis regression

In the previous section, we defined a set of descriptive statistics $\overline{\epsilon^2_{ij}}$ quantifying the magnitude of the local double-mutant epistatic coefficients when introducing mutations at a specific pair of sites i and j. While these statistics can only be evaluated for complete fitness landscapes, we can easily use them to construct a prior distribution for use in inferring complete fitness landscapes from noisy and incomplete data. We call the resulting Bayesian regression method Local Epistasis Regression.

In particular, we consider the prior

$$\mathrm{p}(f) \propto e^{-\sum_{i<j} a_{ij}\overline{\epsilon^2_{ij}}(f)}, \tag{4}$$

where the $\binom{l}{2}$ hyperparameters $a_{ij} > 0$ penalize the size of the local epistatic coefficients between each specific pair of sites i and j. This prior turns out to be an improper Gaussian prior that penalizes the size of local genetic interactions as specified by the a_{ij} but which places a flat prior over the non-epistatic component of the landscape.

In order to better understand the properties of this prior, and in particular why this prior allows epistatic interactions of all orders, it is helpful to re-express this new prior within a more general class of priors previously described by Zhou *et al.* (2022) and Martí-Gómez *et al.* (2026b). This more general prior is a Gaussian random field model parametrized by the variance explained by epistatic interactions between each possible subset of sites U. The prior as a whole is given by

$$f \sim \mathcal{N}\left(0, \sum_U \lambda_U P_U\right), \tag{5}$$

where P_U is the projection matrix into the subspace of landscapes composed solely of $|U|$-way interactions between the sites U

and we constrain the 2^ℓ hyperparameters λ_U to be non-negative, $\lambda_U \geq 0$. This is the most general model in which the covariance between two sequences x, x' only depends on the set of sites at which they differ, and the expected variance explained by $|U|$-way genetic interactions between a set of sites U is given by $(\alpha - 1)^{|U|}\lambda_U$ (see Supplement).

Turning back to our prior in terms of the a_{ij} and $\overline{\epsilon^2_{ij}}(f)$ given in Equation 4, in the Supplement we show that this prior is equivalent to choosing

$$\lambda_U = \begin{cases} \tilde{\lambda}_\emptyset & \text{if } U = \emptyset \\ \tilde{\lambda}_i & \text{if } U = \{i\} \\ \frac{1}{\alpha^2 \sum_{i<j \in U} a_{ij}} & \text{otherwise,} \end{cases} \tag{6}$$

in the limit where $\tilde{\lambda}_\emptyset = \tilde{\lambda}_1 = \ldots = \tilde{\lambda}_\ell \to \infty$. Since we assume $a_{ij} > 0$, we see that $\lambda_U > 0$ for $|U| \geq 2$ indicating that our prior has positive variance for every U with $|U| \geq 2$ and can thus fit genetic interactions of all orders. In practice, instead of taking the limit and working with an improper prior, we retain $\tilde{\lambda}_\emptyset$ and $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_\ell$ as additional hyper-parameters that provide control over the extent to which the non-epistatic component of the model is regularized. Finally, although a_{ij} penalizes the size of $\overline{\epsilon^2_{ij}}(f)$, the expected value of $\overline{\epsilon^2_{ij}}(f)$ under the prior depends not only on a_{ij}, but also on the a_{ik} and a_{jk} for all the $\ell - 2$ other sites k (see Supplement).

In order to perform inference under this prior, we further assume that the experimental measurements y for a series of sequences X have normally distributed errors with known experimental variance y_{var} so that we can use standard Gaussian process results (Rasmussen and Williams 2008) to compute the closed form Gaussian posterior distribution over the complete fitness landscape f with mean

$$\widehat{f} = K_{*X}(K_{XX} + D_{var})^{-1} y \tag{7}$$

and covariance matrix

$$K - K_{*X}(K_{XX} + D_{var})^{-1} K_{X*}, \tag{8}$$

where K_{XX}, K_{*X}, K_{X*} are submatrices of $K = \sum_U \lambda_U P_U$ indexed by sequences X and $*$, where $*$ represents all possible sequences, and D_{var} is a diagonal matrix with the known experimental variances y_{var} along the diagonal. Naively evaluating and computing with these matrices becomes quickly impractical beyond a few thousand measurements. However, we can leverage the mathematical properties of the covariance matrix to derive efficient routines for computing matrix-vector products without explicitly constructing the matrices, which can be implemented as linear operators and used for solving large linear systems using iterative methods in *gpmap-tools* (see Supplement) (Martí-Gómez *et al.* 2026b).

The final remaining issue is how to choose the values for the $1 + \ell + \binom{\ell}{2}$ hyperparameters (specifically, $\tilde{\lambda}_\emptyset, \tilde{\lambda}_1, \ldots, \tilde{\lambda}_\ell$ and the $\binom{\ell}{2}$ a_{ij}). In principle, these hyperparameters can be specified based on prior knowledge, for example if certain sites are known to interact more strongly with each other such as contacting residues within a known structure. Here, however, we adopt an empirical Bayes approach and assume that the statistical properties of the measured sequences generalize to the full sequence space, allowing us to estimate these parameters directly from the data without needing any a priori knowledge about the structure of genetic interactions between sites. Practically speaking, because the covariance in fitness between two sequences under the prior depends only on the set of sites D at which they differ, we can compute the empirical correlation in measured fitness between all pairs of sequences differing exactly at a set of sites D, for each possible subset D. We then choose the hyperparameters so that the correlations implied by the prior match these empirical correlations as closely as possible, using a procedure known as kernel alignment (Wang *et al.* 2015; Zhou *et al.* 2022). This structure reduces the kernel alignment problem to a 2^ℓ-dimensional weighted least squares problem, which can be solved efficiently under the constraint that the a_{ij}, the $\tilde{\lambda}_i$ and $\tilde{\lambda}_\emptyset$ are all positive (see Supplement).

Validation on simulated data

We first evaluate the performance of Local Epistasis Regression using simulated data to illustrate the advantages of this model in comparison with previously proposed approaches. Specifically, we define a random field model inspired by the expected interactions between 8 positions forming an RNA helix, where sites interact more strongly with their neighboring positions as well as with the sites with which they form base-pairs (Figure 2A). Figure 2B shows the correlation under this prior between pairs of sequences differing at every possible subset of sites D within each Hamming distance class and where lines represent distance classes that differ from each other at a single position e.g. $D_1 = \{1, 2\}$ and $D_2 = \{1, 2, 3\}$. We see that while the correlation under the prior generally decays with increasing Hamming distance $d = |D|$, the correlation is not a strict function of d and instead varies based on the specific positions at which two sequences differ.

Having specified a family of random fitness landscapes, we then drew a specific fitness landscape from this prior distribution and evaluated the performance of the hyperparameter estimation procedure in recovering the ground truth hyperparameters in the presence of limited amount of data. In particular, we kept fitness values for only 15% of the sequences, computed the empirical fitness correlation between pairs of sequences for each class of differences D and estimated the hyperparameters for Local Epistasis Regression using kernel alignment. Figure 2C shows that we can accurately infer a prior that approximates the true correlation structure of the landscape using only a limited amount of data ($R^2 = 0.94$).

Since our hyperparameter estimation procedure appeared to be working, we then evaluated the performance of the full Local Epistasis Regression procedure for different random samples of training data ranging from 1% to 99% of the full landscape and compare it with the predictive power of previously proposed models, including classical models like additive and pairwise interaction models, and models that allow interactions of all possible orders, including Minimum Epistasis Interpolation (Zhou and McCandlish 2020), Empirical Variance Component regression (Zhou *et al.* 2022) and a new implementation of Connectedness regression (Zhou *et al.* 2025) in *gpmap-tools* (Martí-Gómez *et al.* 2026b) that uses kernel alignment for hyperparameter inference (Figure 2D, see Supplement). Our results show that all models that allow higher order interactions can accurately reconstruct the true landscape given sufficient amount of data, unlike additive and pairwise interaction models. Moreover, by encoding information about which sites interact with each other and at which sites mutations combine more additively, Local Epistasis Regression outperformed the other methods for any amount of

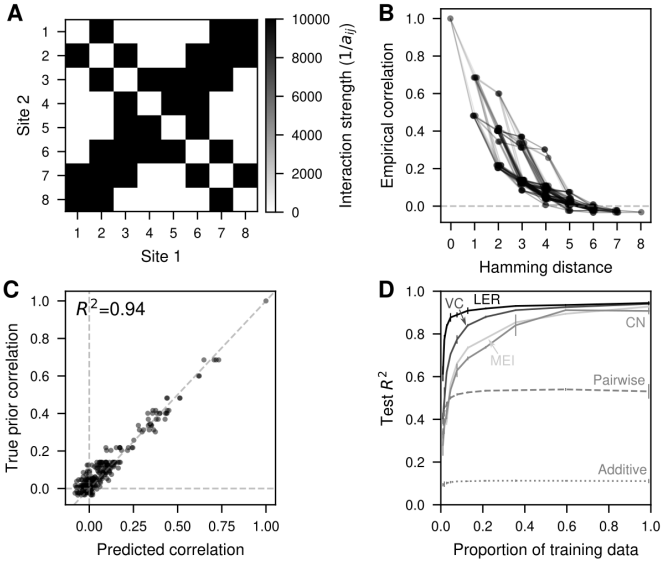

Figure 2 Validation of Local Epistasis Regression. (A) Heatmap representing the model hyperparameters as $1/a_{ij}$ for every pair of sites i, j highlighting the patterns of genetic interactions across sites under the prior. (B) Correlation under the prior for pairs of sequences that differ at each possible subset of sites D arranged according to the Hamming distance $d = |D|$. Each dot represents a single distance class D and are joined by lines whenever the distance classes differ by a single position from each other. (C) Comparison of the true correlation under the prior and the estimated ones using kernel alignment on a simulated dataset comprising 15% of the sequences from a fitness landscape drawn from the prior. (D) Predictive performance evaluated by the R^2 between the predicted and the true fitness values of held-out test sequences when using different amounts of training data for different models (MEI: Minimum Epistasis Interpolation, VC: Variance Component regression, CN: Connectedness Model regression, LER: Local Epistasis Regression). Predicted values are the maximum a posteriori estimate given by each method, which is equal to the posterior mean $\widehat{f}$. Error bars represent the standard deviation across 3 different random samples for each fraction of training data.

training data, and exhibited a particularly strong advantage for low amounts of training data.

Learning the structure of genetic interactions from empirical data

In the previous section, we have shown how Local Epistasis Regression can be used to learn the structure of genetic interactions from incomplete and noisy data, and then used that information to make more accurate inference of complete combinatorial fitness landscapes using simulated data. Here, we investigate the performance of Local Epistasis Regression using data from diverse empirical fitness landscapes.

First, we used data from a high-throughput experiment evaluating the functionality of nearly every possible 5′ splice site sequence of the exon 7 in the Smn1 gene context (Wong *et al.* 2018; Zhou *et al.* 2022) (positions -3 through -1 at the exonic region and +2 through +6 at the intronic region are variable, with position +1 fixed as G, which is necessary for lariat formation during the splicing reaction). We start by computing the correlation between pairs of sequences differing at each possible combination of sites (Figure 3A). This analysis shows that the correlation between pairs of sequences depends, not only on the Hamming distance, but also on the specific combination of sites that are different. For example, pairs of sequences differing only at position +6 show a correlation of 0.83, whereas pairs of sequences that differ only at position +2 show a correlation of 0.12. Next, we estimated the hyperparameters of the prior for our Local Epistasis Regression model and show that the estimated prior can recapitulate the observed correlations almost perfectly (Figure 3B). Figure 3C displays the estimated strength of the penalization for local epistatic coefficients for mutations between each possible pair of sites, showing a rich structure of interactions between positions. First, we find that mutations interact more strongly with mutations at neighboring positions, as expected from its recognition mechanism by the U1 snRNA via base-pair complementarity (Wong *et al.* 2018) because the energetic contributions of a basepair to the thermodynamic stability of an RNA helix depend strongly on the adjacent basepairs (Borer *et al.* 1974). There are some exceptions to this general interaction pattern between positions: i) mutations at positions +2 and +5 tend to interact more strongly with each other than expected under the neighbor interaction model, as previously shown in the 5′splice site fitness landscape inferred from their frequencies in the human genome (Chen *et al.* 2021), and ii) the effects of mutations at position +6 are expected to combine more additively with the effects of other mutations. Interestingly, while mutations at position +2 show strong interactions with mutations at −1 and +5, mutations at −1 and +5 are expected to combine nearly additively, a pattern of interaction that cannot be captured by simpler random field models such as the Connectedness Model (Reddy and Desai 2021; Zhou *et al.* 2025). Finally, we evaluated the performance of Local Epistasis Regression when predicting the fitness of held out sequences for different amounts of training data (Figure 3D) and found that, given sufficient training data, predictions are more accurate than under all previously proposed models.

Second, we used data from a high-throughput experiment measuring the translational activity of over 250,000 9-nucleotide sequences at the Shine-Dalgarno sequence in the 5′UTR of the dmsC gene in *E. coli* (Kuo *et al.* 2020). As before, the correlation in the measurements between pairs of sequence depends not only on the Hamming distance between them, but also on the specific combination of sites at which they differ (Figure 3E). These empirical correlations can be accurately captured by the Local Epistasis Regression prior under the estimated hyperparameters (Figure 3F). The inferred hyperpameters recapitulate the previously characterized pattern of genetic interactions between mutations at different pairs of sites for this landscape, where sites interact more strongly with other sites within a 4-nucleotide window (Figure 3G) (Martí-Gómez *et al.* 2026b). While the Shine-Dalgarno sequence is also recognized via base-pair complementarity, in this case with the 16S rRNA (Shine and Dalgarno 1975), the observed pattern of interaction between positions can be explained by the ability of the 16S rRNA to bind the target sequence at different registers relative to the start codon (Martí-Gómez *et al.* 2026b). In contrast to results in the Smn1 dataset, Local Epistasis Regression is roughly tied as the best performing model together with Variance Component Regression, although the Variance Component Regression model performs better for very small amounts of training data

(Figure 3H).

Finally, we applied Local Epistasis Regression to estimate the structure of genetic interactions across sites in two protein fitness landscapes: (i) a complete combinatorial dataset in which nearly all possible combinations of amino acids at four positions in the binding domain of protein G were measured (GB1) (Wu *et al.* 2016), and (ii) a combinatorial landscape in which combinations of five amino acids at seven positions in the core of the SH3 domain of the FYN tyrosine kinase were quantified (FYN-SH3) (Escobedo *et al.* 2025). These datasets also show heterogeneous fitness correlations between pairs of sequences depending on the specific combination of sites at which they differ, which can be effectively captured by our estimated prior distribution (Figure S1A,B,D,E). The inferred parameters suggest that epistatic interactions are concentrated within specific subsets of sites (Figure S1C,F), consistent with patterns observed in the empirical RNA landscapes described above. For GB1, Local Epistasis Regression and Variance Component Regression are the best performing models and perform essentially identically over the whole range of training data sizes (Figure S1D), whereas for the SH3 domain Variance Component Regression performed better across the whole range of training data set sizes (Figure S1H). Thus, in summary, our analysis shows that empirical fitness landscapes often display heterogeneous correlations between pairs of sequences differing by the same number of mutations depending on the specific sites that are mutated, however, the extent to which incorporating this information into our prior distribution improves predictive performance varies across datasets relative to previous methods.

The fitness landscape of a self-splicing intron

After validating the ability of Local Epistasis Regression to characterize the patterns of genetic interactions between pairs of sites in well-characterized fitness landscapes, we set out to study an 8-nucleotide fitness landscape of a self-splicing type I intron from *Tetrahymena thermophila* (Soo *et al.* 2021) that we have not previously analyzed. In particular, these 8 mutagenized nucleotides can form an extension of the P1 helix that is involved in recognition of the 5′splice site (Figure 4A). This P1 helix must then dissociate to form an alternative P10 helix with the 3′ splice site for catalyzing the splicing reaction between the 5′ and 3′ splice sites (Figure 4A). Thus, we expect a particularly rich structure of genetic interactions across sites, given the heterogeneous structural roles played by these nucleotides at different stages of the splicing reaction.

We start again by computing the correlation in the measured fitness values for pairs of sequences differing at each possible combination of sites. The correlations, as before, depend not only on the number of sites at which two sequence differ, but more specifically on the combination of sites at which they do so, suggesting that different mutations have different effects on the predictability of other mutations (Figure 4B). We next estimate the parameters of the Local Epistasis Regression prior and find that the learned prior can again accurately recapitulate the observed correlations in the data (Figure 4C). The estimated a_{ij} values indicate that epistatic coefficients are not identically distributed across pairs of mutations at different combinations of sites, but that pairs of mutations at different sites tend to interact more strongly with each other (Figure 4D). The strongest signal comes from mutations at sites 2 and 21 which are by far the most strongly interacting positions, consistent with the P1 helix extension mechanism wherein position 2 basepairs with position 21 (Figure 4A).

Next, we perform inference of the complete fitness landscape under the learned prior while taking into account the estimated experimental error. These estimates accurately recapitulate the measured fitness in 0.5% (327) random held-out sequences (Figure 4E). Importantly, our Gaussian process formulation allows us to obtain uncertainty estimates for the fitness values of the held-out sequences showing good coverage properties, as the 95% posterior credible interval included the measured fitness for 91.1% of the held-out sequences. We found that all the regression models allowing higher-order interactions made similar predictions and had similar predictive performance (Figure S2A), and that in particular Local Epistasis Regression and Variance Component Regression both produced very similar fits (Figure S2B) and predictions for the set of held-out sequences (Figure S2C).

