Inference and visualization of complex genotype-phenotype maps with *gpmap-tools*

Carlos Martí-Gómez[1], Juannan Zhou[2], Wei-Chia Chen[3], Arlin Stoltzfus[4], Justin B. Kinney[1], and David M. McCandlish[1]

[1]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724
[2]Department of Biology, University of Florida, Gainesville, FL, 32611
[3]Department of Physics, National Chung Cheng University, Chiayi 62102, Taiwan, Republic of China
[4]Institute for Bioscience and Biotechnology Research, Rockville, USA

Understanding how biological sequences give rise to observable traits, that is, how genotype maps to phenotype, is a central goal in biology. Yet our knowledge of genotype-phenotype maps in natural systems is limited due to the high dimensionality of sequence space and the context-dependent effects of mutations. The emergence of Multiplex assays of variant effect (MAVEs), along with large collections of natural sequences, offer new opportunities to empirically characterize these maps at an unprecedented scale. However, tools for statistical and exploratory analysis of these high-dimensional data are still needed. To address this gap, we developed *gpmap-tools* (https://github.com/cmarti/gpmap-tools), a *python* library that integrates a series of models for inference, phenotypic imputation, and error estimation from MAVE data or collections of natural sequences in the presence of genetic interactions of every possible order. *gpmap-tools* also provides methods for summarizing patterns of epistasis and visualization of genotype-phenotype maps containing up to millions of genotypes. To demonstrate its utility, we used *gpmap-tools* to infer genotype-phenotype maps containing 262,144 variants of the Shine-Dalgarno sequence from both genomic 5'UTR sequences and experimental MAVE data. Visualization of the inferred landscapes consistently revealed high-fitness ridges that link core motifs at different distances from the start codon. In summary, *gpmap-tools* provides a flexible, interpretable framework for studying complex genotype-phenotype maps, opening new avenues for understanding the architecture of genetic interactions and their evolutionary consequences.

genotype-phenotype map | fitness landscape | epistasis | Gaussian process | Shine-Dalgarno sequence

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

Introduction

The genotype-phenotype map describes how changes in biological sequences, such as DNA, RNA or proteins, give rise to variation in observable traits. Understanding this relationship is essential across many areas of biology, ranging from evolutionary theory (Wright, 1932; Kondrashov *et al.*, 2002; Phillips, 2008; Weinreich *et al.*, 2013; De Visser and Krug, 2014; Sailer and Harms, 2017; Bank, 2022; Johnson *et al.*, 2023) and human disease (Moore and Williams, 2009; Dasari *et al.*, 2021; Moulana *et al.*, 2023a), to synthetic biology, protein engineering applications (Yang *et al.*, 2019; Freschlin *et al.*, 2022; Lipsh-Sokolik and Fleishman, 2024) and plant and animal breeding (De Los Campos *et al.*, 2013; Sackton and Hartl, 2016; Soyk *et al.*, 2020; Dwivedi *et al.*, 2024). However, our current understanding of genotype-phenotype maps in nature remains limited due to two fundamental challenges. First, the number of possible genotypes is astronomically large, making it impossible to explore more than a tiny fraction of this space in any empirical setting. Second, the phenotypic effect of a mutation often depends on the genetic background in which it occurs, a phenomenon known as epistasis (Starr and Thornton, 2016; Domingo *et al.*, 2019; Miton *et al.*, 2021; Bank, 2022). Addressing this context dependence requires experimental and computational approaches that are capable of capturing these complex genetic interactions.

One powerful approach to study empirical genotype-phenotype maps is to experimentally construct sequence variants and measure their functional consequences. Historically, these studies were limited by the difficulty of engineering a large number of genetic variants and quantifying their phenotypes at scale, restricting most empirical genotype-phenotype maps to small numbers of genotypes, typically ranging from tens to a few hundreds (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). The development of Multiplex Assays of Variant Effects (MAVEs) (Kinney *et al.*, 2010; Fowler and Fields, 2014; Kinney and McCandlish, 2019) has increased our phenotyping throughput by several orders of magnitude, enabling the simultaneous measurement of libraries containing thousands to millions of genotypes in a single experiment. These techniques have been used to characterize the phenotypic landscapes for short regulatory elements (Noderer *et al.*, 2014; Bonde *et al.*, 2016; Wong *et al.*, 2018; Kuo *et al.*, 2020; Komarova *et al.*, 2020; Westmann *et al.*, 2024b,a; Kuo *et al.*, 2025; Chattopadhyay *et al.*, 2025), RNAs (Domingo *et al.*, 2018; Baeza-Centurion *et al.*, 2019; Bendixsen *et al.*, 2019; Soo *et al.*, 2021; Rotrattanadumrong and Yokobayashi, 2022) and 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; Moulana *et al.*, 2023b; Papkou *et al.*, 2023; Sundar *et al.*, 2024; Zarin and Lehner, 2024; Escobedo *et al.*, 2024; Johnston *et al.*, 2024; Herrera-Álvarez *et al.*, 2025), and combinatorial gene interactions (Bakerlee *et al.*, 2022). Yet the highly combinatorial nature of these data poses significant challenges, and accurate inference typically relies on complex latent-variable models (Bloom, 2015; Otwinowski *et al.*, 2018; Tareen *et al.*,

2020; Tonner *et al.*, 2022; Faure and Lehner, 2024) or neural networks (Bryant *et al.*, 2021; Gelman *et al.*, 2021; Sethi and Zhou, 2024). Gaussian processes offer an alternative class of flexible models that both capture high order interactions and provide accurate uncertainty quantification (Romero *et al.*, 2013; Yang *et al.*, 2019; Zhou and McCandlish, 2020; Zhou *et al.*, 2022, 2025; Petti *et al.*, 2025). Moreover, the mathematical tractability of Gaussian processes means that they can be designed and interpreted in terms of existing genetic concepts, for example by explicitly learning the variance explained by epistatic interactions of different orders, which was shown to allow accurate inference and statistical analysis of complex genotype-phenotype maps from MAVE data (Zhou *et al.*, 2022).

An alternative approach to study genotype-phenotype maps consists in analyzing collections of natural sequences. Since natural selection tends to preserve functional sequences, we can assume that the probability of observing a given sequence in nature depends on how well it performs its function. Thus, the probability distribution over sequences with a shared function can be interpreted as a genotype-phenotype map, where the phenotype is the probability of observing a sequence. Independent site models, such as Position-Weight Matrices (PWMs), estimate this sequence probability distribution by assuming that positions are independent from each other (Stormo, 2013), whereas pairwise interaction models, also known as Potts models, relax this assumption by allowing interactions between pairs of positions (Sly, 2011; Morcos *et al.*, 2011; Ekeberg *et al.*, 2013; Stein *et al.*, 2015; Haldane and Levy, 2021). These models have proven very effective in predicting structural contacts in proteins (Marks *et al.*, 2012; Haldane *et al.*, 2016, 2018) and between proteins (Bitbol *et al.*, 2016; Malinverni and Babu, 2023), as well as for predicting the effects of mutations in human proteins (Hopf *et al.*, 2017). Pairwise interaction models have also been useful for identifying novel functional proteins (Russ *et al.*, 2020) and regulatory sequences (Yeo and Burge, 2004), and for quantifying the strength of selection at the gene level (Vigué and Tenaillon, 2023). A recently proposed Gaussian process model further generalizes these methods by inferring sequence probability distributions under a prior that controls the magnitude of local epistatic coefficients (Chen *et al.*, 2021, 2024). These Gaussian process based generalizations of pairwise interaction models open the opportunity to study complex genotype-phenotype maps containing higher-order epistatic interactions using readily available collections of natural sequences.

Another important challenge is the interpretation of complex genotype-phenotype maps. One way to develop an intuitive understanding of complex datasets is through data visualization tools. An approach to visualizing empirical genotype-phenotype maps is to embed the Hamming graph representing it, where each node is a genotype and edges represent single-point mutations, into a low-dimensional space. For instance, one can embed the graph by placing genotypes according to their Hamming distance to a reference sequence on one axis and their phenotype on the other (Wright, 1932; Brouillet *et al.*, 2015; Domingo *et al.*, 2018; Bendixsen *et al.*, 2019; Baeza-Centurion *et al.*, 2019; Escobedo *et al.*, 2024), or applying spectral and force-directed layouts (Starr *et al.*, 2017; Fragata *et al.*, 2019; Martin and Ahnert, 2022; Herrera-Álvarez *et al.*, 2025). However, how these representations relate to the conceptual framework of fitness peaks, valleys, and plateaus, which has shaped much of our theoretical understanding of genotype-phenotype maps (Wright, 1932; de Visser *et al.*, 2018), remains unclear. A different strategy is to construct a low-dimensional representation that reflects the evolutionary dynamics induced by the genotype-phenotype map of interest. This can be done, for example, by having the distances between genotypes represent the expected time to evolve between them for a population under selection for high phenotypic values (McCandlish, 2011). This property is very useful because it places sets of functional sequences that are inaccessible to each other i.e. peaks, far apart in the visualization, naturally displaying the key genetic interactions separating them, i.e. valleys. This technique has been successfully applied to uncover qualitative features of multiple genotype-phenotype maps (Zhou and McCandlish, 2020; Zhou *et al.*, 2022; Chen *et al.*, 2021; Weinstein *et al.*, 2023; Avizemer *et al.*, 2025), as well as to study how the structure of the genetic code influences protein evolution (Rozhoňová *et al.*, 2024), illustrating its potential as a general framework for interpreting and comparing complex genotype-phenotype relationships.

Here, we present *gpmap-tools*, a *python* library that offers an integrated interface to methods for inference, statistical analysis and visualization of large, complex genotype-phenotype maps (Figure 1). Among its new features and other improvements, *gpmap-tools* is built on a new computational back-end that represents large matrices as linear operators, enabling memory-efficient computation. This design also allows users to simulate large genotype-phenotype maps with different types and magnitudes of epistasis, and to perform statistical analysis of quantities of interest, such as mutational effects and epistatic coefficients, even in the presence of missing data and experimental noise. *gpmap-tools* also enables calculation of the variance explained by genetic interactions involving specific sets of sites, providing a powerful tool to characterize the structure and complexity of genetic interactions. Finally, *gpmap-tools* provides an extended interface for visualizing genotype-phenotype maps with varying numbers of alleles per site, new tools for investigating the sequence features that distinguish different regions of the representation, and accelerated rendering of plots containing millions of genotypes, enabling exploration of complex genotype-phenotype maps at unprecedented scale and resolution.

We demonstrate the capabilities of *gpmap-tools* by inferring the fitness landscape of the Shine-Dalgarno (SD) sequence from two fundamentally different types of data: (i) natural sequence diversity in genomic 5' untranslated regions (UTRs) and (ii) MAVE data (Kuo *et al.*, 2020). The inferred landscapes reveal a shared structure consisting of peaks corresponding to 16S rRNA binding at different distances rel-

ative to the start codon. These peaks are connected by extended ridges of functional sequences when the corresponding binding sites are spaced three nucleotides apart. These ridges arise from overlapping SD motifs in the same sequence, a consequence of the motif's quasi-repetitive structure, which allows new binding sites to emerge at offset positions without disrupting functionality. Building on this qualitative understanding of the genotype-phenotype map, we fit a simplified mechanistic model with parameters that have clear biophysical interpretations. This model allows us to disentangle the effects of mutations on binding at different registers relative to the start codon *in vivo*, while capturing the key structural features of the empirical landscape. Taken together, our analysis illustrates how *gpmap-tools* enables the inference and characterization of genotype-phenotype maps from diverse data sources, facilitating the discovery of simple molecular mechanisms capable of generating the observed architecture of epistatic interactions and offering insights into the evolutionary consequences induced by these complex genotype-phenotype maps.

Approach

In this section, we provide a brief technical overview of the methods for inference and interpretation of genotype-phenotype maps implemented in *gpmap-tools* (Figure 1), for application see Results.

A genotype-phenotype map is a function that assigns a phenotype, typically a scalar value, to every possible sequence of length ℓ on α alleles (where e.g. $\alpha = 4$ for DNA and $\alpha = 20$ for proteins). This function can be represented by an α^ℓ-dimensional vector f containing the phenotype for every possible genotype. We begin by discussing several methods for quantifying the amount and type of epistasis present in any particular vector f.

Epistasis in genotype-phenotype maps. *gpmap-tools* implements two different methods for measuring the amount and pattern of epistasis in a given genotype-phenotype map: one based on the typical magnitude of local epistatic coefficients across all possible subsets of mutations and the other based on the proportion of phenotypic variance explained by genetic interactions of different orders or involving specific subsets of sites.

Local epistatic coefficients. The traditional epistatic coefficient quantifies how much the effect of a mutation $A \to a$ in one site changes in the presence of an additional mutation $B \to b$ in a second site in an otherwise identical genetic background C:

$$\epsilon = (f_{aBC} - f_{ABC}) - (f_{abC} - f_{AbC}).$$

The average squared epistatic coefficient $\overline{\epsilon^2}$ across all possible pairs of mutations and genetic backgrounds provides a measure of the variability in mutational effects between neighboring genotypes across the whole genotype-phenotype map (Zhou and McCandlish, 2020). $\overline{\epsilon^2}$ can be efficiently calculated as a quadratic form $\overline{\epsilon^2} = \frac{1}{s} f^T \Delta^{(2)} f$, where s is the number of epistatic coefficients and $\Delta^{(2)}$ is a previously described sparse $\alpha^\ell \times \alpha^\ell$ positive semi-definite matrix (Zhou and McCandlish, 2020). This statistic can be generalized to characterize the typical size of local P-way epistatic interactions (Chen *et al.*, 2021) and to a setting in which alleles are naturally ordered e.g. copy number (Chen *et al.*, 2024).

Variance components. A genotype-phenotype map f can be decomposed into the contributions of $\ell + 1$ orthogonal subspaces $f = \sum_k f_k$, where f_k represents a function containing epistatic interactions solely of order k. These orthogonal components f_k are obtained by projecting f onto the k-th order subspace using the projection matrix P_k. This decomposition enables quantification of the variance explained by interactions of different orders, providing a global summary of the complexity of genetic interactions in a genotype-phenotype map (Stadler, 1996; Happel and Stadler, 1996; Stadler and Happel, 1999; Stadler, 2002; Zhou *et al.*, 2022).

Here we show that each f_k can be further decomposed into the contribution of $\binom{\ell}{k}$ smaller orthogonal subspaces $f_k = \sum_{U:|U|=k} f_U$, where f_U represents a function containing genetic interactions only among the k sites in U. These orthogonal components f_U are obtained by projecting the function f into the corresponding subspace using the orthogonal projection matrix P_U given by:

$$P_U(x,x') = \alpha^{-\ell} \prod_{\substack{p \in U \\ x_p = x'_p}} (\alpha - 1) \prod_{\substack{p \in U \\ x_p \neq x'_p}} (-1),$$

for any pair of sequences x, x' (see Supplementary Information A). A similar decomposition was shown for certain models of sequence-function relationships parametrized using a specific scheme of interaction terms (Park *et al.*, 2024; Posfai *et al.*, 2025), but this formulation allows the direct decomposition of the function f rather than relying on a particular parameterization.

These projection matrices allow us to quantify not only the variance explained by interactions of different order, but also the variance explained by all possible subsets of sites U. Although there are 2^ℓ such subsets in total, they can be aggregated to yield informative, low-dimensional summary statistics. For instance, we can compute the variance explained by order-k epistatic interactions involving a specific site or pair of sites, or the variance explained by interactions of all orders involving each pair of sites (see Supplementary Information B) (Crawford *et al.*, 2017; Reddy and Desai, 2021). Through these capabilities, *gpmap-tools* enables a fine-grained decomposition of epistatic variance, offering new insights into the structure and complexity of genetic interactions across sites.

Gaussian process inference of genotype-phenotype maps. Gaussian process models are a class of Bayesian non-parametric models that place a multivariate Gaussian prior distribution over all possible functions and compute the posterior distribution given observed data (Rasmussen and

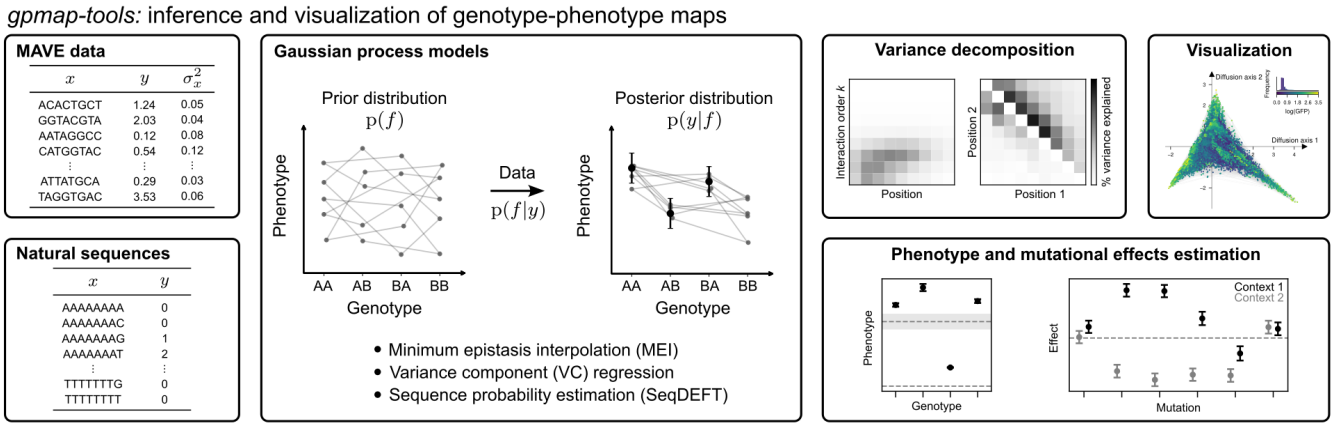

Fig. 1. Overview of the functionality provided by *gpmap-tools*. The software uses data from Multiplex Assays of Variant Effects (MAVEs) or natural sequence variation to fit Gaussian process models for inference of empirical genotype-phenotype maps. It enables interpretation of the inferred maps by decomposing phenotypic variance into components associated with interactions of different orders and specific subsets of sites, visualizing the full genotype-phenotype landscape, and computing posterior distributions for specific genotypes or mutational effects of interest.

Williams, 2008). In our case, we assign a zero-mean Gaussian prior distribution over genotype-phenotype maps $\mathrm{p}(f)$ characterized by either its covariance matrix K or precision matrix C. The covariance matrix K is most often defined through a kernel function that returns the prior covariance between any pair of sequences. Then, given some data y and using a likelihood function $\mathrm{p}(y|f)$, we update the probability distribution of plausible genotype-phenotype maps to be consistent with these observations by computing the posterior distribution $\mathrm{p}(f|y)$.

