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

Carlos Martí-Gómez[1], Juannan Zhou[2], Wei-Chia Chen[3], 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

Multiplex assays of variant effect (MAVEs) allow the functional characterization of an unprecedented number of sequence variants in both gene regulatory regions and protein coding sequences. This has enabled the study of nearly complete combinatorial libraries of mutational variants and revealed the widespread influence of higher-order genetic interactions that arise when multiple mutations are combined. However, the lack of appropriate tools for exploratory analysis of this high-dimensional data limits our overall understanding of the main qualitative properties of complex genotype-phenotype maps. To fill this gap, we have developed *gpmap-tools* (https://github.com/cmarti/gpmap-tools), a *python* library that integrates Gaussian process models for inference, phenotypic imputation, and error estimation from incomplete and noisy MAVE data and collections of natural sequences, together with methods for summarizing patterns of higher-order epistasis and non-linear dimensionality reduction techniques that allow visualization of genotype-phenotype maps containing up to millions of genotypes. Here, we used *gpmap-tools* to study the genotype-phenotype map of the Shine-Dalgarno sequence, a motif that modulates binding of the 16S rRNA to the 5' untranslated region (UTR) of mRNAs through base pair complementarity during translation initiation in prokaryotes. We inferred full combinatorial landscapes containing 262,144 different sequences from the sequences of 5,311 5'UTRs in the *E. coli* genome and from experimental MAVE data. Visualizations of the inferred landscapes were largely consistent with each other, and unveiled a simple molecular mechanism underlying the highly epistatic genotype-phenotype map of the Shine-Dalgarno sequence.

Gaussian process | genotype-phenotype map | fitness landscape
Correspondence: ***mccandlish@cshl.edu***

Introduction

The genotype-phenotype map is a fundamental concept in genetics and evolutionary biology, encapsulating the relationship between sequence variation and its phenotypic outcome. Understanding this relationship is crucial for evolution (1–8), human genetic and infections disease as well as cancer (9–11), and also for synthetic biology and protein engineering (12–14) and plant and animal breeding (15–18). This task is inherently challenging because the effects of mutations often depend on the genetic background in which they occur due to the presence of genetic interactions within and between genes (19–22).

One approach to study empirical genotype-phenotype maps is by constructing sequences of interest and evaluating their biological function experimentally. Historically, our limited ability to engineer large number of genetic variants and measure their phenotypes has constrained our knowledge to phenotypic landscapes containing only a small number of genotypes (23–26). However, Multiplex Assays of Variant Effects (MAVEs) have emerged as a powerful tool, enabling the simultaneous measurement of molecular phenotypes for vast libraries of genetic variants in a single experiment (27, 28). These techniques have allowed quantifying the biological functionality of a large fraction of possible sequences for short regulatory elements (29–36), at specific positions in RNAs (37–41) and proteins (21, 42–52), or for specific combinations of mutations across different genes (53). The highly combinatorial nature of this data makes interpretation challenging, often requiring specialized software powered by sophisticated latent variable models and neural networks to accurately fit the data and make phenotypic predictions (54–58). Another promising family of approaches for modeling complex sequence-function relationships are Gaussian process models (59). Following this line of work, recently developed Gaussian process models naturally incorporate genetic interactions of any order by specifying priors that control the type and magnitude of genetic interactions, and which allow inference of complete genotype-phenotype maps from MAVE data, achieving state-of-the-art predictive power (60, 61).

An alternative approach to study genotype-phenotype maps is through collections of natural sequences. Because natural selection often acts to preserve functionality, we can assume that the probability of observing a sequence in nature depends on how well it performs its function. Thus, the probability distribution from which sequences with a specific function are drawn can be viewed as a genotype-phenotype map in which the phenotype is the probability of observing a sequence. Independent site models, such as Position-Weight Matrices, learn this probability distribution over possible sequences by assuming that positions are independent (62), whereas pairwise interaction models, also known as Potts models (63–66), relax this strong assumption by allowing interactions between pairs of positions. These models have been very successful in predicting structural contacts in proteins (67), mutational effects (68), novel functional proteins (69) and specific regulatory sequences, such as the splice sites (70). A recently proposed bayesian nonparametric model further generalizes these models by defining a prior distribution controlling the magnitude of local

epistatic coefficients and inferring the complete probability distribution over sequence spaces under that prior, opening up the opportunity to study genotype-phenotype maps containing high-order epistatic interactions from readily available collections of natural sequences (71).

Another important challenge is the interpretation of these complex genotype-phenotype maps, particularly when the number of possible genotypes is large. A way to develop an intuitive understanding of complex datasets is through data visualization tools. Various strategies to represent genotype-phenotype maps have been proposed. The genotype-phenotype map has often been represented as a topographic map, where genotypes are points in a 2D space and the height represents the phenotype (1, 72). This conceptual representation is very intuitive and has provided us with language to qualitatively describe genotype-phenotype maps e.g. fitness peaks, valleys or plateaus. However, the space of possible sequences is a discrete space, which is more realistically represented as a Hamming graph, in which nodes represent genotypes and edges represent single-point mutations between them. For genotype-phenotype maps containing few genotypes, it often suffices to embed this graph in a two dimensional space using the distance to a reference sequence and the phenotype as coordinates (1, 37, 41, 48, 73) or specific embeddings of the Hamming graph (74) or subgraph induced by the set of highly functional sequences (44). A different strategy is to construct a low-dimensional representation that reflects the evolutionary dynamics induced by the genotype-phenotype map of interest, for example by having the distances between genotypes represent the expected time to evolve between them for a population evolving under selection for high phenotypic values (75). This is a very useful property for visualizing genotype-phenotype maps, as sets of functional sequences that are inaccessible to each other (peaks) are represented far apart in the visualization, naturally displaying the key genetic interactions separating them (valleys). This technique has been effectively applied to understand the qualitative features of a number of genotype-phenotype maps (60, 61, 71, 76, 77) and the constraints imposed by the structure of the genetic code in protein evolution (78). However, its broader applicability has been limited by the lack of accessible software packages implementing it.

Here, we present *gpmap-tools*, a *python* library that provides an integrated and accessible interface to methods for inference and visualization of large complex genotype-phenotype maps. Among other improved features, *gpmap-tools* incorporates a new computational back-end in which large matrices are represented as linear operators, enabling efficient computation. This allows us to easily sample from the prior distribution to simulate genotype-phenotype maps with different types and amounts of epistasis and to do statistical analysis of specific features of the genotype-phenotype maps in the presence of missing data through computation of the posterior distribution of arbitrary linear combinations of phenotypes e.g. mutational effects and epistatic coefficients across genetic backgrounds. Moreover, using the Laplace approximation, *gpmap-tools*, allows these same calculations to be conducted when using non-Gaussian likelihood functions, such as when inferring genotype-phenotype maps from collections of natural sequences. We also present a new projection operator that enables calculation of the variance explained by interactions between sets of sites in a complete genotype-phenotype map, providing useful statistics to understand the patterns and complexity of genetic interactions across different positions. In addition, *gpmap-tools* provides an extended interface for visualizing genotype-phenotype maps with different number of alleles per site, new functionality to investigate the sequence features that characterize different regions of the representation, tools for interactive visualization, and accelerated rendering of plots containing up to millions of genotypes.

We demonstrate the capabilities of *gpmap-tools* by inferring the fitness landscape of the Shine-Dalgarno (SD) sequence from two different types of data: i) sequence diversity across the 5' untranslated regions (UTRs) in bacterial genomes and ii) MAVE data (31). The inferred landscapes show a common structure consisting of peaks corresponding to the 16S rRNA binding at different distances relative to the start codon, with registers separated by 3 nucleotides forming extended ridges of functional sequences due to the quasi-repetitive nature of the canonical SD motif. Using this knowledge about the qualitative properties of the genotype-phentoype map, we fit a simplified mechanistic model whose parameters have clear biophysical interpretations, allowing us to disentangle the effects of mutations on binding at different registers in vivo, while recapitulating the observed main qualitative features.

Approach

Epistasis in genotype-phenotype maps. 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). The genotype-phenotype map can be represented by an α^ℓ-dimensional vector f containing the phenotype for every genotype. *gpmap-tools* implements two different methods for measuring the amount and pattern of epistasis in a genotype-phenotype map, one based on the typical size of local epistatic interactions and the other based on the fraction of variance explained by different 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 an otherwise identical genetic background C:

$$\epsilon = (f_{aBC} - f_{ABC}) - (f_{abC} - f_{AbC}). \quad \textbf{(1)}$$

The average squared epistatic coefficient $\overline{\epsilon^2}$ across all possible pairs of mutations and across all possible genetic backgrounds C provides a measure of the variability in mutational effects between neighboring genotypes across the whole genotype-phenotype map (60). This measure $\overline{\epsilon^2}$ can be computed efficiently as a positive semi-definite 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 $\alpha^\ell \times \alpha^\ell$ matrix (60). This statistic can also be generalized to $\overline{\epsilon_P{}^2} = \frac{1}{s_P} f^T \Delta^{(P)} f$ for local epistatic coefficients of any order P to characterize the typical size of local P-way epistatic interactions in a mean square sense (71).

Variance components. Any genotype-phenotype map f can be decomposed into the contribution of $\ell+1$ orthogonal subspaces $f = \sum_k f_k$, where f_k represent a function containing epistatic interactions solely of order k. These orthogonal components f_k can be obtained by projecting the function f into the k-th order subspace using the known orthogonal projection matrix P_k with entries given by the Krawtchauk polynomials and can be used to compute the relative contribution of the different orders of interactions to a genotype-phenotype map f (61, 79–82). Moreover, here we show that each k-th order subspace can be further decomposed into $\binom{\ell}{k}$ smaller subspaces defined by the genetic interactions involving k specific sites. For instance, the 3rd order subspace can be broken down into the relative contributions of interactions among every possible combination of 3 sites. If we define U to be a subset of k sites and P_U the projection matrix into the subspace defined by that subset of sites U, we can express the function as the sum of contributions of 2^ℓ components ($f = \sum_U P_U f$), where the P_U are given by:

$$P_U(x,y) = \alpha^{-\ell} \prod_{\substack{p \in U \\ x_p = y_p}} (\alpha - 1) \prod_{\substack{p \in U \\ x_p \neq y_p}} (-1). \qquad \textbf{(2)}$$

These projection matrices allow us to identify, not only the contribution of interactions of different order, but also which sites and subsets of sites are involved in those interactions, providing a finer-grained characterization of the sequence-function relationship. We can aggregate these components in different ways to compute other low dimensional summary statistics, such as the total variance explained by epistatic interactions of a specific order k, or across all orders k involving a site or a subset of sites (83, 84).