Next, we further explore the structure of epistasis in this dataset by computing a series of informative summary statistics from the maximum a posteriori (MAP) reconstruction of the complete landscape. First, we compute the percentage of variance explained by epistatic interactions of every possible order (Zhou *et al.* 2022) and find that the reconstructed landscape is highly epistatic, with only 42.3% of the variance explained by the additive component, 22.2% by pairwise interactions, and the remaining 35.4% explained by higher-order genetic interactions. To investigate how much each site contributes to interactions of different orders, we compute the percentage of variance explained by genetic interactions of order k that involve each site i (Figure 4F) (Martí-Gómez *et al.* 2026b). These statistics reveal substantial heterogeneity across sites. Some sites contribute little across all interaction orders, such as sites 4 and 18. Other sites contribute more strongly through their additive effects, such as position 5, whereas sites like 2 and 21 contribute primarily through pairwise and higher-order interactions. Next, we characterize the structure of genetic interactions between sites by computing the root mean square local double-mutant epistatic coefficient $\sqrt{\overline{\epsilon_{ij}^2}}$ for every pair of sites (Figure 4G). These statistics reveal widespread epistatic interactions across pairs of sites, particularly among sites $2, 3, 20$ and 21, and within the groups of positions 2–5 and 18–21, consistent with the regularization parameters (Figure 4D). However, as shown by Eq. 2, these quantities reflect contributions from both pairwise and higher-order interactions. To disentangle these contributions, we compute the percentage of pairwise ($\mathrm{Var}_{k=2}^{(\{i,j\})}[\hat{f}] = \mathrm{Var}[P_{\{i,j\}}\hat{f}]$) and higher-order variance ($\mathrm{Var}_{k>2}^{(\{i,j\})}[\hat{f}]$) explained by interactions involving each pair of sites (Figure 4H). This analysis shows that a large fraction of pairwise epistatic variance is explained by interactions between positions 2 and 21 alone (Figure 4A, lower triangular part), whereas higher-order epistatic variance is more broadly distributed across sites, indicating that local double mutant epistatic coefficients for essentially all pairs of sites vary substantially across genetic backgrounds (Eq.3).

Structure of the self-spliced intron fitness landscape

In order to better understand the qualitative structure of this fitness landscape, we apply a visualization method for producing low-dimensional representations of fitness landscapes where distances between genotypes in the representation reflect the expected waiting time to evolve from one genotype to another under a model of molecular evolution that includes mutation, selection, and drift (McCandlish 2011). Applying this method as implemented in the software package *gpmap-tools* resulted

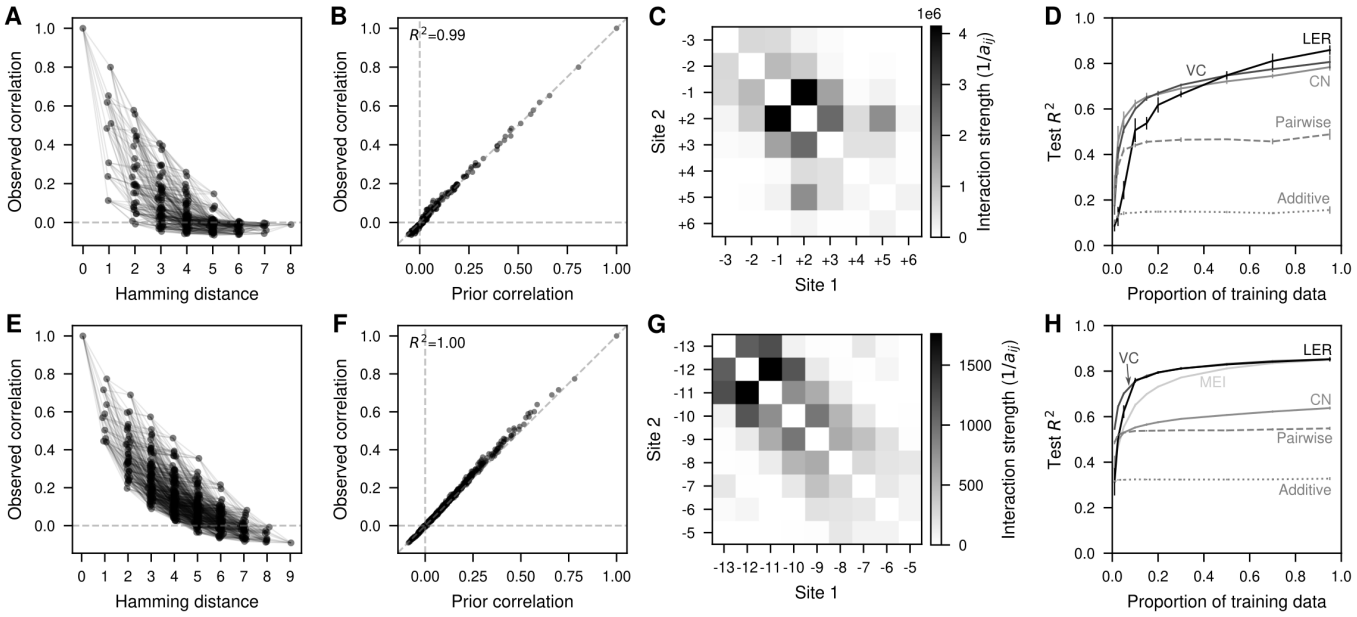

Figure 3 Application of Local Epistasis Regression to estimate the structure of epistatic interactions across sites in empirical fitness landscapes. (A,E) Correlation in the measured fitness values for pairs of sequences differing at each possible subset of sites D arranged according to the Hamming distance $d = |D|$. Each dot represents a single distance class D and are joined by lines whenever the distance classes differ by a single position from each other. (B,F) Comparison of the observed correlation values in the data and the values under the estimated prior ones using Local Epistasis Regression for every possible distance class D (each dot represents a different D). Correlations were estimated using 80% of the data for training. (C,G) Heatmap representing the inferred model hyperparameters as $1/a_{ij}$ for every pair of sites i, j highlighting the patterns of genetic interactions across sites under the prior. (D,H) Predictive performance evaluated by the R^2 between the predicted and the measured fitness of held-out test sequences when using different amounts of training data for different models (MEI: Minimum Epistasis Interpolation, VC: Variance Component regression, CN: Connectedness Model regression, LER: Local Epistasis Regression). Predicted values are the maximum a posteriori estimate given by each method, which is equal to the posterior mean $\widehat{f}$. Error bars represent the standard deviation across 3 different random samples for each fraction of training data. Each row represents a fitness landscape: Smn1 exon 7 5′ splice site (A,B,C,D); dmsC Shine-Dalgarno sequence (E,F,G,H).

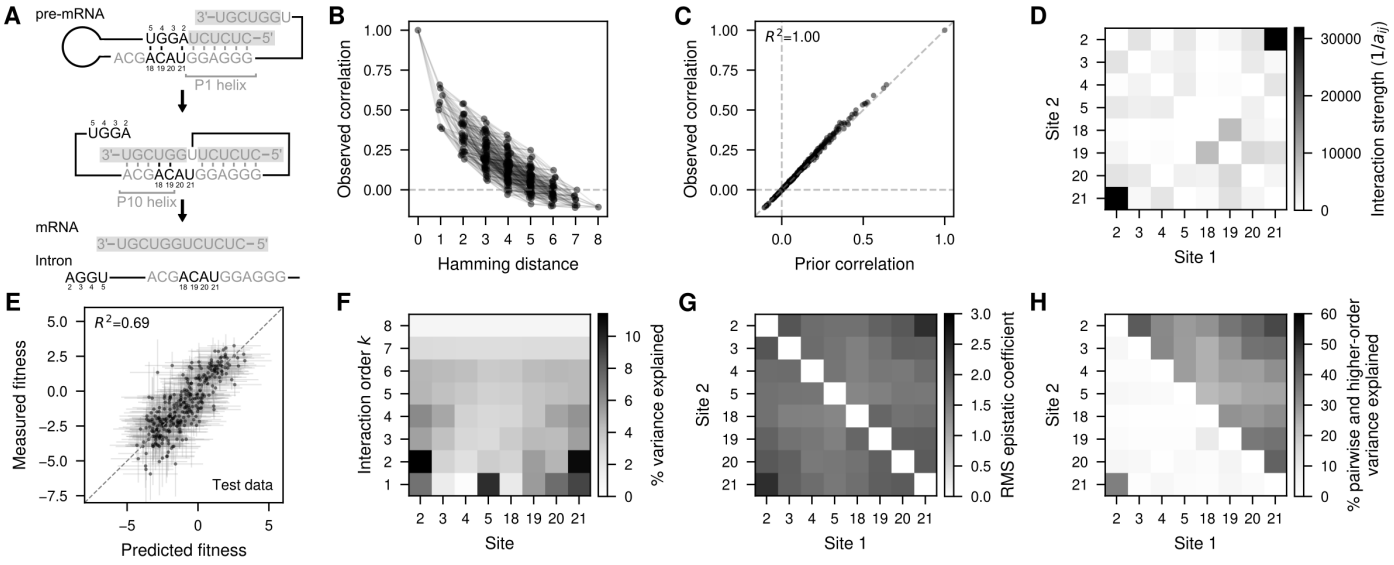

Figure 4 Inference and summary statistics of the fitness landscape of a self-spliced intron. (A) Schematic representation of the molecular mechanism of self-splicing in the model intron from *Tetrahymena thermophila* (Soo *et al.* 2021). Positions considered in the fitness landscape are highlighted in black and are numbered by their relative position in the intron. Other relevant sequences that are fixed in the background are shown in grey. Exonic sequences are shown with gray background. (B) Correlation in the measured fitness values for pairs of sequences differing at each possible subset of sites D arranged according to the Hamming distance $d = |D|$. Each dot represents a single distance class D and are joined by lines whenever the distance classes differ by a single position from each other. (C) Comparison of the observed correlation values in the data and the values under the estimated prior using Local Epistasis Regression. One correlation is shown for each possible set of positions D where two sequences may be differ (each dot represents a different D). (D) Heatmap representing the inferred model hyperparameters as $1/a_{ij}$ for every pair of sites i, j highlighting the patterns of genetic interactions across sites under the prior. (E) Measured values for held-out test sequences versus Local Epistasis Regression predictions. Horizontal error bars represent the 95% credible interval, whereas vertical error bars correspond to the 95% confidence interval under each measurement's variance. (F) Heatmap representing the percentage of variance in the maximum a posteriori Local Epistasis Regression reconstruction explained by interactions of order k involving each position. (G) Root mean squared local double-mutant epistatic coefficient magnitude between mutations at each possible pair of predictions for the maximum a posteriori reconstruction. The plot indicates that relatively large local epistatic coefficients occur between mutations at essentially all pairs of positions. (H) Heatmap representing the percentage of variance in the maximum a posteriori reconstruction explained by pairwise (lower triangle) and higher-order (upper triangle) interactions that is explained by interactions involving pairs of positions.

in Figure 5A (see Methods, Martí-Gómez *et al.* (2026b)). In the previous section we noted a particularly strong pattern of both pairwise and higher-order interactions involving positions 2 and 21 (Figure 4F), and we see that the visualization in Figure 5A largely separates sequences based on the nucleotides present at this pair of positions. Because the axes in such visualizations tend to highlight key barriers that make it difficult for a population to diffuse from one area of sequence space to another, we examined the mean fitness conferred by different combinations of alleles at these two sites for potential allelic incompatibilities (Figure 5B), revealing a consistent pattern wherein having both positions 2 and 21 occupied by pyrimidines (or also C_2A_{21}) results in low mean fitness. This observation suggests that the P1 helix extension by base-pairing of positions 2 and 21 is not strictly necessary for functionality.

Because the visualizations tend to spread apart high-fitness sequences that are separated by these types of incompatibilities, they are also useful for understanding how mutational effects and local epistatic coefficients vary across sequence space. Figures 5D and E show how *G*2*C* and *C*21*G* mutations are deleterious in the G_2C_{21} cluster on the left-hand (negative) side of Diffusion Axis 1, but the same mutations become advantageous as one moves to the right-hand (positive) side. In fact, looking across all 8 positions, Figure S4 shows a widespread tendency for allelic preferences to change in a coherent manner across different clusters of sequences. Similarly, we can see how local double-mutant epistatic coefficients change in different regions of sequence space. For example, in the C_2G_{21} cluster on the right-hand side of Figure 5A, *G*3*C* and *C*20*G* are strongly negatively epistatic (Figure 5F) whereas in many other backgrounds such as U_2A_{21} background (Figure 5C), their interaction is strongly positive. Likewise the local double-mutant epistatic coefficient between *A*3*U* and *U*20*A* are strongly negative in the C_2G_{21} cluster but typically neutral or positive elsewhere (Figure 5G). Note that Figures 5F and 5G provide a simple illustration of the geometry of a 4th-order interaction, as we can see that the epistatic interactions at position 3 and 20 differ across the clusters of sequences defined by the base identities at positions 2 and 21.

The visualization also separates sequences within the C_2G_{21} background into three clusters, depending on the presence of *C* at positions 19 and 20 (Figure 5A). Moreover, Figure S4 shows that while C_{19} and C_{20} are generally preferred across the landscape, consistent with base pairing with cognate bases in the P10 helix in the second step of the splicing reaction (Figure 4A), the effects of alleles *C* and *G* at these two positions are strongly dependent on genetic background. To better understand these dependencies, we examined the average fitness of the four combinations of *C* and *G* alleles at positions 19 and 20 across different genetic contexts (Figure 5H). We find that in some genetic contexts, like C_2G_{21}, U_2A_{21}, and C_2A_{21}, the base-pair–breaking mutations *C*19*G* and *C*20*G* are on average neutral or only mildly deleterious individually, but become substantially more deleterious when combined together. In contrast, in other genetic contexts, such as G_2C_{21}, A_2U_{21}, and A_2C_{21}, these mutations are more strongly deleterious individually and combine more additively. These epistatic interactions limit the accessible mutational paths between $C_{19}G_{20}$ and $G_{19}C_{20}$ sequences, and explain why these clusters appear separated in the visualization (Figure 5A). The background dependence of this interaction provides an additional illustration of how higher-order interactions in this system can be largely understood in terms of the identities at positions 2 and 21, and how the alleles at these two positions modulate epistatic interactions between the other sites.

Discussion

In this work, we propose a new method for inferring empirical fitness landscapes from experimental data. Our method is based on the idea that genetic interactions are not uniformly distributed across sites; rather, some sites tend to interact more strongly with specific other sites. This information about sets of sites exhibiting stronger epistatic interactions can be extracted from how correlations between measured fitness values depend on the specific combinations of sites at which sequences differ. We then incorporate this information into a prior distribution defined over all possible fitness landscapes for Bayesian inference of complete combinatorial landscapes from incomplete and noisy data. Importantly, although our prior allows the incorporation of different degrees of regularization for epistatic coefficients between different pairs of sites, it still allows these coefficients to vary across genetic backgrounds, thereby preserving the benefits of other methods that contain genetic interactions of all orders (Zhou and McCandlish 2020; Zhou *et al.* 2022). Applying this method to three experimental datasets revealed that mutations do tend to interact more strongly with other specific mutations in empirical landscapes (Figures 3C,F, S1C,G and 4D), often reflecting known molecular interactions between positions (Figures 4A,D). Finally, we performed an in-depth analysis of the inferred landscape for a self-splicing intron, where we found that the nucleotide identities at positions 2 and 21, which can potentially form an extension of the P1 helix, determine the nature of genetic interactions at many other pairs of positions. This shows how higher-order interactions, which may at first seem mysterious, can be understood in terms of local pairwise interactions that change coherently as one moves from one region of the fitness landscape to another.