Interpretable priors. *gpmap-tools* implements two families of priors based on the two approaches to quantify epistasis in genotype-phenotype maps described above: one family that is defined in terms of local epistatic coefficients and a second that is defined in terms of variance components. The first prior is parametrized by its precision matrix $C = \frac{a}{s}\Delta^{(P)}$ and assigns a prior probability to f depending on its average squared epistatic coefficient of order P, i.e. $\log \mathrm{p}(f|a) \propto -\frac{a}{2s} f^T \Delta^{(P)} f$ (Zhou and McCandlish, 2020; Chen *et al.*, 2021). This prior implicitly leaves genetic interactions of order $k < P$ unconstrained, and hence correspond to the use of an improper Gaussian prior. For examples, for $P = 2$ additive effects are not penalized, for $P = 3$ additive and pairwise effects are not penalized, etc. For fixed P, this family of priors has a single hyperparameter a that is inversely proportional to the expected average squared local epistatic coefficient under the prior. As $a \to 0$, we assign the same prior probability to every possible genotype-phenotype map. On the other hand, as $a \to \infty$, we decrease the prior probability of genotype-phenotype maps with non-zero local P-epistatic coefficients (Chen *et al.*, 2021).

The second family of priors are the variance component priors, which are parametrized by their covariance matrix $K = \sum_{k=0}^{\ell} \lambda_k K_k$, where K_k represents the covariance matrix for genotype-phenotype maps with only kth order interactions. The $\ell + 1$ hyperparameters λ_k control the variance explained by genetic interactions of order k (Neidhart *et al.*, 2013) and equivalently the decay in the predictability of mutational effects and epistatic coefficients as the number of mutations separating two genetic backgrounds increases (Zhou *et al.*, 2022). The formal relationship between the two sets of priors is that the priors based on the $\Delta^{(P)}$ operators can be obtained as limits of the variance component prior (Zhou *et al.*, 2022).

These prior distributions for f have hyperparameters with clear biological interpretations in terms of the expected magnitude and type of epistasis. This allows users to define the corresponding priors in a principled and interpretable way. Moreover, under the assumption that the structure of epistasis observed in the data generalizes to the full genotype-phenotype map, *gpmap-tools* can infer these hyperparameters using either cross-validation or kernel alignment (Rasmussen and Williams, 2008; Wang *et al.*, 2015), providing estimates that are both data-driven and biologically meaningful.

Likelihood functions. *gpmap-tools* implements two likelihood functions for inference of genotype-phenotype maps from different types of data. For experimental data, the vector y contains the measurements associated to a subset of sequences x with known Gaussian measurement variance σ_x^2. Thus, the likelihood function is given by

$$\mathrm{p}(y|f) = \mathrm{N}(f_x, D_{\sigma_x^2}),$$

where $D_{\sigma_x^2}$ is a diagonal matrix with σ_x^2 along the diagonal.

For observations of natural sequences, data consists of the number of times N_i a given sequence i was observed out of a total of $N_T = \sum_i N_i$ observations. In this case, the likelihood function is given by the multinomial distribution

$$\mathrm{p}(N|\pi, N_T) = \mathrm{Multinomial}(\pi, N_T),$$

where π is the vector representing the sequence probability distribution, such that π_i is corresponds to the probability of observing sequence i.

Posterior distributions. *gpmap-tools* enables the computation of the posterior distribution over the space of possible

genotype-phenotype maps using both Gaussian and multinomial likelihood functions. Under a Gaussian likelihood, the posterior distribution is also a Gaussian $\mathrm{p}(f|y, D_{\sigma^2}) = \mathrm{N}(\hat{f}, \Sigma)$ with closed form analytical solutions for the mean $\hat{f}$ and covariance matrix Σ. *gpmap-tools* implements the classical solution expressed in terms of the prior covariance matrix K (Rasmussen and Williams, 2008) but also the solution when the prior is defined by its precision matrix C, which are equivalent when $C = K^{-1}$ (see Supplementary Information F).

Under non-Gaussian likelihood functions, such as the multinomial likelihood, the posterior distribution has no closed form analytical solution. However, given that $\log \mathrm{p}(\mathrm{f}|\mathrm{y})$ is proportional to $\log \mathrm{p}(\mathrm{y}|\mathrm{f}) + \log \mathrm{p}(\mathrm{f})$, which can be efficiently computed for any f, *gpmap-tools* leverages optimization methods to find the maximum a posteriori (MAP) $\hat{f}$ and compute an approximate Gaussian posterior using the Laplace approximation, where the posterior covariance matrix Σ is defined by the inverse Hessian of the posterior at its mean $\hat{f}$ (Rasmussen and Williams, 2008).

These solutions are completely general, in the sense that they hold for arbitrary valid prior covariance of precision matrices. However, as we will explain below, *gpmap-tools* implements highly optimized versions of these calculations that take advantage of the structure of sequence space and our specific choices for C and K.

Inference of genotype-phenotype maps with *gpmap-tools*. *gpmap-tools* combines prior distributions with likelihood functions into a number of Gaussian process models for inference of complete genotype-phenotype maps.

Minimum epistasis interpolation. The minimum epistasis interpolation (MEI) method was originally proposed in terms of finding the f_z at unobserved sequences z given the known phenotype f_x at sequences x by minimizing the average squared epistatic coefficient $\overline{\epsilon^2}$ over the complete genotype-phenotype map (Zhou and McCandlish, 2020). *gpmap-tools* provides a generalization to local epistatic coefficients of any order P, (by minimizing $f^T \Delta^{(P)} f$ (Chen *et al.*, 2021)), incorporates known Gaussian measurement noise through σ_x^2 and, by re-framing minimum epistasis interpolation as a Gaussian process model, enables uncertainty quantification via the posterior covariance (see Supplementary Information E,F).

Empirical variance component regression. Empirical variance component regression (VC regression), proposed in (Zhou *et al.*, 2022), combines a variance component prior parameterized by the variance λ_k associated to interactions of each possible order k with a Gaussian likelihood with known noise variance σ_x^2 to compute the exact Gaussian posterior distribution over f. The hyperparameters λ_k controlling the expected variance explained by interactions of order k under the prior are optimized through kernel alignment (Wang *et al.*, 2015). This procedure minimizes the squared distance between the covariance under the prior and the empirical distance-covariance function computed from the incomplete data. While a naive kernel alignment implementation requires computation with large covariance matrices, the prior covariance between two sequences depends only on the Hamming distance between them, resulting in only $\ell + 1$ different values. As a consequence, the problem of kernel alignment can be efficiently solved as a simpler $\ell + 1$-dimensional constrained weighted least squares problem (Zhou *et al.*, 2022).

Sequence probability distribution estimation. *gpmap-tools* also implements the SeqDEFT method for estimating probability distributions π over sequence space (Chen *et al.*, 2021). SeqDEFT parametrizes the probability distribution as

$$\pi_i = \frac{e^{-\phi_i}}{\sum_j e^{-\phi_j}}$$

and defines an improper prior distribution over the latent phenotype ϕ that penalizes local epistatic coefficients of order P given by $\log \mathrm{p}(\phi|a) \propto -\frac{a}{2s} \phi^T \Delta^{(P)} \phi$, which is combined with a multinomial likelihood function to compute an approximate posterior distribution over $\phi = -\log \pi$. The hyperparameter a is optimized by maximizing the cross-validated log-likelihood under the MAP estimate over a one-dimensional grid search. This also enables to examination of the behavior of the model towards the two limiting solutions i.e. when local epistatic coefficients are unconstrained ($a = 0$) or forced to be zero ($a \to \infty$) (Chen *et al.*, 2021).

Visualization of genotype-phenotype maps. Genotype-phenotype maps are inherently high-dimensional objects, and thus difficult to visualize in an intuitive manner. *gpmap-tools* implements a previously proposed strategy for visualizing fitness landscapes (McCandlish, 2011) that computes embedding coordinates for genotypes such that squared distances between pairs of genotypes in the low-dimensional representation approximate the expected times to evolve from one to another under selection for high phenotypic values. This layout highlights regions of sequence space containing highly functional genotypes that are nevertheless poorly accessible to each other e.g. fitness peaks separated by valleys, or sets of sequences where the intermediates are functional but the order of the intervening mutations is highly constrained.

Evolutionary model. We assume a weak mutation model of evolution in haploid populations, such that mutations are always fixed or lost before a new mutation arises (McCandlish, 2011, 2018; Zhou and McCandlish, 2020; Chen *et al.*, 2021). Under this model, the evolutionary rate $Q(i,j)$ from genotype i to j depends on the mutation rate $M(i,j)$ (which we assume is taken from a time-reversible mutational model) and the probability of fixation relative to a neutral mutation (Bulmer, 1991; McCandlish and Stoltzfus, 2014):

$$Q(i,j) = \begin{cases} M(i,j) \frac{S(i,j)}{1-e^{S(i,j)}} & \text{if } i \text{ and } j \text{ are neighbors} \\ -\sum_{k \neq i} Q(i,k) & \text{if } i = j \\ 0 & \text{otherwise,} \end{cases}$$

where $S(i,j)$ is the scaled selection coefficient of genotype j relative to genotype i. For the purposes of constructing a

useful visualization, we then assume that this scaled selection coefficient is proportional to the phenotypic differences between the two genotypes i.e. $S(i,j) = c(f(j) - f(i)$, where the constant c can be interpreted as the scaled selection coefficient ($2N_e s$, for a Haploid Wright-Fisher population) associated with a phenotypic difference of 1. Unless specifically studying the role of mutational biases on evolution on empirical landscapes, we would typically assume that $M(i,j) = 1$ for any i,j pair (i.e. measuring time in units of the inverse mutation rate), and focus on the evolutionary dynamics induced by the structure of the genotype-phenotype map alone. This model assigns a low but non-zero probability of fixation to deleterious mutations and has a unique stationary distribution $\pi(i)$ given by

$$\pi(i) = \frac{\pi_M(i)e^{cf(i)}}{\sum_j \pi_M(j)e^{cf(j)}}.$$

where $\pi_M(i)$ are the time-reversible neutral stationary frequencies, which are uniform in absence of mutational biases (Sella and Hirsh, 2005; McCandlish *et al.*, 2015). The stationary distribution can be used to select a reasonable value of c for our evolutionary process. When representing a probability distribution, such as one inferred using SeqDEFT, setting $f(i) = \log \pi(i)$ and $c = 1$ will result in a stochastic process in which the stationary distribution exactly matches the estimated genotype probabilities, providing a very natural representation of the landscape. When inferring the genotype-phenotype map from MAVE data, c can be adjusted so that the mean phenotype under the stationary distribution aligns with realistic natural values e.g. the phenotype associated to a wild-type or reference sequence. Alternatively, a range of c values can be used to generate a family of visualizations for a single genotype-phenotype map to reflect the evolutionary impact of its structure under different assumptions concerning the relative strengths of selection and drift.

Low-dimensional representation. The right eigenvectors r_k of Q associated to the largest eigenvalues λ_k ($\lambda_1 = 0 > \lambda_2 \geq \lambda_3 \geq \ldots$) can be computed using iterative methods that leverage the sparse structure of Q. When appropriately normalized and re-scaled as $u_k = \frac{1}{\sqrt{-\lambda_k}} \frac{r_k}{r_k^T D_\pi r_k}$, the first few r_k for $k \geq 2$ can be used as embedding coordinates, resulting in a low-dimensional representation in which squared distances between genotypes optimally approximate the commute times i.e. the sum of hitting times $H(i,j)$ from i to j and $H(j,i)$ from j to i, thus separating sets of functional genotypes that are largely inaccessible to each other for a population evolving under selection for high phenotypic values:

$$\sum_{k=2} (u_k(i) - u_k(j))^2 \approx H(i,j) + H(j,i).$$

The eigenvalues λ_k represent the rates at which the associated eigenvectors become less relevant for predicting evolutionary outcomes with time. The associated relaxation times $-\frac{1}{\lambda_k}$ have units of expected number of substitutions and allow us to identify components that decay slower than expected under neutral evolution, where we note that if all mutations occur at rate 1, the neutral relaxation time is given by the reciprocal of the minimum number of alleles across sites. Because u_k captures the $k-1$-th strongest barrier to the movement of a population in sequence space, we refer to u_k as diffusion axis $k-1$ (see McCandlish, 2011; Chen *et al.*, 2021, for more details).

Rendering and visualization. In addition to computing the coordinates u_k, *gpmap-tools* provides functionality at both high and low levels to plot and render the visualizations of genotype-phenotype maps using different backend plotting libraries. This includes the standard plotting library in python, *matplotlib* (Hunter, 2007), for generating highly customized visualizations, and *plotly* (Hossain, 2019), for generating interactive 3D visualizations that display the sequence associated to each node when hovering the mouse over them. Moreover, as rendering large numbers of points and lines becomes limiting in large datasets, the *gpmap-tools* plotting library leverages the power of *datashader* (Bednar *et al.*, 2022) for efficiently rendering plots containing millions of different elements, achieving close to an order of magnitude speed up for large genotype-phenotype maps (Figure S1).

Efficient computation with *gpmap-tools*. We aim to study genotype-phenotype maps with a number of genotypes ranging from a few thousands up to millions. However, all of the described methods require computing with unreasonably large matrices of size $\alpha^\ell \times \alpha^\ell$. For instance, to study a genotype-phenotype map for 9 nucleotides, a naive implementation would need to build a $4^9 \times 4^9$ matrix requiring 512GB of memory using 64 bit floating point numbers and over 100 billion operations to compute matrix-vector products. While some of the necessary matrices are sparse e.g. $\Delta^{(P)}$ and Q, allowing efficient storage and computation (McCandlish, 2011; Zhou and McCandlish, 2020; Chen *et al.*, 2021), other matrices e.g. P_U and K, are dense.

gpmap-tools circumvents these challenges using two strategies. First, we note that every matrix A with entries A_{ij} depending only on the Hamming distance between sequence i and j, such as $\Delta^{(P)}$ as well as the dense matrices P_k and K_k, can be expressed as an ℓ-order polynomial in the Laplacian of the Hamming graph L (Zhou *et al.*, 2022). This enables efficient computation of matrix-vector products $Ab = \sum_i^\ell c_i L^i b$ by multiplying the vector b by L up to ℓ times, e.g. $L^2 b = L(Lb)$, and taking linear combinations of the results without explicitly building the possibly dense matrix A. *gpmap-tools* also implements L as a linear operator (see Supplementary Information D). While this new linear operator-based implementation achieves comparable efficiency to a sparse matrix formulation in nucleotide space, it is an order of magnitude faster in protein spaces (Figure S2). More importantly, the L linear operator requires virtually no time for construction and has much lower memory requirements (Figure S2).

Second, we note that many of the relevant matrices can be obtained as ℓ-Kronecker products of $\alpha \times \alpha$ matrices, such as $P_U = \bigotimes_p^\ell P_p$. By using *scipy*'s (Virtanen *et al.*, 2020) LinearOperators functionality, we can leverage the mixed Kro-

necker matrix-vector product property to enable computation of e.g. $P_U b$ without constructing P_U (see Supplementary Information C, Figure S3). Rather than calculating explicit inverse matrices, we can likewise use these linear operators to find numerical solutions to matrix equations using Conjugate Gradient (CG). By combining multiple linear operators, we are able to compute the posterior variance for a small number of sequences of interest or the posterior covariance for any set of linear combinations of phenotypic outcomes e.g. calculating posterior variance for mutational effects in specific genetic backgrounds and epistatic coefficients of any order, while limiting the number of linear systems to solve with CG to the number of linear combinations of interest.

Results

In this section, we illustrate the power of *gpmap-tools* by studying the genotype-phenotype map of the Shine-Dalgarno (SD) sequence. The SD sequence is a motif located in the 5'UTR of most prokaryotic mRNAs. This motif is recognized by the 3' tail of the 16S rRNA through base pair complementarity with a region known as the anti Shine-Dalgarno (aSD) sequence, promoting translation initiation (Shine and Dalgarno, 1975). Understanding how the SD sequence modulates protein translation *in vivo* is key in synthetic biology applications (Salis *et al.*, 2009; Gilliot and Gorochowski, 2024). Previous studies used existing sequence diversity (Hockenberry *et al.*, 2018; Wen *et al.*, 2020) and MAVE data (Bonde *et al.*, 2016; Kuo *et al.*, 2020) to build models for this genotype-phenotype map. However, these models cannot account for higher-order genetic interactions and provide limited understanding of the structure of the genotype-phenotype map. Thus, *gpmap-tools* offers a new opportunity to model and understand the patterns of genetic interactions and the main qualitative features that define this important regulatory element.

Inferring the probability distribution of the Shine-Dalgarno sequence. Here, we use SeqDEFT to infer the sequence probability distribution for the SD sequence by using the 5' untranslated regions (UTRs) across the whole *E. coli* genome. We extracted the 5' UTR sequence from 5,311 annotated genes and aligned them with respect to the start codon. Figure 2A shows site-specific allele frequencies for up to 20bp upstream of the start codon.

These allele frequencies showed an enrichemnt in purines between positions -13 and -5 that is characteristic of the location of the SD sequence, which classically exhibits the consensus sequence AGGAGGU (Shine and Dalgarno, 1975; Hockenberry *et al.*, 2018; Wen *et al.*, 2020). Thus, we attempted to infer a genotype-phenotype map for this 9 nucleotide region. Out of the total $4^9 = 262,144$ possible sequences, we observed $3,690$ unique sequences, most of them observed a single time. Given that the number of sampled sequences is two orders of magnitude smaller than the number of possible sequences, we expect many unobserved sequences to be functional. Therefore, sharing information across neighboring genotypes through SeqDEFT's

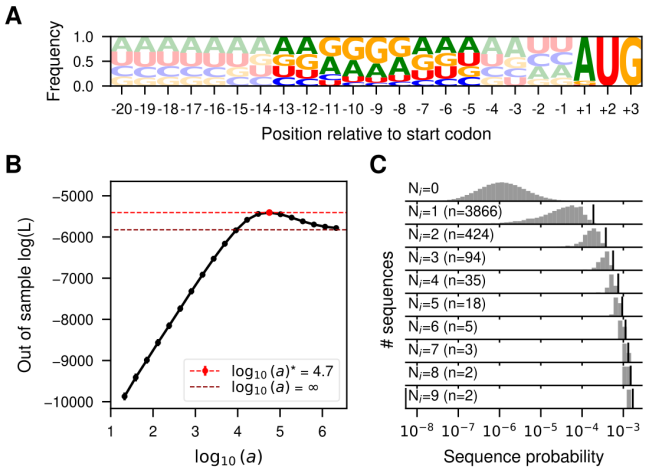

Fig. 2. Inference of the probability distribution of the Shine-Dalgarno sequence. (A) Sequence logo representing the site-specific allele frequencies of 5,311 5'UTRs in the *E. coli* genome aligned with respect to the annotated start codon. The start codon and the 9 nucleotide region 4 bases upstream are highlighted to emphasize the region most relevant for translation initiation. (B) Log-likelihood computed in the 20% held-out sequences in 5-fold cross-validations of a series of SeqDEFT models (P=2) under varying values of the hyperparameter a. The horizontal dashed lines represent the log-likelihood of the limiting maximum entropy model (black), corresponding to the independent sites model shown in panel A, or the best SeqDEFT model (red). (C) Distribution of inferred sequence probabilities depending on the number of times N_i they were present in the *E. coli* genome represented in a logarithmic scale. Vertical black lines represent the empirical frequency N_i/N_T corresponding to each N_i value.

prior distribution could alleviate this limited amount of data. Figure 2B shows that the SeqDEFT model better predicts the frequencies of held-out sequences than either the site-independent model (the maximum entropy model given the site-specific frequency profiles, $a = \infty$) or as we approach the empirical frequencies model ($a \to 0$, which maximizes the likelihood). In particular, this optimum at an intermediate value of a provides strong support for the presence of epistatic interactions (Chen *et al.*, 2021).