Gaussian process inference of genotype-phenotype maps.

Interpretable priors. Gaussian process models are a class of Bayesian models that place a multivariate Gaussian prior distribution over all possible functions and compute the posterior distribution given observed data (85). In our case, the prior distribution is a multivariate Gaussian distribution given by $p(f) = N(0, K)$, typically characterized by its covariance matrix K or precision matrix C, where the covariance matrix K is most often defined through a kernel function $K(x_i, x_j)$ that returns the prior covariance between any pair of sequences x_i, x_j. Our aim is to define prior distributions for f with hyperparameters that have clear biological interpretations in terms of the type and extent of epistasis included in the prior. This not only allows us to better understand inference and prediction under each of these priors but also to set the hyperparameters of these priors in a principled way and to interpret their values when learned from data.

gpmap-tools implements two families of priors, 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 parametrizes the prior distribution through its precision matrix $C = \frac{a}{s}\Delta^{(P)}$ assigning a prior probability to f depending on the average squared epistatic coefficient of order P, i.e. $\log p(f) \propto -\frac{1}{2} f^T C f$ (60, 71). This prior implicitly leaves genetic interactions of order $k < P$ unconstrained, e.g. for $P = 2$ additive effects are not penalized, and hence correspond to the use of an improper Gaussian prior. This family of priors has a single hyperparameter a that is inversely proportional to the expected squared local epistatic coefficient under the prior. As $a \to 0$, we assign the same prior probability to every possible genotype-phenotype map, so that the Maximum a Posteriori (MAP) matches exactly the Maximum likelihood estimate. On the other hand, as $a \to \infty$, we assign zero prior probability to genotype-phenotype maps with non-zero P-epistatic coefficients, which is equivalent to fitting a model with epistatic interactions up to order $P-1$ (71). The second family of priors are the variance component priors, which are parametrized by their covariance matrix $K = \sum_{k=0}^{\ell} \lambda_k K_k$. It has $\ell+1$ hyperparameters λ_k that control the variance explained by genetic interactions of order k (86) and which equivalently control the decay in the predictability of mutational effects and epistatic coefficients in genetic backgrounds separated by an increasing number of mutations (61). 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 (61).

Posterior distributions. Given observations y for some set of sequences x, we update the probability distribution of plausible genotype-phenotype maps to be consistent with these observations by computing the posterior distribution $p(f|y)$. Under a Gaussian likelihood with known sequence-specific variance for the measurement error $p(y|f_x, D_{\sigma^2}) = N(f_x, D_{\sigma^2})$, where D_{σ^2} is a diagonal matrix with the variance σ_x^2 associated to each measurement x along the diagonal, the posterior distribution for the phenotype f is also a multivariate Gaussian, whose mean we write as $\hat{f}$ and whose covariance matrix we write as Σ.

Our approach for calculating $\hat{f}$ and Σ differs depending on whether we are using the prior based on the mean squared local epistatic coefficient or our variance component prior. For the prior based on the mean squared local epistatic coefficient, $\hat{f}$ is given by the solution of:

$$(C + X D_{\sigma^2}^{-1} X^T)\hat{f} = X D_{\sigma^2}^{-1} y \qquad \textbf{(3)}$$

where X is a sparse matrix with $X_{ij} = 1$ if sequence i is the j-th sequence in x and $X_{ij} = 0$ otherwise. Σ is given by:

$$\Sigma = (C + X D_{\sigma^2}^{-1} X^T)^{-1}, \qquad \textbf{(4)}$$

(for derivation, see Supplementary Information). This formulation is useful when the $\alpha^\ell \times \alpha^\ell$ precision or cost matrix

C is sparse, as is the case here, since then Equation 3 can be solved numerically using iterative methods. For the variance component prior, since it is defined in terms of the covariance matrix K, we can use the classical solution for Gaussian processes (85):

$$\hat{f} = KX\left(X^T KX + D_{\sigma^2}\right)^{-1} y \tag{5}$$

$$\Sigma = K - KX\left(X^T KX + D_{\sigma^2}\right)^{-1} X^T K. \tag{6}$$

The above 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* provides a number of methods to infer complete genotype-phenotype maps using either phenotypic measurements of specific genotypes (including the possible user-provided error estimates for each sequence) or from collections of known functional sequences even in the absence of direct phenotypic measurements.

Minimum epistasis interpolation / local epistatic coefficient priors. The minimum epistasis interpolation method was originally proposed (60) in terms of finding the f_z at unobserved sequences z given the known f_x at sequences x by minimizing $\overline{\epsilon^2}$ over the complete genotype-phenotype map and *gpmap-tools* provides a slight generalization to local epistatic coefficients of any order P, (by minimizing $f^T \Delta^{(P)} f$). Through use of the local epistatic coefficient prior, *gpmap-tools* also allows Gaussian process regression under the improper prior with precision matrix C and the presence of Gaussian measurement error given by D_{σ^2}. Using this local epistatic coefficient prior, the value of the hyperparameter a controlling the magnitude of the local P-th order epistatic coefficients is optimized via cross-validation. In addition to the point estimate $\hat{f}$ of the reconstructed genotype-phenotype map, *gpmap-tools* can also provide uncertainty quantification via the posterior covariance given in Equation 4.

Empirical variance component regression. Empirical variance component regression (VC regression), proposed in (61), combines a Variance Component prior parameterized by the variance λ_k associated to interactions of every order k with a Gaussian likelihood with known noise variance D_{σ^2} to compute the exact Gaussian posterior distribution over f using equations 5 and 6. The hyperparameters λ_k controlling the action of the prior are optimized through kernel alignment, this is, by minimizing the squared distance between the covariance under the prior and the empirical distance-covariance function computed from the incomplete data. This can be done very efficiently, because the prior correlation between two sequences depends only on the Hamming distance, reducing it to a lower dimensional constrained weighted least squares problem (61).

Sequence probability distribution estimation. *gpmap-tools* also implements the SeqDEFT method (71) for estimating probability distributions over sequence space. SeqDEFT aims to infer the probability distribution π from which natural sequences are drawn. This problem is similar to the previous models if we consider the $\log \pi_i$ as a phenotype associated to sequence i. Data typically 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 and can be naturally modeled by a multinomial distribution parametrized by the probability π_i of observing every possible sequence i:

$$p(N|\pi) = \text{Multinomial}\,(N, \pi)\,. \tag{7}$$

SeqDEFT parametrizes $\pi_i = \frac{e^{-\phi_i}}{\sum_j e^{-\phi_j}}$ and defines an improper prior distribution over the latent phenotype ϕ that penalizes local epistatic coefficients of order P $\log p(\phi) \propto -\frac{1}{2}\phi^T C \phi$. As there is no analytical solution for the posterior distribution under this non-Gaussian likelihood function

$$\log p(\phi|N, a) \propto -\frac{a}{2s}\phi^T \Delta^{(P)} \phi - \sum_i N_i \phi_i - N_T \sum_i e^{-\phi_i}, \tag{8}$$

we resort to optimization methods to obtain the Maximum a Posteriori (MAP) estimate $\hat{\phi} = \arg\max_\phi \log p(\phi|N, a)$ given a fixed value of the hyperparameter a. We use cross-validation and one-dimensional grid search to characterize how the log-likelihood in held-out data changes as a function of a and select the optimal value a^*(71). In addition to computing the MAP $\hat{\phi}$ under a^*, *gpmap-tools* implements the Laplace approximation (85) to the posterior distribution of ϕ as a multivariate normal distribution with mean $\hat{\phi}$ and covariance matrix given by the inverse of the Hessian evaluated at the MAP $((\nabla\nabla_\phi \log p(\hat{\phi}|N, a))^{-1} = (C + D_{\hat{\phi}})^{-1}$, where $D_{\hat{\phi}}$ is the diagonal matrix with $N_T e^{-\hat{\phi}}$ down its main diagonal).

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 (75) 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 (60, 71, 75).

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 (87, 88):

$$Q(i,j) = \begin{cases} M(i,j)\frac{S(i,j)}{1-e^{S(i,j)}} & \text{if } i \text{ and } j \text{ are neighbors} \\ -\sum_{k\neq i} Q(i,k) & \text{if } i = j \\ 0 & \text{Otherwise,} \end{cases} \quad \textbf{(9)}$$

where $S(i,j)$ is the scaled selection coefficient for mutation from i to j. We can assume that this scaled selection coefficient is proportional to the phenotypic differences between the two genotypes ($S(i,j) = c(f(j) - f(i))$, where the constant c can be interpreted as the scaled selection coefficient ($2N_e s$, for a Haploid Wright-Fisher population) associated to 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. measure 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

$$\pi(i) = \frac{\pi_M(i)e^{cf(i)}}{\sum_j \pi_M(j)e^{cf(j)}}. \quad \textbf{(10)}$$

where $\pi_M(i)$ are the time-reversible neutral stationary frequencies, which are uniform in absence of mutational biases. 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 $f(i) = \log P(i)$ 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(s). 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 the genotype-phenotype map 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:

$$\sum_{k=2} (u_k(i) - u_k(j))^2 \approx H(i,j) + H(j,i). \quad \textbf{(11)}$$

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 $k-1$.

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* (89), for generating highly customized visualizations, but also an equivalent interface to generate interactive 3D visualizations that display the sequence associated to each node of the graph by hovering the mouse over them using *plotly* (90). Moreover, as rendering large numbers of points and lines becomes limiting in large datasets, the *gpmap-tools* plotting library leverages the power of *datashader* (91) 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, studying a genotype-phenotype map for 9 nucleotides, a naive implementation with 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 matrices are sparse e.g. $\Delta^{(P)}$ and Q, allowing efficient storage and computation (60, 71, 75), 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 (61). 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.

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 (92) LinearOperators functionality, we can leverage the mixed Kronecker matrix-vector product property to efficiently compute e.g. $P_U b$ without constructing P_U (see Supplementary Information). 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* to study 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 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 (93). Understanding how the SD sequence modulates protein translation in vivo is key for optimizing protein production (94). Previous studies used existing sequence diversity (95) and MAVE experiments (31, 32) to build models for this genotype-phenotype map. However, these models cannot account for high-order genetic interactions and provided 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 sequence.