Our method relies on a simple summary statistic that quantifies the overall strength of local genetic interactions between a given pair of mutations across across all genetic backgrounds. Specifically, we compute the average squared epistatic coefficient between mutations at every possible set of sites. While this statistic was previously derived in the context of the $\gamma_{i\to j}$ measure (Ferretti *et al.* 2016), here we show that it can be expressed as a quadratic form with a positive semi-definite matrix that admits a Kronecker product decomposition into ℓ smaller matrices. This allowed us to derive (i) the relationship between the total epistatic variance of all orders involving any pair of sites i and j (Reddy and Desai 2021; Martí-Gómez *et al.* 2026b) and the average squared size of local double-mutant epistatic coefficients defined by pairs of mutations at sites i and j, and (ii) the relationship between the higher order epistatic variance and the variance in these epistatic coefficients across genetic backgrounds, clarifying how local epistatic effects and their context dependence connect to the global variance decomposition of fitness landscapes. Interestingly, while the ability to conduct this Kronecker factorization enables efficient computation of these summary statistics in combinatorially complete fitness landscapes (Martí-Gómez *et al.* 2026b,a), it also opens up the possibility of calculating these statistics for astronomically large fitness landscapes for the maximum a posteriori estimate under a certain class of Gaussian process models (Petti *et al.* 2025), which would allow the incorporation of higher-order interactions in applications such as protein contact prediction (Marks *et al.* 2012) and 3D-structure inference from mutagenesis experiments (Schmiedel and Lehner 2019). Finally, the Kronecker

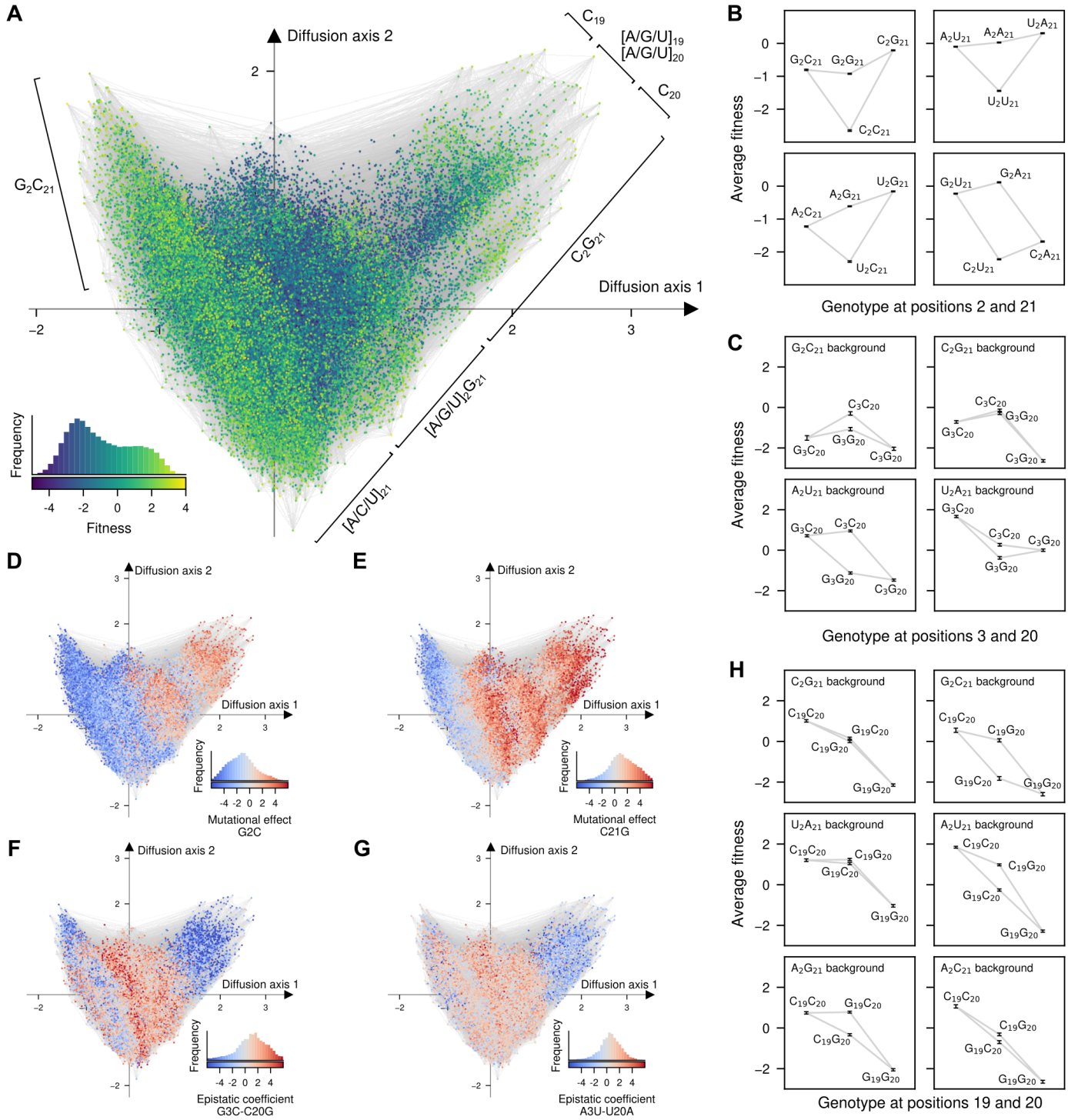

Figure 5 Visualization and characterization of the fitness landscape of a self-spliced intron. (A) Visualization of the inferred fitness landscape using Local Epistasis Regression. Every dot represents one of the possible 4^8 possible sequences and is colored according to the predicted fitness. The inset represents the phenotypic distribution along with their corresponding color in the map. Sequences are laid out according to the first two Diffusion axes and dots are plotted in order according to Diffusion axis 3. (B,C,H) Diagrams representing average fitness for different subsets of sequences at positions 2 and 21 (B), for combinations of G and C alleles at positions 3 and 20 across different genetic contexts at positions 2 and 21 (C) and for combinations of G and C alleles at positions 19 and 20 across different genetic contexts at positions 2 and 21 (H). Indicated subsets of sequences are arranged along the x-axis according to Hamming distance from the leftmost subset. Error bars represent the 95% posterior credible intervals for these average fitness values. (D-G) Visualization of the inferred fitness landscape, as shown in (A), where nodes are colored by the mutational effect of G2C (D) and C21G (E), and the local epistatic coefficients between mutations G3C and C20G (F) and between A3U and U20A (G) when introduced in every possible genetic background throughout the landscape. The inset represents the distribution of the specific mutational effects or epistatic coefficients along with their corresponding color in the map.

factorization form makes it straightforward to generalize these statistics in two directions: (i) computing squared epistatic coefficients between specific mutations within the same pair of sites (see Supplement) and (ii) averaging of the squared coefficients across different genetic backgrounds or over backgrounds drawn from site-factorizable probability distributions (rather than from the uniform distribution as we have done here).

As in previous work (Zhou and McCandlish 2020; Chen *et al.* 2021; Zhou *et al.* 2022; Reddy and Desai 2021; Zhou *et al.* 2025), we use our mathematical understanding of this summary statistic to define random field models of fitness landscapes, which can be used as prior distributions for Gaussian process inference of empirical landscapes from high-throughput experimental data. Our prior allows the same mutation to have different effects on the predictability of different mutations, whereas previous approaches assign all mutations (Zhou and McCandlish 2020; Chen *et al.* 2021) or each individual mutation (Reddy and Desai 2021; Zhou *et al.* 2025) a constant effect on the predictability of other mutations, preventing them from encoding structured interaction patterns between sites. Despite finding widespread evidence that specific pairs of sites tend to interact more than others, incorporating this realism into our prior did not always increase predictive performance. One potential reason for this is that while more flexibly encoding differences between pairs of sites, our model makes assumptions about the form of higher-order epistasis that may be more appropriate for some datasets than others. A second potential reason for this variable performance concerns our kernel alignment procedure for choosing the hyperparameters. We had previously seen a modest improvement in model performance for Empirical Variance Component Regression when the hyperparameters were chosen by evidence maximization rather than kernel alignment (Zhou *et al.* 2025), suggesting that the two modes of inference should generally have similar performance. However, here we see that the connectedness model fit by kernel alignment performed far worse on the GB1 and Smn1 datasets than our previous implementation using evidence maximization (Zhou *et al.* 2025), suggesting that the method of hyperparameter optimization might have a greater impact than previously thought. Unfortunately, unlike kernel alignment, which we were able to implement efficiently using the fact that the covariance under our model only depends on the subset of sites at which two sequences differ, implementing evidence maximization for Local Epistasis Regression as in Zhou *et al.* (2025) is more challenging due to the need to calculate and combine 2^ℓ kernels.

A fundamental aspect of fitness landscapes is their dependence on the environment. Increasingly, datasets measure fitness across multiple environmental conditions, necessitating models that allow mutational effects and epistatic interactions to vary across environments (Nguyen Ba *et al.* 2022; N'Guessan *et al.* 2025; Bakerlee *et al.* 2022; Soo *et al.* 2021; Ishigami *et al.* 2024). Our framework naturally extends to this setting by treating environmental conditions as additional loci, enabling the inference of priors in which mutational effects depend jointly on genetic background and environment. Importantly, our prior can learn and encode that mutations tend to change by different magnitudes when introducing an additional mutation compared to when changing environments, effectively allowing different prior variances for gene-by-gene and gene-by-environment interactions.

Our method shares several limitations with previous Gaussian process approaches (Zhou and McCandlish 2020; Zhou *et al.* 2022, 2025). First, it does not explicitly model non-specific or global epistasis (Otwinowski *et al.* 2018; Domingo *et al.* 2019), but instead relies on specific epistatic interactions to fit these global dependencies, potentially limiting the interpretability of the estimated hyperparameters as the structure of specific genetic interactions between sites. Second, by taking an empirical Bayes approach, in which we first estimate the parameters of the prior and then use those parameters for inference of the fitness landscape, we are not taking into account potential uncertainty in the estimation of these parameters. Our new model may be more sensitive to this limitation than previous approaches because of the larger number of hyperparameters that need to be estimated (Zhou and McCandlish 2020; Zhou *et al.* 2022). Third, our current implementation takes advantage of the mathematical structure of the covariance matrix to allow efficient computation of the posterior distribution (Equations 7 and 8) without explicitly building the covariance matrix for complete fitness landscapes. While this trick allows us to compute the posterior distribution for fitness landscapes containing hundreds of thousands of sequences, it is limited to sequences of relatively short length (about 5 amino acids, 12 nucleotides or 24 bi-allelic loci). Moreover, the need to evaluate 2^ℓ different kernels, even if the number of hyperparameters is much lower, hinders the applicability even under GPU-accelerated frameworks for scalable Gaussian process inference (Gardner *et al.* 2018; Wang *et al.* 2019) that facilitated the application of previous models to fitness landscapes defined over longer sequences (Zhou *et al.* 2025).

Materials and methods

Fitness landscape of a 5′ splice site sequence

Data reported by Wong *et al.* (2018) was processed as previously reported (Zhou *et al.* 2022, 2025). Briefly, we assumed a log-normal distribution of enrichment ratios across 1-7 replicates, for each different 5′ splice site sequence x. The bias corrected geometric mean of the enrichment ratio was used as an estimate of the median enrichment ratio when the enrichment ratio was strictly positive for all replicates. Otherwise, the median of enrichment scores was used to estimate the phenotype y_x. Sequence-specific variance $y_{x,var}$ was estimated as indicated below, where s_x^2 is the sample variance of the log-enrichment ratios if all replicates were strictly positive and were measured in at least two samples or the median of all s_x^2 for sequences x with at least two replicate measurements:

$$y_{x,var}^2 = \left(e^{s_x^2} - 1\right) e^{2y_x + s_x^2}, \tag{9}$$

see (Zhou *et al.* 2022) for more details.

Fitness landscape of the Shine-Dalgarno sequence

Data reported by Kuo *et al.* (2020) was processed as previously reported (Martí-Gómez *et al.* 2026b). Briefly, fitness was estimated as the mean log(GFP) for 257,565 measured sequences with a common measurement variance of $s^2 = 0.058$ using genotypes measured across all three experimental replicates. The squared standard error for each genotype i was computed by dividing this observed experimental variance s^2 by the number of replicates n_x in which each sequence was measured ($y_{x,var} = s^2/n_x$).

Fitness landscape of protein G binding domain

Data was processed as previously described (Zhou and McCandlish 2020; Zhou *et al.* 2022). Briefly, we used the number of

sequencing reads for each sequence x in the input sample (c_x^{input}) and in the selected sample (c_x^{sel}) reported in (Wu *et al.* 2016) to estimate the log-enrichment ratio relative to the wild-type sequence y_x as a measure of the binding strength. Moreover, we estimated the error variance $y_{x,var}$ of this estimate (Rubin *et al.* 2017):

$$y_x = \log\left(\frac{c_x^{sel}+0.5}{c_x^{input}+0.5}\right) - \log\left(\frac{c_{wt}^{sel}+0.5}{c_{wt}^{input}+0.5}\right) \quad (10)$$

$$y_{x,var} = \frac{1}{c_x^{input}+0.5} + \frac{1}{c_x^{sel}+0.5} + \frac{1}{c_{wt}^{input}+0.5} + \frac{1}{c_{wt}^{sel}+0.5}. \quad (11)$$

Fitness landscape of the FYN protein SH3 domain

Data was used as reported by the original study (Escobedo *et al.* 2025). Specifically, we downloaded the processed data from GEO (GSE266299) used the scaled relative fitness measurements and the reported experimental errors for our downstream analysis.

Fitness landscape of a self-splicing intron

The number of reads for each variant across six-replicates in the presence and absence of Kanamycin was collected at 30º from the original study (Soo *et al.* 2021). Following Soo *et al.* (2021), fitness y_x for each sequence x was estimated as the $\log_2$(Fold change) between Kanamycin treated and control samples using PyDESeq2 (Muzellec *et al.* 2023) without shrinkage towards zero, and squared standard errors were kept as measure of experimental variance $y_{x,var}$ for downstream analysis. Local Epistasis Regression was used to estimate the complete combinatorial fitness landscape taking into account the experimental errors for each of the sequence using all available data except 0.5% (327) sequences. We computed the posterior mean and variance for these 327 sequences to evaluate the performance of the model in unobserved sequences. Using the posterior mean for the complete fitness landscape, we computed the variance explained by epistatic interactions of every possible order for each site and between pair of sites using *gpmap-tools* (Martí-Gómez *et al.* 2026b). We then generated a low-dimensional representation in which distances between pairs of sequences reflect the expected time to evolve from one sequence to the other (McCandlish 2011) under an evolutionary model in the weak mutation regime as implemented in *gpmap-tools* (Martí-Gómez *et al.* 2026b). We generated visualization coordinates under different strengths of selection by choosing different values for the expected fitness under long-term mutation-selection-drift (i.e. expected fitness at stationarity, Figure S3). A long-term expected fitness of 1.6 (corresponding to the 87% percentile in the distribution of fitness values) was used for the final visualization.

Data and code availability

The methods presented in this work have been implemented in *gpmap-tools* v0.4.2 (Martí-Gómez *et al.* 2026b), an open-source library available at https://github.com/cmarti/gpmap-tools. Code and data to reproduce the analyses presented in this paper are available at https://github.com/cmarti/deltaU.

Funding

CMG and DMM were supported by the US National Institutes of Health (NIH) grant R35GM133613 and additional funding from the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory. This work was performed with assistance from the NIH Grant S10OD028632.

Conflicts of interest

The authors declare no conflicts of interest.

Literature cited

Agarwal V, Inoue F, Schubach M, Penzar D, Martin BK, Dash PM, Keukeleire P, Zhang Z, Sohota A, Zhao J *et al*. 2025. Massively parallel characterization of transcriptional regulatory elements. Nature. pp. 1–10. Publisher: Nature Publishing Group.

Aguirre L, Hendelman A, Hutton SF, McCandlish DM, Lippman ZB. 2023. Idiosyncratic and dose-dependent epistasis drives variation in tomato fruit size. Science. 382:315–320.

Aita T, Husimi Y. 1998. Fitness Landscape of a Biopolymer Participating in a Multi-step Reaction. Journal of Theoretical Biology. 191:377–390.

Baeza-Centurion P, Miñana B, Schmiedel JM, Valcárcel J, Lehner B. 2019. Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing. Cell. 176:549–563.e23. Publisher: Cell Press.

Bakerlee CW, Nguyen Ba AN, Shulgina Y, Rojas Echenique JI, Desai MM. 2022. Idiosyncratic epistasis leads to global fitness-correlated trends. Science. 376:630–635.

Bank C. 2022. Epistasis and Adaptation on Fitness Landscapes. Annual Review of Ecology, Evolution, and Systematics. 53:457–479. Publisher: Annual Reviews.

Bank C, Matuszewski S, Hietpas RT, Jensen JD. 2016. On the (un)predictability of a large intragenic fitness landscape. Proceedings of the National Academy of Sciences. 113:14085–14090.

Bendixsen DP, Collet J, Østman B, Hayden EJ. 2019. Genotype network intersections promote evolutionary innovation. PLoS Biology. 17. Publisher: Public Library of Science.

Bonde MT, Pedersen M, Klausen MS, Jensen SI, Wulff T, Harrison S, Nielsen AT, Herrgård MJ, Sommer MO. 2016. Predictable tuning of protein expression in bacteria. Nature Methods. 13:233–236.

Borer PN, Dengler B, Tinoco Jr I, Uhlenbeck OC. 1974. Stability of ribonucleic acid double-stranded helices. Journal of molecular biology. 86:843–853.

Bryant DH, Bashir A, Sinai S, Jain NK, Ogden PJ, Riley PF, Church GM, Colwell LJ, Kelsic ED. 2021. Deep diversification of an aav capsid protein by machine learning. Nature Biotechnology. 39:691–696.

Chattopadhyay G, Papkou A, Wagner A. 2025. The fitness landscape of the E.coli lac operator is highly rugged in two different environments. bioRxiv. .

Chen Wc, Zhou J, Sheltzer JM, Kinney JB, Mccandlish DM. 2021. Field-theoretic density estimation for biological sequence space with applications to 5 splice site diversity and aneuploidy in cancer. Proc. Natl. Acad. Sci. USA. .

Chou HH, Chiu HC, Delaney NF, Segrè D, Marx CJ. 2011. Diminishing Returns Epistasis Among Beneficial Mutations Decelerates Adaptation. Science. 332:1190–1192. Publisher: American Association for the Advancement of Science.

Crawford L, Zeng P, Mukherjee S, Zhou X. 2017. Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. PLOS Genetics. 13:e1006869. Publisher: Public Library of Science.

Dasari K, Somarelli JA, Kumar S, Townsend JP. 2021. The somatic molecular evolution of cancer: Mutation, selection, and epistasis. Progress in Biophysics and Molecular Biology. 165:56–65.

de Boer CG, Vaishnav ED, Sadeh R, Abeyta EL, Friedman N, Regev A. 2020. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. Nature Biotechnology. 38:56–65. Publisher: Nature Publishing Group.

De Los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. 2013. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics. 193:327–345.

De Visser JAG, Krug J. 2014. Empirical fitness landscapes and the predictability of evolution. Nature Reviews Genetics. 15:480–490. Publisher: Nature Publishing Group.

Domingo J, Baeza-Centurion P, Lehner B. 2019. The causes and consequences of genetic interactions (epistasis). Annual Review of Genomics and Human Genetics. 20:433–460. Publisher: Annual Reviews Inc.

Domingo J, Diss G, Lehner B. 2018. Pairwise and higher-order genetic interactions during the evolution of a tRNA. Nature. 558:117–121.