We then computed the MAP solution (using all available data) under the value a^* that maximized the likelihood for the held-out sequences and compared the inferred probabilities with the observed frequencies among *E. coli* 5' UTRs (Figure 2C). Sequences that appear more than 2-3 times are always inferred to be highly functional (i.e. high frequency). However, there is a wide range of variability for unobserved sequences, ranging 4 orders of magnitude in their estimated probabilities, many of them with larger probabilities than some sequences that are observed once. The MAP solution yields $\overline{\epsilon^2} = 0.10$, corresponding to a root mean square local epistatic coefficient of 0.32, which is slightly less than half of the size of the root mean squared mutational effect (0.78). This indicates that adding a single mutation to the genetic background often substantially changes the effects of other mutations.

Inferring the genotype-phenotype map of the Shine–Dalgarno sequence from MAVE data. We next used data from a previously published MAVE (Kuo *et al.*, 2020) measuring the expression of a GFP reporter controlled by a sequence library containing nearly all 262,144 possible 9 nucleotide sequences 4 nucleotides upstream of the start codon,

i.e., the same region considered in our previous analysis. We first run MEI to predict the phenotype for all missing genotypes. The imputed genotype-phenotype map had an $\overline{\epsilon^2} = 0.11$. This value is not directly comparable with the results of our SeqDEFT analysis because of the difference in measurement scale (log probability vs. log GFP). Still, we can compare the root mean squared epistatic coefficient, which for MEI takes a value of 0.32, to the root mean squared size of mutational effects, which for MEI is 0.33. These similar magnitudes indicate that there is substantial variability in the effects of mutations across neighboring genotypes, more so than in the genotype-phenotype map inferred with SeqDEFT.

To better capture this high degree of inferred epistasis, we turned to VC regression, where the prior reflects the observed predictability of mutational effects in the training data. We found that the empirical phenotypic correlation between pairs of sequences decayed quite quickly with the number of mutations e.g. pairs of sequences separated by three mutations only showed a correlation of 0.25 between their measured phenotypes (Figure 3A). We next estimated the variance component prior distribution that best matched the observed distance correlation patterns and computed the variance explained by interactions of every possible order under this prior (Figure 3B). The additive and pairwise component explained only 57.6% of the overall variance, suggesting an important influence of higher-order genetic interactions. We then inferred the complete genotype-phenotype map under this prior. These estimates recapitulated the experimental data remarkably well ($R^2 = 0.94$, Figure 3C) and made predictions almost as accurate in held-out test sequences ($R^2 = 0.87$, Figure 3D). Importantly, our estimates of the uncertainty of the phenotypic predictions are well calibrated, as we find approximately the expected fraction of measurements in the test set within posterior credible intervals (Figure S4C). Comparing the predictive performance of MEI against VC regression as a function of the number of sequences used for training, we find that while the two models perform comparably well when the genotype-phenotype map is densely sampled, and MEI performs better with extremely low sampling (likely due to error in the estimation of variance components), overall VC regression exhibited substantially higher performance across a wide range of training data densities (Figure S4A,B) and better calibration of the prediction's uncertainty (Figure S4C).

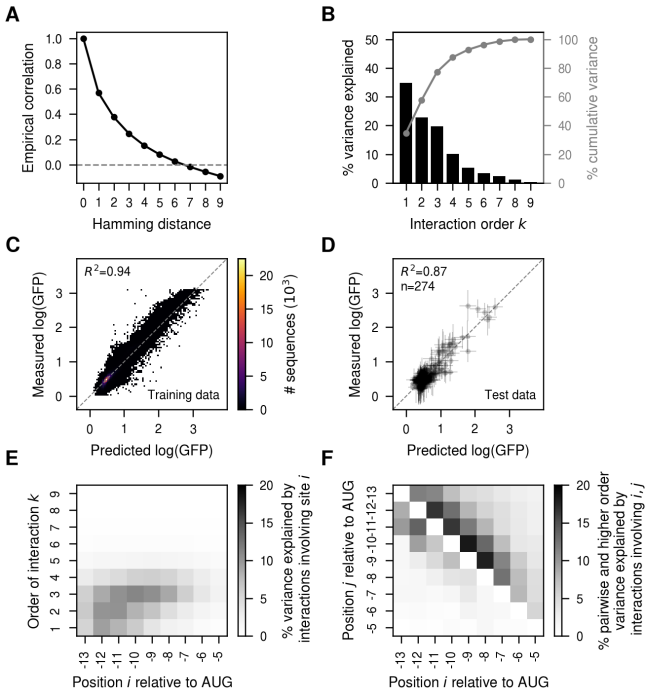

Fig. 3. VC regression analysis of the experimentally measured genotype-phenotype map for the Shine-Dalgarno sequence in the dmsC gene context (Kuo *et al.*, 2020). (A) Empirical distance-correlation function using the measured log(GFP) values in the experimentally evaluated sequences. (B) Percentage of variance explained by interactions of order k in the inferred VC regression prior. Grey lines represent the cumulative percentage of variance explained by interactions up to order k. (C) Two-dimensional histogram showing the comparison of the measured log(GFP) and the MAP estimate under the VC model in sequences used for model fitting. (D) Comparison of the posterior distribution for held-out test sequences and the measured log(GFP) values. Horizontal error bars represent posterior uncertainty represented as the 95% credible interval, whereas vertical error bars correspond to the 95% confidence interval under each measurement's variance. (E) Heatmap representing the percentage of variance explained by interactions of order k involving each position relative to the start codon. (F) Heatmap representing the percentage of variance explained by pairwise (lower triangle) and higher-order (upper triangle) interactions that is explained by interactions involving pairs of positions relative to the start codon. See Supplementary Information B for details on calculating (E,F).

Position-specific contributions to epistasis. An important advance in *gpmap-tools* is its ability to use the P_U matrices to evaluate the contribution of each site to genetic interactions of different orders (Supplementary Information B). Figure 3E shows this analysis for the MAP solution obtained using VC regression. We see that while positions -6 and -5 have an overall weak influence in the measured translational efficiency, and sites -13 to -10 have both strong additive and epistatic contributions, sites -9 to -7 influence the phenotype mostly through higher-order epistatic interactions. Thus, we find that sites in the SD sequence have very heterogenous contributions to genetic interactions of different orders, with some sites having stronger additive and lower order epistatic interactions, whereas other sites influence translation primarily via higher-order interactions. These variances can be further decomposed into variances explained by epistatic interactions of any order involving each possible pair of sites (Supplementary Information B). This decomposition reveals that pairwise interactions are largely confined to sites within 3 nucleotides of each other and are strongest between positions -13 to -10 (Figure 3F, lower triangle). Higher-order interactions extend to sites separated by up to 4 nucleotides, with the most prominent effects involving positions -9 to -7 (Figure 3F, upper triangle). In contrast, interactions between sites separated by 5 or more nucleotides are rare across all orders (Figure 3F). These findings indicate that the effect of a mutation at a given site depends primarily on nearby sites and becomes nearly independent of mutations beyond a 4-nucleotide range. Overall, the ability to quantify the variance explained by interactions of different orders and positional combinations offers a powerful framework for characterizing the nature and strength of epistasis, and for identifying communities of interacting sites within genotype-phenotype

maps.

Visualizing the probability distribution of the SD sequence. In order to understand the main qualitative properties of this highly epistatic genotype-phenotype map, we generated a low-dimensional representation using our visualization technique. Figure 4A shows that the genotype-phenotype maps consists of at least three largely isolated peaks. These peaks correspond to the canonical SD motif AGGAG located at three consecutive positions relative to the start codon, with a fourth central peak corresponding to a shift of the canonical motif one additional base upstream appearing along Diffusion axes 3 in a 3-dimensional representation (Figure S5). This shows that not only can the aSD sequence bind at different distances from the start codon to induce efficient translation initiation, consistent with the interaction neighborhoods shown in Figure 3D, but also that it is hard to evolve a sequence with a shifted SD motif by one or two positions through single point mutations without losing translational efficiency. In contrast, sequences with an SD motif shifted by three positions remain largely connected by extended ridges of functional sequences. In these extended ridges, a second binding site can evolve through a sequence of point mutations without destroying the first. Specifically, within each trinucleotide sequence around the central AGG common to two binding registers, mutations can accumulate in diverse orders, opening up many different evolutionary paths only subject to the constraint of evolving a second SD motif before destroying the first one. Figure 4A highlights two examples of such paths.

Comparing sequence probability across different species. To investigate whether the structure of the genotype-phenotype map is the same across distant species, we performed the same analysis using 5'UTR sequences from 4,328 annotated genes in the genome of *B. subtilis*, whose most recent common ancestor with *E. coli* dates back to $\sim$2 billion years ago (Feng *et al.*, 1997). We first found that the AG bias marking the location of the SD sequence in the 5'UTR is located approximately 2 bp further upstream from the start codon compared to its location in *E. coli* (Figure S6A), as previously reported (Hockenberry *et al.*, 2018). We then extracted the 9 nucleotides sequences 6 bp upstream of the start codon and inferred the sequence probability distribution using SeqDEFT. The estimated log-probabilities were highly correlated with those obtained from the *E. coli* genome (Spearman $\rho = 0.94$, Figure S6B), but more importantly, the inferred genotype-phenotype map displayed a similar structure, with peaks corresponding to different binding registers of the aSD sequence and extended ridges connecting sets of sequences with overlapping binding sequences separated by 3 positions (Figure S6B). Overall, the probability distributions of the SD sequences are quantitatively very similar across distant species and show the same main qualitative features.

Comparing sequence probability and functional measurements. We next compared the genotype-phenotype maps based on genomic sequences with the genotype-phenotype map obtained with MAVE data. First, we directly compared the estimated sequence probability across the *E. coli* genome with the inferred translational efficiency from MAVE data (Figure 4B) for every possible sequence. We found a moderate non-linear relationship between these two independently inferred quantities (Spearman $\rho = 0.55$). Sequences with very low estimated probability ($P < 10^{-8}$) consistently showed low translational efficiency ($\log(\text{GFP}) < 1.0$), whereas sequences with high sequence probability ($P > 10^{-4}$) had consistently higher but variable translational efficiencies (mean=1.84, standard deviation=0.63).

To investigate whether this modest degree of agreement is due to noise in the estimates for individual sequences or to having inferred qualitatively different genotype-phenotype maps, we applied the visualization technique to the empirical genotype-phenotype map inferred with VC regression (Figure 4C and S7). Despite the much more skewed phenotypic distribution of estimated translational efficiencies, this low-dimensional representation has essentially the same structure with isolated peaks corresponding to different distances of the SD motif to the start codon and extended ridges connecting sequences with SD motifs shifted by 3 positions separated along several Diffusion axes (Figure 4C and S8). In addition to the previous structure, we identify an additional extended ridge of functional sequences with sequences starting by GAG. This subsequence, together with the upstream G from the fixed genetic context in which the experiment was performed, forms a binding site for the aSD sequence. In contrast, the probability distribution of SD sequences was inferred from genomic sequences with different flanking nucleotides such that genotypes starting with GAG, on average, are not as functional. Thus, we can conclude that, despite showing only a moderate quantitative agreement, the two inference procedures using different types of data are able to recover genotype-phenotype maps with the same qualitative features and expected long-term evolutionary dynamics.

Uncertainty quantification for genetic interactions and phenotypic predictions. Visualization of genotype-phenotype maps based on our MAP estimates enabled the identification of key shared qualitative features and the genetic interactions underlying them. However, we can also evaluate the strength of evidence supporting these interactions by leveraging the uncertainty quantification capabilities of our Gaussian process models as implemented in *gpmap-tools* e.g. we can compute the posterior distribution of the effects of specific mutations in different backgrounds. As an illustration of this strategy, we first validated the incompatibilities separating peaks in the SD genotype-phenotype map by computing the posterior distribution for mutational effects in the two backgrounds UUA**AGGAG**C (grey) and UA**AGGAG**CA (black), which contain the same AGGAG motif shifted by one position (Figure 4D). In absence of epistasis, mutational effects are expected to be exactly the same in the two genetic backgrounds. While this is true for some mutations e.g. C-5A (Figure 4D), the three mutations that allow shifting the SD motif one position up-

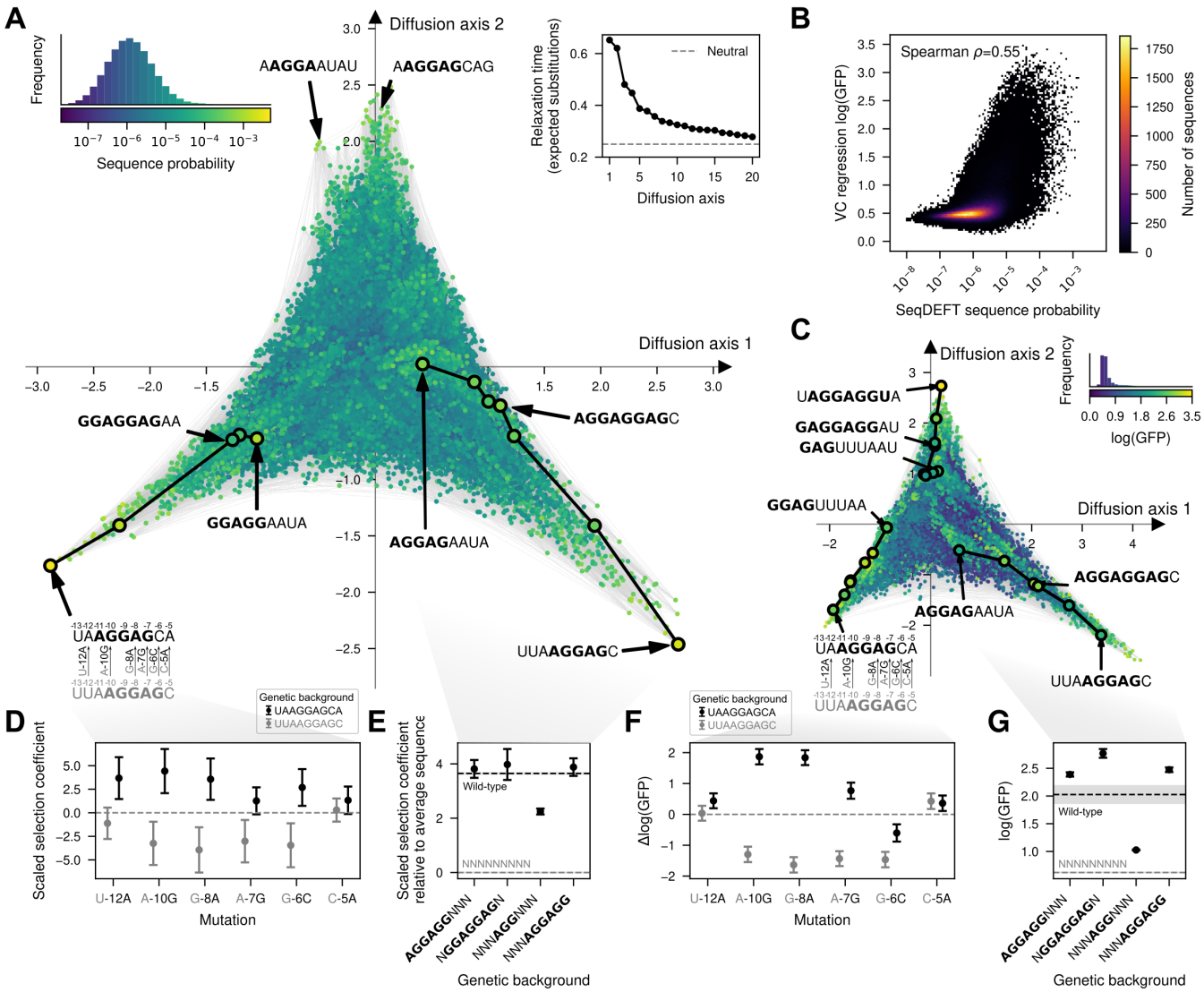

Fig. 4. Visualization of the genotype-phenotype map of the Shine-Dalgarno sequence. (A,C) Low-dimensional representation of the *E. coli* Shine-Dalgarno sequence probability distribution inferred with SeqDEFT (A) and the translational efficiencies inferred with VC regression (C). Every dot represents one of the possible 4^9 possible sequences and is colored according to their inferred probability (A) or log(GFP) values (C). The inset represents the distribution of inferred sequence probabilities or log(GFP) values along with their corresponding color in the map. Inset in the upper right corner of (A) shows the relaxation times associated to the 20 most relevant Diffusion axes, showing that the first two Diffusion axes have much longer relaxation times than the rest. Sequences are laid out with coordinates given by these first two Diffusion axes and dots are plotted in the order of the 3rd Diffusion axis. (B) Two-dimensional histogram representing the relationship between the inferred sequence probabilities from their frequency in the *E. coli* genome and the estimated translational efficiencies inferred with VC regression from MAVE data. (D) Posterior distribution inferred by SeqDEFT for the scaled selection coefficient of specific mutations when introduced in two genetic contexts, UUAAGGAGC (grey) and UAAGGAGCA (black), representing a shift of the AGGAG motif by one nucleotide. Mutational effects are reported in units of scaled selection coefficients. (E) Posterior distribution inferred by SeqDEFT for the average scaled selection coefficient, relative to the average across all possible sequences, for genotypes containing the AGGAGG motif at positions separated by three nucleotides, along with their potential mutational intermediates. (F) Posterior distribution inferred with VC regression for the effects of specific mutations on log(GFP) when introduced in two genetic contexts, UUAAGGAGC (grey) and UAAGGAGCA (black), representing a shift of the AGGAG motif by one nucleotide.(G) Posterior distribution inferred with VC regression for the average log(GFP) of genotypes containing the AGGAGG motif at positions separated by three nucleotides, along with their potential intermediates. (E,G) Horizontal dashed lines represent posterior mean of the average phenotype across all possible sequences (grey) or wild-type sequences, given by the genomic sequences in (E) and AAGGAGGUG in (G) (black). Shaded areas represent the 95% credible intervals. (D-G) Points represent the maximum a posteriori (MAP) estimates and error bars represent the 95% credible intervals.

stream in the UUA**AGGAG**C context (grey) i.e. A-10G, G-8A and A-7G, are strongly deleterious in that context, but beneficial when introduced in a UA**AGGAG**CA background (black, Figure 4D). Importantly, the posterior distributions are concentrated around the means, showing that the data strongly supports that mutations needed to shift the SD motif by one position are substantially deleterious in that speific context, creating the valleys that separate the main peaks of this genotype-phenotype map.