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 1A shows site-specific allele frequencies for up to 20bp upstream of the start codon. This shows only a small bias towards increased G content between positions -13 and -5. While this observation suggests the location of the SD sequence relative to the start codon, the limited expressivity of this site-independent model is likely insufficient to capture the genotype-phenotype map underlying the sequence diversity of the SD sequence. Thus, we focused on the genotype-phenotype map of this 9 nucleotide region. Out of the total $4^9 = 262,144$ possible sequences, we observe 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 and that sharing information across neighboring genotypes through SeqDEFT's prior distribution could alleviate this limited amount of data. Figure 1B shows that the model predicts much better the frequencies of held-out sequences than either the site-independent model $a = \infty$ or the empirical frequencies model $a = 0$, providing strong support for the presence of epistatic interactions. 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 (Figure 1C). Sequences that appear more than 2-3 times in the genome are inferred to always be highly functional. 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 shows a $\overline{\epsilon^2} = 0.10$, 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 making one mutation in the genetic background will often substantially change the effects of other mutations.

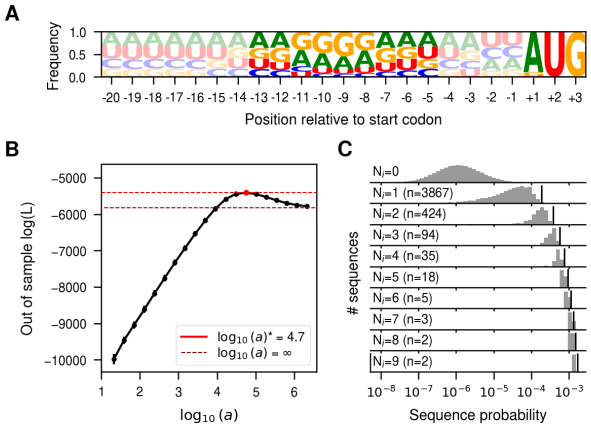

Fig. 1. Inference of the probability distribution of the Shine-Dalgarno sequence. (A) Sequence logo representing the site-specific allele frequencies of 5,311 5'UTRs in the *E. coli* genome aligned with respect to the annotated start codon. The start codon and the 9 nucleotide region 4 bases upstream 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 case maximum entropy model, corresponding to the independent sites model shown in panel A (black) 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.

Inferring the genotype-phenotype map of the Shine–Dalgarno sequence from MAVE. We next use data from a previously published MAVE (31) measuring the expression of a GFP reporter controlled by a library of sequences containing nearly all 262,144 possible 9 nucleotide sequences 4 nucleotides upstream the start codon, as in our previous analysis. We first run MEI to predict the phenotype for all missing genotypes. The imputed genotype-phenotype map had an $\overline{\epsilon^2} = 0.11$. While 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), we can again 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, indicating that there is more epistasis in this dataset than inferred by our SeqDEFT analysis. Overall the relatively large amount of epistasis means that there is substantial variability in the effects of mutations across neighboring genotypes.

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 2A). 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 2B). 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 extremely well ($R^2 = 0.94$, Figure 2C) and made predictions almost as accurate in held-out test sequences ($R^2 = 0.87$, Figure 2D). 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 S2C). 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 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 S2A,B).

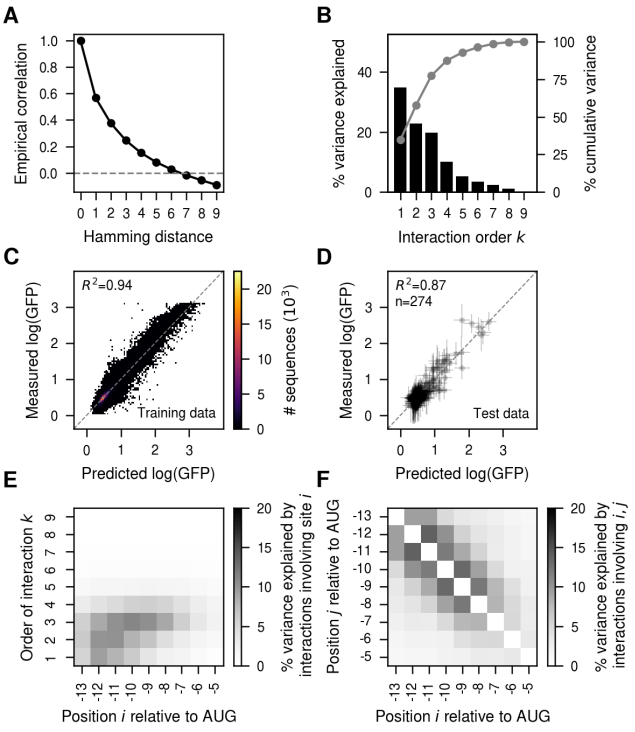

Fig. 2. VC regression analysis of the experimentally measured genotype-phenotype map for the Shine-Dalgarno sequence in the dmsC gene context (31). (A) Empirical distance-correlation function using the measured log(GFP) values in the experimentally evaluated sequences. (B) Percentage of variance explained 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 interactions of orders 2 or greater involving pairs of positions relative to the start codon.

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 order. Figure 2E 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, sites -13 to -10 have both strong additive and epistatic contributions, whereas 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. To investigate the presence of communities of interacting sites, we evaluated the variance explained by epistatic interactions of any order involving each possible pair of sites (Figure 2F). We found that all sites strongly interact with neighboring sites up to a distance of 4 nucleotides, a pattern that is compatible with communities of 4 consecutive sites across the 9 nucleotide sequence.

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 3A 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 S3). This shows that not only the aSD sequence can bind at different distances from the start codon to induce efficient translation initiation, consistent with the interaction neighborhoods shown in Figure 2D, but also that it is hard to evolve a sequence with a shifted SD motif by one or

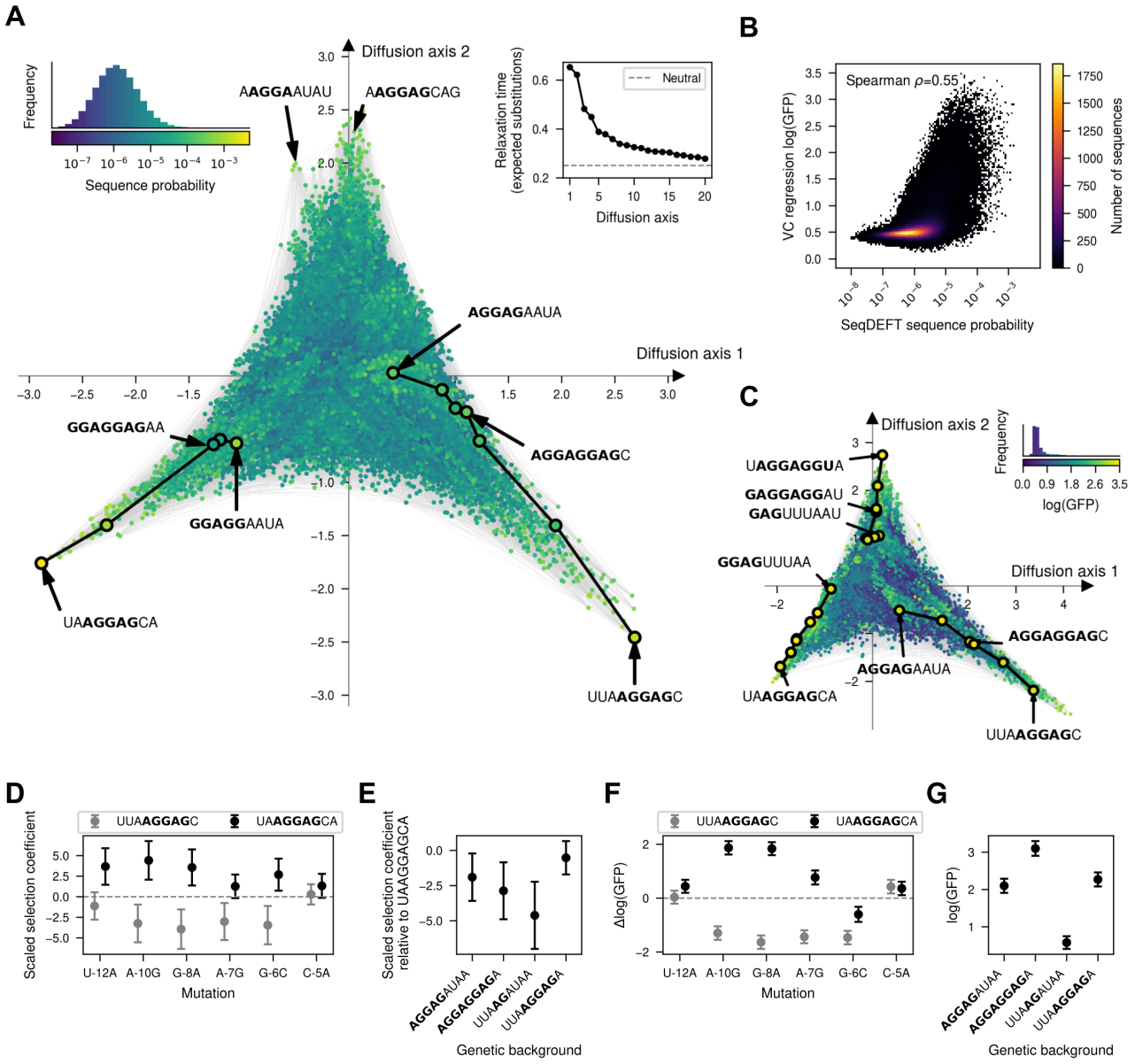

Fig. 3. Visualization of the genotype-phenotype map of the Shine-Dalgarno sequence. (A,C) Low dimensional representation of the *E. coli* Shine-Dalgarno sequence probability distribution inferred with SeqDEFT (A) and the translational efficiencies inferred with VC regression (C). Every dot represents 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 by 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,F) Posterior distribution of the effects of specific mutations when introduced in two genetic contexts, UUAAGGAGC and UAAGGAGCA, which represent a shift in the position of the AGGAG motif. Posterior distributions from SeqDEFT (D) and VC regression models (F). (E,G) Posterior distribution of phenotypes associated to genotypes representing the shift of the AGGAG motif by 3 positions using SeqDEFT (E) or VC regression models (G). Points represent the Maximum a posteriori (MAP) estimates and error bars represent the 95% credible intervals.

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 which a second binding site can evolve through a sequence of point mutations paths without destroying the first. Specifically, within each trinucleotide sequence around the central AGG common to the 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 3A 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 the distant *B. subtilis*. We first found that the AG bias marking the location of the SD sequence in the 5'UTR is located about 2 bp further upstream from the start codon compared to its location in *E. coli* (Figure S4A), as previously reported (95). 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 $r = 0.94$, Figure S4B), but more importantly, the inferred genotype-phenotype map displayed the same type of 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 S4B). Overall, the probability distributions of the SD sequences are quantitatively very similar across distant species and shows the same main qualitative features.