Dwivedi SL, Heslop-Harrison P, Amas J, Ortiz R, Edwards D. 2024. Epistasis and pleiotropy-induced variation for plant breeding. Plant Biotechnology Journal. 22:2788–2807. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/pbi.14405.

Escobedo A, Voigt G, Faure AJ, Lehner B. 2025. Genetics, energetics, and allostery in proteins with randomized cores and surfaces. Science. 389:eadq3948. Publisher: American Association for the Advancement of Science.

Evfratov SA, Osterman IA, Komarova ES, Pogorelskaya AM, Rubtsova MP, Zatsepin TS, Semashko TA, Kostryukova ES, Mironov AA, Burnaev E *et al*. 2017. Application of sorting and next generation sequencing to study 5'-UTR influence on translation efficiency in Escherichia coli. Nucleic Acids Research. 45:3487–3502.

Faure AJ, Lehner B, Miró Pina V, Serrano Colome C, Weghorn D. 2024a. An extension of the walsh-hadamard transform to calculate and model epistasis in genetic landscapes of arbitrary shape and complexity. PLoS Computational Biology. 20:e1012132.

Faure AJ, Martí-Aranda A, Hidalgo-Carcedo C, Beltran A, Schmiedel JM, Lehner B. 2024b. The genetic architecture of protein stability. Nature. 634:995–1003. Publisher: Nature Publishing Group.

Ferretti L, Schmiegelt B, Weinreich D, Yamauchi A, Kobayashi Y, Tajima F, Achaz G. 2016. Measuring epistasis in fitness landscapes: The correlation of fitness effects of mutations. Journal of Theoretical Biology. 396:132–143.

Ferretti L, Weinreich D, Tajima F, Achaz G. 2018. Evolutionary constraints in fitness landscapes. Heredity. 121:466–481. Publisher: Nature Publishing Group.

Flynn KM, Cooper TF, Moore FB, Cooper VS. 2013. The Environment Affects Epistatic Interactions to Alter the Topology of an Empirical Fitness Landscape. PLoS Genetics. 9:1003426.

Fowler DM, Fields S. 2014. Deep mutational scanning: a new style of protein science. Nature Methods. 11:801–807. Publisher: Nature Publishing Group.

Fragata I, Blanckaert A, Dias Louro MA, Liberles DA, Bank C. 2019. Evolution in the light of fitness landscape theory. Trends in Ecology and Evolution. 34:69–82. Publisher: Elsevier Ltd.

Gao Y, Lin KT, Jiang T, Yang Y, Rahman M, Gong S, Bai J, Wang L, Sun J, Sheng L *et al*. 2022. Systematic characterization of short intronic splicing-regulatory elements in *SMN2* pre-mRNA. Nucleic Acids Research. 50:731–749.

Gardner J, Pleiss G, Weinberger KQ, Bindel D, Wilson AG. 2018. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. Advances in neural information processing systems. 31.

Gavrilets S. 2004. *Fitness Landscapes and the Origin of Species*. Princeton University Press.

Haldane A, Flynn WF, He P, Levy RM. 2018. Coevolutionary Landscape of Kinase Family Proteins: Sequence Probabilities and Functional Motifs. Biophysical Journal. 114:21–31.

Haldane A, Flynn WF, He P, Vijayan R, Levy RM. 2016. Structural propensities of kinase family proteins from a Potts model of residue co-variation. Protein Science : A Publication of the Protein Society. 25:1378–1384.

Haseman JK, Elston RC. 1972. The investigation of linkage between a quantitative trait and a marker locus. Behavior Genetics. 2:3–19.

Herrera-Álvarez S, Patton JEJ, Thornton JW. 2025. Ancient biases in phenotype production drove the functional evolution of a protein family.

Hwang S, Schmiegelt B, Ferretti L, Krug J. 2018. Universality Classes of Interaction Structures for NK Fitness Landscapes. Journal of Statistical Physics. 172:226–278. arXiv: 1708.06556 Publisher: Springer US.

Ishigami Y, Wong MS, Martí-Gómez C, Ayaz A, Kooshkbaghi M, Hanson SM, McCandlish DM, Krainer AR, Kinney JB. 2024. Specificity, synergy, and mechanisms of splice-modifying drugs. Nature Communications. 15:1880.

Jalal AS, Tran NT, Stevenson CE, Chan EW, Lo R, Tan X, Noy A, Lawson DM, Le TB. 2020. Diversification of DNA-Binding Specificity by Permissive and Specificity-Switching Mutations in the ParB/Noc Protein Family. Cell Reports. 32:107928.

Johnson MS, Reddy G, Desai MM. 2023. Epistasis and evolution: recent advances and an outlook for prediction. BMC Biology. 21:120.

Johnston KE, Almhjell PJ, Watkins-Dulaney EJ, Liu G, Porter NJ, Yang J, Arnold FH. 2024. A combinatorially complete epistatic fitness landscape in an enzyme active site. Proceedings of the National Academy of Sciences of the United States of America. 121:e2400439121.

Kauffman S, Levin S. 1987. Towards a general theory of adaptive walks on rugged landscapes. Journal of Theoretical Biology. 128:11–45.

Kauffman SA, Weinberger ED. 1989. The NK model of rugged fitness landscapes and its application to maturation of the immune response. Journal of Theoretical Biology. 141:211–245.

Khan AI, Dinh DM, Schneider D, Lenski RE, Cooper TF. 2011. Negative epistasis between beneficial mutations in an evolving bacterial population. Science. 332:1193–1196.

Kingman JFC. 1978. A simple model for the balance between selection and mutation. Journal of Applied Probability. 15:1–12.

Kinney JB, McCandlish DM. 2019. Massively Parallel Assays and Quantitative Sequence–Function Relationships. Annual Review of Genomics and Human Genetics. 20:annurev–genom–083118–014845.

Kinney JB, Murugan A, Callan CG, Cox EC. 2010. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proceedings of the National Academy of Sciences of the United States of America. 107:9158–9163.

Komarova ES, Chervontseva ZS, Osterman IA, Evfratov SA, Rubtsova MP, Zatsepin TS, Semashko TA, Kostryukova ES, Bogdanov AA, Gelfand MS *et al*. 2020. Influence of the spacer region between the Shine–Dalgarno box and the start codon for fine-tuning of the translation efficiency in Escherichia coli. Microbial Biotechnology. 13:1254–1261.

Kondrashov AS, Sunyaev S, Kondrashov FA. 2002. Dobzhansky-Muller incompatibilities in protein evolution. Proceedings of the National Academy of Sciences of the United States of America. 99:14878–14883.

Kuo ST, Chang JK, Chang C, Shen WY, Hsu C, Lai SW, Chou HHD. 2025. Unraveling the start element and regulatory divergence of core promoters across the domain Bacteria.

Kuo ST, Jahn RL, Cheng YJ, Chen YL, Lee YJ, Hollfelder F, Wen JD, Chou HHD. 2020. Global fitness landscapes of the Shine-Dalgarno sequence. Genome Research. 30:711–723.

Kvitek DJ, Sherlock G. 2011. Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. PLoS genetics. 7:e1002056.

Liao SE, Sudarshan M, Regev O. 2023. Deciphering RNA splicing logic with interpretable machine learning. Proceedings of the National Academy of Sciences. 120:e2221165120.

Lite TLV, Grant RA, Nocedal I, Littlehale ML, Guo MS, Laub MT. 2020. Uncovering the basis of protein-protein interaction specificity with a combinatorially complete library. eLife. 9:1–57.

Manrubia S, Cuesta JA, Aguirre J, Ahnert SE, Altenberg L, Cano AV, Catalán P, Diaz-Uriarte R, Elena SF, García-Martín JA *et al*. 2021. From genotypes to organisms: State-of-the-art and perspectives of a cornerstone in evolutionary dynamics. Physics of Life Reviews. 38:55–106. arXiv: 2002.00363 Publisher: Elsevier B.V.

Marks DS, Hopf TA, Sander C. 2012. Protein structure prediction from sequence variation. Nature Biotechnology. 30:1072–1080.

Martí-Gómez C, McCandlish DM, Kinney JB. 2026a. GaugeFixer: overcoming parameter non-identifiability in models of sequence-function relationships. ISSN: 2692-8205 Pages: 2025.12.08.693054 Section: New Results.

Martí-Gómez C, Zhou J, Chen WC, Stoltzfus A, Kinney JB, McCandlish DM. 2026b. Inference and Visualization of Complex Genotype–Phenotype Maps. Molecular Biology and Evolution. 43:msag023.

Matsui T, Mullis MN, Roy KR, Hale JJ, Schell R, Levy SF, Ehrenreich IM. 2022. The interplay of additivity, dominance, and epistasis on fitness in a diploid yeast cross. Nature Communications. 13:1463.

McCandlish DM. 2011. Visualizing fitness landscapes. Evolution. 65:1544–1558. Publisher: John Wiley & Sons, Ltd.

Moore JH, Williams SM. 2009. Epistasis and Its Implications for Personal Genetics. The American Journal of Human Genetics. 85:309–320. Publisher: Elsevier.

Moulana A, Dupic T, Phillips AM, Chang J, Roffler AA, Greaney AJ, Starr TN, Bloom JD, Desai MM. 2023. The landscape of antibody binding affinity in SARS-CoV-2 Omicron BA.1 evolution. eLife. 12:e83442.

Muzellec B, Teleńczuk M, Cabeli V, Andreux M. 2023. PyDESeq2: a python package for bulk RNA-seq differential expression analysis. Bioinformatics. 39:btad547.

Neidhart J, Szendro IG, Krug J. 2013. Exact results for amplitude spectra of fitness landscapes. Journal of Theoretical Biology. 332:218–227. arXiv: 1301.1923 Publisher: Elsevier.

Neidhart J, Szendro IG, Krug J. 2014. Adaptation in Tunably Rugged Fitness Landscapes: The Rough Mount Fuji Model. Genetics. 198:699–721.

Nguyen Ba AN, Lawrence KR, Rego-Costa A, Gopalakrishnan S, Temko D, Michor F, Desai MM. 2022. Barcoded bulk QTL mapping reveals highly polygenic and epistatic architecture of complex traits in yeast. eLife. 11:e73983.

Noderer WL, Flockhart RJ, Bhaduri A, Diaz De Arce AJ, Zhang J, Khavari PA, Wang CL. 2014. Quantitative analysis of mammalian translation initiation sites by facs-seq. Molecular Systems Biology. 10:748.

N'Guessan A, Tong WY, Heydari H, Ba ANN. 2025. Refining the resolution of the yeast genotype-phenotype map using single-cell RNA-sequencing. eLife. 13. Publisher: eLife Sciences Publications Limited.

Ogbunugafor CB, Wylie CS, Diakite I, Weinreich DM, Hartl DL. 2016. Adaptive Landscape by Environment Interactions Dictate Evolutionary Dynamics in Models of Drug Resistance. PLOS Computational Biology. 12:e1004710. Publisher: Public Library of Science.

O'Maille PE, Malone A, Dellas N, Andes Hess B, Smentek L, Sheehan I, Greenhagen BT, Chappell J, Manning G, Noel JP. 2008. Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. Nature Chemical Biology. 4:617–623.

Otwinowski J, McCandlish DM, Plotkin JB. 2018. Inferring the shape of global epistasis. Proceedings of the National Academy of Sciences. 115:E7550–E7558.

Papkou A, Garcia-Pastor L, Escudero JA, Wagner A. 2023. A rugged yet easily navigable fitness landscape. Science. 382:eadh3860. Publisher: American Association for the Advancement of Science.

Petti S, Martí-Gómez C, Kinney JB, Zhou J, McCandlish DM. 2025. On learning functions over biological sequence space: relating gaussian process priors, regularization, and gauge fixing. arXiv preprint arXiv:2504.19034. .

Poelwijk FJ, Socolich M, Ranganathan R. 2019. Learning the pattern of epistasis linking genotype and phenotype in a protein. Nature Communications. 10:1–11. Publisher: Springer US.

Pokusaeva VO, Usmanova DR, Putintseva EV, Espinar L, Sarkisyan KS, Mishin AS, Bogatyreva NS, Ivankov DN, Akopyan AV, Avvakumov SY *et al*. 2019. An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. PLos Genet.. 15:e1008079.

Rabani M, Pieper L, Chew GL, Schier AF. 2017. A Massively Parallel Reporter Assay of 3 UTR Sequences Identifies In Vivo Rules for mRNA Degradation. Molecular Cell. 68:1083–1094.e5. Publisher: Elsevier.

Rasmussen CE, Williams CKI. 2008. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press. Cambridge, Mass.. third edition.

Reddy G, Desai MM. 2021. Global epistasis emerges from a generic model of a complex trait. eLife. 10:1–36.

Rojas Echenique JI, Kryazhimskiy S, Nguyen Ba AN, Desai MM. 2019. Modular epistasis and the compensatory evolution of gene deletion mutants. PLoS genetics. 15:e1007958.

Rosenberg AB, Patwardhan RP, Shendure J, Seelig G. 2015. Learning the sequence determinants of alternative splicing from millions of random sequences. Cell. 163:698–711.

Rotrattanadumrong R, Yokobayashi Y. 2022. Experimental exploration of a ribozyme neutral network using evolutionary algorithm and deep learning. Nature Communications. pp. 1–14. Publisher: Springer US ISBN: 4146702232.

Rubin AF, Gelman H, Lucas N, Bajjalieh SM, Papenfuss AT, Speed TP, Fowler DM. 2017. A statistical framework for analyzing deep mutational scanning data. Genome Biology. 18:1–15. Publisher: Genome Biology.

Sackton TB, Hartl DL. 2016. Genotypic Context and Epistasis in Individuals and Populations. Cell. 166:279–287.

Schmiedel JM, Lehner B. 2019. Determining protein structures using deep mutagenesis. Nature Genetics. 51:1177–1186.

Schmiegelt B, Krug J. 2014. Evolutionary Accessibility of Modular Fitness Landscapes. Journal of Statistical Physics. 154:334–355.

Shine J, Dalgarno L. 1975. Determinant of cistron specificity in bacterial ribosomes. Nature. 254:34–38.

Somermeyer LG, Fleiss A, Mishin AS, Bozhanova NG, Igolkina AA, Meiler J, Pujol MEA, Putintseva EV, Sarkisyan KS, Kondrashov FA. 2022. Heterogeneity of the gfp fitness landscape and data-driven protein design. Elife. 11:e75842.

Soo VW, Swadling JB, Faure AJ, Warnecke T. 2021. Fitness landscape of a dynamic RNA structure. PLoS Genetics. 17:e1009353. ISBN: 1111111111.

Soyk S, Benoit M, Lippman ZB. 2020. New Horizons for Dissecting Epistasis in Crop Quantitative Trait Variation. Annual Review of Genetics. 54:287–307. Publisher: Annual Reviews.

Stadler PF. 1996. Landscapes and their correlation functions. Journal of Mathematical Chemistry. 20:1–45.

Stadler PF. 2002. Fitness landscapes, In: Lässig M, Valleriani A, editors, *Biological Evolution and Statistical Physics*, Springer. Berlin, Heidelberg. pp. 183–204.

Stadler PF, Happel R. 1999. Random field models for fitness landscapes. Journal of Mathematical Biology. 38:435–478.

Starr TN, Picton LK, Thornton JW. 2017. Alternative evolutionary histories in the sequence space of an ancient protein. Nature. 549:409–413.

Sundar V, Tu B, Guan L, Esvelt K. 2024. A NEW ULTRA-HIGH-THROUGHPUT ASSAY FOR MEASURING PROTEIN FITNESS. Proceedings of the Generalization and Epistemic Measures (GEM) Workshop at the International Conference on Learning Representations (ICLR). .

Szendro IG, Schenk MF, Franke J, Krug J, de Visser JAGM. 2013. Quantitative analyses of empirical fitness landscapes. Journal of Statistical Mechanics: Theory and Experiment. 2013:P01005. Publisher: IOP Publishing and SISSA.

Vaishnav ED, de Boer CG, Molinet J, Yassour M, Fan L, Adiconis X, Thompson DA, Levin JZ, Cubillos FA, Regev A. 2022. The evolution, evolvability and engineering of gene regulatory DNA. Nature. Publisher: Springer US.

Wang K, Pleiss G, Gardner J, Tyree S, Weinberger KQ, Wilson AG. 2019. Exact Gaussian Processes on a Million Data Points. NeurIPS. .

Wang T, Zhao D, Tian S. 2015. An overview of kernel alignment and its applications. Artificial Intelligence Review. 43:179–192.

Weinberger E. 1990. Correlated and uncorrelated fitness landscapes and how to tell the difference. Biological Cybernetics. 63:325–336.

Weinreich DM, Delaney NF, DePristo MA, Hartl DL. 2006. Darwinian evolution can follow only very few mutational paths to fitter proteins. Science. 312:111–114.

Weinreich DM, Lan Y, Jaffe J, Heckendorn RB. 2018. The Influence of Higher-Order Epistasis on Biological Fitness Landscape Topography. Journal of Statistical Physics. 172:208–225. Publisher: Springer US.

Weinreich DM, Watson RA, Chao L. 2005. Perspective: sign epistasis and genetic costraint on evolutionary trajectories. Evolution. 59:1165–1174.

Westmann CA, Goldbach L, Wagner A. 2024a. Entangled adaptive landscapes facilitate the evolution of gene regulation by exaptation. Pages: 2024.11.10.620926 Section: New Results.

Westmann CA, Goldbach L, Wagner A. 2024b. The highly rugged yet navigable regulatory landscape of the bacterial transcription factor TetR. Nature Communications. 15:10745. Publisher: Nature Publishing Group.