We next evaluated the evidence supporting the existence of the extended ridges connecting sequences with an SD motif shifted by 3 positions. To do so, we computed the posterior distribution for the average phenotype (scaled selection coefficient relative to the average fitness across all sequences) of genotypes containing two overlapping binding registers (NGGAGGAGN), only one (AGGAGGNNN and NNNAGGAGG), or none (NNNAGGNNN). Whereas NGGAGGAGN, AGGAGGNNN and NNNAGGAGG are highly functional, sequences with a central AGG have on average roughly half as large of a scaled selection coefficient than sequences containing full motifs in either or both registers (Figure 4E).

The posterior distributions for these same sets of genotypes and mutations estimated from sequence data from the *B. subtilis* genome (Figure S6D,E) and from VC regression analysis on MAVE data (Figure 4F,G) are largely concordant. The agreement between these independent data sources, together with uncertainty quantification in each case, provides strong support for a common landscape structure with distinct peaks at the different binding registers of the aSD sequence, connected by extended ridges of functional sequences linking registers offset by three nucleotides in aSD binding position.

A biophysical model recapitulates the qualitative properties of empirical SD genotype-phenotype maps. Although the inferred genotype-phenotype map exhibited extensive epistasis, the visualization revealed that its complexity could be largely explained by a simple underlying mechanism in which the aSD sequence can bind at varying distances from the start codon. We hypothesize that this mechanism alone explains both the existence of isolated peaks and, together with the quasi-repetitive nature of the aSD sequence, the extended ridges. Moreover, despite our ability to estimate mutational effects in different contexts, inference of the actual binding preferences of the aSD from the data is hindered by the convolution of the effects of mutations on the binding affinities at different registers. To tackle these issues, we fit a simple mechanistic model, in which GFP protein abundance is linearly dependent on the fraction of mRNA bound by the aSD at thermodynamic equilibrium at different positions p relative to the start codon, where the binding energy ΔG of the aSD is an additive function of the sequence at that position x_p (Figure 5A, see Methods).

We fit this biophysical model (Figure S9A) by maximum likelihood to the MAVE dataset and achieved good predictive performance in both training ($R^2 = 0.59$, Figure S9B) and held-out sequences ($R^2 = 0.64$, Figure S9C). Importantly, this model contains only 27 free parameters with clear biophysical interpretations e.g. in terms of mutational effects on binding energies. The model also includes a parameter β_0 that specifies the background fluorescence in absence of aSD binding to the 5'UTR, which we estimated as $\hat{\beta}_0 = 0.47$. This model allowed us to deconvolve the effects of mutations on binding at different registers and to infer the allele and position specific energetic contributions to binding (Figure 5B). As expected, the reverse complement of the aSD is the most stable binder, but different mutations have substantially variable effects in the binding energy. Not only do some positions have stronger energetic contributions in general (positions 2-5 within the SD sequence), but different missmatches with the aSD in the same position have different energetic effects e.g. A4G is only slightly destabilizing ($\Delta\Delta G = 0.84$ kcal/mol), whereas A4C is highly destabilizing ($\Delta\Delta G = 1.98$ kcal/mol). Importantly, predictions of this simple model for all 4^9 SD sequences recapitulate the main structure of the genotype-phenotype map with isolated peaks and extended ridges corresponding to different registers of binding, as expected (Figure 5C). We can verify that the peaks correspond to different binding registers by computing the binding energy of every sequence at specific positions relative to the start codon and color the visualization by those energies (Figure 5D). Likewise, this representation also shows that the extended ridges of functional sequences correspond to sequences that are strongly bound at positions separated by three nucleotides. We next compared our thermodynamic model to an alternative thermodynamic model based on base-pair stacking interaction energies from classical RNA folding algorithms (Lorenz *et al.*, 2011; Salis *et al.*, 2009). Interestingly, our model achieved higher predictive accuracy (Figure S10A,B, $R^2 = 0.44$) and more faithfully captured the overall structure of the empirical genotype-phenotype map (Figure S10C). These findings suggest that the SD:aSD interaction *in vivo* cannot be fully explained by RNA thermodynamics alone, highlighting the importance of molecular context in shaping RNA-RNA interactions. More broadly, this analysis shows how visualization can guide the construction of simplified, biophysically interpretable models that reproduce the key qualitative features of genotype-phenotype maps.

Discussion

In this paper, we present *gpmap-tools*, an extensively documented software library with tools for the inference, visualization and interpretation of empirical genotype-phenotype maps containing arbitrarily complex higher-order genetic interactions. By providing a framework for the analysis of complex genetic interactions, *gpmap-tools* has the potential to reveal the simple qualitative properties of these complex mappings and to aid in development of biophysical and mechanistic hypotheses for these observed features.

The first step in this framework is the inference of the complete genotype-phenotype map comprising all possible sequences from either experimental MAVE data or sequence counts. Taking into consideration the noise in the data (due

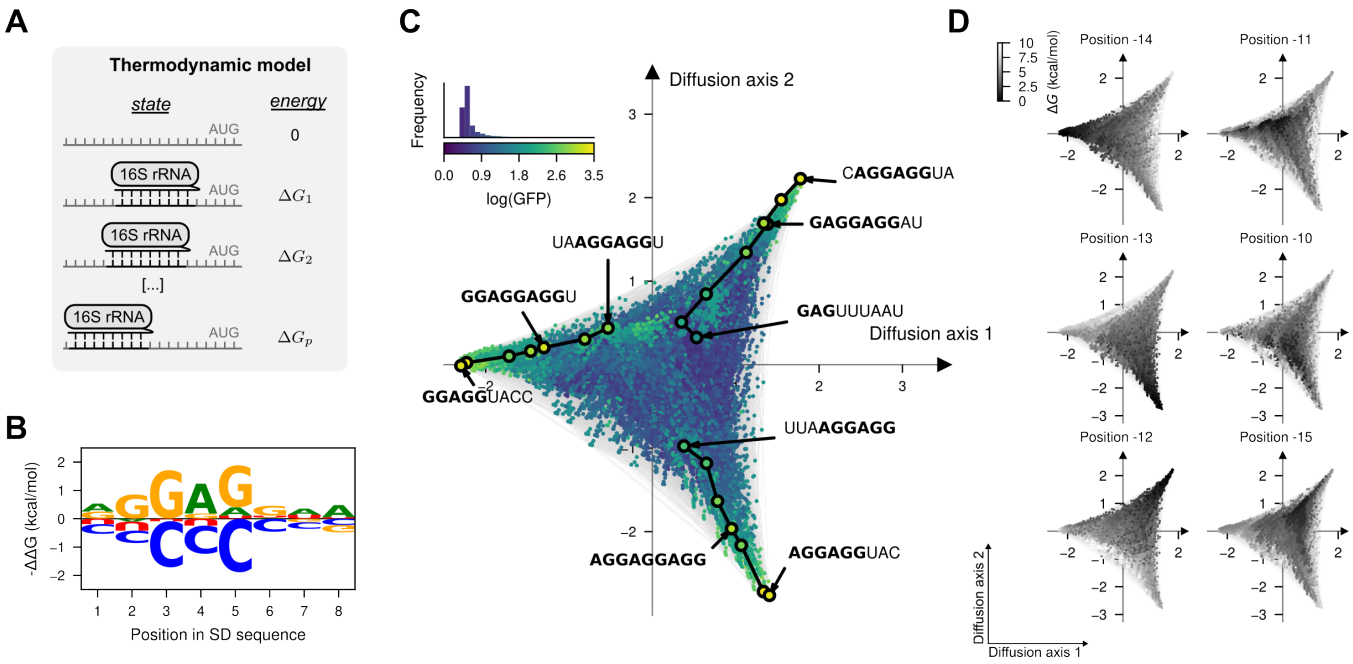

Fig. 5. Biophysical model of sequence-dependent translational efficiency. (A) Thermodyanmic model of binding of the 16S rRNA 3'tail to the 5'UTR of mRNAs at different positions p relative to the start codon AUG. (B) Sequence logo representing the site-specific but register-independent allelic contributions to the binding energy, where the size of the letter represents the difference in binding energy to the average across nucleotides. (C) Visualization of the genotype-phenotype map that results from predicting the phenotype of every possible sequence under the inferred thermodynamic model. Every dot represents one of the possible 4^9 possible sequences and is colored according to the predicted log(GFP). 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. (D) Visualization of the genotype-phenotype map under the inferred thermodynamic model representing the binding energies at positions -15 to -10 relative to the start codon showing that the peaks in the visualization correspond to the strongest binding at different positions and extended ridges correspond to sequences that are bound in two registers separated by 3 nucleotide positions. Binding energies are relative to the strongest binding sequence AGGAGGAA under the inferred model and are reported in units of kcal/mol assuming a temperature of 37ºC. Dots are plotted in reverse order of binding energy in the corresponding register.

either to sampling noise or experimental error), *gpmap-tools* is capable of computing the high-dimensional posterior distribution over all possible genotype-phenotype maps under a variety of priors. This allows us to obtain the MAP estimate, that is, the most probable genotype-phenotype map given the observed data. However, in contrast to other expressive models able to capture complex genetic interactions, such as neural networks (Bryant *et al.*, 2021; Gelman *et al.*, 2021; Sethi and Zhou, 2024), our inference methods provide rigorous uncertainty quantification of phenotypes, mutational effects or any linear combination of phenotypic values. This is important, as it tells the user which phenotypic predictions, mutational effects or genetic interactions can be trusted and to what extent, given the data.

The second step in this framework is the interpretation of the inferred genotype-phenotype maps. *gpmap-tools* provides a powerful method for visualizing fitness landscapes (McCandlish, 2011) that allows exploratory data analysis, interpretation and comparison of complex datasets and models. Thus, rather than interpreting the results through an explicit parametric model allowing high-order genetic interactions (Poelwijk *et al.*, 2016; Tareen *et al.*, 2020; Faure and Lehner, 2024; Park *et al.*, 2024; Faure *et al.*, 2024) or descriptive statistics like the number of peaks or adaptive walks (Szendro *et al.*, 2013; Ferretti *et al.*, 2018; Papkou *et al.*, 2023; Westmann *et al.*, 2024b; Li and Zhang, 2025; Chattopadhyay *et al.*, 2025), this method leverages the evolutionary dynamics on the genotype-phenotype map to highlight its main, potentially unexpected, qualitative features. Thus, we can use the visualization to generate hypotheses for how mutational effects change across genetic backgrounds and test these predictions by computing the corresponding background-dependent posterior distributions (Figure 4 and S6). Identifying the main features of the genotype-phenotype map can be crucial for defining an appropriate mechanistic or biophysical model. Here, visualization of the Shine-Dalgarno landscapes allowed us to define a thermodynamic model in which the binding energy depends only additively on the sequence at each register, and to verify that this simple model recapitulated the main qualitative features of the landscape (Figure 5). Additionally, this technique enabled a detailed comparison of genotype-phenotype maps inferred with different methods and data sources. In contrast to broadly used metrics, like Pearson or Spearman coefficients, this method shows the extent to which different landscapes have the same structure and qualitative features. In this study, it showed that genotype-phenotype maps inferred from two distantly related species *E. coli* and *B. subtilis* (Figure 4 and S6), as well as from entirely independent data sources (MAVE experiments versus natural sequence data), exhibited strikingly similar structure despite only moderate quantitative agreement (Figure 4). Identifying consistent structures across different data types and sources is essential for linking experimentally measured landscapes to the evolutionary forces shaping regulatory sequences, given that true fitness values in natural populations are typically unknown. More

broadly, our visualization technique enables comparison of genotype-phenotype maps across different classes of genetic elements, such as regulatory sequences, protein-protein interactions and enzymes, by revealing shared landscape features that may reflect similar evolutionary dynamics, despite differences in biological context.

The methods implemented in *gpmap-tools* scale to genotype-phenotype maps with millions of sequences by making several modeling assumptions, which also entail certain limitations. First, MEI and VC regression are phenomenological models. As such, they do not explicitly model global or non-specific epistasis that often arises from nonlinear dependencies between the underlying quantities affected by mutations and our measurements (Bloom, 2015; Otwinowski *et al.*, 2018; Tareen *et al.*, 2020; Tonner *et al.*, 2022; Faure and Lehner, 2024). Instead, these models rely on learning the pervasive genetic interactions induced by these global nonlinearities to nonetheless make accurate phenotypic predictions. Second, SeqDEFT assumes that observed sequences are drawn independently from the underlying probability distribution. While this assumption may hold for a few specific regulatory sequences that are repeated many times along the genome of a single species e.g. the Shine-Dalgarno sequence or the 5' splice site (Chen *et al.*, 2021), it remains unclear how robust it is to the known challenge of using phylogenetically related sequences from widespread multiple sequence alignments of protein families (Hockenberry and Wilke, 2019; Rodriguez Horta and Weigt, 2021; Dietler *et al.*, 2023). Third, both inference and visualization methods still require storing all possible sequences and their phenotypes in memory. The number of such sequences grows exponentially with sequence length, limiting the applicability of *gpmap-tools* to spaces of sequences of a constant and relatively short length (5 amino acids, 12 nucleotides, 24 biallelic sites). Despite these limitations, *gpmap-tools* provides a unique set of tools for studying the genotype-phenotype maps of short genetic elements. By combining nuanced analysis of epistasis, rigorous uncertainty quantification, and the capacity to infer landscapes containing millions of genotypes, it serves as a necessary stepping stone towards understanding the vastly larger genotype-phenotype maps arising at the gene, protein, and genome-wide scale.

Methods

Sequence diversity of the Shine-Dalgarno sequence. We downloaded the *E. coli* genome and annotation from Ensembl bacteria release 51, built on assembly version ASM160652v1, and *B. subtilis* assembly ASM904v1 from GeneBank. We extracted the 5'UTR sequence for every annotated gene using *pysam* (Li *et al.*, 2009; Bonfield *et al.*, 2021) and kept the 5,311 and 4,328 sequences, respectively, for which we could extract 20 bp upstream of the start codon without any ambiguous character 'N'. These sequences were aligned with respect to the start codon and used for computing site-frequency logos using *logomaker* (Tareen and Kinney, 2020) and estimating the complex probability distribution using the *gpmap-tools* implementation of SeqDEFT (Chen *et al.*, 2021). The MAP estimate was used to compute the coordinates of a low-dimensional representation assuming that the stationary distribution of the evolutionary random walk matches the estimated sequence probabilities by selecting a proportionality constant of $c = 1$ and uniform mutation rates.

Analysis of the experimental fitness landscape of the Shine-Dalgarno sequence. Phenotype data was computed from the processed data for independent replicates conducted in the dmsC genetic background as reported in the original manuscript (Kuo *et al.*, 2020). The mean and standard error was computed for all the 257,565 measured sequences. We estimated a common measurement variance of $\hat{\sigma}^2 = 0.058$ using genotypes measured across all 3 experimental replicates. The squared standard error for each genotype i was computed by dividing the overall experimental variance $\hat{\sigma}^2$ by the number of replicates n_i in which each sequence was measured ($\hat{\sigma}_i^2 = \hat{\sigma}^2/n_i$). We kept 0.1% of the sequences as test set, and use the remaining sequences for fitting different models to infer the complete genotype-phenotype map while evaluating their performance on the held-out test data. We estimated the variance components from the empirical distance-correlation function and used them to define a Gaussian process prior for inference of the complete combinatorial landscape containing all 4^9 genotypes, taking into account the known experimental variance $\hat{\sigma}_x^2$ for every sequence. We also computed the posterior mean and variances across all test sequences to assess the accuracy of the predictions and the calibration of the posterior probabilities in held-out data. We used the MAP estimate to compute the coordinates of the visualization assuming several different average values of log(GFP) under the stationary distribution that ranged from 1 to 2.5 (Figure S7). An average log(GFP) of 2 at stationarity was selected and used for all subsequent visualizations, similar to our MAP estimate of a log(GFP) of 2.03 for the wild-type reference.

Thermodynamic model of the Shine-Dalgarno genotype-phenotype map. We assume that translation is limited by the initiation step, which is itself modulated by the binding of the 16S rRNA to the 5'UTR of the mRNA, where we assume that the mRNA concentration is independent of the identity of the Shine-Dalgarno sequence. Binding and dissociation are assumed to be much faster than the rate at which translation is effectively initiated, so that the protein abundance is proportional to the fraction of mRNA bound by the 16S rRNA across all registers p at thermodynamic equilibrium, where we assume that binding occurs in at most one register at a time. The fraction of mRNA bound at thermodynamic equilibrium depends on the binding energy ΔG of the 16S rRNA to the mRNA to the sequence x_p starting at each position p, the temperature, which is assumed to the 37ºC (310K), and the universal gas constant $R = 1.9872 \times 10^{-3}$ kcal/mol K^{-1}. The overall GFP concentration for a sequence x depends on the fraction of bound mRNA and the translation rate when bound β:

$$[GFP](x) = \beta \left(\frac{\sum_p e^{-\frac{\Delta G(x_p)}{RT}}}{1+\sum_p e^{-\frac{\Delta G(x_p)}{RT}}} \right), \quad (1)$$

where $\Delta G(x_p)$ is the energy of binding of the 16S rRNA to the 8-nucleotide subsequence x_p at position p. The binding energy ΔG is independent of the position p at which binding occurs relative to the start codon and depends additively on the sequence x_p alone given by $\Delta G(x_p) = \Delta G_0 + \sum_i \sum_c x_p(i,c)\Delta\Delta G_{ic}$, where $x_p(i,c)$ takes value 1 if sequence x_p has allele c at position i and 0 otherwise, ΔG_0 represents the average binding energy across every possible sequence and $\Delta\Delta G_{i,c}$ is the energetic contribution of allele c at position i, subject to the constraint $\sum_c \Delta\Delta G_{i,c} = 0$ across all positions i. In order to incorporate the effect of mutations in binding registers spanning both fixed and variable regions of the sequence, we extended the variable 9 nucleotide sequences with the fixed upstream and downstream sequences CCG and UGAG from the dmsC genetic context.