Comparing sequence probability and functional measurements. We next compared the genotype-phenotype maps inferred based on observed 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 3B) 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 modest 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 3C and S5). 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 3C and S6). 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 functional binding site for the aSD sequence. In contrast, the probability distribution of SD sequences was inferred from genomic sequences with different flanking sequences in which sequences 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.

Evaluating the confidence of genetic interactions and phenotypic predictions. Visualizations of the inferred genotype-phenotype maps have enabled the identification of their main qualitative features and the potential genetic interactions underlying them, but they rely on a point estimate of the genotype-phenotype map that does not take into account uncertainty. However, we can complement these analyses by leveraging the uncertainty quantification capabilities of our Gaussian process models as implemented in *gpmap-tools*. For example, we can compute the posterior distribution of the effects of specific mutations in different backgrounds in order to evaluate the strength of evidence in the data supporting different hypotheses suggested by the visualizations of MAP estimates. As an illustration of this strategy, we first validated the incompatibilities separating peaks by computing the posterior distribution for mutational effects in the two backgrounds UUA**AGGAG**C and UA**AGGAG**CU, which contain the same AGGAG motif shifted by one position (Figure 3D). Mutations affecting sites outside the SD motifs in the two registers e.g. U-12A and C-5A, showed small and similar effects in the two genetic backgrounds (Figure 3D). In contrast, the three mutations that allow shifting the SD motif one position upstream (A-10G, G-8A, A-7G) have strong effects with opposite signs in the two genetic contexts (Figure 3D). 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 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 ridge of functional sequences that shifts the SD motif by 3 positions. To do so, we computed the posterior distribution at four specific genotypes that contain the two binding registers, only one, or none. Whereas **AGGAGG**UAA, UAA**AGGAGG** and **AGGAGGAGG** are highly functional sequences, as they allow binding of the aSD sequence at either or both positions, if the first three nucleotides are mutated first as in UUA**AGG**UAA the first

SD motif is destroyed before evolving the second one, resulting in low translational efficiency (Figure 3E). The posteriors for these same genotypes and mutations from the *B. subtilis* genome (Figure S4D,E) and from VC regression analysis on MAVE data (Figure 3F,G) are largely concordant and allow us to conclude that the evidence for this particular high fitness ridge is stronger from the MAVE data than it is from the *E. coli* or *B. subtilis* genomic sequence data.

A biophysical model recapitulates the qualitative properties of empirical SD genotype-phenotype maps. Despite inferring a highly epistatic genotype-phenotype map from the experimental data, the visualization revealed that it can be explained by a rather simple underlying mechanism consisting on the ability of the aSD sequence to bind at different distances from the start codon. We hypothesize that this mechanism alone explains both the existence of isolated peaks and, together with the quasi-repetive 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 at different registers. To tackle these issues, we fit a simple mechanistic model, in which the measured 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 (see Methods).

We fit this biophysical model by maximum likelihood (Figure S7A) to the MAVE dataset and achieved good predictive performance in both training ($R^2 = 0.59$, Figure S7B) and held-out sequences ($R^2 = 0.64$, Figure S7C). Importantly, this model contains only 34 parameters that all have clear biophysical interpretations e.g. in terms of mutational effects on binding energies. The model also includes a parameter α that specifies the background fluorescence in absence of aSD binding to the 5'UTR, which we estimated as $\hat{\alpha} = 0.47$. Likewise, the maximal protein abundance, obtained when the mRNA is saturated with aSD, is estimated to be 3.85. These estimates suggest that the equilibrium occupancy of mRNA's 5'UTR by aSD is low for most sequences, since the dataset has a median protein abundance of 0.511 and a maximum of 3.12. Moreover, position specific binding energies ΔG_p are lower (more stable) further from the start codon (Figure 4A). This can be explained by either SD:aSD complex being stabilized by other trans elements binding further from the start codon, or by the aSD hindering the binding of Met-tRNA:AUG when located too close to each other. Moreover, we were able to deconvolve the effects of mutations on binding at different registers and inferred allele and position specific energetic contributions to binding (Figure 4B). 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 ($\Delta\Delta G = 0.79$ kcal/mol), whereas A4C is highly destabilizing ($\Delta\Delta G = 1.77$ 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 4C). 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 4D). Overall, the visualization allowed us to develop a simplified biophysically interpretable model and to verify that this model recapitulates the main qualitative features of the genotype-phenotype map of the Shine-Dalgarno sequence.

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 comprising all possible sequences from either experimental MAVE data or sequence counts. Taking into consideration the noise in the data (due 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 maximum a posteriori (MAP) estimate, this is, the most probable genotype-phenotype map given the observed data. However, in contrast to other expressive models able to capture complex genetic interactions e.g. neural networks (57, 96), our inference methods provide a rigorous quantification of the uncertainty about the phenotypes of specific sequences, mutational effects across genetic backgrounds and, more generally, any linear combination of the phenotypes of a number of genotypes. 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 information provided by the data.

gpmap-tools re-implements a powerful method for visualizing fitness landscapes (75) that allows exploratory data analysis, interpretation and comparison of these complex datasets. Thus, rather than interpreting the results through an explicit parametric model allowing high-order genetic interactions or descriptive statistics like the number of peaks or adaptive walks (26, 35, 97, 98), this method leverages the evolutionary dynamics on the genotype-phenotype map to highlight its main, potentially unexpected, qualitative features. Moreover, we can use it to generate hypotheses for how mutational effects change with genetic backgrounds,

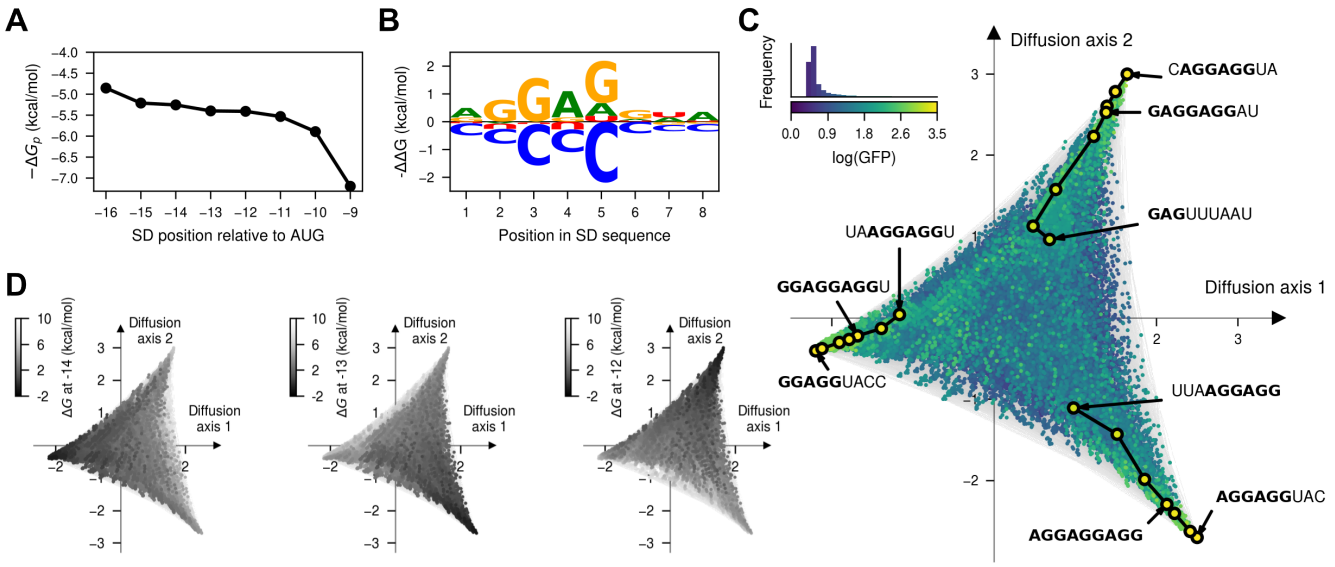

Fig. 4. Thermodynamic model of sequence-dependent translational efficiency. (A) Estimated average free energy of binding at position p relative to the start codon across all possible sequences. (B) Sequence logo representing the site-specific but register-independent allelic contributions to the binding 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 the predicted log(GFP). (D) Visualization of the genotype-phenotype map under the inferred thermodynamic model representing the binding energies at positions -14, -13 and -12 relative to the start codon showing that the peaks in the visualization correspond to the strongest binding at different positions. Binding energies 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.

which can then be evaluated through the posterior distribution (Figure 3 and S4). Identifying the main features of the genotype-phenotype map can be crucial for defining an appropriate mechanistic or biophysical model. For instance, visualization of the Shine-Dalgarno genotype-phenotype map allowed us to define a thermodynamic model in which the binding energy depended only additively on the sequence at each register, while recapitulating the peaks observed in the data. Additionally, this technique enabled a detailed comparison of genotype-phenotype maps inferred with different methods and data sources and the extent to which they had the same structure, in contrast to broadly used metrics, like Pearson or Spearman correlation coefficients. In this work, we used it to show genotype-phenotype maps with essentially the same structure inferred from data from distant species like *E. coli* and *B. subtilis* and completely independent data sources (experimental MAVE data and observations of natural sequences). Finding consistent structures across different data types and sources is particularly relevant for understanding the role of the fitness landscape on evolution of these regulatory sequences because we do not have access to the true fitness values. More generally, the visualization technique opens up the opportunity to compare genotype-phenotype maps of different genetic elements e.g. regulatory sequences, protein-protein interactions, and enzymes, by for example finding shared fitness landscape structures that induce similar evolutionary dynamics despite the differing biological substrates.

gpmap-tools enables inference and interpretation of complex genotype-phenotype maps comprising millions of sequences by making a number of assumptions that introduce some limitations. First, MEM and VC regression are phenomenological models that do not explicitly account for non-specific epistasis e.g. biophysical models. While these models can still make highly accurate phenotypic predictions in the presence of global epistasis through pervasive specific interactions, this limits our ability to distinguish specific from non-specific genetic interactions (54–56, 99). 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 (71), it remains unclear how robust it is to the known challenge of using phylogenetically related sequences from widespread multiple sequence alignments of protein families (100–102). Third, both inference and visualization methods still require storing all possible sequences and their phenotypes in memory, whose number grow exponentially with sequence length, thus limiting the applicability of *gpmap-tools* to spaces of sequences of a constant and relatively short length. Despite these limitations, *gpmap-tools* provides a unique set of tools for studying the structure of short sequence genotype-phenotype maps at an unprecedented scale, which is a necessary stepping stone towards understanding genotype-phenotype maps at the gene, protein or 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 the assembly version