Wong MS, Kinney JB, Krainer AR. 2018. Quantitative Activity Profile and Context Dependence of All Human 5′ Splice Sites. Molecular Cell. 71:1012–1026.e3. Publisher: Elsevier Inc.

Wright S. 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proceedings of the Sixth International Congress of Genetics. pp. 356–366.

Wu NC, Dai L, Olson CA, Lloyd-Smith JO, Sun R. 2016. Adaptation in protein fitness landscapes is facilitated by indirect paths. eLife. 5:1–21.

Zarin T, Lehner B. 2024. A complete map of specificity encoding for a partially fuzzy protein interaction.

Zebell SG, Martí-Gómez C, Fitzgerald B, Cunha CP, Lach M, Seman BM, Hendelman A, Sretenovic S, Qi Y, Bartlett M *et al*. 2025. Cryptic variation fuels plant phenotypic change through hierarchical epistasis. Nature. pp. 1–9. Publisher: Nature Publishing Group.

Zhou J, Martí-Gómez C, Petti S, McCandlish DM. 2025. Learning sequence-function relationships with scalable, interpretable Gaussian processes. eLife. 14. Publisher: eLife Sciences Publications Limited.

Zhou J, McCandlish DM. 2020. Minimum epistasis interpolation for sequence-function relationships. Nature Communications. 11.

Zhou J, Wong MS, Chen Wc, Krainer AR, Justin B, Mccandlish DM. 2022. Higher-order epistasis and phenotypic prediction. Proc. Natl. Acad. Sci. USA. 119.

Supplementary Information

Epistasis on fitness landscapes

Let f be an α^ℓ-dimensional vector encoding the fitness associated with each genotype in the space of possible haploid sequences with α alleles and ℓ sites $S = \{1, 2, \ldots, \ell\}$. In this section, we review two common ways to quantify epistasis in a fitness landscape: one based on local epistatic coefficients (Zhou and McCandlish 2020; Chen *et al.* 2021) and a second one based on the variance explained by genetic interactions of different orders (Stadler and Happel 1999; Zhou *et al.* 2022) or across subsets of sites (Martí-Gómez *et al.* 2026b).

Variance components Any fitness landscape f can be decomposed into orthogonal components f_k corresponding to genetic interactions of order k

$$f = \sum_{k=0}^{\ell} f_k. \tag{S1}$$

These components can be obtained by projecting the f into the kth-subspace using the orthogonal projection matrix P_k given by the Krawtchouk polynomials (Stadler and Happel 1999; Zhou *et al.* 2022):

$$P_k(x, x') = \alpha^{-\ell} \sum_{q=0}^{k} (-1)^q (\alpha - 1)^{k-q} \binom{d(x,x')}{q} \binom{\ell - d(x,x')}{k-q}. \tag{S2}$$

Moreover, each k-th order subspace can be further decomposed into orthogonal components corresponding to the contributions of interactions between specific subsets of sites $U \subseteq S$ of size k:

$$f = \sum_{U \in \mathcal{P}(\mathcal{S})} f_U, \tag{S3}$$

where $\mathcal{P}(\mathcal{S})$ is the set of all subsets of S (i.e. the power set of S) and $f_U = P_U f$, where P_U is an orthogonal projection matrix onto the corresponding subspace. Each such P_U can be expressed as a Kronecker product of a series of site-specific projection matrices onto the constant subspace, defined by $P_{\text{con}} = \frac{1}{\alpha} 1 1^T$, and the orthogonal or additive subspace, given by $P_{\text{add}} = I - P_{\text{con}}$. In particular

$$P_U = \bigotimes_{i=1}^{\ell} P_{i \in U}, \tag{S4}$$

where $P_{i \in U} = P_{\text{add}}$ when site $i \in U$ and $P_{i \in U} = P_{\text{con}}$ otherwise (Martí-Gómez *et al.* 2026b):

$$P_U(x, x') = \alpha^{-\ell} \prod_{i \in U : x_i = x'_i} (\alpha - 1) \prod_{i \in U : x_i \neq x'_i} (-1). \tag{S5}$$

Because U-components are orthogonal to each other, the total variance in a fitness landscape can be expressed as a sum of variances explained by genetic interactions of order k (i.e. summing over all U with $|U| = k$) or by interactions between the sites U, respectively (Martí-Gómez *et al.* 2026b):

$$\text{Var}[f] = \sum_{k=1}^{\ell} \text{Var}[f_k] = \sum_{U \in \mathcal{P}(S)} \text{Var}[f_U]. \tag{S6}$$

Local epistatic coefficients The classical way to quantify epistasis is through the definition of the epistatic coefficient ϵ for a pair of mutations at different loci $A \to a$ and $B \to b$, which quantifies the difference in the effect of mutation $A \to a$ in the presence of allele B compared to that in the presence of allele b at the other locus:

$$\epsilon = (f_{AB} - f_{aB}) - (f_{Ab} - f_{ab}). \tag{S7}$$

This epistatic coefficient is defined locally, as it depends only on four specific genotypes sharing the same genetic background. Still, we can compute the average magnitude of local epistatic interactions $\overline{\epsilon^2}$ in a given fitness landscape by averaging the its squared values across every possible pair of mutations at every possible genetic background:

$$\overline{\epsilon^2} = \frac{1}{s} f^T \Delta^{(2)} f, \tag{S8}$$

where $s = \binom{\ell}{2}\binom{\alpha}{2}^2 \alpha^{\ell-2}$ is the number of epistatic coefficients and $\Delta^{(2)}$ is a positive semi-definite sparse matrix (Zhou and McCandlish 2020; Chen *et al.* 2021). These results can be generalized to quantify local epistatic coefficients of any order P using

$$\Delta^{(P)}(x, x') = \begin{cases} 0 & \text{if } d(x,x') > P \\ (-1)^{d(x,x')} (\alpha - 1)^{P - d(x,x')} \binom{\ell - P}{P - d(x,x')} & \text{if } d(x,x') \leq P, \end{cases} \tag{S9}$$

where $d(x, x')$ represents the Hamming distance i.e. number of single point mutations, separating sequences x and x', such that the sum of the squared P-th order local epistatic coefficients is given by $f^T \Delta^{(2)} f$ and the mean sqaured P-th order local epistatic coefficient

is obtained by dividing by the number of such coefficients $\binom{\ell}{P}\binom{\alpha}{2}^P \alpha^{\ell-P}$ (Chen *et al.* 2021). Interestingly, $\Delta^{(P)}$ can also be expressed as a weighted sum of the projection operators into the k-th order subspaces given by P_k:

$$\Delta^{(P)} = \sum_{k=0}^{\ell} \lambda_k P_k = \sum_{k=P}^{\ell} \lambda_k P_k, \tag{S10}$$

where $\lambda_k = \alpha^P \binom{k}{P}$ (Chen *et al.* 2021) correspond to the eigenvalues of the $\Delta^{(P)}$ operator. Note that the second equality arises because $\binom{k}{P} = 0$ for $k < P$, corresponding to the fact that all local P-th order epistatic coefficients are zero for any function f with maximal order less than P (so that all such functions are in the null space of $\Delta^{(P)}$) and that moreover alteration of any component of order less than P for an arbitrary f leaves all P-th order local epistatic coefficients unchanged.

Epistatic coefficients among subsets of sites

In this section, we describe how the $\Delta^{(P)}$ operator that extracts the sum of squared epistatic interactions in a fitness landscape can be decomposed into the sum of simpler operators that we call $\Delta^{(U)}$, corresponding to local epistatic interactions only among a subset of sites $U \subseteq S$. These results generalize the main text results focused on local epistatic interactions between pairs of sites i and j i.e. $U = \{i, j\}$.

Let E_U be a $s_U \times \alpha^\ell$ matrix such that the entries in the $E_U f$ vector encode all the local $|U|$-th order epistatic coefficient for a subset of positions U. s_U is the number of different epistatic coefficients corresponding to local $|U|$-th order mutant epistatic coefficients within sites U, and is given by the product of the number of genetic backgrounds at sites not in U $\alpha^{\ell-|U|}$ and the number of possible combinations of mutations across the sites U $\binom{\alpha}{2}^{|U|}$:

$$s_U = \alpha^{\ell-|U|}\binom{\alpha}{2}^{|U|} = \alpha^{\ell-|U|}\left(\frac{\alpha(\alpha-1)}{2}\right)^{|U|} = \frac{\alpha^\ell}{2^{|U|}}(\alpha-1)^{|U|}. \tag{S11}$$

Thus,

$$\overline{\epsilon_U^2} = \frac{1}{s_U}(E_U f)^T E_U f = \frac{1}{s_U} f^T \Delta^{(U)} f, \tag{S12}$$

where the entries of $\Delta^{(U)}$ for a pair of sequences x and x' can be obtained by summing over all possible local $|U|$-mutant epistatic coefficients between sites in U

$$\Delta^{(U)}(x, x') = \sum_{m=1}^{s_U} E_U(m, x) E_U(m, x'). \tag{S13}$$

$E_U(m, x') = 0$ if sequence x' is not involved in epistatic coefficient m, and takes values -1 or 1 otherwise. Thus, we only need to sum overlocal epistatic coefficients involving both x and x'. If $x_i \neq x'_i$ for any position $i \notin U$, then sequences x and x' cannot be involved in any epistatic coefficient and thus $\Delta^{(U)}(x, x') = 0$. Otherwise, $E_U(m, x)E_U(m, x') = (-1)^{d(x,x')}$ depending on the Hamming distance between them $d(x, x')$. Moreover, the number of local $|U|$-mutant epistatic coefficients that involve both sx and x' is given by $(\alpha-1)^{|U|-d(x,x')}$. Thus

$$\Delta^{(U)}(x, x') = \begin{cases} (-1)^{d(x,x')}(\alpha-1)^{|U|-d(x,x')} & \text{if } x_i = x'_i \ \forall i \notin U \\ 0 & \text{otherwise.} \end{cases} \tag{S14}$$

We can verify that we can recover the $\Delta^{(P)}$ operator by summing $\Delta^{(U)}$ over all possible subsets of sites U of size P. If $d(x, x') > P$, there is no single set of sites U for which the context at sequences x and x' can match and thus $\sum_{U:|U|=P} \Delta^{(U)}(x, x') = 0$. For $d(x, x') \leq P$, the entries of $\Delta^{(U)}(x, x')$ are the same but they are summed over multiple U. As we only sum over U's such that x and x' share the context, the number of times we are summing them corresponds to the number of ways we can choose $P - d(x, x')$ sites that match out of the $\ell - P$ sites in the shared context between them. Thus,

$$\sum_{U:|U|=P} \Delta^{(U)}(x, x') = \begin{cases} (-1)^{d(x,x')}(\alpha-1)^{P-d(x,x')}\binom{\ell-P}{P-d(x,x')} & \text{if } d(x, x') \leq P \\ 0 & \text{if } d(x, x') > P, \end{cases} \tag{S15}$$

which exactly matches the $\Delta^{(P)}$ operator (Zhou and McCandlish 2020; Chen *et al.* 2021).

Relationship between local epistatic coefficients and variance components

In this section, we describe some properties of the new $\Delta^{(U)}$ operator and the relationships with the projection operators into the subspaces corresponding to interactions of different orders and subsets of sites U.

One useful property of the $\Delta^{(U)}$ operator is that it can be expressed as a Kronecker product of site-specific matrices

$$\Delta^{(U)} = \bigotimes_{i=1}^{\ell} \Delta_{i \in U} = \bigotimes_{i=1}^{\ell} \begin{cases} \alpha P_{\text{add}} & \text{if } i \in U \\ I & \text{if } i \notin U, \end{cases} \tag{S16}$$

with entry-wise formula given by:

$$\Delta^{(U)}(x, x') = \prod_{\substack{i \in U \\ x_i = x'_i}} (\alpha - 1) \prod_{\substack{i \in U \\ x_i \neq x'_i}} (-1) \prod_{\substack{i \notin U \\ x_i = x'_i}} (1) \prod_{\substack{i \notin U \\ x_i \neq x'_i}} 0. \tag{S17}$$

Thus, we can use the mixed-product property to show that the columns of $P_{U'}$ are eigenvectors of $\Delta^{(U)}$ with eigenvalue $\alpha^{|U|}$ if $U \subseteq U'$ or are in the null space of $\Delta^{(U)}$ whenever $U \not\subseteq U'$.

$$\Delta^{(U)} P_{U'} = \bigotimes_{i=1}^{\ell} \Delta_{i \in U} \bigotimes_{i=1}^{\ell} P_{i \in U'} = \bigotimes_{i=1}^{\ell} \Delta_{i \in U} P_{i \in U'}, \tag{S18}$$

where

$$\Delta_{i \in U} P_{i \in U'} = \begin{cases} P_{\text{con}} & \text{if } i \notin U \wedge i \notin U' \\ P_{\text{add}} & \text{if } i \notin U \wedge i \in U' \\ 0 & \text{if } i \in U \wedge i \notin U' \\ \alpha P_{\text{add}} & \text{if } i \in U \wedge i \in U'. \end{cases} \tag{S19}$$

Therefore

$$\Delta^{(U)} P_{U'} = \begin{cases} \alpha^{|U|} P_{U'} & \text{if } U \subseteq U' \\ 0 & \text{if } U \not\subseteq U'. \end{cases} \tag{S20}$$

Because the projection operators $P_{U'}$ are orthogonal to each other, we can express $\Delta^{(U)}$ as a linear combination of $P_{U'}$ weighted by their eigenvalues (either 0 or $\alpha^{|U|}$):

$$\Delta^{(U)} = \alpha^{|U|} \sum_{U' : U \subseteq U'} P_{U'}, \tag{S21}$$

so that, using the fact that sums of projection matrices into orthogonal subspaces are themselves projection matrices, we see that $\Delta^{(U)}$ is itself just a $|U|$-dependent constant times a specific projection matrix. Moreover, this relation between $\Delta^{(U)}$ and the $P_{U'}$ with $U \subseteq U'$ can be used to derive the relationship between the $\Delta^{(P)}$ operator and the projection operator into the kth order subspace P_k:

$$\Delta^{(P)} = \sum_{U : |U| = P} \Delta^{(U)} = \sum_{U : |U| = P} \alpha^{|U|} \sum_{U' : U \subseteq U'} P_{U'} = \alpha^P \sum_{U : |U| = P} \sum_{U' : U \subseteq U'} P_{U'}. \tag{S22}$$

It is now easy to see that the number of times we are summing each $P_{U'}$ depends on the size of U'. In particular, it corresponds to the number of $U : |U| = P$ subsets in U', which can be calculated as the number of ways of choosing $|U| = P$ sites out of $|U'|$. Then, we can use the fact that $P_k = \sum_{U : |U| = k} P_U$ (Martí-Gómez *et al.* 2026b) to recover the eigendecomposition of the $\Delta^{(P)}$ operator (Chen *et al.* 2021):

$$\Delta^{(P)} = \alpha^P \sum_{U' : |U'| \geq P} \binom{|U'|}{P} P_{U'} = \sum_{k \geq P} \alpha^P \binom{k}{P} \sum_{U' : |U'| = k} P_{U'} = \sum_{k \geq P} \alpha^P \binom{k}{P} P_k. \tag{S23}$$

Relationship between local $|U|$-way epistatic coefficients and the epistatic variance explained by the subset U

A useful low dimensional summary statistic to describe the patterns of genetic interactions across subsets of sites in a fitness landscape is the epistatic variance explained by a set of sites U (Crawford *et al.* 2017; Reddy and Desai 2021; Martí-Gómez *et al.* 2026b). We define this variance $\text{Var}^{(U)}[f]$ to include not only the variance explained by $|U|$-way interactions between mutations at the sites U, but also all other interactions of order higher than $|U|$ where all sites in U are involved, so that $\text{Var}^{(U)}[f]$ provides an overall measure of the amount of $|U|$-way and higher epistasis that the subset of sites U is involved in. This quantity can be computed by projecting the fitness landscape f onto the 2^ℓ subspaces defined by every possible U' given by $f_{U'} = P_{U'} f$, summing over the $f_{U'}$'s of interest and computing the inner product.

$$\text{Var}^{(U)}[f] = \left(\sum_{U' : U \subseteq U'} f_{U'} \right)^T \left(\sum_{U' : U \subseteq U'} f_{U'} \right) = \left(\sum_{U' : U \subseteq U'} P_{U'} f \right)^T \left(\sum_{U' : U \subseteq U'} P_{U'} f \right) = f^T \left(\sum_{U' : U \subseteq U'} P_{U'} \right) f, \tag{S24}$$

where

$$\sum_{U' : U \subseteq U'} P_{U'} = \sum_{U' : U \subseteq U'} \bigotimes_{i=1}^{\ell} P_{i \in U'}. \tag{S25}$$

As the Kronecker product is commutative up to row and column permutations given by Q_1 and Q_2, we can write the factors in order depending on whether the sites are in U or not:

$$\begin{aligned} \sum_{U' : U \subseteq U'} P_{U'} &= Q_1 \left(\sum_{U' : U \subseteq U'} \bigotimes_{i \in U} P_{i \in U'} \bigotimes_{i \notin U} P_{i \in U'} \right) Q_2 = Q_1 \left(\bigotimes_{i \in U} P_{\text{add}} \sum_{U' : U \subseteq U'} \bigotimes_{i \notin U} P_{i \in U'} \right) Q_2 \\ &= Q_1 \left(\bigotimes_{i \in U} P_{\text{add}} \bigotimes_{i \notin U} I \right) Q_2 = \bigotimes_{i=1}^{\ell} \begin{cases} P_{\text{add}} & \text{if } i \in U \\ I & \text{if } i \notin U \end{cases} = \alpha^{-|U|} \Delta^{(U)}. \end{aligned} \tag{S26}$$

Thus, this shows that the sum of squared epistatic coefficients among a subset of sites U given by $f^T\Delta^{(U)}f$ is equal to the epistatic variance explained by those sites $\text{Var}^{(U)}[f]$ multiplied by a factor of $\alpha^{|U|}$. Noting that $f^T\Delta^{(U)}f = \overline{\epsilon_U^2}s_U$ and simplifying $s_U/\alpha^{|U|}$ yields Equation 2 in the main text. Similarly, the sum of squared local P-epistatic coefficients across the complete fitness landscape given by $f^T\Delta^{(P)}f$ can also be interpreted as the sum the epistatic variances of all possible combinations of P sites multiplied by a factor of α^P.