Following previous work (Kuo *et al.*, 2020), we assume that occupancy at thermodynamic equilibrium is low so that $\sum_p e^{-\frac{\Delta G(x_p)}{RT}} \ll 1$, and thus $[GFP](x) \approx \beta \sum_p e^{-\frac{\Delta G(x_p)}{RT}}$. We also model a background fluorescence signal β_0 due to cells auto-fluorescence in the GFP channel even in absence of GFP, which is independent of the variable 5'UTR sequence in the experiment. Finally, we consider that experimental errors lie on the log-scale, such that the measured log(GFP) y for sequence x is observed with known noise variance $\hat{\sigma}_x^2$ and an extra or uncharacterized variance σ^2 under a Gaussian likelihood function given by

$$\mathrm{p}(y|x) = N\left(\mu_x, \hat{\sigma}_x^2 + \sigma^2\right), \quad (2)$$

where μ_x is the expected log(GFP) under the model given by

$$\mu_x = \log\left(\beta_0 + \sum_p e^{-\frac{\theta_0 + \sum_i \sum_c x_p(i,c)\Delta\Delta G_{i,c}}{RT}}\right) \quad (3)$$

and $\theta_0 = RT\beta + \Delta G_0$. We used PyTorch to encode the model and used the Adam optimizer with a learning rate of 0.02 for 1500 iterations, while monitoring for convergence (Figure S9A), to find the maximum likelihood estimates of the model parameters.

Additionally, we fit a 4-parameter calibration model using ensemble binding energies ΔG_x computed with a thermodynamic model of RNA folding and interaction (Lorenz *et al.*, 2011), where $\mu_x = \log\left(\beta_0 + e^{\theta_0 + \theta \Delta G_x}\right)$. Specifically, we used RNAcofold v2.4.9 with -p0 option for computing the binding energy of each SD variant, embedded between the CCG and UGAG flanking sequences, with the anti-SD sequence ACCUCCU across all possible binding configurations. Several candidate anti-SD sequences, ranging from 5 to 9 nucleotides, were tested; ACCUCCU was selected due to the higher predictive power of GFP abundance of the resulting model.

Code availability

gpmap-tools is an open-source library with source code available at https://github.com/cmarti/gpmap-tools. It is thoroughly documented with several tutorials and explanations of the provided functionalities at https://gpmap-tools.readthedocs.io. Code to reproduce the analyses of the Shine-Dalgarno landscapes is available at https://github.com/cmarti/shine_dalgarno.

Acknowledgements

We thank Bryan Gitschlag, Álvaro Serrano-Navarro, Víctor Jiménez-Jiménez and Alejandra Laguillo-Diego for providing feedback during the preparation of this manuscript. CMG and DMM were supported by NIH grant R35GM133613, JBK and DMM were supported by NIH grant R01HG011787, JBK was supported by NIH grant R35GM133777, and CMG, JBK, and DMM were supported by additional funding from the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory. JZ was supported by NIH grant R35GM154908. WCC was supported by the National Science and Technology Council of Taiwan, R.O.C., under Grant No. NSTC 111-2112-M-194-008-MY3. This work was performed with assistance from the US National Institutes of Health Grant S10OD028632.

References

Aguirre, L., Hendelman, A., Hutton, S. F., McCandlish, D. M., and Lippman, Z. B. 2023. Idiosyncratic and dose-dependent epistasis drives variation in tomato fruit size. *Science*, 382(6668): 315–320.

Avizemer, Z., Martí-Gómez, C., Hoch, S. Y., McCandlish, D. M., and Fleishman, S. J. 2025. Evolutionary paths that link orthogonal pairs of binding proteins. *Cell Systems*, 16(5).

Baeza-Centurion, P., Miñana, B., Schmiedel, J. M., Valcárcel, J., and Lehner, B. 2019. Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing. *Cell*, 176(3): 549–563.e23. Publisher: Cell Press.

Bakerlee, C. W., Nguyen Ba, A. N., Shulgina, Y., Rojas Echenique, J. I., and Desai, M. M. 2022. Idiosyncratic epistasis leads to global fitness-correlated trends. *Science*, 376(6593): 630–635.

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

Bank, C., Matuszewski, S., Hietpas, R. T., and Jensen, J. D. 2016. On the (un)predictability of a large intragenic fitness landscape. *Proceedings of the National Academy of Sciences*, 113(49): 14085–14090.

Bednar, J. A., Crail, J., Crist-Harif, J., Rudiger, P., Brener, G., B, C., Thomas, I., Mease, J., Signell, J., Liquet, M.,

Stevens, J.-L., Collins, B., Thorve, A., Bird, S., thuydotm, esc, kbowen, Abdennur, N., Smirnov, O., Hansen, S. H., maihde, Hawley, A., Oriekhov, A., Ahmadia, A., Jr, B. A. B., Brandt, C. H., Tolboom, C., G, E., Welch, E., and Bourbeau, J. 2022. holoviz/datashader: Version 0.14.3.

Bendixsen, D. P., Collet, J., Østman, B., and Hayden, E. J. 2019. Genotype network intersections promote evolutionary innovation. *PLoS Biology*, 17(5). Publisher: Public Library of Science.

Bitbol, A.-F., Dwyer, R. S., Colwell, L. J., and Wingreen, N. S. 2016. Inferring interaction partners from protein sequences. *Proceedings of the National Academy of Sciences*, 113(43): 12180–12185. Publisher: Proceedings of the National Academy of Sciences.

Bloom, J. D. 2015. Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinformatics*, 16(1): 168.

Bonde, M. T., Pedersen, M., Klausen, M. S., Jensen, S. I., Wulff, T., Harrison, S., Nielsen, A. T., Herrgård, M. J., and Sommer, M. O. 2016. Predictable tuning of protein expression in bacteria. *Nature Methods*, 13(3): 233–236.

Bonfield, J. K., Marshall, J., Danecek, P., Li, H., Ohan, V., Whitwham, A., Keane, T., and Davies, R. M. 2021. HTSlib: C library for reading/writing high-throughput sequencing data. *GigaScience*, 10(2): giab007.

Brouillet, S., Annoni, H., Ferretti, L., and Achaz, G. 2015. MAGELLAN: a tool to explore small fitness landscapes. Pages: 031583 Section: New Results.

Bryant, D. H., Bashir, A., Sinai, S., Jain, N. K., Ogden, P. J., Riley, P. F., Church, G. M., Colwell, L. J., and Kelsic, E. D. 2021. Deep diversification of an AAV capsid protein by machine learning. *Nature Biotechnology*. Publisher: Springer US.

Bulmer, M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129(3): 897–907.

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

Chen, W.-c., Zhou, J., Sheltzer, J. M., Kinney, J. B., and Mccandlish, D. M. 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*.

Chen, W.-C., Zhou, J., and McCandlish, D. M. 2024. Density estimation for ordinal biological sequences and its applications. *Physical Review E*, 110(4): 044408. Publisher: American Physical Society.

Chou, H.-H., Chiu, H.-C., Delaney, N. F., Segrè, D., and Marx, C. J. 2011. Diminishing Returns Epistasis Among Beneficial Mutations Decelerates Adaptation. *Science*, 332(6034): 1190–1192. Publisher: American Association for the Advancement of Science.

Crawford, L., Zeng, P., Mukherjee, S., and Zhou, X. 2017. Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLOS Genetics*, 13(7): e1006869. Publisher: Public Library of Science.

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

De Los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. L. 2013. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics*, 193(2): 327–345.

De Visser, J. A. G. and Krug, J. 2014. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, 15(7): 480–490. Publisher: Nature Publishing Group.

de Visser, J. A. G., Elena, S. F., Fragata, I., and Matuszewski, S. 2018. The utility of fitness landscapes and big data for predicting evolution. *Heredity*, 121(5): 401–405. Publisher: Springer US.

Dietler, N., Lupo, U., and Bitbol, A.-F. 2023. Impact of phylogeny on structural contact inference from protein sequence data. arXiv:2209.13045 [physics, q-bio].

Domingo, J., Diss, G., and Lehner, B. 2018. Pairwise and higher-order genetic interactions during the evolution of a tRNA. *Nature*, 558(7708): 117–121.

Domingo, J., Baeza-Centurion, P., and Lehner, B. 2019. The causes and consequences of genetic interactions (epistasis). *Annual Review of Genomics and Human Genetics*, 20(1): 433–460. Publisher: Annual Reviews Inc.

Dwivedi, S. L., Heslop-Harrison, P., Amas, J., Ortiz, R., and Edwards, D. 2024. Epistasis and pleiotropy-induced variation for plant breeding. *Plant Biotechnology Journal*, 22(10): 2788–2807. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/pbi.14405.

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., and Aurell, E. 2013. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 87(1): 1–16. arXiv: 1211.1281.

Escobedo, A., Voigt, G., Faure, A. J., and Lehner, B. 2024. Genetics, energetics and allostery during a billion years of hydrophobic protein core evolution.

Faure, A. J. and Lehner, B. 2024. Mochi: neural networks to fit interpretable models and quantify energies, energetic couplings, epistasis, and allostery from deep mutational scanning data. *Genome Biology*, 25(1): 303.

Faure, A. J., Lehner, B., Miró Pina, V., Serrano Colome, C., and Weghorn, D. 2024. An extension of the walsh-hadamard transform to calculate and model epistasis in genetic landscapes of arbitrary shape and complexity. *PLOS Computational Biology*, 20(5): e1012132.

Feng, D.-F., Cho, G., and Doolittle, R. F. 1997. Determining divergence times with a protein clock: Update and reevaluation. *Proceedings of the National Academy of Sciences*, 94(24): 13028–13033.

Ferretti, L., Weinreich, D., Tajima, F., and Achaz, G. 2018. Evolutionary constraints in fitness landscapes. *Heredity*, 121(5): 466–481. Publisher: Nature Publishing Group.

Flynn, K. M., Cooper, T. F., Moore, F. B., and Cooper, V. S. 2013. The Environment Affects Epistatic Interactions to Alter the Topology of an Empirical Fitness Landscape. *PLoS Genetics*, 9(4): 1003426.

Fowler, D. M. and Fields, S. 2014. Deep mutational scanning: a new style of protein science. *Nature Methods*, 11(8): 801–807. Publisher: Nature Publishing Group.

Fragata, I., Blanckaert, A., Dias Louro, M. A., Liberles, D. A., and Bank, C. 2019. Evolution in the light of fitness landscape theory. *Trends in Ecology and Evolution*, 34(1): 69–82. Publisher: Elsevier Ltd.

Freschlin, C. R., Fahlberg, S. A., and Romero, P. A. 2022. Machine learning to navigate fitness landscapes for protein engineering. *Current Opinion in Biotechnology*, 75: 102713.

Gao, Y., Lin, K.-T., Jiang, T., Yang, Y., Rahman, M., Gong, S., Bai, J., Wang, L., Sun, J., Sheng, L., Krainer, A., and Hua, Y. 2022. Systematic characterization of short intronic splicing-regulatory elements in *SMN2* pre-mRNA. *Nucleic Acids Research*, 50(2): 731–749.

Gelman, S., Fahlberg, S. A., Heinzelman, P., Romero, P. A., and Gitter, A. 2021. Neural networks to learn protein sequence–function relationships from deep mutational scanning data. *Proceedings of the National Academy of Sciences*, 118(48): e2104878118.

Gilliot, P.-A. and Gorochowski, T. E. 2024. Transfer learning for cross-context prediction of protein expression from 5'UTR sequence. *Nucleic Acids Research*, 52(13): e58.

Haldane, A. and Levy, R. M. 2021. Mi3-GPU: MCMC-based inverse Ising inference on GPUs for protein covariation analysis. *Computer Physics Communications*, 260: 107312.

Haldane, A., Flynn, W. F., He, P., Vijayan, R., and Levy, R. M. 2016. Structural propensities of kinase family proteins from a Potts model of residue co-variation. *Protein Science : A Publication of the Protein Society*, 25(8): 1378–1384.

Haldane, A., Flynn, W. F., He, P., and Levy, R. M. 2018. Coevolutionary Landscape of Kinase Family Proteins: Sequence Probabilities and Functional Motifs. *Biophysical Journal*, 114(1): 21–31.

Happel, R. and Stadler, P. F. 1996. Canonical approximation of fitness landscapes. *Complexity*, 2(1): 53–58.

Herrera-Álvarez, S., Patton, J. E. J., and Thornton, J. W. 2025. Ancient biases in phenotype production drove the functional evolution of a protein family.

Hockenberry, A. J. and Wilke, C. O. 2019. Phylogenetic Weighting Does Little to Improve the Accuracy of Evolutionary Coupling Analyses. *Entropy*, 21(10): 1000.

Hockenberry, A. J., Stern, A. J., Amaral, L. A., and Jewett, M. C. 2018. Diversity of translation initiation mechanisms across bacterial species is driven by environmental conditions and growth demands. *Molecular Biology and Evolution*, 35(3): 582–592.

Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P. I., Springer, M., Sander, C., and Marks, D. S. 2017. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2): 128–135.

Hossain, S. 2019. Visualization of Bioinformatics Data with Dash Bio. *scipy*.

Hunter, J. D. 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3): 90–95. Conference Name: Computing in Science & Engineering.

Jalal, A. S., Tran, N. T., Stevenson, C. E., Chan, E. W., Lo, R., Tan, X., Noy, A., Lawson, D. M., and Le, T. B. 2020. Diversification of DNA-Binding Specificity by Permissive and Specificity-Switching Mutations in the ParB/Noc Protein Family. *Cell Reports*, 32(3): 107928.

Johnson, M. S., Reddy, G., and Desai, M. M. 2023. Epistasis and evolution: recent advances and an outlook for prediction. *BMC Biology*, 21(1): 120.

Johnston, K. E., Almhjell, P. J., Watkins-Dulaney, E. J., Liu, G., Porter, N. J., Yang, J., and Arnold, F. H. 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(32): e2400439121.

Khan, A. I., Dinh, D. M., Schneider, D., Lenski, R. E., and Cooper, T. F. 2011. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science*, 332(6034): 1193–1196.

Kinney, J. B. and McCandlish, D. M. 2019. Massively Parallel Assays and Quantitative Sequence–Function Relationships. *Annual Review of Genomics and Human Genetics*, 20(1): annurev–genom–083118–014845.

Kinney, J. B., Murugan, A., Callan, C. G., and Cox, E. C. 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(20): 9158–9163.

Komarova, E. S., Chervontseva, Z. S., Osterman, I. A., Evfratov, S. A., Rubtsova, M. P., Zatsepin, T. S., Semashko, T. A., Kostryukova, E. S., Bogdanov, A. A., Gelfand, M. S., Dontsova, O. A., and Sergiev, P. V. 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(4): 1254–1261.

Kondrashov, A. S., Sunyaev, S., and Kondrashov, F. A. 2002. Dobzhansky-Muller incompatibilities in protein evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 99(23): 14878–14883.

Kuo, S. T., Jahn, R. L., Cheng, Y. J., Chen, Y. L., Lee, Y. J., Hollfelder, F., Wen, J. D., and Chou, H. H. D. 2020. Global fitness landscapes of the Shine-Dalgarno sequence. *Genome Research*, 30(5): 711–723.

Kuo, S.-T., Chang, J. K., Chang, C., Shen, W.-Y., Hsu, C., Lai, S.-W., and Chou, H.-H. D. 2025. Unraveling the start element and regulatory divergence of core promoters across the domain Bacteria.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16): 2078–2079.

Li, Y. and Zhang, J. 2025. On the Probability of Reaching High Peaks in Fitness Landscapes by Adaptive Walks. *Molecular Biology and Evolution*, 42(4): msaf066.

Lipsh-Sokolik, R. and Fleishman, S. J. 2024. Addressing epistasis in the design of protein function. *Proceedings of the National Academy of Sciences*, 121(34): e2314999121. Publisher: Proceedings of the National Academy of Sciences.

Lite, T. L. V., Grant, R. A., Nocedal, I., Littlehale, M. L., Guo, M. S., and Laub, M. T. 2020. Uncovering the basis of protein-protein interaction specificity with a combinatorially complete library. *eLife*, 9(ii): 1–57.

Lorenz, R., Bernhart, S. H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. 2011. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1): 26.

Malinverni, D. and Babu, M. M. 2023. Data-driven design of orthogonal protein-protein interactions. *Science Signaling*, 16(774): eabm4484.

Marks, D. S., Hopf, T. A., and Sander, C. 2012. Protein structure prediction from sequence variation. *Nature Biotechnology*, 30(11): 1072–1080.

Martin, N. S. and Ahnert, S. E. 2022. Thermodynamics and neutral sets in the RNA sequence-structure map. *Europhysics Letters*, 139(3): 37001. Publisher: EDP Sciences, IOP Publishing and Società Italiana di Fisica.

McCandlish, D. M. 2011. Visualizing fitness landscapes. *Evolution*, 65(6): 1544–1558. Publisher: John Wiley & Sons, Ltd.

McCandlish, D. M. 2018. Long-term evolution on complex fitness landscapes when mutation is weak. *Heredity*, 121(5): 449–465.

McCandlish, D. M. and Stoltzfus, A. 2014. Modeling Evolution Using the Probability of Fixation: History and Implications. *The Quarterly Review of Biology*, 89(3): 225–252.

McCandlish, D. M., Otwinowski, J., and Plotkin, J. B. 2015. Detecting epistasis from an ensemble of adapting populations. *Evolution*, 69(9): 2359–2370.

Miton, C. M., Buda, K., and Tokuriki, N. 2021. Epistasis and intramolecular networks in protein evolution. *Current Opinion in Structural Biology*, 69: 160–168.

Moore, J. H. and Williams, S. M. 2009. Epistasis and Its Implications for Personal Genetics. *The American Journal of Human Genetics*, 85(3): 309–320. Publisher: Elsevier.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49): E1293–E1301. Publisher: Proceedings of the National Academy of Sciences.

Moulana, A., Dupic, T., Phillips, A. M., and Desai, M. M. 2023a. Genotype–phenotype landscapes for immune–pathogen coevolution. *Trends in Immunology*, 44(5): 384–396. Publisher: Elsevier.

Moulana, A., Dupic, T., Phillips, A. M., Chang, J., Roffler, A. A., Greaney, A. J., Starr, T. N., Bloom, J. D., and Desai, M. M. 2023b. The landscape of antibody binding affinity in SARS-CoV-2 Omicron BA.1 evolution. *eLife*, 12: e83442.

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

Noderer, W. L., Flockhart, R. J., Bhaduri, A., Diaz De Arce, A. J., Zhang, J., Khavari, P. A., and Wang, C. L. 2014. Quantitative analysis of mammalian translation initiation sites by facs-seq. *Molecular Systems Biology*, 10(8): 748.

Ogbunugafor, C. B., Wylie, C. S., Diakite, I., Weinreich, D. M., and Hartl, D. L. 2016. Adaptive Landscape by Environment Interactions Dictate Evolutionary Dynamics in Models of Drug Resistance. *PLOS Computational Biology*, 12(1): e1004710. Publisher: Public Library of Science.

O'Maille, P. E., Malone, A., Dellas, N., Andes Hess, B., Smentek, L., Sheehan, I., Greenhagen, B. T., Chappell, J., Manning, G., and Noel, J. P. 2008. Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. *Nature Chemical Biology*, 4(10): 617–623.