ASM160652v1, and *B. subtilis* assembly ASM904v1 from GeneBank. We extrated the 5'UTR sequence for every annotated gene using *pysam* (103, 104) 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* (105) and estimating the complex probability distribution using *gpmap-tools* implementation of SeqDEFT (71). 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 for 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 (31). 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 $\hat{\sigma}_i^2$ for each 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 S5). An average log(GFP) of 2 at stationarity was selected and used for all subsequent visualizations, similar to our best estimate 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, whose concentration is assumed to be independent 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 in any configuration or register p at thermodynamic equilibrium. This quantity depends on the binding energy ΔG of the 16S rRNA to the mRNA to the sequence x_p located at 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}. Moreover, we assume that there is a minimal value α that is independent of the variable 5'UTR sequence in the experiment, which can represent a background translation level that depends on a different mechanism of initiation, or background signal in the assay e.g. cells auto-fluorescence in the GFP channel. Thus, the overall protein abundance $f(x)$ for a sequence x depends on α and the fraction of bound mRNA multiplied by the translation rate when bound β, which also encodes for the maximal protein output through this initiation mechanism:

$$f(x) = \alpha + \beta \left(\frac{\sum_p e^{-\frac{\Delta G_p(x_p)}{RT}}}{1 + \sum_p e^{-\frac{\Delta G_p(x_p)}{RT}}} \right). \quad \textbf{(12)}$$

We next assumed that the binding energy at position p depends only on the 8 nucleotide subsequence at that position, such that each position has an intrinsic preference of binding ΔG_p^0 and a site independent contribution of the allele c at each position i of the sequence $x_p(i,c)$. Importantly, we extended the variable 9 nucleotide sequences with the fixed upstream and downstream sequences CCG and UGAG from the dmsC genetic context to incorporate the effect of mutations in binding registers spanning both fixed and variable regions of the sequence:

$$\Delta G(x_p) = \Delta G_p^0 + \sum_i \sum_c x_p(i,c) \Delta\Delta G_{ic}. \quad \textbf{(13)}$$

Finally, we assume that the measurement y for sequence x is observed with known noise variance σ^2 and an extra or uncharacterized variance τ^2 under a Gaussian likelihood function given by $p(y|x) = N(f(x), \sigma^2 + \tau^2)$. We used PyTorch to encode the model and the Adam optimizer with a learning rate of 0.01 for 2500 iterations, while monitoring for convergence (Figure S7A), to find the maximum likelihood estimates of the model parameters.

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 can is available at https://github.com/cmarti/shine_dalgarno.

Funding

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

1. Sewall Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the Sixth International Congress of Genetics*, pages 356–366, 1932.
2. J. Arjan G.M. De Visser and Joachim Krug. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, 15(7):480–490, June 2014. ISSN 14710064. doi: 10.1038/nrg3744. Publisher: Nature Publishing Group.
3. Alexey S. Kondrashov, Shamil Sunyaev, and Fyodor A. Kondrashov. Dobzhansky-Muller incompatibilities in protein evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 99(23):14878–14883, November 2002. ISSN 0027-8424. doi: 10.1073/pnas.232565499.
4. Daniel M. Weinreich, Yinghong Lan, C. Scott Wylie, and Robert B. Heckendorn. Should evolutionary geneticists worry about higher-order epistasis? *Current Opinion in Genetics & Development*, 23(6):700–707, December 2013. ISSN 1879-0380. doi: 10.1016/j.gde.2013.10.007.
5. Zachary R. Sailer and Michael J. Harms. High-order epistasis shapes evolutionary trajectories. *PLoS computational biology*, 13(5):e1005541, May 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005541.
6. Claudia Bank. Epistasis and Adaptation on Fitness Landscapes. *Annual Review of Ecology, Evolution, and Systematics*, 53(Volume 53, 2022):457–479, November 2022. ISSN 1543-592X, 1545-2069. doi: 10.1146/annurev-ecolsys-102320-112153. Publisher: Annual Reviews.
7. Patrick C. Phillips. Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews. Genetics*, 9(11):855–867, November 2008. ISSN 1471-0064. doi: 10.1038/nrg2452.
8. Milo S. Johnson, Gautam Reddy, and Michael M. Desai. Epistasis and evolution: recent advances and an outlook for prediction. *BMC Biology*, 21(1):120, May 2023. ISSN 1741-7007. doi: 10.1186/s12915-023-01585-3.
9. Alief Moulana, Thomas Dupic, Angela M. Phillips, and Michael M. Desai. Genotype–phenotype landscapes for immune–pathogen coevolution. *Trends in Immunology*, 44(5):384–396, May 2023. ISSN 1471-4906, 1471-4981. doi: 10.1016/j.it.2023.03.006. Publisher: Elsevier.
10. Jason H. Moore and Scott M. Williams. Epistasis and Its Implications for Personal Genetics. *The American Journal of Human Genetics*, 85(3):309–320, September 2009. ISSN 0002-9297, 1537-6605. doi: 10.1016/j.ajhg.2009.08.006. Publisher: Elsevier.
11. Krishna Dasari, Jason A. Somarelli, Sudhir Kumar, and Jeffrey P. Townsend. The somatic molecular evolution of cancer: Mutation, selection, and epistasis. *Progress in Biophysics and Molecular Biology*, 165:56–65, October 2021. ISSN 0079-6107. doi: 10.1016/j.pbiomolbio.2021.08.003.
12. Chase R Freschlin, Sarah A Fahlberg, and Philip A Romero. Machine learning to navigate fitness landscapes for protein engineering. *Current Opinion in Biotechnology*, 75:102713, June 2022. ISSN 0958-1669. doi: 10.1016/j.copbio.2022.102713.
13. Kevin K. Yang, Zachary Wu, and Frances H. Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature Methods*, 16(8):687–694, August 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0496-6.
14. Rosalie Lipsh-Sokolik and Sarel J. Fleishman. Addressing epistasis in the design of protein function. *Proceedings of the National Academy of Sciences*, 121(34):e2314999121, August 2024. doi: 10.1073/pnas.2314999121. Publisher: Proceedings of the National Academy of Sciences.
15. Timothy B. Sackton and Daniel L. Hartl. Genotypic Context and Epistasis in Individuals and Populations. *Cell*, 166(2):279–287, July 2016. ISSN 1097-4172. doi: 10.1016/j.cell.2016.06.047.
16. Gustavo De Los Campos, John M Hickey, Ricardo Pong-Wong, Hans D Daetwyler, and Mario P L Calus. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics*, 193(2):327–345, February 2013. ISSN 1943-2631. doi: 10.1534/genetics.112.143313.
17. Sebastian Soyk, Matthias Benoit, and Zachary B. Lippman. New Horizons for Dissecting Epistasis in Crop Quantitative Trait Variation. *Annual Review of Genetics*, 54 (Volume 54, 2020):287–307, November 2020. ISSN 0066-4197, 1545-2948. doi: 10.1146/annurev-genet-050720-122916. Publisher: Annual Reviews.
18. Sangam L. Dwivedi, Pat Heslop-Harrison, Junrey Amas, Rodomiro Ortiz, and David Edwards. Epistasis and pleiotropy-induced variation for plant breeding. *Plant Biotechnology Journal*, 22(10):2788–2807, 2024. ISSN 1467-7652. doi: 10.1111/pbi.14405. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/pbi.14405.
19. Tyler N. Starr and Joseph W. Thornton. Epistasis in protein evolution. *Protein Science*, 25(7):1204–1218, 2016. ISSN 1469-896X. doi: 10.1002/pro.2897. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.2897.
20. Júlia Domingo, Pablo Baeza-Centurion, and Ben Lehner. The causes and consequences of genetic interactions (epistasis). *Annual Review of Genomics and Human Genetics*, 20(1):433–460, August 2019. ISSN 1527-8204. doi: 10.1146/annurev-genom-083118-014857. Publisher: Annual Reviews Inc.
21. Claudia Bank, Sebastian Matuszewski, Ryan T. Hietpas, and Jeffrey D. Jensen. On the (un)predictability of a large intragenic fitness landscape. *Proceedings of the National Academy of Sciences*, 113(49):14085–14090, December 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1612676113.
22. Charlotte M. Miton, Karol Buda, and Nobuhiko Tokuriki. Epistasis and intramolecular networks in protein evolution. *Current Opinion in Structural Biology*, 69:160–168, August 2021. ISSN 1879-033X. doi: 10.1016/j.sbi.2021.04.007.
23. Aisha I. Khan, Duy M. Dinh, Dominique Schneider, Richard E. Lenski, and Tim F. Cooper. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science*, 332(6034):1193–1196, 2011. ISSN 00368075. doi: 10.1126/science.1203801.
24. Kenneth M Flynn, Tim F. Cooper, Francisco B.G. Moore, and Vaughn S Cooper. The Environment Affects Epistatic Interactions to Alter the Topology of an Empirical Fitness Landscape. *PLoS Genetics*, 9(4):1003426, 2013. ISSN 15537390. doi: 10.1371/journal.pgen.1003426.
25. Daniel M. Weinreich, Yinghong Lan, Jacob Jaffe, and Robert B. Heckendorn. The Influence of Higher-Order Epistasis on Biological Fitness Landscape Topography. *Journal of Statistical Physics*, 172(1):208–225, 2018. ISSN 00224715. doi: 10.1007/s10955-018-1975-3. Publisher: Springer US.
26. Ivan G Szendro, Martijn F Schenk, Jasper Franke, Joachim Krug, and J Arjan G M de Visser. Quantitative analyses of empirical fitness landscapes. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(01):P01005, January 2013. ISSN 1742-5468. doi: 10.1088/1742-5468/2013/01/P01005. Publisher: IOP Publishing and SISSA.
27. Douglas M. Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature Methods*, 11(8):801–807, August 2014. ISSN 1548-7105. doi: 10.1038/nmeth.3027. Publisher: Nature Publishing Group.
28. Justin B. Kinney and David M. McCandlish. Massively Parallel Assays and Quantitative Sequence–Function Relationships. *Annual Review of Genomics and Human Genetics*, 20(1):annurev–genom–083118–014845, 2019. ISSN 1527-8204. doi: 10.1146/annurev-genom-083118-014845.
29. Justin B. Kinney, Anand Murugan, Curtis G. Callan, and Edward C. Cox. 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, 2010. ISSN 00278424. doi: 10.1073/pnas.1004290107.
30. Mandy S. Wong, Justin B. Kinney, and Adrian R. Krainer. Quantitative Activity Profile and Context Dependence of All Human 5 Splice Sites. *Molecular Cell*, 71(6):1012–1026.e3, 2018. ISSN 10974164. doi: 10.1016/j.molcel.2018.07.033. Publisher: Elsevier Inc.
31. Syue Ting Kuo, Ruey Lin Jahn, Yuan Ju Cheng, Yi Lan Chen, Yun Ju Lee, Florian Hollfelder, Jin Der Wen, and Hsin Hung David Chou. Global fitness landscapes of the Shine-Dalgarno sequence. *Genome Research*, 30(5):711–723, 2020. ISSN 15495469. doi: 10.1101/gr.260182.119.
32. Mads T. Bonde, Margit Pedersen, Michael S. Klausen, Sheila I. Jensen, Tune Wulff, Scott Harrison, Alex T. Nielsen, Markus J. Herrgård, and Morten O.A. Sommer. Predictable tuning of protein expression in bacteria. *Nature Methods*, 13(3):233–236, 2016. ISSN 15487105. doi: 10.1038/nmeth.3727.
33. William L Noderer, Ross J Flockhart, Aparna Bhaduri, Alexander J Diaz De Arce, Jiajing Zhang, Paul A Khavari, and Clifford L Wang. Quantitative analysis of mammalian translation initiation sites by facs-seq. *Molecular Systems Biology*, 10(8):748, August 2014. ISSN 1744-4292, 1744-4292. doi: 10.15252/msb.20145136.
34. Ekaterina S. Komarova, Zoya S. Chervontseva, Ilya A. Osterman, Sergey A. Evfratov, Maria P. Rubtsova, Timofei S. Zatsepin, Tatiana A. Semashko, Elena S. Kostryukova, Alexey A. Bogdanov, Mikhail S. Gelfand, Olga A. Dontsova, and Petr V. Sergiev. 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, 2020. ISSN 17517915. doi: 10.1111/1751-7915.13561.
35. Cauã Antunes Westmann, Leander Goldbach, and Andreas Wagner. The highly rugged yet navigable regulatory landscape of the bacterial transcription factor TetR. *Nature Communications*, 15(1):10745, December 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-54723-y. Publisher: Nature Publishing Group.
36. Caua Antunes Westmann, Leander Goldbach, and Andreas Wagner. Entangled adaptive landscapes facilitate the evolution of gene regulation by exaptation, November 2024. Pages: 2024.11.10.620926 Section: New Results.
37. Pablo Baeza-Centurion, Belén Miñana, Jörn M. Schmiedel, Juan Valcárcel, and Ben Lehner. Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing. *Cell*, 176(3):549–563.e23, January 2019. ISSN 10974172. doi: 10.1016/j.cell.2018.12.010. Publisher: Cell Press.
38. Júlia Domingo, Guillaume Diss, and Ben Lehner. Pairwise and higher-order genetic interactions during the evolution of a tRNA. *Nature*, 558(7708):117–121, 2018. ISSN 14764687. doi: 10.1038/s41586-018-0170-7.
39. Rachapun Rotrattanadumrong and Yohei Yokobayashi. Experimental exploration of a ribozyme neutral network using evolutionary algorithm and deep learning. *Nature Communications*, pages 1–14, 2022. doi: 10.1038/s41467-022-32538-z. Publisher: Springer US ISBN: 4146702232.
40. Valerie W.C. Soo, Jacob B. Swadling, Andre J. Faure, and Tobias Warnecke. Fitness landscape of a dynamic RNA structure. *PLoS Genetics*, 17(2):e1009353, 2021. ISSN 15537404. doi: 10.1371/JOURNAL.PGEN.1009353. ISBN: 1111111111.
41. Devin P. Bendixsen, James Collet, Bjørn Østman, and Eric J. Hayden. Genotype network intersections promote evolutionary innovation. *PLoS Biology*, 17(5), May 2019. ISSN 15457885. doi: 10.1371/JOURNAL.PBIO.3000300. Publisher: Public Library of Science.
42. Thúy Lan V. Lite, Robert A. Grant, Isabel Nocedal, Megan L. Littlehale, Monica S. Guo, and Michael T. Laub. Uncovering the basis of protein-protein interaction specificity with a combinatorially complete library. *eLife*, 9(ii):1–57, 2020. ISSN 2050084X. doi: 10.7554/eLife.60924.
43. Nicholas C. Wu, Lei Dai, C. Anders Olson, James O. Lloyd-Smith, and Ren Sun. Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife*, 5(July):1–21, 2016. ISSN 2050084X. doi: 10.7554/eLife.16965.
44. Tyler N. Starr, Lora K. Picton, and Joseph W. Thornton. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature*, 549(7672):409–413, September 2017. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature23902.
45. Alief Moulana, Thomas Dupic, Angela M Phillips, Jeffrey Chang, Anne A Roffler, Allison J Greaney, Tyler N Starr, Jesse D Bloom, and Michael M Desai. The landscape of antibody binding affinity in SARS-CoV-2 Omicron BA.1 evolution. *eLife*, 12:e83442, February 2023. ISSN 2050-084X. doi: 10.7554/eLife.83442.
46. Vikram Sundar, Boqiang Tu, Lindsey Guan, and Kevin Esvelt. A NEW ULTRA-HIGH-