Relationship between mean and variance of local $|U|$-way epistatic coefficients and the variance components

In the previous sections, we have shown the relationships between the average squared epistatic coefficients between mutations at sites U and the total variance explained by interactions of order $|U|$ or larger involving all sites in U. In this section, we describe how these quantities relate to the mean and variance of the distribution of $|U|$-way epistatic coefficients for a set of mutations within sites U. To do so, we first decompose the epistatic coefficients between mutations at sites U into the epistatic coefficients for each possible combination of mutations at sites U. Using this more granular statistic, we derive the mean of the epistatic coefficients between a given set of mutations, and use its square, together with the average squared epistatic coefficient to compute the variance. Finally, we take the average of these variances over all possible combinations of mutations at sites U and derive its relationship to the variance components defined over the subsets of sites U.

Let $C = \{\{c_{11}, c_{12}\}, \ldots, \{c_{k1}, c_{k2}\}\}$ be the set of k pairs of characters at a set of k positions U and ϵ_{UC} be the $\alpha^{\ell-k}$-dimensional vector of k-way epistatic coefficients between the mutations specified by C across every possible genetic background. Let E_{UC} be an $\alpha^{\ell-k} \times \alpha^\ell$ such that $\epsilon_{UC} = E_{UC}f$ and note that this matrix that can be expressed as a Kronecker product of site specific factors:

$$E_{UC} = \bigotimes_i^\ell \begin{cases} E_i^{c_{i1},c_{i2}} & \text{if } i \in U \\ I & \text{if } i \notin U, \end{cases} \tag{S27}$$

where $E_i^{c_1,c_2}$ is a row vector with entries given by

$$E_i^{c_1,c_2}(x) = \begin{cases} 1 & \text{if } x_i = c_1 \\ -1 & \text{if } x_i = c_2 \\ 0 & \text{otherwise.} \end{cases} \tag{S28}$$

Using these expressions, we can compute the average squared epistatic coefficients restricted to the sets of mutations defined by C as

$$\overline{\epsilon_{UC}^2} = \frac{1}{\alpha^{\ell-|U|}}(E_{UC}f)^T(E_{UC}f) = \frac{1}{\alpha^{\ell-|U|}}f^TE_{UC}^TE_{UC}f = \frac{1}{\alpha^{\ell-|U|}}f^T\Delta^{(UC)}f, \tag{S29}$$

where $\Delta^{(UC)}$ can be easily computed by using the Kronecker factorization of E_{UC}

$$\Delta^{(UC)} = \bigotimes_i^\ell \begin{cases} \Delta_i^{c_1,c_2} & \text{if } i \in U \\ I & \text{if } i \notin U. \end{cases} \tag{S30}$$

The factor $\Delta_i^{c_1,c_2} = (E_i^{c_1,c_2})^T E_i^{c_1,c_2}$ takes the form

$$\Delta_i^{c_1,c_2}(x,x') = \begin{cases} 1 & \text{if } (x_i = c_1 \wedge x_i' = c_1) \vee (x_i = c_2 \wedge x_i' = c_2) \\ -1 & \text{if } (x_i = c_1 \wedge x_i' = c_2) \vee (x_i = c_2 \wedge x_i' = c_1) \\ 0 & \text{otherwise.} \end{cases} \tag{S31}$$

It is easy to see that $\sum_{c_1,c_2}\Delta_i^{c_1,c_2} = \alpha P_{\text{add}}$ and that $\Delta^{(U)}$ can be expressed as a sum of $\Delta^{(UC)}$ for all possible combinations of pairs of alleles C at sites U.

$$\sum_C \Delta^{(UC)} = \sum_C \bigotimes_i^\ell \begin{cases} \Delta_i^{c_{i1},c_{i2}} & \text{if } i \in U \\ I & \text{if } i \notin U \end{cases} = \bigotimes_i^\ell \begin{cases} \sum_C \Delta_i^{c_{i1},c_{i2}} & \text{if } i \in U \\ I & \text{if } i \notin U \end{cases} = \Delta^{(U)}. \tag{S32}$$

Next, we use the E_{UC} to compute the average epistatic coefficients $\overline{\epsilon_{UC}}$ between the set of mutations defined by C at sites U across all possible genetic backgrounds:

$$\overline{\epsilon_{UC}} = \frac{1}{\alpha^{\ell-|U|}}1^TE_{UC}f. \tag{S33}$$

This quantity corresponds to the previously proposed background-averaged epistatic coefficient for parametrizing and describing sequence-function relationships (Faure *et al.* 2024a; Petti *et al.* 2025). As the sign of $\overline{\epsilon_{UC}}$ depends on the ordering of alleles c_{i1} and c_{i2} for each position i, we can use its squared value $\overline{\epsilon_{UC}}^2$ to describe the magnitude of the average epistatic coefficients independently of the choice of reference allele at each position:

$$\overline{\epsilon_{UC}}^2 = \left(\frac{1}{\alpha^{\ell-|U|}}1^TE_{UC}f\right)^2 = \frac{1}{\alpha^{2\ell-2|U|}}f^TE_{UC}^T11^TE_{UC}f. \tag{S34}$$

Let $A_{UC} = \frac{1}{\alpha^{\ell-|U|}} E_{UC}^T 1 1^T E_{UC}$, so that $\overline{\epsilon_{UC}}^2$ can be expressed as a function of a quadratic form with this matrix given by $\frac{1}{\alpha^{\ell-|U|}} f^T A_{UC} f$. To derive a simple expression for A_{UC} and understand its relationship with other objects, we use the fact that all of the matrices E_{UC} and $1 1^T$ can be expressed as Kronecker products of site specific factors:

$$A_{UC} = \frac{1}{\alpha^{\ell-|U|}} E_{UC}^T 1 1^T E_{UC} = \frac{1}{\alpha^{\ell-|U|}} \bigotimes_i^\ell \begin{cases} (E_i^{c_{i1},c_{i2}})^T E_i^{c_{i1},c_{i2}} & \text{if } i \in U \\ I 1 1^T I & \text{if } i \notin U \end{cases} = \bigotimes_i^\ell \begin{cases} \Delta_i^{c_{i1},c_{i2}} & \text{if } i \in U \\ \frac{1}{\alpha} 1 1^T & \text{if } i \notin U. \end{cases} \tag{S35}$$

It is easy to see that A_{UC} is in the subspace defined by genetic interactions of order $|U|$ between sites U since $P_U A_{UC} = A_{UC}$. In fact, if we sum over all possible combinations of mutations C, we obtain a matrix that is proportional to the projection operator into the U-subspace P_U:

$$\sum_C A_{UC} = \sum_C \bigotimes_i^\ell \begin{cases} \Delta_i^{c_{i1},c_{i2}} & \text{if } i \in U \\ \frac{1}{\alpha} 1 1^T & \text{if } i \notin U. \end{cases} = \bigotimes_i^\ell \begin{cases} \sum_{c_{i1},c_{i2}} \Delta_i^{c_{i1},c_{i2}} & \text{if } i \in U \\ \frac{1}{\alpha} 1 1^T & \text{if } i \notin U. \end{cases} = \bigotimes_i^\ell \begin{cases} \alpha P_{\text{add}} & \text{if } i \in U \\ P_{\text{con}} & \text{if } i \notin U. \end{cases} = \alpha^{|U|} P_U. \tag{S36}$$

Using this equivalence, we can show that the average squared mean epistatic coefficient for mutations at sites U is proportional to the variance explained by genetic interactions of order $|U|$ between those sites:

$$\begin{aligned} \frac{1}{\binom{\alpha}{2}^{|U|}} \sum_C (\overline{\epsilon_{UC}})^2 &= \frac{1}{\binom{\alpha}{2}^{|U|}} \sum_C \frac{1}{\alpha^{\ell-|U|}} f^T A_{UC} f = \frac{1}{\binom{\alpha}{2}^{|U|} \alpha^{\ell-|U|}} f^T \left(\sum_C A_{UC} \right) f = \frac{1}{\binom{\alpha}{2}^{|U|} \alpha^{\ell-|U|}} f^T \left(\sum_C A_{UC} \right) f \\ &= \frac{1}{\binom{\alpha}{2}^{|U|} \alpha^{\ell-|U|}} \alpha^{|U|} f^T P_U f = \frac{2^{|U|}}{\alpha^\ell} \left(\frac{\alpha}{\alpha - 1} \right)^{|U|} \text{Var}[f_U] \end{aligned} \tag{S37}$$

Finally, we can derive the variance over local epistatic coefficients between mutations at sites U by using the relationship between the variance and the raw second moments $\text{Var}[\epsilon_{UC}] = \overline{\epsilon_{UC}^2} - \overline{\epsilon_{UC}}^2$ as well as the average variance across all possible sets of mutations C at sites U and show that it is proportional to the variance explained by genetic interactions of order higher than $|U|$ involving sites U:

$$\begin{aligned} \text{Var}[\epsilon_U] &= \frac{1}{\binom{\alpha}{2}^{|U|}} \sum_C \text{Var}[\epsilon_{UC}] = \frac{1}{\binom{\alpha}{2}^{|U|}} \sum_C \left(\overline{\epsilon_{UC}^2} - \overline{\epsilon_{UC}}^2 \right) = \frac{1}{\binom{\alpha}{2}^{|U|}} \sum_C \overline{\epsilon_{UC}^2} - \frac{1}{\binom{\alpha}{2}^{|U|}} \sum_C \overline{\epsilon_{UC}}^2 \\ &= \frac{2^{|U|}}{\alpha^\ell} \left(\frac{\alpha}{\alpha - 1} \right)^{|U|} \left(\text{Var}^{(U)}[f] - \text{Var}[f_U] \right) = \frac{2^{|U|}}{\alpha^\ell} \left(\frac{\alpha}{\alpha - 1} \right)^{|U|} \text{Var}^{(U)}_{k>|U|}[f]. \end{aligned} \tag{S38}$$

Covariance function between sequences differing at subsets of sites

In addition to characterizing interaction structure through squared epistatic coefficients or epistatic variances of specific subsets of sites U, it is often useful to summarize an *arbitrary* landscape f by how similar fitness values are for sequences that differ at particular sets of sites. In particular, let $H(x, x')$ be the subset of sites at which two sequences x and x' differ

$$H(x, x') = \{i \in S : x_i \neq x_i'\}. \tag{S39}$$

We define the covariance function for a mismatch set D as

$$C_f(D) = \frac{1}{N_D} \sum_{(x,x'):H(x,x')=D} \left(f(x) - \bar{f} \right) \left(f(x') - \bar{f} \right), \tag{S40}$$

where $\bar{f} = \alpha^{-\ell} \sum_x f(x)$ and $N_D = \alpha^\ell (\alpha - 1)^{|D|}$ is the number of sequence pairs that differ at sites D.

$C_f(D)$ can be expressed as a quadratic form $C_f(D) = \frac{1}{N_D} (f - \bar{f})^T A_D (f - \bar{f})$, where A_D is an $\alpha^\ell \times \alpha^\ell$ matrix given by

$$A_D(x, x') = \begin{cases} 1 & \text{if } H(x, x') = D \\ 0 & \text{if } H(x, x') \neq D. \end{cases} \tag{S41}$$

This matrix can be written as a Kronecker product of single-site matrices, enabling efficient computation of this quantity without explicit evaluation of the matrices A_D (Martí-Gómez *et al.* 2026b):

$$A_D = \bigotimes_{i=1}^\ell \begin{cases} 1 1^T - I & \text{if } i \in D \\ I & \text{if } i \notin D. \end{cases} \tag{S42}$$

Moreover, if we consider the decomposition of f into its orthogonal components $f_U = P_U f$ such that $f = \sum_{U \in \mathcal{P}(S)} f_U$, we can express the covariance function as a sum of covariance over all the possible U-components:

$$C_f(D) = \frac{1}{N_D} \left(\sum_{U \in \mathcal{P}(S)} P_U f \right)^T A_D \left(\sum_{U \in \mathcal{P}(S)} P_U f \right) = \frac{1}{N_D} \sum_{U \in \mathcal{P}(S)} f_U^T P_U A_D P_U f_U = \sum_{U \in \mathcal{P}(S)} C_{f_U}(D). \tag{S43}$$

Next, we note that $P_U A_D P_U$ can be easily calculated through multiplication of their Kronecker factors individually:

$$P_U^T A_D P_U = \bigotimes_{i=1}^{\ell} \begin{cases} -P_{\text{add}} & \text{if } i \in D \wedge i \in U \\ (\alpha-1)P_{\text{con}} & \text{if } i \in D \wedge i \notin U \\ P_{\text{add}} & \text{if } i \notin D \wedge i \in U \\ P_{\text{con}} & \text{if } i \notin D \wedge i \notin U. \end{cases} \tag{S44}$$

In fact, this can be summarized as $P_U^T A_D P_U = (-1)^{|D\cap U|}(\alpha-1)^{|D|-|D\cap U|}P_U$. Since f_U is in the U-subspace, $P_U f_U = f_U$ and thus

$$\begin{aligned} C_{f_U}(D) &= \frac{1}{N_D} f_U^T P_U A_D P_U f_U = \left(\frac{1}{\alpha^\ell(\alpha-1)^{|D|}}\right)\left(f_U^T(-1)^{|D\cap U|}(\alpha-1)^{|D|-|D\cap U|}P_U f_U\right) \\ &= \alpha^{-\ell}(-1)^{|D\cap U|}(\alpha-1)^{-|D\cap U|} f_U^T f_U = \frac{||f_U||^2}{(\alpha-1)^{|U|}}\alpha^{-\ell}(-1)^{|D\cap U|}(\alpha-1)^{|U|-|D\cap U|} \\ &= \frac{||f_U||^2}{(\alpha-1)^{|U|}} w_U(D). \end{aligned} \tag{S45}$$

Here, we denote by $w_U(D)$ the contribution of genetic interactions among sites U to the covariance between sequences that differ at sites D. We refer to these quantities as subset-resolved covariance weights and they are related to the classical Krawtchouk polynomials $w_k(d)$ arising in the Fourier analysis of fitness landscapes (Stadler 1996; Zhou *et al.* 2022). In particular, the functions $w_k(d)$, which depend only on the Hamming distance $d = |D|$, are recovered by summing $w_U(D)$ over all subsets U of size k,

$$w_k(d) = \sum_{U:|U|=k} w_U(D), \qquad \text{for any } D \text{ such that } |D| = d.$$

Gaussian random field landscapes

In the previous sections, we have explained different ways to characterize the patterns of epistatic interactions in a given fitness landscape f. Here we aim to define a probabilistic ensemble of fitness landscapes via a Gaussian distribution $\mathrm{p}(f)$. This Gaussian distribution can be used as a theoretical random fitness landscape model, similar to the classical random fitness landscape models e.g. Rough Mount Fuji, NK or House of Cards landscapes (Kingman 1978; Kauffman and Weinberger 1989; Aita and Husimi 1998), or as a prior distribution for Gaussian process inference

$$\mathrm{p}(f) = \mathcal{N}(\mu, K), \tag{S46}$$

as previously proposed (Zhou *et al.* 2022, 2025). Here, we generally assume that the prior Gaussian distribution has zero mean i.e. $\mu = 0$. To define this prior, let Q_U be an $\alpha^\ell \times (\alpha-1)^{|U|}$-dimensional matrix containing an orthonormal basis for the subspace defined by $|U|$-way interactions among the sites U (i.e. the column space of P_U). We assume that the coefficients for these basis vectors are drawn independently from a zero-mean Gaussian with variance λ_U. In particular, let b_U be the $(\alpha-1)^{|U|}$-dimensional vector of coefficients:

$$b_U \sim \mathcal{N}(0, \lambda_U I). \tag{S47}$$

If we then define $f = \sum_{U\in\mathcal{P}(S)} Q_U b_U$, it is easy to see that f is also a zero-mean Gaussian distribution with covariance matrix K given by

$$K = \mathbb{E}_f[ff^T] = \mathbb{E}_b\left[\left(\sum_{U\in\mathcal{P}(S)} Q_U b_U\right)\left(\sum_{U'} Q_{U'} b_{U'}\right)^{\mathsf{T}}\right] = \sum_{U,U'} Q_U \,\mathbb{E}_b\left[b_U b_{U'}^{\mathsf{T}}\right] Q_{U'}^{\mathsf{T}} = \sum_{U\in\mathcal{P}(S)} \lambda_U Q_U Q_U^{\mathsf{T}} = \sum_{U\in\mathcal{P}(S)} \lambda_U P_U. \tag{S48}$$

Since the columns of Q_U are orthonormal, P_U corresponds to the projection matrix into the function subspace spanned by interactions between sites in U, as given by Eq. S4. Here, we note that the entries of $P_U(x,x')$ only depend on the Hamming distance between sequences x, x' at sites in U, but not on the alleles at which they match or differ. Thus, the covariance matrix $K(x,x')$ only depends on the subset of sites at which x and x' differ.