Otwinowski, J., McCandlish, D. M., and Plotkin, J. B. 2018. Inferring the shape of global epistasis. *Proceedings of the National Academy of Sciences*, 115(32): E7550–E7558.

Papkou, A., Garcia-Pastor, L., Escudero, J. A., and Wagner, A. 2023. A rugged yet easily navigable fitness landscape. *Science*, 382(6673): eadh3860. Publisher: American Association for the Advancement of Science.

Park, Y., Metzger, B. P. H., and Thornton, J. W. 2024. The simplicity of protein sequence-function relationships. *Nature Communications*, 15(1): 7953. Publisher: Nature Publishing Group.

Petti, S., Martí-Gómez, C., Kinney, J. B., Zhou, J., and McCandlish, D. M. 2025. On learning functions over biological sequence space: relating gaussian process priors, regularization, and gauge fixing. *bioRxiv*, pages 2025–04.

Phillips, P. C. 2008. Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews. Genetics*, 9(11): 855–867.

Poelwijk, F. J., Krishna, V., and Ranganathan, R. 2016. The context-dependence of mutations: a linkage of formalisms. *PLoS computational biology*, 12(6): e1004771.

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

Posfai, A., Zhou, J., McCandlish, D. M., and Kinney, J. B. 2025. Gauge fixing for sequence-function relationships. *PLOS Computational Biology*, 21(3): e1012818.

Rasmussen, C. E. and Williams, C. K. I. 2008. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass., 3. print edition.

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

Rodriguez Horta, E. and Weigt, M. 2021. On the effect of phylogenetic correlations in coevolution-based contact prediction in proteins. *PLOS Computational Biology*, 17(5): e1008957.

Romero, P. A., Krause, A., and Arnold, F. H. 2013. Navigating the protein fitness landscape with Gaussian processes. *Proceedings of the National Academy of Sciences*, 110(3).

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

Rozhoňová, H., Martí-Gómez, C., McCandlish, D. M., and Payne, J. L. 2024. Robust genetic codes enhance protein evolvability. *PLOS Biology*, 22(5): e3002594.

Russ, W. P., Figliuzzi, M., Stocker, C., Barrat-Charlaix, P., Socolich, M., Kast, P., Hilvert, D., Monasson, R., Cocco, S., Weigt, M., and Ranganathan, R. 2020. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369(6502): 440–445. Publisher: American Association for the Advancement of Science.

Sackton, T. B. and Hartl, D. L. 2016. Genotypic Context and Epistasis in Individuals and Populations. *Cell*, 166(2): 279–287.

Sailer, Z. R. and Harms, M. J. 2017. High-order epistasis shapes evolutionary trajectories. *PLoS computational biology*, 13(5): e1005541.

Salis, H. M., Mirsky, E. A., and Voigt, C. A. 2009. Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology*, 27(10): 946–950.

Sella, G. and Hirsh, A. E. 2005. The application of statistical physics to evolutionary biology. *Proceedings of the National Academy of Sciences*, 102(27): 9541–9546.

Sethi, P. and Zhou, J. 2024. Importance of higher-order epistasis in large protein sequence-function relationships. *bioRxiv*.

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

Sly, L. 2011. Reconstruction for the potts model. *Annals of Probability*, 39(4): 1365–1406.

Soo, V. W., Swadling, J. B., Faure, A. J., and Warnecke, T. 2021. Fitness landscape of a dynamic RNA structure. *PLoS Genetics*, 17(2): e1009353. ISBN: 1111111111.

Soyk, S., Benoit, M., and Lippman, Z. B. 2020. New Horizons for Dissecting Epistasis in Crop Quantitative Trait Variation. *Annual Review of Genetics*, 54(Volume 54, 2020): 287–307. Publisher: Annual Reviews.

Stadler, P. F. 1996. Landscapes and their correlation functions. *Journal of Mathematical Chemistry*, 20(1): 1–45.

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

Stadler, P. F. and Happel, R. 1999. Random field models for fitness landscapes. *Journal of Mathematical Biology*, 38(5): 435–478.

Stadler, P. F., Happel, R., *et al.* 1994. Canonical approximation of landscapes. *Santa Fe Institute Preprint*, pages 94–09.

Starr, T. N. and Thornton, J. W. 2016. Epistasis in protein evolution. *Protein Science*, 25(7): 1204–1218. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.2897.

Starr, T. N., Picton, L. K., and Thornton, J. W. 2017. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature*, 549(7672): 409–413.

Stein, R. R., Marks, D. S., and Sander, C. 2015. Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. *PLOS Computational Biology*, 11(7): e1004182.

Stormo, G. D. 2013. Modeling the specificity of protein-DNA interactions. *Quantitative Biology*, 1(2): 115–130.

Sundar, V., Tu, B., Guan, L., and 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, I. G., Schenk, M. F., Franke, J., Krug, J., and de Visser, J. A. G. M. 2013. Quantitative analyses of empirical fitness landscapes. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(01): P01005. Publisher: IOP Publishing and SISSA.

Tareen, A. and Kinney, J. B. 2020. Logomaker: beautiful sequence logos in Python. *Bioinformatics*, 36(7): 2272–2274.

Tareen, A., Ireland, W. T., Posfai, A., McCandlish, D. M., and Kinney, J. B. 2020. MAVE-NN: Quantitative modeling of genotype-phenotype maps as information bottlenecks. *bioRxiv*, pages 1–15. Publisher: Cold Spring Harbor Laboratory.

Tonner, P. D., Pressman, A., and Ross, D. 2022. Interpretable modeling of genotype–phenotype landscapes with state-of-the-art predictive power. *Proceedings of the National Academy of Sciences*, 119(26): e2114021119.

Vigué, L. and Tenaillon, O. 2023. Predicting the effect of mutations to investigate recent events of selection across 60,472 Escherichia coli strains. *Proceedings of the National Academy of Sciences*, 120(31): e2304177120. Publisher: Proceedings of the National Academy of Sciences.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17: 261–272.

Wang, T., Zhao, D., and Tian, S. 2015. An overview of kernel alignment and its applications. *Artificial Intelligence Review*, 43(2): 179–192.

Weinreich, D. M., Lan, Y., Wylie, C. S., and Heckendorn, R. B. 2013. Should evolutionary geneticists worry about higher-order epistasis? *Current Opinion in Genetics & Development*, 23(6): 700–707.

Weinreich, D. M., Lan, Y., Jaffe, J., and Heckendorn, R. B. 2018. The Influence of Higher-Order Epistasis on Biological Fitness Landscape Topography. *Journal of Statistical Physics*, 172(1): 208–225. Publisher: Springer US.

Weinstein, J. Y., Martí-Gómez, C., Lipsh-Sokolik, R., Hoch, S. Y., Liebermann, D., Nevo, R., Weissman, H., Petrovich-Kopitman, E., Margulies, D., Ivankov, D., McCandlish, D. M., and Fleishman, S. J. 2023. Designed active-site library reveals thousands of functional GFP variants. *Nature Communications*, 14(1): 2890.

Wen, J. D., Kuo, S. T., and Chou, H. H. D. 2020. The diversity of Shine-Dalgarno sequences sheds light on the evolution of translation initiation. *RNA Biology*, 00(00): 1–12. Publisher: Taylor & Francis.

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

Westmann, C. A., Goldbach, L., and Wagner, A. 2024b. The highly rugged yet navigable regulatory landscape of the bacterial transcription factor TetR. *Nature Communications*, 15(1): 10745. Publisher: Nature Publishing Group.

Wong, M. S., Kinney, J. B., and Krainer, A. R. 2018. Quantitative Activity Profile and Context Dependence of All Human 5' Splice Sites. *Molecular Cell*, 71(6): 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*, pages 356–366.

Wu, N. C., Dai, L., Olson, C. A., Lloyd-Smith, J. O., and Sun, R. 2016. Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife*, 5(July): 1–21.

Yang, K. K., Wu, Z., and Arnold, F. H. 2019. Machine-learning-guided directed evolution for protein engineering. *Nature Methods*, 16(8): 687–694.

Yeo, G. and Burge, C. B. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. In *Journal of Computational Biology*, volume 11, pages 377–394. Issue: 2-3 ISSN: 10665277.

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

Zebell, S. G., Martí-Gómez, C., Fitzgerald, B., Cunha, C. P., Lach, M., Seman, B. M., Hendelman, A., Sretenovic, S., Qi, Y., Bartlett, M., Eshed, Y., McCandlish, D. M., and Lippman, Z. B. 2025. Cryptic variation fuels plant phenotypic change through hierarchical epistasis. *Nature*, pages 1–9. Publisher: Nature Publishing Group.

Zhou, J. and McCandlish, D. M. 2020. Minimum epistasis interpolation for sequence-function relationships. *Nature Communications*, 11(1).

Zhou, J., Wong, M. S., Chen, W.-c., Krainer, A. R., Justin, B., and Mccandlish, D. M. 2022. Higher-order epistasis and phenotypic prediction. *Proc. Natl. Acad. Sci. USA*, 119(39).

Zhou, J., Martí-Gómez, C., Petti, S., and McCandlish, D. M. 2025. Learning sequence-function relationships with scalable, interpretable gaussian processes. *bioRxiv*.

Supplementary Figures

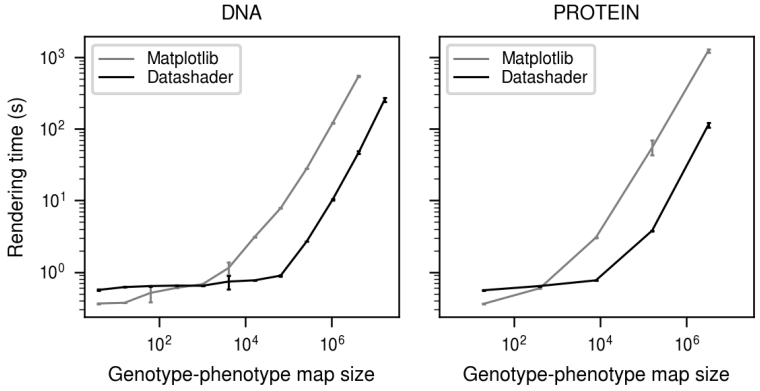

Fig. S1. Visualization rendering times using two different back-end libraries for plotting as a function of the size of DNA and protein genotype-phenotype maps.

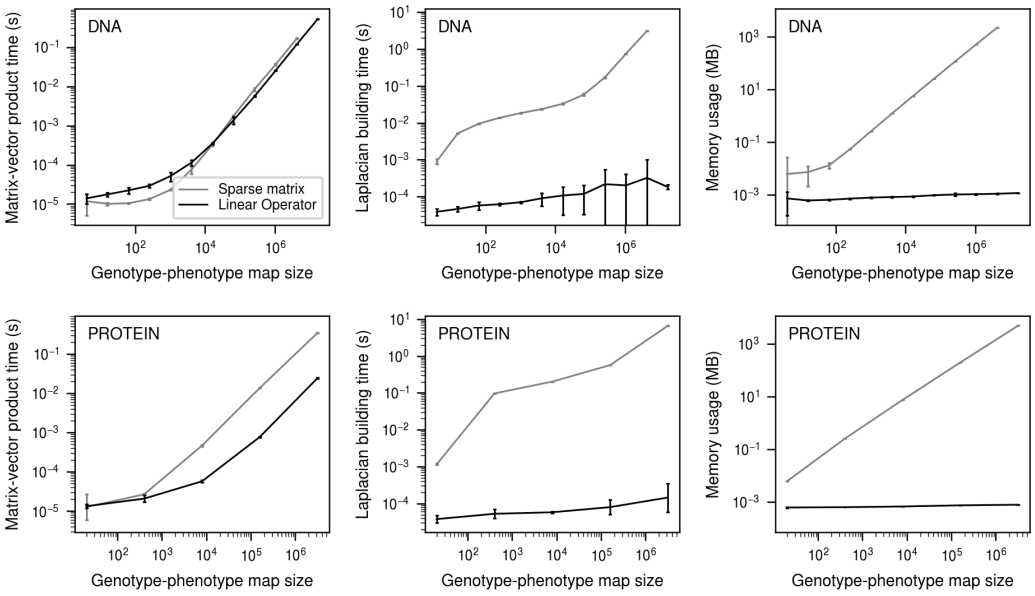

Fig. S2. Comparison of the running times and memory requirements for computation of matrix-vector products with the Laplacian of the Hamming graph using our new Linear Operator or our previous sparse matrix formulation.

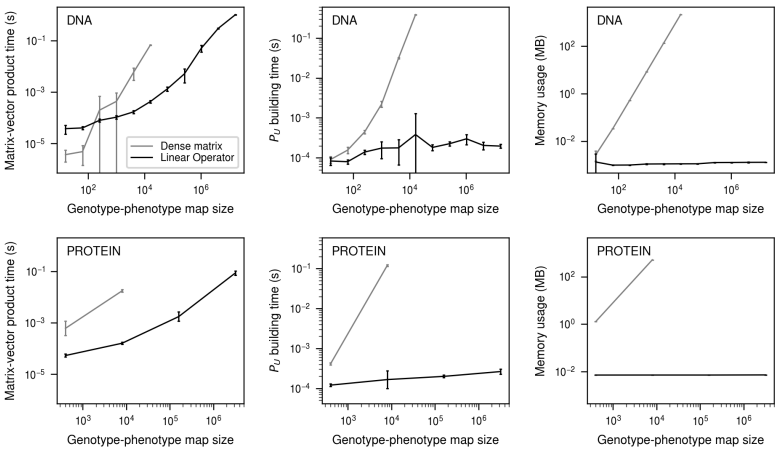

Fig. S3. Comparison of the running times and memory requirements for computation of P_U matrix-vector products using our Linear Operator or the corresponding dense matrix.

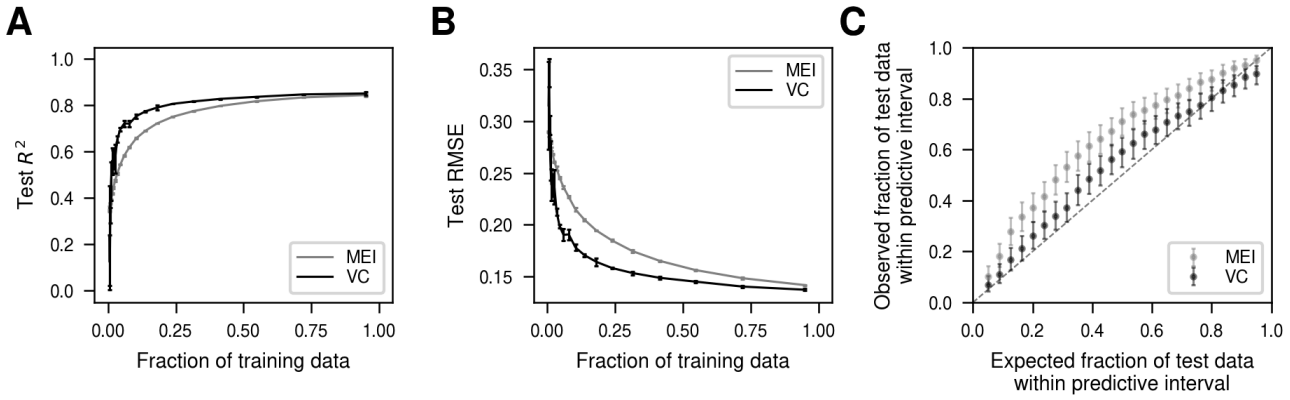

Fig. S4. Predictive performance of Minimum Epistasis Interpolation (MEI) and Variance Component (VC) regression in held-out data. (A,B) Model predictive performance measured by the R^2 (A) and RMSE (B) in held-out data as a function of the fraction of data used for training and phenotypic prediction. Error bars represent the standard deviation over 3 independent subsets of sequences used for training at each proportion. (C) Evaluation of the models calibration by comparing the expected fraction of times a predictive interval will contain the real phenotypic value compared to the fraction of times it actually contained the measured phenotype across 274 test data points. Error bars represent the 95% Jeffreys confidence interval for the estimated fraction of data points laying within the corresponding predictive interval. Diagonal dashed gray line shows the expectation under perfect model calibration.

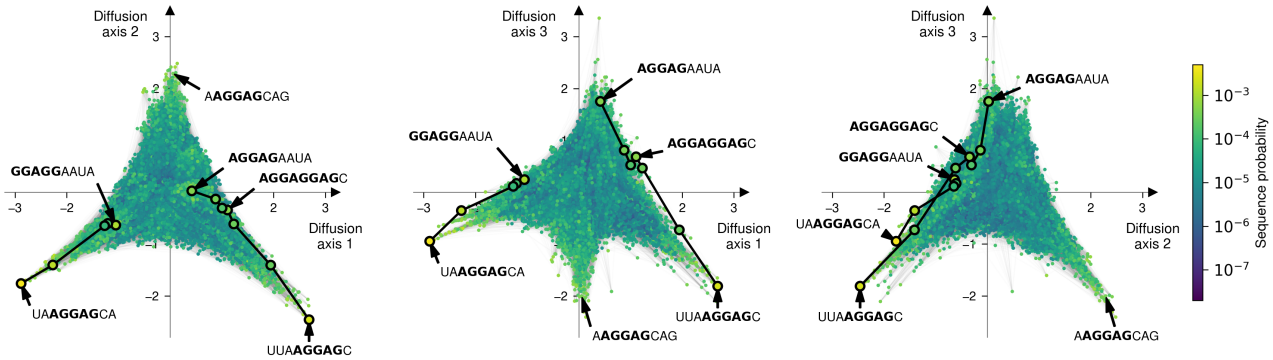

Fig. S5. Low-dimensional representation of the Shine-Dalgarno probability distribution inferred with SeqDEFT along Diffusion axes 1, 2 and 3. Every dot represents one of the possible 4^9 possible sequences and is colored according to their inferred probability. Sequences are laid out according to the indicated Diffusion axes and dots are plotted in order according to the missing Diffusion axis.