THROUGHPUT ASSAY FOR MEASURING PROTEIN FITNESS. 2024.
47. Taraneh Zarin and Ben Lehner. A complete map of specificity encoding for a partially fuzzy protein interaction, April 2024.
48. Albert Escobedo, Gesa Voigt, Andre J Faure, and Ben Lehner. Genetics, energetics and allostery during a billion years of hydrophobic protein core evolution, May 2024.
49. Adam S.B. Jalal, Ngat T. Tran, Clare E. Stevenson, Elliot W. Chan, Rebecca Lo, Xiao Tan, Agnes Noy, David M. Lawson, and Tung B.K. Le. Diversification of DNA-Binding Specificity by Permissive and Specificity-Switching Mutations in the ParB/Noc Protein Family. *Cell Reports*, 32(3):107928, July 2020. ISSN 22111247. doi: 10.1016/j.celrep.2020.107928.
50. Frank J. Poelwijk, Michael Socolich, and Rama Ranganathan. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nature Communications*, 10(1): 1–11, 2019. ISSN 20411723. doi: 10.1038/s41467-019-12130-8. Publisher: Springer US.
51. Paul E. O'Maille, Arthur Malone, Nikki Dellas, B. Andes Hess, Lidia Smentek, Iseult Sheehan, Bryan T. Greenhagen, Joe Chappell, Gerard Manning, and Joseph P. Noel. Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. *Nature Chemical Biology*, 4(10):617–623, October 2008. ISSN 1552-4469. doi: 10.1038/nchembio.113.
52. Kadina E. Johnston, Patrick J. Almhjell, Ella J. Watkins-Dulaney, Grace Liu, Nicholas J. Porter, Jason Yang, and Frances H. Arnold. 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, August 2024. ISSN 1091-6490. doi: 10.1073/pnas.2400439121.
53. Christopher W. Bakerlee, Alex N. Nguyen Ba, Yekaterina Shulgina, Jose I. Rojas Echenique, and Michael M. Desai. Idiosyncratic epistasis leads to global fitness-correlated trends. *Science*, 376(6593):630–635, May 2022. ISSN 1095-9203. doi: 10.1126/science.abm4774.
54. Ammar Tareen, William T. Ireland, Anna Posfai, David M. McCandlish, and Justin B. Kinney. MAVE-NN: Quantitative modeling of genotype-phenotype maps as information bottlenecks. *bioRxiv*, pages 1–15, July 2020. ISSN 26928205. doi: 10.1101/2020.07.14.201475. Publisher: Cold Spring Harbor Laboratory.
55. Andre J. Faure and Ben Lehner. 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, December 2024. ISSN 1474-760X. doi: 10.1186/s13059-024-03444-y.
56. Peter D. Tonner, Abe Pressman, and David Ross. Interpretable modeling of genotype–phenotype landscapes with state-of-the-art predictive power. *Proceedings of the National Academy of Sciences*, 119(26):e2114021119, June 2022. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2114021119.
57. Sam Gelman, Sarah A. Fahlberg, Pete Heinzelman, Philip A. Romero, and Anthony Gitter. Neural networks to learn protein sequence–function relationships from deep mutational scanning data. *Proceedings of the National Academy of Sciences*, 118(48):e2104878118, November 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2104878118.
58. Jesse D Bloom. Software for the analysis and visualization of deep mutational scanning data. *BMC Bioinformatics*, 16(1):168, December 2015. ISSN 1471-2105. doi: 10.1186/s12859-015-0590-4.
59. Philip A. Romero, Andreas Krause, and Frances H. Arnold. Navigating the protein fitness landscape with Gaussian processes. *Proceedings of the National Academy of Sciences*, 110(3), January 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1215251110.
60. Juannan Zhou and David M. McCandlish. Minimum epistasis interpolation for sequence-function relationships. *Nature Communications*, 11(1), 2020. ISSN 20411723. doi: 10.1038/s41467-020-15512-5.
61. Juannan Zhou, Mandy S Wong, Wei-chia Chen, Adrian R Krainer, B Justin, and David M Mccandlish. Higher-order epistasis and phenotypic prediction. *Proc. Natl. Acad. Sci. USA*, 119(39), 2022. doi: https://doi.org/10.1073/pnas.2204233119.
62. Gary D. Stormo. Modeling the specificity of protein-DNA interactions. *Quantitative Biology*, 1(2):115–130, June 2013. ISSN 2095-4689, 2095-4697. doi: 10.1007/s40484-013-0012-4.
63. Llan Sly. Reconstruction for the potts model. *Annals of Probability*, 39(4):1365–1406, 2011. ISSN 00911798. doi: 10.1214/10-AOP584.
64. Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 87(1):1–16, 2013. ISSN 15393755. doi: 10.1103/PhysRevE.87.012707. arXiv: 1211.1281.
65. Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S. Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, December 2011. doi: 10.1073/pnas.1111471108. Publisher: Proceedings of the National Academy of Sciences.
66. Allan Haldane and Ronald M. Levy. Mi3-GPU: MCMC-based inverse Ising inference on GPUs for protein covariation analysis. *Computer Physics Communications*, 260:107312, March 2021. ISSN 0010-4655. doi: 10.1016/j.cpc.2020.107312.
67. Debora S Marks, Thomas A Hopf, and Chris Sander. Protein structure prediction from sequence variation. *Nature Biotechnology*, 30(11):1072–1080, November 2012. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt.2419.
68. Thomas A Hopf, John B Ingraham, Frank J Poelwijk, Charlotta P I Schärfe, Michael Springer, Chris Sander, and Debora S Marks. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2):128–135, February 2017. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt.3769.
69. William P. Russ, Matteo Figliuzzi, Christian Stocker, Pierre Barrat-Charlaix, Michael Socolich, Peter Kast, Donald Hilvert, Remi Monasson, Simona Cocco, Martin Weigt, and Rama Ranganathan. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369(6502):440–445, July 2020. doi: 10.1126/science.aba3304. Publisher: American Association for the Advancement of Science.
70. Gene Yeo and Christopher B Burge. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. In *Journal of Computational Biology*, volume 11, pages 377–394, March 2004. doi: 10.1089/1066527041410418. Issue: 2-3 ISSN: 10665277.
71. Wei-chia Chen, Juannan Zhou, Jason M Sheltzer, Justin B Kinney, and David M Mccandlish. Field-theoretic density estimation for biological sequence space with applications to 5 splice site diversity and aneuploidy in cancer. 2021. doi: 10.1073/pnas.2025782118/-/DCSupplemental.y.
72. J. Arjan G.M. de Visser, Santiago F. Elena, Inês Fragata, and Sebastian Matuszewski. The utility of fitness landscapes and big data for predicting evolution. *Heredity*, 121(5): 401–405, 2018. ISSN 13652540. doi: 10.1038/s41437-018-0128-4. Publisher: Springer US.
73. S. Brouillet, H. Annoni, L. Ferretti, and G. Achaz. MAGELLAN: a tool to explore small fitness landscapes, November 2015. Pages: 031583 Section: New Results.
74. Inês Fragata, Alexandre Blanckaert, Marco António Dias Louro, David A. Liberles, and Claudia Bank. Evolution in the light of fitness landscape theory. *Trends in Ecology and Evolution*, 34(1):69–82, 2019. ISSN 01695347. doi: 10.1016/j.tree.2018.10.009. Publisher: Elsevier Ltd.
75. David M. Mccandlish. Visualizing fitness landscapes. *Evolution*, 65(6):1544–1558, June 2011. ISSN 00143820. doi: 10.1111/j.1558-5646.2011.01236.x. Publisher: John Wiley & Sons, Ltd.
76. Jonathan Yaacov Weinstein, Carlos Martí-Gómez, Rosalie Lipsh-Sokolik, Shlomo Yakir Hoch, Demian Liebermann, Reinat Nevo, Haim Weissman, Ekaterina Petrovich-Kopitman, David Margulies, Dmitry Ivankov, David M. McCandlish, and Sarel J. Fleishman. Designed active-site library reveals thousands of functional GFP variants. *Nature Communications*, 14(1):2890, May 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-38099-z.
77. Ziv Avizemer, Carlos Martí-Gómez, Shlomo Yakir Hoch, David M. McCandlish, and Sarel J. Fleishman. Evolutionary paths that link orthogonal pairs of binding proteins, April 2023.
78. Hana Rozhoňová, Carlos Martí-Gómez, David M. McCandlish, and Joshua L. Payne. Robust genetic codes enhance protein evolvability. *PLOS Biology*, 22(5):e3002594, May 2024. ISSN 1545-7885. doi: 10.1371/journal.pbio.3002594.
79. Peter F. Stadler. Landscapes and their correlation functions. *Journal of Mathematical Chemistry*, 20(1):1–45, March 1996. ISSN 1572-8897. doi: 10.1007/BF01165154.
80. Robert Happel and Peter F. Stadler. Canonical approximation of fitness landscapes. *Complexity*, 2(1):53–58, 1996. ISSN 10990526. doi: 10.1002/(SICI)1099-0526(199609/10)2:1<53::AID-CPLX11>3.0.CO;2-W.
81. Peter F. Stadler and Robert Happel. Random field models for fitness landscapes. *Journal of Mathematical Biology*, 38(5):435–478, 1999. ISSN 03036812. doi: 10.1007/s002850050156.
82. Peter F. Stadler. Fitness landscapes. In Michael Lässig and Angelo Valleriani, editors, *Biological Evolution and Statistical Physics*, pages 183–204. Springer, Berlin, Heidelberg, 2002. ISBN 978-3-540-45692-6. doi: 10.1007/3-540-45692-9_10.
83. Lorin Crawford, Ping Zeng, Sayan Mukherjee, and Xiang Zhou. Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLOS Genetics*, 13 (7):e1006869, July 2017. ISSN 1553-7404. doi: 10.1371/journal.pgen.1006869. Publisher: Public Library of Science.
84. Gautam Reddy and Michael M. Desai. Global epistasis emerges from a generic model of a complex trait. *eLife*, 10:1–36, 2021. ISSN 2050084X. doi: 10.7554/eLife.64740.
85. Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass., 3. print edition, 2008. ISBN 978-0-262-18253-9.
86. Johannes Neidhart, Ivan G. Szendro, and Joachim Krug. Exact results for amplitude spectra of fitness landscapes. *Journal of Theoretical Biology*, 332:218–227, 2013. ISSN 00225193. doi: 10.1016/j.jtbi.2013.05.002. arXiv: 1301.1923 Publisher: Elsevier.
87. David M. McCandlish and Arlin Stoltzfus. Modeling Evolution Using the Probability of Fixation: History and Implications. *The Quarterly Review of Biology*, 89(3):225–252, 2014.
88. M. Bulmer. The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129(3):897–907, November 1991. ISSN 0016-6731. doi: 10.1093/genetics/129.3.897.
89. John D. Hunter. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95, May 2007. ISSN 1558-366X. doi: 10.1109/MCSE.2007.55. Conference Name: Computing in Science & Engineering.
90. Shammamah Hossain. Visualization of Bioinformatics Data with Dash Bio. *scipy*, June 2019. doi: 10.25080/Majora-7ddc1dd1-012.
91. James A. Bednar, Joseph Crail, Jim Crist-Harif, Philipp Rudiger, Greg Brener, Chris B, Ian Thomas, Jon Mease, Julia Signell, Maxime Liquet, Jean-Luc Stevens, Brendan Collins, Ajay Thorve, Sarah Bird, thuydotm, esc, kbowen, Nezar Abdennur, Oleg Smirnov, Simon Høxbro Hansen, maihde, Adam Hawley, Andrii Oriekhov, Aron Ahmadia, Barry A. Bragg Jr, Carlos H. Brandt, Clemens Tolboom, Enno G, Erik Welch, and James Bourbeau. holoviz/datashader: Version 0.14.3, November 2022.
92. Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.
93. J Shine and L Dalgarno. Determinant of cistron specificity in bacterial ribosomes. *Nature*, 254:34–38, 1975.
94. Pierre-Aurélien Gilliot and Thomas E Gorochowski. Transfer learning for cross-context prediction of protein expression from 5'UTR sequence. *Nucleic Acids Research*, 52(13): e58, July 2024. ISSN 0305-1048. doi: 10.1093/nar/gkae491.
95. Adam J. Hockenberry, Aaron J. Stern, Luıs A.N. Amaral, and Michael C. Jewett. Diversity of translation initiation mechanisms across bacterial species is driven by environmental conditions and growth demands. *Molecular Biology and Evolution*, 35(3):582–592, 2018. ISSN 15371719. doi: 10.1093/molbev/msx310.
96. Drew H. Bryant, Ali Bashir, Sam Sinai, Nina K. Jain, Pierce J. Ogden, Patrick F. Riley, George M. Church, Lucy J. Colwell, and Eric D. Kelsic. Deep diversification of an AAV