Moreover, the variance of the regression coefficients λ_U for each subset of sites U determines the expected variance explained by interactions between exactly sites in U and is given by $\mathbb{E}_{f\sim\mathcal{N}(0,K)}\left[||f_U||^2\right] = (\alpha-1)^{|U|}\lambda_U$.

$$\mathbb{E}_f\left[||f_U||^2\right] = \mathbb{E}_f\left[f_U^T f_U\right] = \mathbb{E}_f\left[(P_U f)^T (P_U f)\right] = \mathbb{E}_f\left[f^T P_U f\right]. \tag{S49}$$

We note that samples from the random field model $f \sim \mathcal{N}(0,K)$ can be drawn by first drawing $z \sim \mathcal{N}(0,I)$ and then computing $f = K^{\frac{1}{2}} z$, such that

$$\mathbb{E}_f\left[f^T P_U f\right] = \mathbb{E}_z\left[(K^{\frac{1}{2}}z)^T P_U (K^{\frac{1}{2}}z)\right]. \tag{S50}$$

Moreover, as the columns of P_U are eigenvectors of K with eigenvalue λ_U and P_U matrices are orthogonal to each other, $K^{\frac{1}{2}} = \sum_{U\in\mathcal{P}(S)} \sqrt{\lambda_U} P_U$ and thus

$$\mathbb{E}_f\left[||f_U||^2\right] = \mathbb{E}_z\left[\lambda_U z^T P_U z\right] = \lambda_U \operatorname{tr}(P_U) = \lambda_U \sum_x P_U(x,x) = \lambda_U \alpha^\ell\left(\alpha^{-\ell}(\alpha-1)^{|U|}\right) = (\alpha-1)^{|U|}\lambda_U. \tag{S51}$$

A prior distribution for fitness landscapes

Because for sequences of length ℓ these priors are defined by there 2^ℓ values λ_U, defining and interpreting this large number of the parameters can be challenging. Here, we use the relationship between the $\Delta^{(U)}$ and the projection operators into the U-subspace shown in Eq. S21 to define a simplified prior where the λ_U corresponding to interactions of order higher than P are parametrized via the different $\binom{\ell}{P}$ values of a_U:

$$\lambda_U = \begin{cases} \tilde{\lambda}_U & \text{if } |U| < P \\ \frac{1}{\alpha^{|U|} \sum_{U' \subseteq U} a_{U'}} & \text{if } |U| >= P. \end{cases} \tag{S52}$$

To understand the role of the parameters a_U in the prior, we can write it as

$$\mathrm{p}(f) \propto e^{-f^T K^{-1} f} = e^{-\sum_{U:|U|<P} \frac{1}{\tilde{\lambda}_U} f^T P_U f - \sum_{U:|U|=P} \frac{a_U}{s_U} f^T \Delta^{(U)} f}, \tag{S53}$$

where s_U is the number of local epistatic coefficients between sites in U and the value of a_U modulates how much the prior penalizes these coefficients for a given f. Interestingly, one can see that as $a_U \to \infty$, λ_U approaches 0 for all function subspaces explained by interactions involving the sites in U, essentially removing interactions of order equal or higher than $|U|$ involving the set of sites U. For instance, for $P = 2$, setting $a_{ij} = \infty$ enforces the assumption that there are no genetic interactions between sites i and j or, equivalently, that the effects of mutations at site i never change when introducing an additional mutation at site j. This property allows us to define prior distributions where genetic interactions of order higher than $|U|$ are allowed, but are constrained to specific sets of sites. Similarly, if we set all the $\tilde{\lambda}_U = c$ for $|U| < P$ we see that the first sum in the exponent of Equation S53 becomes proportional to the squared norm of the projection of f into the null space of $\Delta^{(P)}$. Taking the limit $c \to \infty$ is then equivalent to imposing a uniform prior on these directions, and for the special case $P = 2$ we recover Equation 4 in the main text.

Next, we consider the expected average squared epistatic coefficients involving sites U under the prior distribution $f \sim \mathcal{N}(0, K)$, parameterized by the lower-order variances $\tilde{\lambda}_U$ and a_U for $|U| \geq P$. To draw samples from this prior, we can first sample $z \sim \mathcal{N}(0, I)$ and then set $f = K^{1/2} z$, so that $\mathrm{Cov}[f] = K^{1/2}(K^{1/2})^\intercal = K$ (for any matrix square root $K^{1/2}$). One way to define the matrix square root is via its known eigendecomposition $K^{1/2} = \sum_{U \in \mathcal{P}(S)} \sqrt{\lambda_U} P_U$. Then, we consider the average squared epistatic coefficient for any f and compute its expectation when f are drawn from the prior distribution as follows

$$\mathbb{E}_f\left[\frac{1}{s_U} f^T \Delta^{(U)} f\right] = \mathbb{E}_z\left[\frac{1}{s_U} z^T (K^{1/2})^\intercal \Delta^{(U)} K^{1/2} z\right] = \frac{1}{s_U} \mathrm{tr}\left((K^{1/2})^\intercal \Delta^{(U)} K^{1/2}\right). \tag{S54}$$

We compute the $K^{1/2} \Delta^{(U)} K^{1/2}$ product by using the known eigendecompositions of K and $\Delta^{(U)}$:

$$\begin{aligned} \mathbb{E}_f\left[\frac{1}{s_U} f^T \Delta^{(U)} f\right] &= \frac{1}{s_U} \mathrm{tr}\left(\sum_{U': U \subseteq U'} \lambda_{U'} \alpha^U P_U\right) = \frac{1}{s_U} \sum_{U': U \subseteq U'} \lambda_{U'} \alpha^{|U|} \mathrm{tr}(P_{U'}) \\ &= \frac{1}{s_U} \sum_{U': U \subseteq U'} \frac{1}{\alpha^{|U'|} \sum_{U'' \subseteq U'} a_{U''}} \alpha^{|U|} (\alpha - 1)^{|U'|} = \frac{\alpha^{|U|}}{s_U} \sum_{U': U \subseteq U'} \left(\frac{\alpha - 1}{\alpha}\right)^{|U'|} \frac{1}{\sum_{U'' \subseteq U'} a_{U''}} \\ &= \frac{\alpha^{|U|} 2^{|U|}}{\alpha^\ell (\alpha - 1)^{|U|}} \sum_{U': U \subseteq U'} \left(\frac{\alpha - 1}{\alpha}\right)^{|U'|} \frac{1}{\sum_{U'' \subseteq U'} a_{U''}} \\ &= \left(\frac{\alpha^\ell}{2^{|U|}} \left(\frac{\alpha - 1}{\alpha}\right)^{|U|}\right)^{-1} \sum_{U': U \subseteq U'} \left(\frac{\alpha - 1}{\alpha}\right)^{|U'|} \frac{1}{\sum_{U'' \subseteq U'} a_{U''}}. \end{aligned} \tag{S55}$$

Thus, we can see that the expected average squared local epistatic coefficient for sites U does not depend only on the value a_U, but also on all other $a_{U'}$ such that both sets have at least one site in common ($|U \cap U'| > 0$). For instance, if we let $a_U \to \infty$ for a particular U, $\lambda_{U'}$ for all U' that include the whole set of sites in U will be set to zero, decreasing the expected average squared epistatic coefficients for sites involving sites in U'.

Relationship with the Connectedness Model

In this section, we review the Connectedness Model and its relationship to Local Epistasis Regression. The Connectedness Model was first introduced as a Gaussian random field model by Reddy and Desai (2021) to allow different loci to have different probability of being involved in epistatic interactions with mutations at other sites. Then, it was used as a prior distribution in a Gaussian process model, uncovering the set of sites that are more strongly involved in epistatic interactions and using that information for inference of high-dimensional empirical fitness landscapes (Zhou *et al.* 2025). Here, we show that the Connectedness Model can be derived as a particular case of Eq. S48 as a function of the variance associated to the constant component $\tilde{\lambda}_0$ and the variance associated to the additive contribution of each site $\tilde{\lambda}_i$

$$\lambda_U = \prod_{i \in U} \tilde{\lambda}_i \prod_{i \notin U} \tilde{\lambda}_0. \tag{S56}$$

If we plug this into Eq. S48, we can see that the resulting kernel can be expressed as a product of site-specific kernels

$$\begin{aligned} K &= \sum_{U\in\mathcal{P}(S)} \lambda_U P_U = \sum_{U\in\mathcal{P}(S)} \prod_{i\in U} \tilde{\lambda}_i \prod_{i\notin U} \tilde{\lambda}_0 \bigotimes_{i=1}^{\ell} P_{i\in U} = \sum_{U\in\mathcal{P}(S)} \bigotimes_{i=1}^{\ell} \begin{cases} \tilde{\lambda}_0 P_{\text{con}} & i \notin U \\ \tilde{\lambda}_i P_{\text{add}} & i \in U \end{cases} \\ &= \bigotimes_{i=1}^{\ell} (\tilde{\lambda}_0 P_{\text{con}} + \tilde{\lambda}_i P_{\text{add}}) = \bigotimes_{i=1}^{\ell} K_i, \end{aligned} \tag{S57}$$

with entry-wise formula given by

$$k(x,x') = \alpha^{-\ell} \prod_{i:x_i=x'_i} (\tilde{\lambda}_0 + \tilde{\lambda}_i(\alpha-1)) \prod_{i:x_i\neq x'_i} (\tilde{\lambda}_0 - \tilde{\lambda}_i). \tag{S58}$$

This kernel can be reparametrized as a function of the prior variance σ^2 and the correlation under this kernel parametrized by $\mu_i = \tilde{\lambda}_i/\tilde{\lambda}_0$ assuming $\tilde{\lambda}_0 > 0$ as in Zhou *et al.* (2025):

$$k(x,x') = \sigma^2 \prod_{i:x_i\neq x'_i} \frac{1-\mu_i}{1+\mu_i(\alpha-1)}. \tag{S59}$$

This construction shows that the variance explained by epistatic interactions under the Connectedness Model is completely specified by the variance explained by the additive contribution of each individual site. In contrast, Local Epistasis Regression allows arbitrary relationships between the additive and pairwise contribution of individual sites and pairs of sites, but the variance for higher-order interactions is fully specified by the the variance associated to pairwise interactions between specific pairs of sites (Eq. S52). These models also make different assumptions on how the variances associated to lower-order interactions combine to specify the variances associated to higher-order interactions. In particular, in the Connectedness Model these variances combine multiplicatively (Eq. S56), whereas in Local Epistasis Regression they are proportional to the harmonic mean of the variances associated to pairwise interactions between each possible pair of sites in U

$$\lambda_U = \frac{1}{\sum_{i<j\in U} \frac{1}{\tilde{\lambda}_{ij}}}, \tag{S60}$$

where $\tilde{\lambda}_{ij} = \frac{1}{\alpha^2 a_{ij}}$.

Inference of fitness landscapes under the prior

In this section, we review how to do Gaussian process inference of a complete fitness landscape f from high throughput experimental data (Martí-Gómez *et al.* 2026b). We start by defining a Gaussian prior distribution over the space of possible fitness landscapes $\mathrm{p}(f) \sim \mathcal{N}(0,K)$ that assigns higher probability to fitness landscapes that we believe are more plausible *a priori*.

Let y be an n-dimensional vector of measurements with known experimental Gaussian error given by the variance n-dimensional vector y_{var} for a subset of $n \leq \alpha^\ell$ sequences X. As both the prior distribution and the likelihood function are Gaussian, the posterior distribution is also multivariate Gaussian with closed form analytical solution (Rasmussen and Williams 2008) for the mean

$$\widehat{f} = K_{*X}(K_{XX} + D_{var})^{-1} y \tag{S61}$$

and covariance matrix

$$K - K_{*X}(K_{XX} + D_{var})^{-1} K_{X*}, \tag{S62}$$

where K_{XX}, K_{*X}, K_{X*} are submatrices of K indexed by sequences X and $*$, where $*$ represents all possible sequences, and D_{var} is a diagonal matrix with the known experimental variances y_{var} along the diagonal.

Despite the simplicity of the solution, practical evaluation of these expressions becomes challenging as the number of observations increases. Traditional approaches rely to computation of the Cholesky decomposition of the $K_{XX} + D_{var}$ matrix to then use efficient triangular solves to compute the solutions to the linear systems rather than using direct matrix inversion for higher numerical stability. However, the algorithm for computing this decomposition is $O(n^3)$ and cannot be parallelized, which has traditionally limited the applicability of Gaussian process models to datasets with at most few thousand data points (Rasmussen and Williams 2008). In previous work, we have circumvented this limitation by leveraging the mathematical properties of the specific precision or kernel matrices, which could be expressed as polynomials in the Laplacian of the Hamming graph representing sequence space (Zhou and McCandlish 2020; Zhou *et al.* 2022; Chen *et al.* 2021; Martí-Gómez *et al.* 2026b). This property allowed us to encode these matrices as linear operators that allow computing matrix-vector products efficiently without explicitly storing them in memory, and use these operators together with iterative methods for solving systems of linear equations to scale these methods up to a few million data points.

Here we use a similar strategy by finding a representation of the kernel matrices that enables efficient computation of matrix-vector products without explicitly constructing them in memory, even if the kernel matrices presented here cannot be represented as polynomials in the Laplacian of the Hamming graph.

In the case of the Connectedness Model (Zhou *et al.* 2025), the kernel can in fact be expressed as a Kronecker product of ℓ site-specific $\alpha \times \alpha$ matrices (Eq. S57), which enables fast computation of matrix-vector products by leveraging the mixed-product property, which reduces the complexity to that of computing ℓ products of an $\alpha \times \alpha$ matrix with an $\alpha \times \alpha^{\ell-1}$ matrix (Martí-Gómez *et al.* 2026b,a).

In the case of Local Epistasis Regression, the kernel matrix can be expressed as the sum of 2^ℓ matrices, each of which is Kronecker factorizable

$$K = \sum_{U \in \mathcal{P}(S)} \lambda_U \bigotimes_{i=1}^{\ell} P_{i \in U}. \tag{S63}$$

While this enables computation of matrix-vector products with a total of $2^\ell \ell$ $\alpha \times \alpha$ by $\alpha \times \alpha^{\ell-1}$ matrix products, the computational burden of these calculations is much larger compared with previous approaches e.g. requires 2^ℓ times the computation needed in the Connecntedness model. However, here we note that all of the 2^ℓ kernel matrices represent different combinations of only two different Kronecker factors P_{con} and P_{add}, which allow us to re-use parts of the computation. In particular, we can decompose any P_U as follows

$$P_U = \bigotimes_{i=1}^{\ell} P_{i \in U} = (P_{1 \in U} \otimes I \otimes \cdots \otimes I)(I \otimes P_{2 \in U} \otimes \cdots \otimes I) \ldots (I \otimes I \otimes \cdots \otimes P_{\ell \in U}) = M_{1 \in U} M_{2 \in U} \ldots M_{\ell \in U}. \tag{S64}$$

While the M matrices are still Kronecker products of ℓ matrices, $\ell - 1$ of the factors correspond to the identity matrix and thus leave the matrices they act on unchanged, reducing the computation to a single $\alpha \times \alpha$ by $\alpha \times \alpha^{\ell-1}$ matrix products. Importantly, P_U matrices differing only at the factor at the first site can be computed with a single additional operation from the same intermediate result

$$P_U v = M_{1 \in U}(M_{2 \in U} \ldots M_{\ell \in U} v). \tag{S65}$$

The intermediate results can also be computed in the same fashion

$$M_{2 \in U} M_{3 \in U} \ldots M_{\ell \in U} v = M_{2 \in U}(M_{3 \in U} \ldots M_{\ell \in U} v) \tag{S66}$$

so that the computation can be shared with products of the same vector with other $P_{U'}$ for a different subset of sites U'. These computational dependencies can be represented by a bifurcating tree where nodes represent α^ℓ-dimensional vectors and edges represent matrix-vector products with specific $M_{i \in U}$ matrices. The vector v is located at the root of the tree, allowing computation of the intermediate vectors at each of the nodes of the tree up to the tips containing all the $P_U v$ for every possible U. This algorithm reduces the total number of operations from $2^\ell \ell$ to $2^\ell \sum_{i=0}^{\ell} 2^{-i}$. This series converges relatively fast to $2^{\ell+1}$, resulting in an approximate $\ell/2$-fold speedup even when ℓ is small. For instance, for $\ell = 8$, the number of matrix-matrix products goes from $8 \times 2^8 = 2048$ under the naive approach to 510, nearly achieving the expected 4-fold increase in computational efficiency.

Hyperparameter optimization

In order to infer a fitness landscape under a given prior distribution, we must first choose the parameters that define the properties of the prior, also known as hyperparameters. Here, we generally consider prior distributions defined over the space of possible fitness landscapes f defined by a kernel function $k(x, x')$ that returns the covariance for any pair of sequences x and x' given by

$$k(x, x') = \sum_{U \in \mathcal{P}(S)} \lambda_U P_U(x, x'), \tag{S67}$$

where $P_U(x, x')$ is the covariance due to interactions between exactly U sites between sequences x and x', which depends only on the combination of sites at which they differ $P_U(x, x') = w_U(H(x, x'))$; and where the parameters λ_U can be free or a function of a smaller set of parameters generally called θ ($\lambda_U = g(\theta)$) e.g. Eq. S52 in Local Epistasis Regression and Eq. S56 in the Connectedness Model.