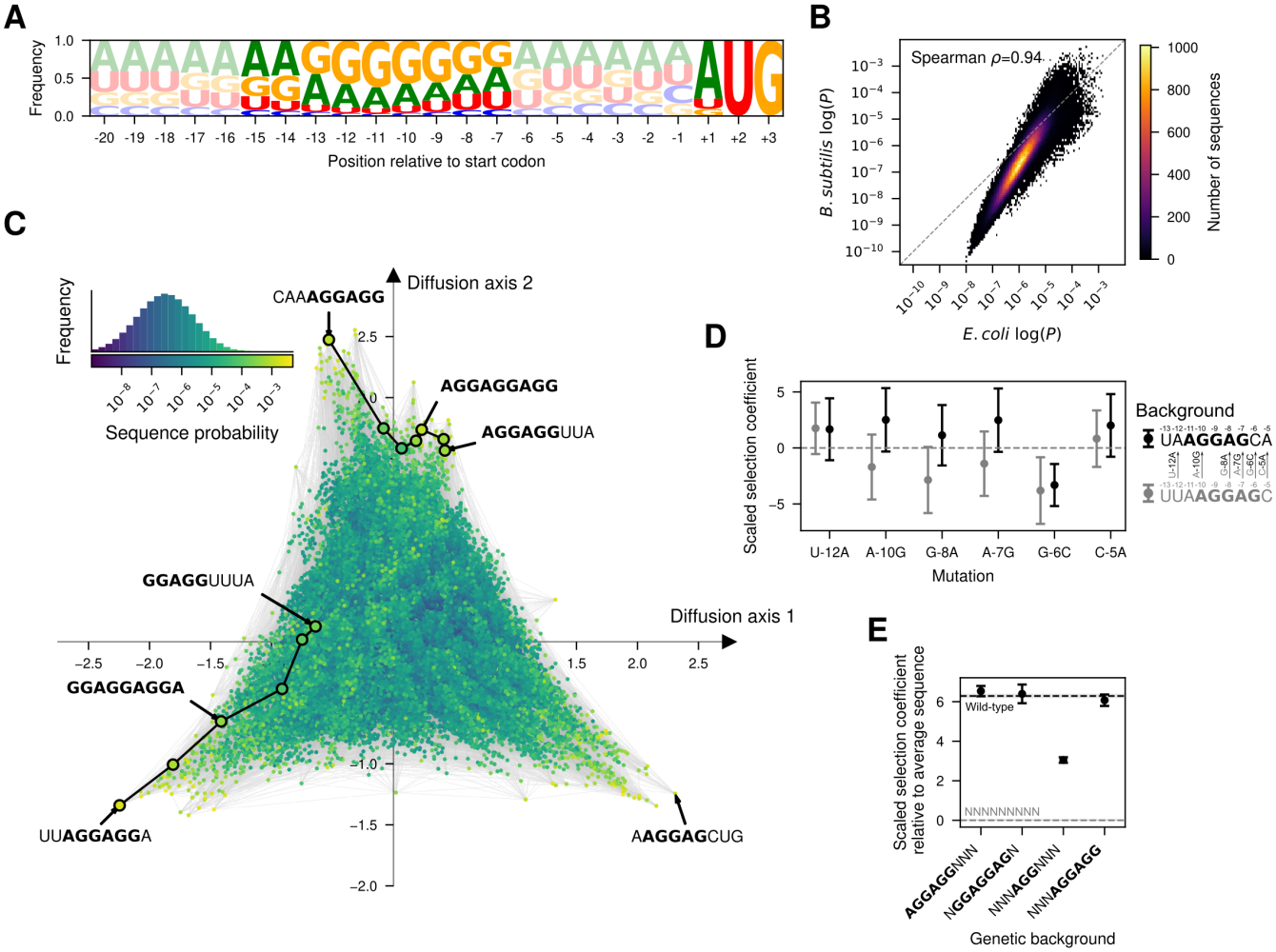

Fig. S6. The structure of the genotype-phenotype map inferred from *B. subtilis* is conserved. (A) Sequence logo representing the site-specific allele frequencies of 4,328 5'UTRs in the *B. subtilis* genome aligned with respect to the annotated start codon. The start codon and the 9 nucleotide sequences 6 bases upstream are highlighted to emphasize the most relevant cis-regulatory sequences for translation initiation. (B) Two-dimensional histogram representing the relationship between the inferred sequence probabilities from their frequency in the *E. coli* and *B. subtilis* genomes. (C) Low-dimensional representation of the Shine-Dalgarno probability distribution inferred with SeqDEFT. Every dot represents one of the possible 4^9 possible sequences and is colored according to its inferred probability. The inset represents the distribution of inferred sequence probabilities along with their corresponding color in the visualization. Sequences are laid out according to the first two Diffusion axes and dots are plotted in order according to the 3rd Diffusion axis. (D) Posterior distribution inferred by SeqDEFT for the scaled selection coefficient of specific mutations when introduced in two genetic contexts, UUAAGGAGC and UAAGGAGCA, representing a shift of the AGGAG motif by one nucleotide. Note that estimated mutational effects, except for G-6C, are largely compatible with those estimated from the *E. coli* genome shown in Figure 4D in the UAAGGAGCA context. (E) Posterior distribution inferred by SeqDEFT for the average scaled selection coefficient, relative to the average across all possible sequences, for genotypes containing the AGGAGG motif at positions separated by three nucleotides, along with their potential mutational intermediates. Horizontal dashed lines represent posterior mean of the average phenotype across all possible sequences (grey) or wild-type genomic sequences (black). Shaded areas represent the 95% credible intervals. (D,E) Points represent the maximum a posteriori (MAP) estimates and error bars represent the 95% credible intervals.

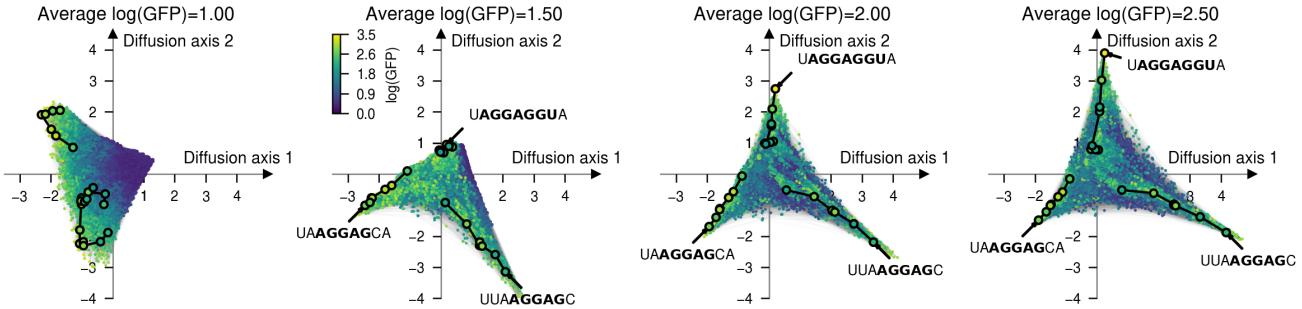

Fig. S7. Low-dimensional representation of the Shine-Dalgarno genotype-phenotype map inferred with VC regression from MAVE data along Diffusion axes 1 and 2 as a function of the assumed average log(GFP) at the stationary distribution (as determined by tuning the strength of selection parameter c). Every dot represents one of the 4^9 possible sequences and is colored according to its inferred log(GFP) values. Dots are plotted in order according to Diffusion axes 3.

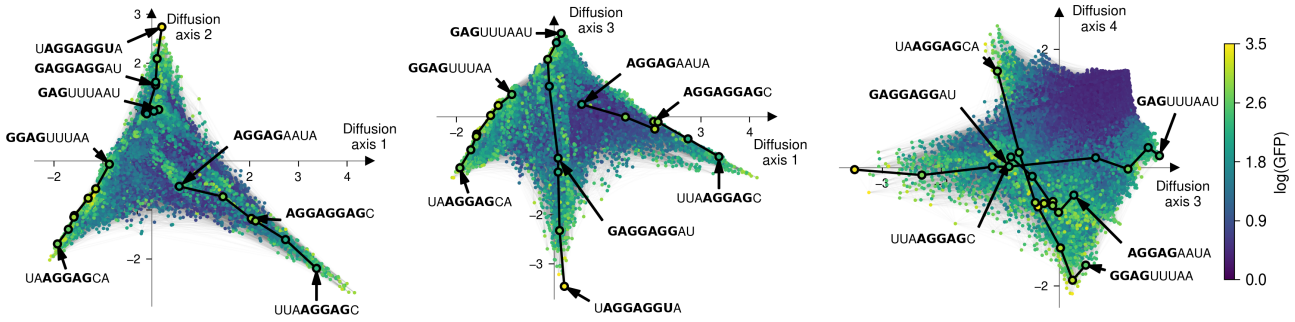

Fig. S8. Low-dimensional representation of the Shine-Dalgarno genotype-phenotype map inferred with VC regression from MAVE data along Diffusion axes 1, 2, 3 and 4. Every dot represents one of the possible 4^9 possible sequences and is colored according to its inferred log(GFP) value. Sequences are laid out according to the indicated Diffusion axes and dots are plotted in order according to Diffusion axes 3, 2 and 1, respectively for each panel.

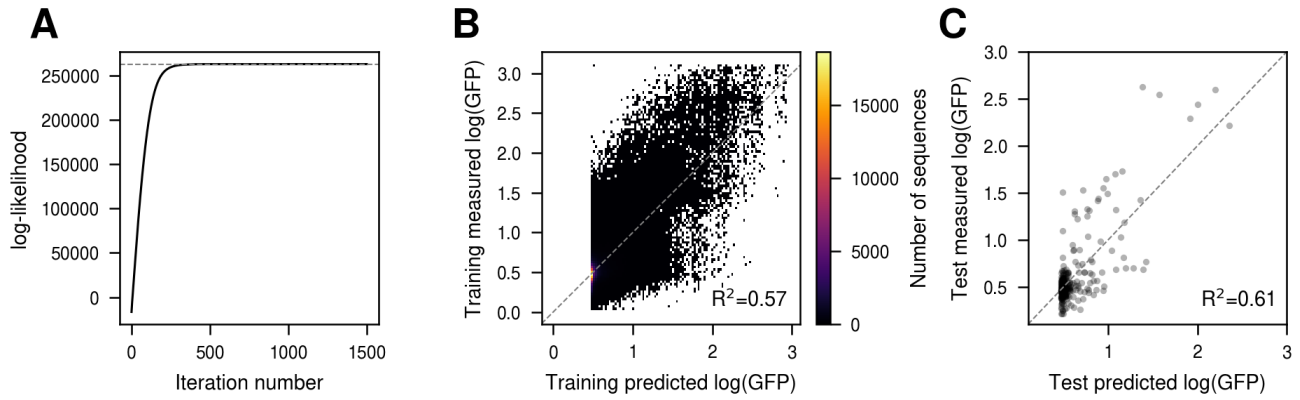

Fig. S9. Fitting a thermodynamic model to the Shine-Dalgarno genotype-phenotype map using MAVE data. (A) Training curve showing the evolution of the log-likelihood as a function of the number of iterations of the Adam optimizer. (B) Comparison of measured log(GFP) in the training data with the predicted values under the estimated thermodynamic model. (C) Comparison of measured log(GFP) in the test data with the predicted values under the estimated thermodynamic model.

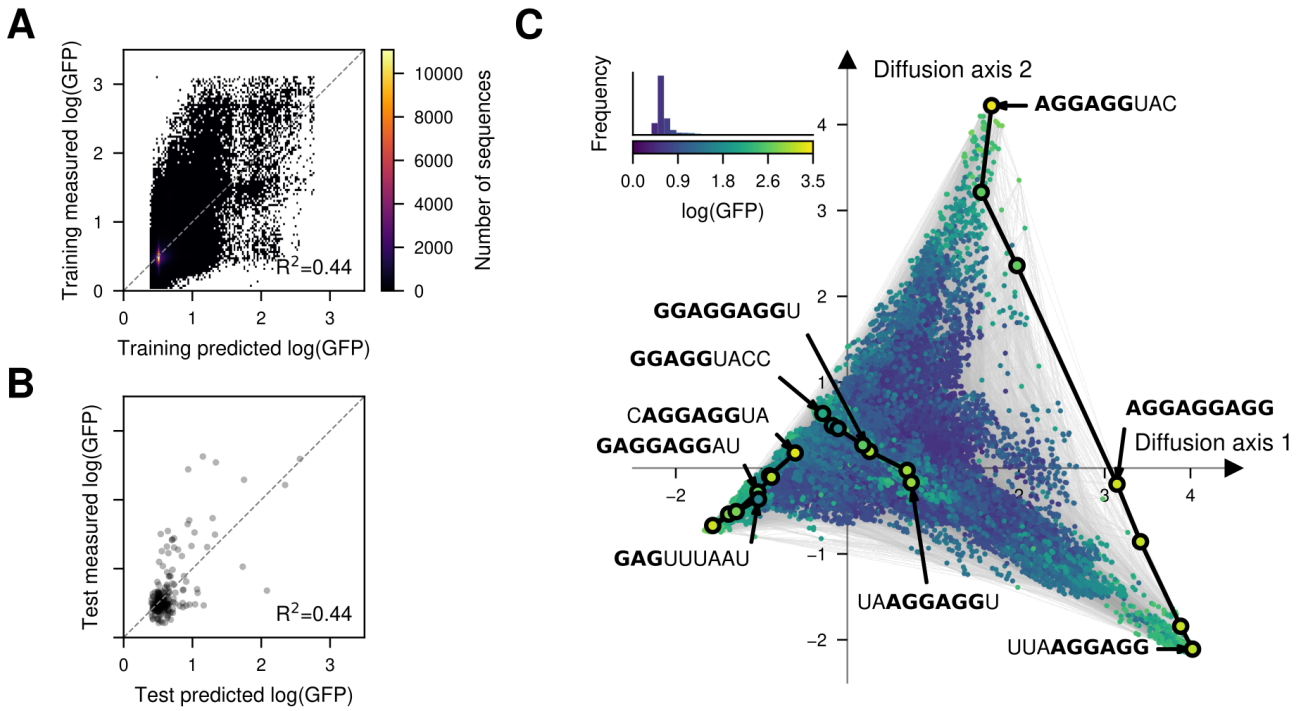

Fig. S10. A general thermodynamic model for RNA folding for predicting MAVE data for the Shine-Dalgarno sequence. (A,B) Comparison of measured log(GFP) in the training (A) and test data (B) with the predicted values under the calibration model using RNAcofold ensemble binding energies with the anti-SD sequence. (C) Visualization of the genotype-phenotype map that results from predicting the phenotype of every possible sequence RNAcofold ensemble binding energies to the anti-SD sequence. Every dot represents one of the possible 4^9 possible sequences and is colored according to the predicted log(GFP). 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.

Supplementary Information

A. Properties of the P_U projection operators. Let f represent a genotype-phenotype map on the space of sequences with a single site and α different alleles. The function f can be projected into a subspace spanned by any vector b using the orthogonal projection matrix $P = b(b^T b)^{-1} b$. Letting V_{con} be the constant subspace spanned by the α-dimensional vector of ones $b = \vec{1}$, the orthogonal projection matrix into V_{con} is thus given by $P_{con} = \frac{1}{\alpha}\vec{1}\vec{1}^T$. Let V_{add} be the orthogonal subspace ($V_{con} \perp V_{add}$), defined by its projection matrix $P_{add} = I - P_{con} = I - \frac{1}{\alpha}\vec{1}\vec{1}^T$. Thus, for any pair of sequences x, y, $P_{con}(x,y) = \frac{1}{\alpha}$ and

$$P_{add}(x,y) = \begin{cases} \frac{-1}{\alpha} & \text{if } x \neq y \\ \frac{\alpha-1}{\alpha} & \text{if } x = y. \end{cases}$$

For a genotype-phenotype map f in the space of sequences of length ℓ, these elementary subspaces can be combined through tensor products into 2^ℓ different $V_U = \bigotimes_p^\ell V_p$ subspaces defined by the set of sites U, such that $V_p = V_{add}$ for $p \in U$ and $V_p = V_{con}$ for $p \notin U$. Thus, the projection operator into the subspaces defined by U are obtained through the Kronecker product $P_U = \bigotimes_p^\ell P_p$, the elements of which are

$$P_U(x,y) = \alpha^{-\ell} \prod_{\substack{p \in U \\ x_p \neq y_p}} (-1) \prod_{\substack{p \in U \\ x_p = y_p}} (\alpha - 1). \tag{4}$$

It is easy to show that the resulting subspaces V_U are orthogonal to each other using the mixed product property of the Kronecker product $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$. As a consequence, if either $AC = 0$ or $BD = 0$, then $A \otimes B$ and $C \otimes D$ are orthogonal. If we consider two subspaces defined by different subsets of sites U and U', there will be at least one position p at which the two site-specific elementary subspaces are orthogonal to each other, i.e. $P_p P'_p = 0$. Consequently, $P_U P'_U = (\bigotimes_p^\ell P_p)(\bigotimes_p^\ell P'_p) = \bigotimes_p^\ell (P_p P'_p) = 0$.

Next, we consider the subspaces defined by the direct sum of subspaces V_U defined by exactly k sites $V_k = \bigoplus_{U:|U|=k} V_U$ and derive the corresponding projection operator P_k:

$$\begin{aligned} P_k(x,y) &= \sum_{U:|U|=k} P_U(x,y) = \sum_{U:|U|=k} \alpha^{-\ell} \prod_{\substack{p \in U \\ x_p \neq y_p}} (-1) \prod_{\substack{p \in U \\ x_p = y_p}} (\alpha - 1) \\ &= \alpha^{-\ell} \sum_{U:|U|=k} \prod_{\substack{p \in U \\ x_p \neq y_p}} (-1) \prod_{\substack{p \in U \\ x_p = y_p}} (\alpha - 1). \end{aligned} \tag{5}$$

We note that the elements in the sum are obtained by multiplying the factors $(\alpha - 1)$ and (-1) together a total of k times. These products can take only $k+1$ possible values: $(-1)^q(\alpha-1)^{\ell-q}$ for $q = 0, 1, \ldots, \ell$, where q represents the number of missmatches between sequences x and y within sets of sites U. Therefore, P_k can be expressed by:

$$P_k(x,y) = \alpha^{-\ell} \sum_q^k (-1)^q (\alpha-1)^{k-q} n_q(x,y), \tag{6}$$

where $n_q(x,y)$ is the number of times each unique value appears when summing through the corresponding P_U matrices. Because we are summing over all possible U of the same size, $n_q(x,y)$ does not depend on the specific sites or alleles at which x and y differ, but only on the Hamming distance between them $d(x,y)$. Specifically $n_q(x,y)$ is obtained by multiplying the number of ways to select q sites within the set of $d(x,y)$ different sites by the number of ways to select the remaining $k-q$ sites within the set of $\ell - d(x,y)$ matching sites:

$$P_k(x,y) = \alpha^{-\ell} \sum_q^k (-1)^q (\alpha-1)^{k-q} \binom{d(x,y)}{q} \binom{\ell - d(x,y)}{k-q} = \alpha^{-\ell} \mathcal{K}_k(d(x,y); \ell, \alpha), \tag{7}$$

where $\mathcal{K}_k(d(x,y); \ell, \alpha)$ denote the Krawtchouk polynomials (Stadler *et al.*, 1994; Zhou *et al.*, 2022). Note that this expression corresponds to the projection operator P_k into the subspace of pure k-th order interactions. Therefore, this shows that this k-th order interaction subspace V_k can be decomposed into smaller orthogonal subspaces V_U corresponding to pure k-th order interactions involving specific subsets of sites.