capsid protein by machine learning. *Nature Biotechnology*, 2021. ISSN 1087-0156. doi: 10.1038/s41587-020-00793-4. Publisher: Springer US.

97. Luca Ferretti, Daniel Weinreich, Fumio Tajima, and Guillaume Achaz. Evolutionary constraints in fitness landscapes. *Heredity*, 121(5):466–481, November 2018. ISSN 1365-2540. doi: 10.1038/s41437-018-0110-1. Publisher: Nature Publishing Group.
98. Andrei Papkou, Lucia Garcia-Pastor, José Antonio Escudero, and Andreas Wagner. A rugged yet easily navigable fitness landscape. *Science*, 382(6673):eadh3860, November 2023. doi: 10.1126/science.adh3860. Publisher: American Association for the Advancement of Science.
99. Jakub Otwinowski, David M. McCandlish, and Joshua B. Plotkin. Inferring the shape of global epistasis. *Proceedings of the National Academy of Sciences*, 115(32):E7550–E7558, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1804015115.
100. Adam J. Hockenberry and Claus O. Wilke. Phylogenetic Weighting Does Little to Improve the Accuracy of Evolutionary Coupling Analyses. *Entropy*, 21(10):1000, October 2019. ISSN 1099-4300. doi: 10.3390/e21101000.
101. Edwin Rodriguez Horta and Martin Weigt. On the effect of phylogenetic correlations in coevolution-based contact prediction in proteins. *PLOS Computational Biology*, 17(5): e1008957, May 2021. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008957.
102. Nicola Dietler, Umberto Lupo, and Anne-Florence Bitbol. Impact of phylogeny on structural contact inference from protein sequence data, January 2023. arXiv:2209.13045 [physics, q-bio].
103. Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009. ISSN 1367-4811, 1367-4803. doi: 10.1093/bioinformatics/btp352.
104. James K Bonfield, John Marshall, Petr Danecek, Heng Li, Valeriu Ohan, Andrew Whitwham, Thomas Keane, and Robert M Davies. HTSlib: C library for reading/writing high-throughput sequencing data. *GigaScience*, 10(2):giab007, January 2021. ISSN 2047-217X. doi: 10.1093/gigascience/giab007.
105. Ammar Tareen and Justin B Kinney. Logomaker: beautiful sequence logos in Python. *Bioinformatics*, 36(7):2272–2274, April 2020. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btz921.

Supplementary Figures

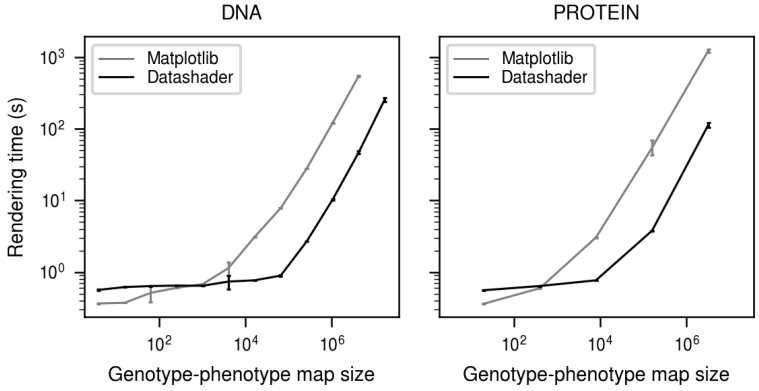

Fig. S1. Visualization rendering times using two different back-end libraries for plotting as a function of the size of DNA and protein genotype-phenotype maps.