In this work, we use a strategy known as kernel alignment (Wang *et al.* 2015; Zhou *et al.* 2022) or Haseman-Elston regression (Haseman and Elston 1972), in which the prior covariance approximates as closely as possible the patterns of covariance in the empirical data. Specifically, this is done by finding the parameter values $\widehat{\theta}$ that minimize the Frobenius norm of the difference between the prior predictive covariance $K_{XX} + D_{var}$ and the empirical second moment matrix yy^T:

$$\widehat{\theta} = \arg\min_{\theta} \left\| yy^T - (K_{XX}(\theta) + D_{var}) \right\|_F^2. \tag{S68}$$

Naively solving this minimization problem is challenging, as we need to work with $n \times n$ matrices, where the number of measured sequences n can be in the order of hundreds of thousands to millions. However, as the prior covariance between two sequences only depends on the set of sites at which they differ, the dimensionality of the problem can be reduced to a more manageable 2^ℓ-dimensional weighted least squares problem as follows:

$$\begin{aligned}
\left\| yy^T - (K_{XX}(\theta) + D_{var}) \right\|_F^2 &= \sum_{x \in X} \left[(y_x^2 - y_{x,var}) - k_\theta(\varnothing) \right]^2 + \sum_{D \neq set} \sum_{x,x' : H(x,x') = D} \left[y_x y_{x'} - k_\theta(D) \right]^2 \\
&= N_\varnothing \left[t(\varnothing) - k_\theta(\varnothing) \right]^2 + \sum_{x \in X} \left[(y_x^2 - y_{x,var}) - t(\varnothing) \right]^2 + \sum_{D \neq \varnothing} N_D \left[t(D) - k_\theta(D) \right]^2 + \sum_{D \neq \varnothing} \sum_{x,x' : H(x,x') = D} \left[y_x y_{x'} - t(D) \right]^2 \\
&= \sum_{D} N_D \left[t(D) - k_\theta(D) \right]^2 + \sum_{x \in X} \left[(y_x^2 - y_{x,var}) - t(\varnothing) \right]^2 + \sum_{D \neq \varnothing} \sum_{x,x' : H(x,x') = D} \left[y_x y_{x'} - t(D) \right]^2.
\end{aligned} \tag{S69}$$

Since the last two terms are independent of θ, we can find by optimal hyperparameter values $\widehat{\theta}$ simply as

$$\widehat{\theta} = \arg\min_{\theta} \sum_{D \in \mathcal{P}(S)} N_D \left[t(D) - \sum_{U \in \mathcal{P}(S)} \lambda_U(\theta) w_U(D) \right]^2, \tag{S70}$$

where $t(D)$ corresponds to the second moment, related with the empirical autocovariance function $c(D)$ ($t(D) = c(D) + \bar{y}^2$) given by

$$c(D) = \begin{cases} \frac{1}{n} \sum_{x \in X} (y_x - \bar{y})^2 - \overline{y_{var}} & D = \emptyset \\ \frac{1}{N_D} \sum_{x,x' \in X : H(x,x') = D} (y_x - \bar{y})(y_{x'} - \bar{y}) & D \neq \emptyset. \end{cases} \tag{S71}$$

Supplementary figures

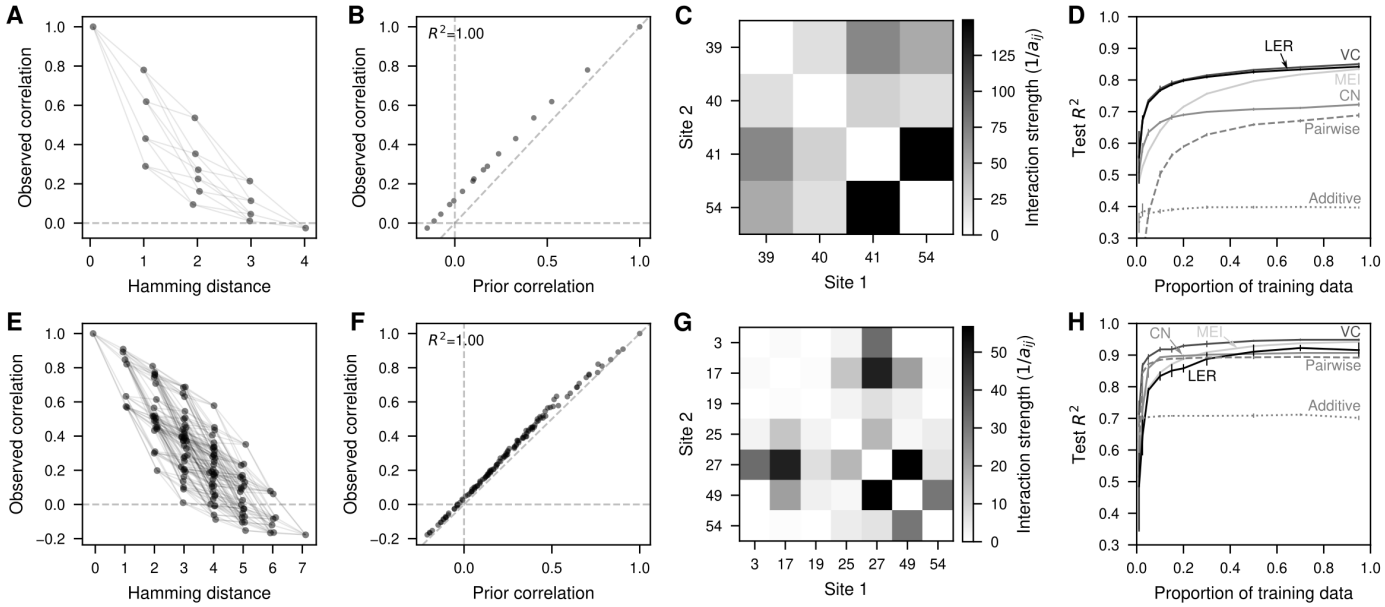

Figure S1 Application of Local Epistasis Regression to protein datasets. (A,E) Correlation in the measured fitness values for pairs of sequences differing at each possible subset of sites D arranged according to the Hamming distance $d = |D|$. Each dot represents a single distance class D and are joined by lines whenever the distance classes differ by a single position from each other. (B,F) Comparison of the observed correlation values in the data and the values under the estimated prior ones using Local Epistasis Regression for every possible distance class D (each dot represents a different D). Correlations were estimated using 80% of the data for training. (C,G) Heatmap representing the inferred model hyperparameters as $1/a_{ij}$ for every pair of sites i, j highlighting the patterns of genetic interactions across sites under the prior. (D,H) Predictive performance evaluated by the R^2 between the predicted and the measured fitness of held-out test sequences when using different amounts of training data for different models (MEI: Minimum Epistasis Interpolation, VC: Variance Component regression, CN: Connectedness Model regression, LER: Local Epistasis Regression). Predicted values are the maximum a posteriori estimate given by each method, which is equal to the posterior mean $\widehat{f}$. Error bars represent the standard deviation across 3 different random samples for each fraction of training data. Each row represents a fitness landscape: GB1 (A,B,C,D); FYN-SH3 (E,F,G,H).

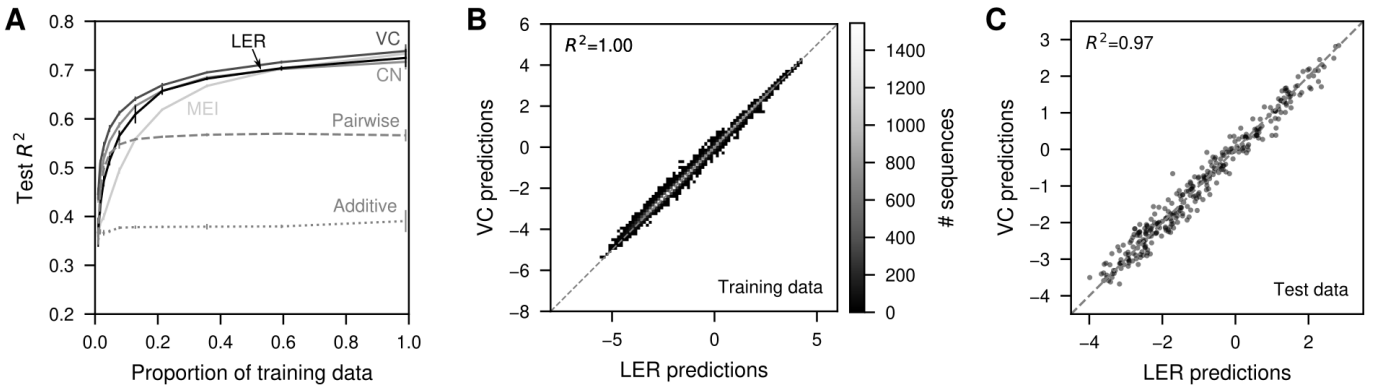

Figure S2 Comparison of model predictions on the self-splicing intron dataset. (A) Predictive performance evaluated by the R^2 between the predicted and the true fitness values of held-out test sequences when using different amounts of training data for different models (MEI: Minimum Epistasis Interpolation, VC: Variance Component regression, CN: Connectedness Model regression, LER: Local Epistasis Regression). Predicted values are the maximum a posteriori estimate given by each method, which is equal to the posterior mean $\widehat{f}$. Error bars represent the standard deviation across 3 different random samples for each fraction of training data. (B) 2D histogram comparing the fitness landscape reconstructions of the self-splicing intron dataset under Local Epistasis Regression (LER) and Variance Component regression (VC). (C) Scatterplot comparing the predictions in held-out sequences of the self-splicing intron dataset under Local Epistasis Regression (LER) and Variance Component Regression (VC).

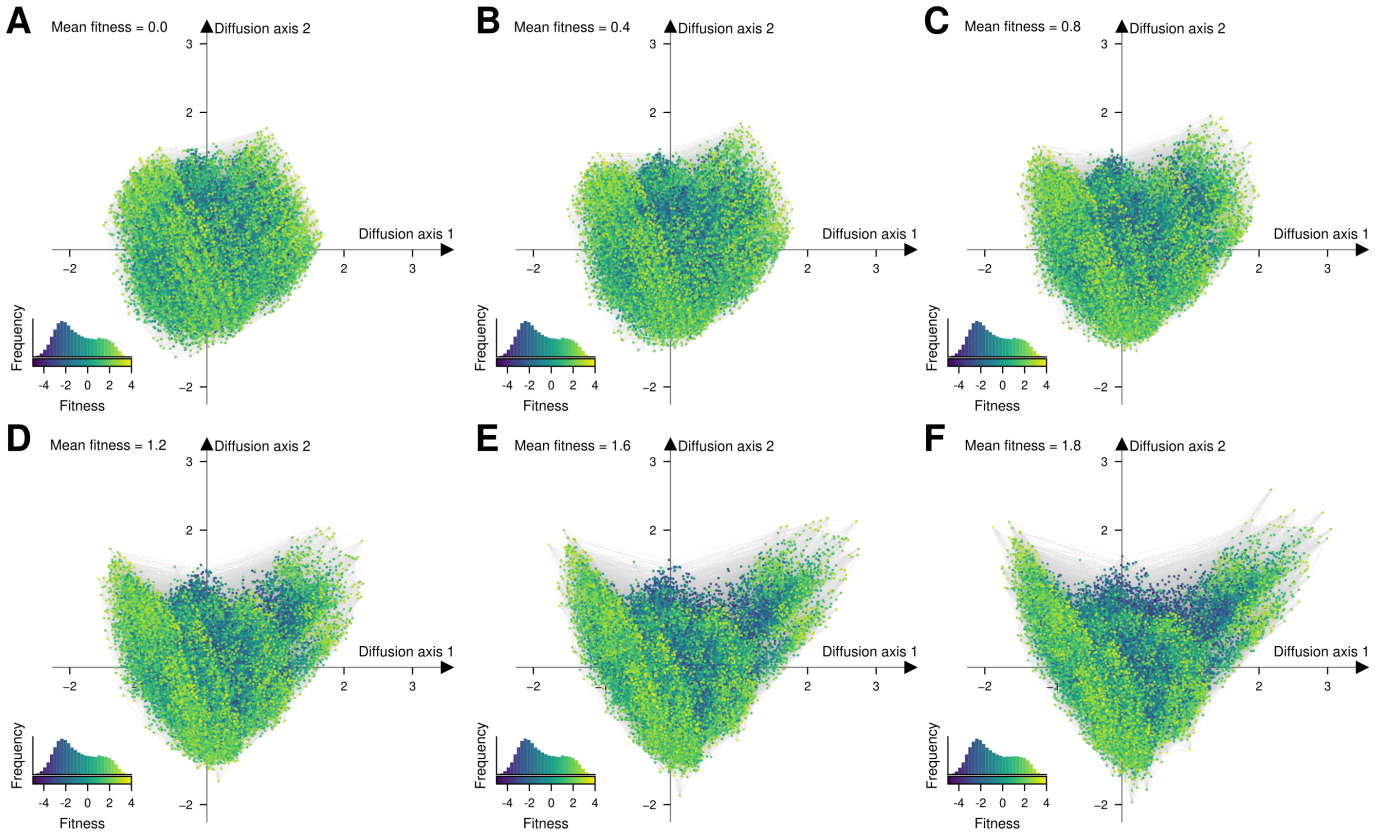

Figure S3 Visualization of the inferred fitness landscape using Local Epistasis Regression under different strengths of selection, where here we quantify the strength of selection by the mean fitness achieved at stationarity, i.e. under long-term purifying selection. Every dot represents one of the possible 4^8 possible sequences and is colored according to the predicted fitness. The inset represents the phenotypic distribution along with their corresponding color in the map. Sequences are laid out according to the first two Diffusion axes and dots are plotted in order according to Diffusion axis 3.

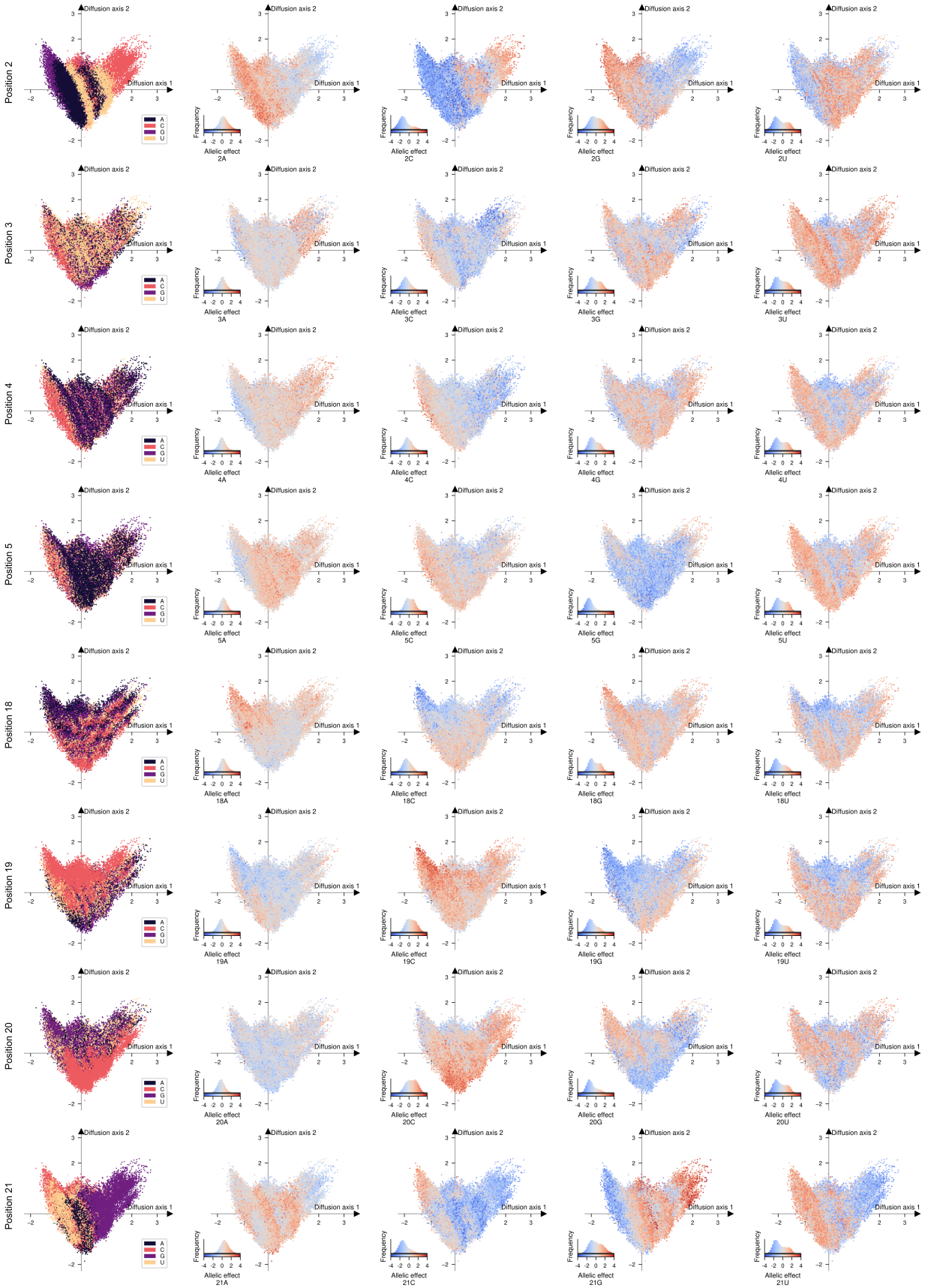

Figure S4 Visualizing alleles and allelic preferences across the fitness landscape visualization. Visualization of the inferred fitness landscape using Local Epistasis Regression. Every dot represents one of the possible 4^8 possible sequences and is colored according to the allele (first column) or the difference in the fitness of the sequence obtained when placing a specific allele at an specifc position relative to the average fitness of the four possible alleles (four last columns). The inset represents the allelic effect distribution along with their corresponding color in the map. Sequences are laid out according to the first two Diffusion axes and dots are plotted in order according to Diffusion axis 3.