It is also easy to show that the columns of P_U are in the k-th eigenspace of the graph Laplacian by taking its projection and using the orthogonality properties of the P_U:

$$P_k P_U = \sum_{U':|U'|=k} P_{U'} P_U = \begin{cases} P_U & \text{if } |U| = k \\ 0 & \text{if } |U| \neq k. \end{cases} \tag{8}$$

Using the same argument, we can show that $P_U P_k = P_U \sum_{U':|U'|=k} P_{U'} = P_U$ when $|U| = k$ and 0 otherwise.

B. Variance explained by interactions among sets of sites. In section A, we introduced the projection matrix P_U. This enables decomposing a genotype-phenotype map f into its 2^ℓ components f_U, each representing pure $|U|$-order interactions among sites in a set U. Here, we present two coarse-grained views of these components by quantifying the percentage variance explained by genetic interactions involving specific sites in a genotype-phenotype map f, where the total variance is given by

$$\mathrm{Var}[f] = \alpha^{-\ell}(f - \bar{f})^T (f - \bar{f}), \tag{9}$$

where $\bar{f} = \alpha^{-\ell} \sum_i f_i$. In the first approach, we consider the variance explained by interactions of order k involving each site p in the sequence:

$$\mathrm{Var}_k^p[f] = \alpha^{-\ell}(f_k^p)^T f_k^p, \tag{10}$$

where

$$f_k^p = \sum_{\substack{U:p\in U \\ |U|=k}} P_U f. \tag{11}$$

The values $100 \times \frac{\mathrm{Var}_k^p[f]}{\mathrm{Var}[f]}$ for every combination of k and p can be arranged in an $\ell \times \ell$-dimensional matrix to reveal which sites contribute most to variance, and how much through additive, low-order or higher-order genetic interactions.

In the second approach, we consider the variance explained by interactions of order $k = 2$ and $k > 2$ involving pairs of sites p, q:

$$\mathrm{Var}_2^{p,q}[f] = \alpha^{-\ell}(f_2^{p,q})^T f_2^{p,q}, \tag{12}$$

$$\mathrm{Var}_{k>2}^{p,q}[f] = \alpha^{-\ell}(f_{k>2}^{p,q})^T f_{k>2}^{p,q}, \tag{13}$$

$$f_2^{p,q} = \sum_{\substack{U:\{p,q\}\subseteq U \\ |U|=2}} P_U f, \tag{14}$$

$$f_{k>2}^{p,q} = \sum_{\substack{U:\{p,q\}\subseteq U \\ |U|>2}} P_U f, \tag{15}$$

These values can be summarized in an $\ell \times \ell$-dimensional matrix M

$$M_{p,q}(f) = \begin{cases} 100 \times \frac{\mathrm{Var}_2^{p,q}[f]}{\mathrm{Var}_2[f]} & \text{if } p < q \\ 0 & \text{if } p = q \\ 100 \times \frac{\mathrm{Var}_{k>2}^{p,q}[f]}{\sum_{k>2} \mathrm{Var}_k[f]} & \text{if } p > q, \end{cases} \tag{16}$$

where $\mathrm{Var}_k[f] = f_k^T f_k$. This matrix reveals the interaction patterns between pairs of sites, distinguishing pairwise from higher-order effects.

C. Linear operator for Kronecker products. In section A, we describe the projection matrix P_U that projects a function f into the subspace corresponding to pure interactions between sites in a set U. Here we describe an efficient method for calculating the product $P_U f$ by noting that P_U can be written as a Kronecker product of ℓ matrices ($P_U = \bigotimes_p^\ell P_p$).

Let A be an arbitrary matrix obtained through an ℓ-Kronecker product, i.e. $A = \bigotimes_p^\ell A_p$ with entries given by $A(x, y) = \prod_p A_p(x_p, y_p)$ and $\mathbf{B} \in \mathbb{R}^{(\alpha\times\alpha\times\ldots\times\alpha)}$ be a tensor with ℓ dimensions such that $\mathbf{B}_{x_1,x_2,\ldots,x_\ell} = b_x$ for any vector $b \in \mathbb{R}^{\alpha^\ell}$. Thus,

any matrix vector product Ab can be computed without explicitly constructing A using tensor dot products as follows:

$$\begin{aligned}(Ab)_x &= \sum_y A(x,y)b_y \\ &= \sum_{y_1,\ldots,y_\ell} \left(\prod_{p=1}^{\ell} A_p(x_p,y_p)\right) \mathbf{B}_{y_1,\ldots,y_\ell} \\ &= \sum_{y_2,\ldots,y_\ell} \left(\prod_{p=2}^{\ell} A_p(x_p,y_p)\right) \sum_{y_1} A_1(x_1,y_1)\mathbf{B}_{y_1,\ldots,y_\ell}.\end{aligned} \tag{17}$$

Let $\mathbf{B}^{(1)}_{x_1,y_2,\ldots,y_\ell} = \sum_{y_1} A_1(x_1,y_1)\mathbf{B}_{y_1,\ldots,y_\ell}$ and repeat the same operation:

$$\begin{aligned}(Ab)_x &= \sum_{y_2,\ldots,y_\ell} \left(\prod_{p=2}^{\ell} A_p(x_p,y_p)\right) \mathbf{B}^{(1)}_{x_1,\ldots,y_\ell} \\ &= \sum_{y_3,\ldots,y_\ell} \left(\prod_{p=3}^{\ell} A_p(x_p,y_p)\right) \sum_{y_2} A_2(x_2,y_2)\mathbf{B}^{(1)}_{x_1,y_2,\ldots,y_\ell} \\ &= \sum_{y_3,\ldots,y_\ell} \left(\prod_{p=3}^{\ell} A_p(x_p,y_p)\right) \mathbf{B}^{(2)}_{x_1,x_2,y_3,\ldots,y_\ell}.\end{aligned} \tag{18}$$

Thus, it is easy to see that $(Ab)_x$ corresponds to $\mathbf{B}^{(\ell)}_{x_1,\ldots,x_\ell}$ under the general recursion

$$\mathbf{B}^{(i)}_{x_1,\ldots,x_i,y_{i+1},\ldots,x_\ell} = \sum_{y_i} A_i(x_i,y_i)\mathbf{B}^{(i-1)}_{x_1,\ldots,x_{i-1},y_i,\ldots,y_\ell}, \tag{19}$$

where $\mathbf{B}^{(0)}_{y_1,\ldots,y_\ell} = \mathbf{B}_{y_1,\ldots,y_\ell} = b_y$. Note that each of the ℓ steps in the recursion can be computed efficiently as a tensor dot product between an $\alpha \times \alpha$ matrix and an $\alpha_1 \times_2 \alpha \times \cdots \times \alpha_\ell$ tensor requiring $\alpha^{\ell-1} \times \alpha^2$ operations. The total number of required operations, $\ell\alpha^{\ell+1}$, is much smaller than the $\alpha^{2\ell}$ operations required for naively computing Ab, but more importantly, this strategy reduces memory requirements from the prohibitive scaling with $\alpha^{2\ell}$ for storing A to only α^ℓ for storing $\mathbf{B}$, enabling practical computation for α^ℓ in the order of millions.

D. Linear operator for Laplacian of the Hamming graph. The space of possible sequences of length ℓ and α different alleles can be represented by a Hamming graph, in which nodes represent genotypes and edges represent single point mutations. The Laplacian matrix L of this graph is given by

$$L(i,j) = \begin{cases} -1 & \text{if } i \text{ and } j \text{ are neighbors,} \\ \ell(\alpha-1) & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

This matrix is sparse and can be stored in Compressed Sparse Row (CSR) format to efficiently compute matrix vector products. Despite the sparsity, there are still $\alpha^\ell \times (1+\ell(\alpha-1))$ non-zero entries. For a the space of sequences of length 9 with 4 alleles with 64 bits floating point values, only storing the non-zero entries would require about 450MB. Here, we develop a matrix-free function to compute matrix-vector products with the Laplacian matrix Lb by leveraging the highly regular structure of this matrix and tensor broadcasting, resulting in memory requirements that scale only with the size of sequence space α^ℓ. In particular, we can express Lb as

$$Lb = (\ell(\alpha-1)I - A)b = (\ell\alpha I - (\ell I + A))b = \ell\alpha b - (\ell I + A)b = \ell\alpha b - w,$$

where A is the adjacency matrix and w is the product $(\ell I + A)b$. For a fixed choice of b, let $\mathbf{B} \in \mathbb{R}^{(\alpha\times\alpha\times\ldots\times\alpha)}$ be a tensor with ℓ dimensions such that $\mathbf{B}_{x_1,x_2,\ldots,x_\ell} = b_x$, where x represents a sequence and x_i the allele at position i. Then for the same choice of b, the tensor $\mathbf{W}$ having elements $\mathbf{W}_{x_1,x_2,\ldots,x_\ell} = w_x$ can be easily computed in tensor form using broadcasting by using the trick

$$\mathbf{W} = \sum_i^{\ell} \mathbf{B}^{(i)}. \tag{21}$$

$\mathbf{B}^{(i)}_{x_1,\ldots,x_{i-1},*,x_{i+1},\ldots,x_\ell} = \sum_c^\alpha \mathbf{B}_{x_1,\ldots,x_{i-1},c,x_{i+1},\ldots,x_\ell}$, where the '*' character indicates broadcasting, i.e. all characters at position i lead to the same value. Thus, this can be efficiently computed by summing the entries of tensor $\mathbf{B}$ over axis i. We can then use $\mathbf{B}$ and $\mathbf{W}$ to calculate $w = (\ell I + A)b$ as:

$$\begin{aligned} w_x = \mathbf{W}_{x_1,x_2,\ldots,x_\ell} &= \sum_i^\ell \sum_c^\alpha \mathbf{B}_{x_1,\ldots,x_{i-1},c,x_{i+1},\ldots,x_\ell} \\ &= \ell \mathbf{B}_{x_1,x_2,\ldots,x_\ell} + \sum_i^\ell \sum_{c \neq x_i}^\alpha \mathbf{B}_{x_1,\ldots,x_{i-1},c,x_{i+1},\ldots,x_\ell} \\ &= \ell b_x + (Ab)_x = ((\ell I + A)b)_x. \end{aligned}$$

E. Minimum epistasis interpolation solution and posterior distribution. In this section, we derive the minimum epistasis interpolation solution as the maximum a posteriori (MAP) estimate of a Gaussian process model under a prior distribution on local epistatic coefficients. Let us consider a complete genotype-phenotype map given by an α^ℓ-dimensional vector f and define an improper prior distribution defined by the precision matrix $C = \frac{a}{s}\Delta^P$, such that $\log \mathrm{p}(f) \propto -f^T C f$.

Assuming we know exactly the phenotypes f_x for a subset of sequences x, we aim to compute the posterior probability of the phenotypes f_z at all remaining unobserved sequences z given by $\mathrm{p}(f_z|f_x)$. Let us define the joint log-probability distribution over $[f_x, f_z]$ given by

$$\begin{aligned} \log \mathrm{p}(f) = \log p\left(\begin{bmatrix} f_x \\ f_z \end{bmatrix}\right) \propto -\frac{1}{2} f^T C f = -\frac{1}{2}\begin{bmatrix} f_x^T & f_z^T \end{bmatrix} \begin{bmatrix} C_{xx} & C_{xz} \\ C_{xz} & C_{zz} \end{bmatrix} \begin{bmatrix} f_x \\ f_z \end{bmatrix} \\ = -\frac{1}{2}\left(f_x^T C_{xx} f_x + 2 f_x^T C_{xz} f_z + f_z^T C_{zz} f_z\right). \end{aligned}$$

Since C_{zz} is a principal submatrix of the positive semidefinite matrix C, it is also positive semidefinite for any z, and thus $\log \mathrm{p}(f_z|f_x)$ is convex. We can now take the gradient $\nabla_{f_z} \log \mathrm{p}(f_z|f_x) = f_x^T C_{xz} f_z + C_{zz} f_z$ and find the MAP $\hat{f}_z$ as the solution to the equation $C_{zz}\hat{f}_z + C_{zx} f_x = 0$, which is unique if and only if the corresponding $P-1$-th order model is uniquely determined (Zhou and McCandlish, 2020) and is given by $\hat{f}_z = -C_{zz}^{-1} C_{zx} f_x$. In this case, it is easy to verify that the posterior covariance is given by C_{zz}^{-1} as

$$\begin{aligned} \log \mathrm{p}(f_z|f_x) &\propto (f_z + C_{zz}^{-1} C_{zx} f_x)^T C_{zz} (f_z + C_{zz}^{-1} C_{zx} f_x) = \\ &= f_z^T C_{zz} f_z + 2 f_x^T C_{xz} C_{zz}^{-1} C_{zz} f_z + f_x^T C_{xz} C_{zz}^{-1} C_{zz} C_{zz}^{-1} C_{zx} f_x \\ &= f_z^T C_{zz} f_z + 2 f_x^T C_{xz} f_z + f_x^T C_{xz} C_{zz}^{-1} C_{zx} f_x \propto f_z^T C_{zz} f_z + 2 f_x^T C_{xz} f_z. \end{aligned} \tag{22}$$

In the case in which $C = \frac{a}{s}\Delta^{(P)}$, the posterior mean $\hat{f}_z = -(\Delta^{(P)}_{zz})^{-1}\Delta^{(P)}_{zx} f_x$ is independent of a. The posterior covariance, on the other hand, is independent of f_x, but depends on a and the pattern of observations. Thus, if we want to compute the variance of the phenotypic predictions of specific sequences, we define the a^* such that the expected sum of squared local epistatic coefficients $\mathbb{E}[\overline{\epsilon_P^2}] = \frac{1}{a}\mathrm{rank}(\Delta^{(P)})$ under the prior (Chen *et al.*, 2021) matches the one under the posterior mean given by $\frac{1}{s_P}\hat{f}^T \Delta^{(P)} \hat{f}$.

F. Posterior distribution computation with the cost matrix. Given a set of n measurements y in a subset of sequences x with measurement variances arranged along the diagonal of an $n \times n$ matrix D, we aim to obtain the complete genotype-phenotype map represented by the α^ℓ-dimensional vector $\hat{f}$ that maximizes the posterior probability of f given the observations y, i.e. we wish to find

$$\hat{f} = \arg\max_f \log \mathrm{p}(f|y).$$

We begin by defining an $\alpha^\ell \times n$ matrix X relating the points in the complete space with the n observed values, such that $f_x = X^T f$

$$X_{ij} = \begin{cases} 1 & \text{if observation } j \text{ corresponds to sequence } i \\ 0 & \text{otherwise.} \end{cases}$$

Using X, we can then write the posterior log-probability as a function of f:

$$\log \mathrm{p}(f|y) \propto -\frac{1}{2} f^T C f - \frac{1}{2}(y - X^T f)^T D^{-1}(y - X^T f). \tag{23}$$

We can then expand this expression to separate factors that depend on f from those that depend only on the data y.

$$\begin{aligned}\log \mathrm{p}(f|y) &\propto -\frac{1}{2} f^T C f - \frac{1}{2} y^T D^{-1} y + f^T X D^{-1} y - \frac{1}{2} f^T X D^{-1} X^T f \\ &= -\frac{1}{2} f^T (C + X D^{-1} X^T) f + f^T X D^{-1} y - \frac{1}{2} y^T D^{-1} y.\end{aligned}$$

We next take the gradient with respect to f

$$\nabla_f \log \mathrm{p}(f|y) = -(C + X D^{-1} X^T) f + X D^{-1} y,$$

and solve for $\hat{f}$ by setting $\nabla_f \log \mathrm{p}(f|y) = 0$, which yields

$$\hat{f} = (C + X D^{-1} X^T)^{-1} X D^{-1} y. \tag{24}$$

If C is invertible, then we can define a kernel matrix $K = C^{-1}$ over the complete genotype-phenotype map and verify that this is equivalent to the classical solution for the posterior mean of a Gaussian process model using Woodbury's identity. Specifically, for a subset of sequences z and the α^ℓ by $|z|$ matrix Z defined by:

$$Z_{ij} = \begin{cases} 1 & \text{if sequence } i \text{ is the } j\text{-th member of } z \\ 0 & \text{otherwise,} \end{cases}$$

we find that

$$\begin{aligned}\hat{f}_z &= Z^T \hat{f} \\ &= Z^T (C + X D^{-1} X^T)^{-1} X D^{-1} y \\ &= Z^T \left(K - K X (X^T K X + D)^{-1} X^T K \right) X D^{-1} y \\ &= Z^T K \left(I - X (X^T K X + D)^{-1} X^T K \right) X D^{-1} y \\ &= Z^T K \left(X - X (X^T K X + D)^{-1} X^T K X \right) D^{-1} y \\ &= Z^T K X \left(I - (X^T K X + D)^{-1} X^T K X \right) D^{-1} y.\end{aligned}$$

Then we can use the fact that $I = (X^T K X + D)^{-1}(X^T K X + D)$ to obtain the identity:

$$\begin{aligned}&I - (X^T K X + D)^{-1} X^T K X \\ &= (X^T K X + D)^{-1}(X^T K X + D) - (X^T K X + D)^{-1} X^T K X \\ &= (X^T K X + D)^{-1}(X^T K X + D - X^T K X) \\ &= (X^T K X + D)^{-1} D.\end{aligned}$$

Substituting this identity into our previous expression for $\hat{f}_z$, we now recover the classical maximum a posteriori solution for Gaussian process regression

$$\hat{f}_z = Z^T K X \left(X^T K X + D \right)^{-1} D D^{-1} y = Z^T K X \left(X^T K X + D \right)^{-1} y,$$

as desired.

Turning to the covariance matrix for the posterior, knowing that the posterior distribution is multivariate Gaussian implies that the posterior covariance matrix is given by the inverse of the Hessian matrix of the log-posterior probability. Thus the covariance matrix of the posterior distribution is given by:

$$\Sigma = \left(\nabla \nabla_f log(f|y) \right)^{-1} = \left(C + X D^{-1} X^T \right)^{-1}, \tag{25}$$

which we note depends on our observations only through the pattern of observed sequence as encoded in X and not on the observed phenotypes y.

Based on the marginalization property of multivariate Gaussian distributions, the posterior covariance at a subset of points z can be obtained simply by taking the submatrix $\Sigma_{zz} = Z^T \Sigma Z$. We can verify that this also matches the classical solution for Gaussian process posterior covariance when $K = C^{-1}$ using Woodbury's identity:

$$\begin{aligned}
\Sigma_{zz} &= Z^T \left(C + X D^{-1} X^T \right)^{-1} Z \\
&= Z^T \left(C^{-1} - C^{-1} \left(X^T C^{-1} X + D \right)^{-1} C^{-1} \right) Z \\
&= Z^T \left(K - K \left(X^T K X + D \right)^{-1} K \right) Z \\
&= Z^T K Z - Z^T K \left(X^T K X + D \right)^{-1} K Z \\
&= K_{zz} - K_{zx} \left(K_{xx} + D \right)^{-1} K_{xz},
\end{aligned} \tag{26}$$

as desired.