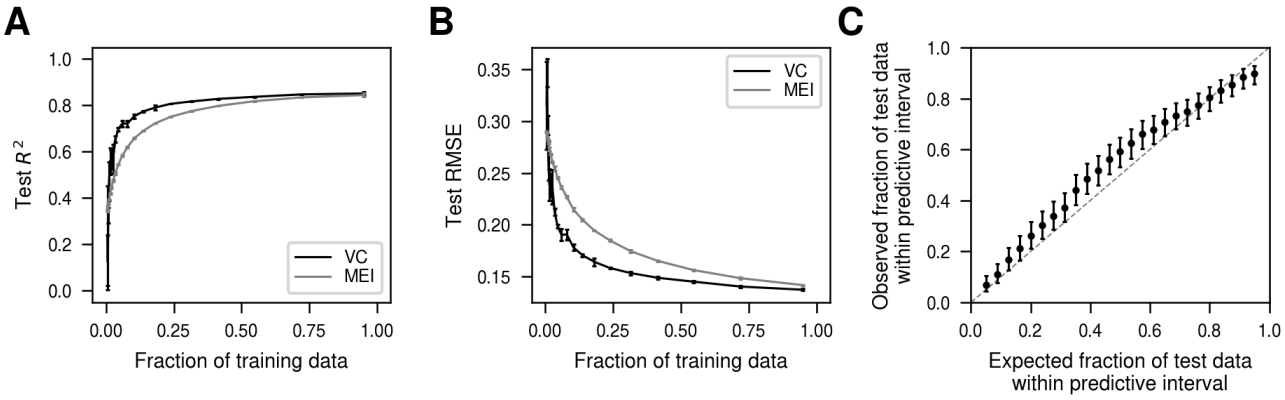

Fig. S2. Predictive performance of Minimum Epistasis Interpolation and Variance Component regression in held-out data. (A,B) Model predictive performance measured by the R^2 (A) and RMSE (B) in held-out data as a function of the fraction of data used for training and phenotypic prediction. Error bars represent the standard deviation over 3 independent subsets of sequences used for training at each proportion. (C) Evaluation of Variance Component regression model 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.

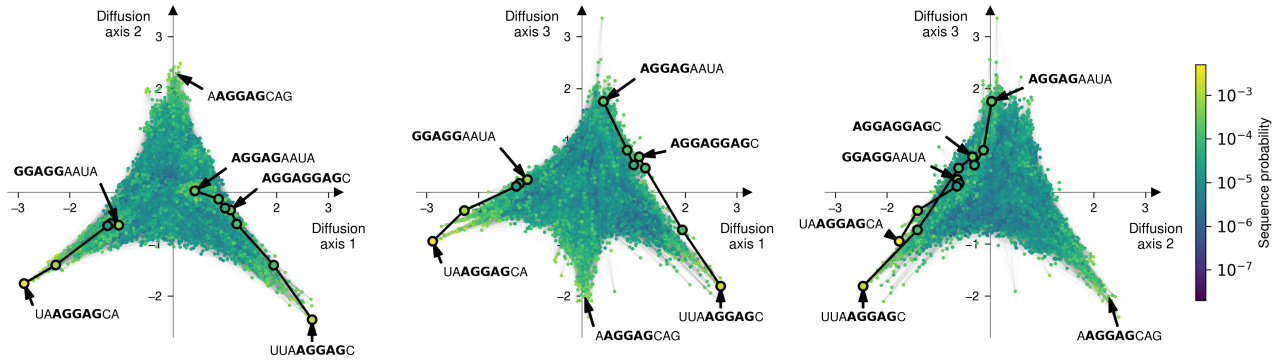

Fig. S3. Low dimensional representation of the Shine-Dalgarno probability distribution inferred with SeqDEFT along Diffusion axes 1, 2 and 3. Every dot represents one of the possible 4^9 possible sequences and is colored according to their inferred probability. Sequences are laid out according to the indicated Diffusion axes and dots are plotted in order according to the missing Diffusion axis.

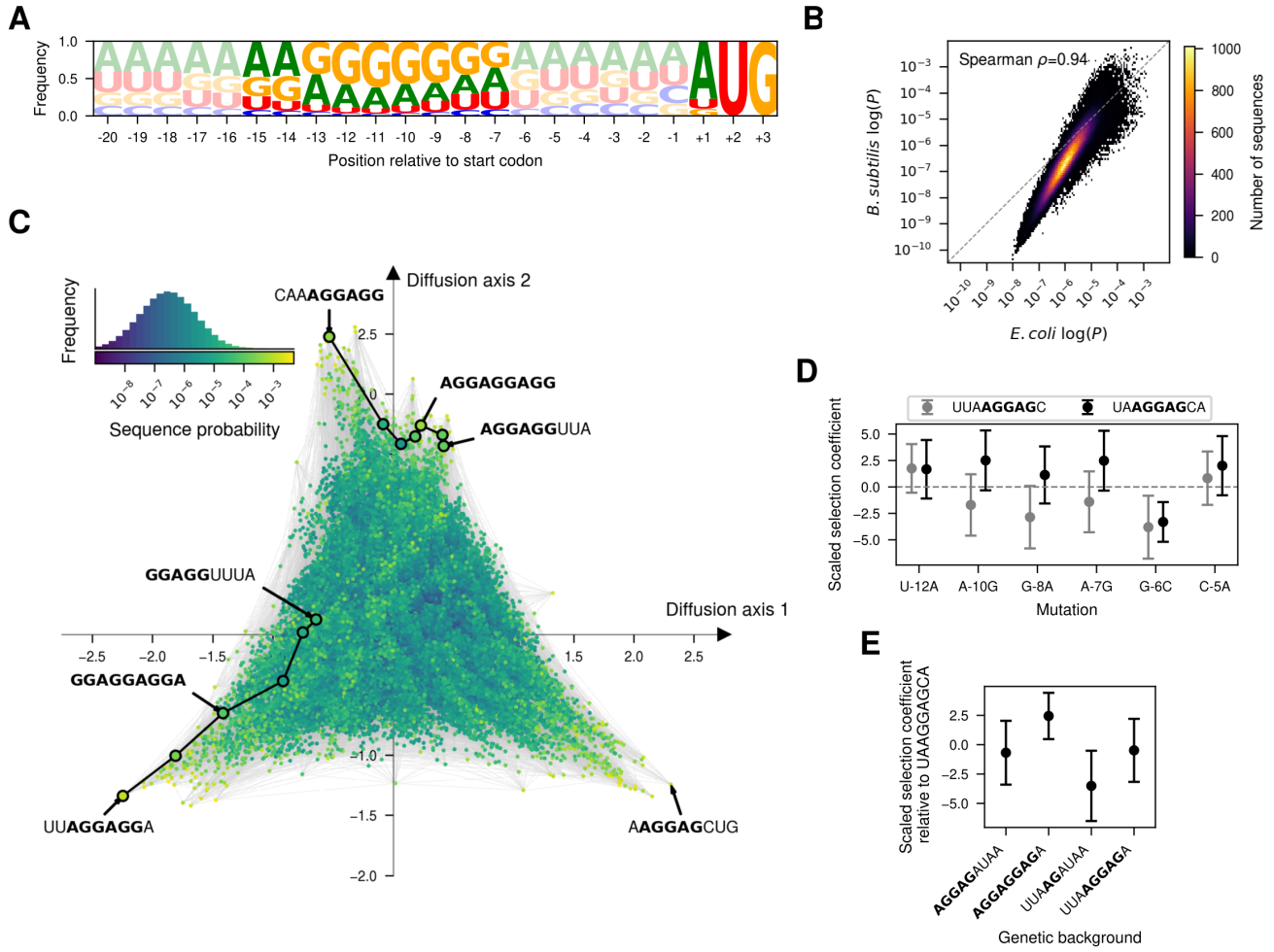

Fig. S4. The structure of the genotype-phenotype map inferred from *B. subtilis* is conserved. (A) Sequence logo representing the site-specific allele frequencies of 4,328 5'UTRs in the *B. subtilis* genome aligned with respect to the annotated start codon. The start codon and the 9 nucleotide sequences 4 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 of the effects of specific mutations when introduced in two genetic contexts UUAAGGAGC and UAAGGAGCA representing a shift by one position of the AGGAG motif. (E) Posterior distribution of phenotypes associated to genotypes representing the shift of the AGGAG motif by 3 positions. Points represent the Maximum a posteriori (MAP) estimates and error bars represent the 95% credible intervals.

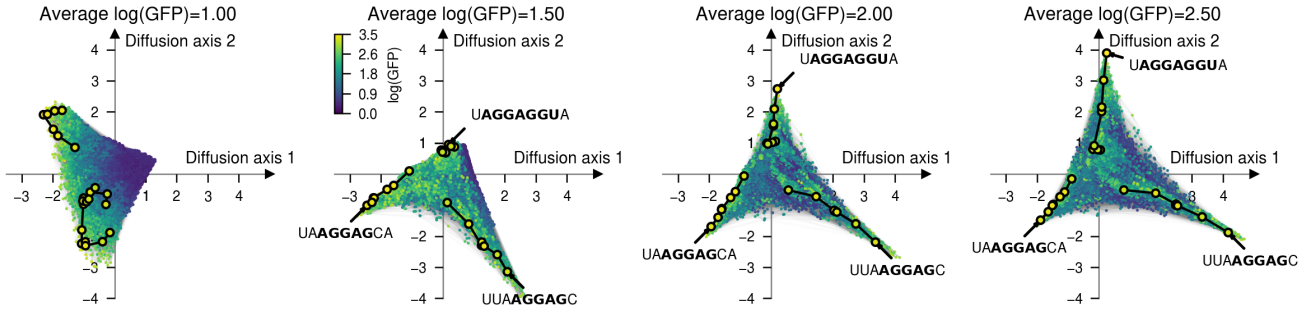

Fig. S5. Low-dimensional representation of the Shine-Dalgarno genotype-phenotype map inferred with VC regression from MAVE data along Diffusion axes 1 and 2 as a function of the assumed average log(GFP) at the stationary distribution (as determined by tuning the strength of selection parameter c). Every 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.

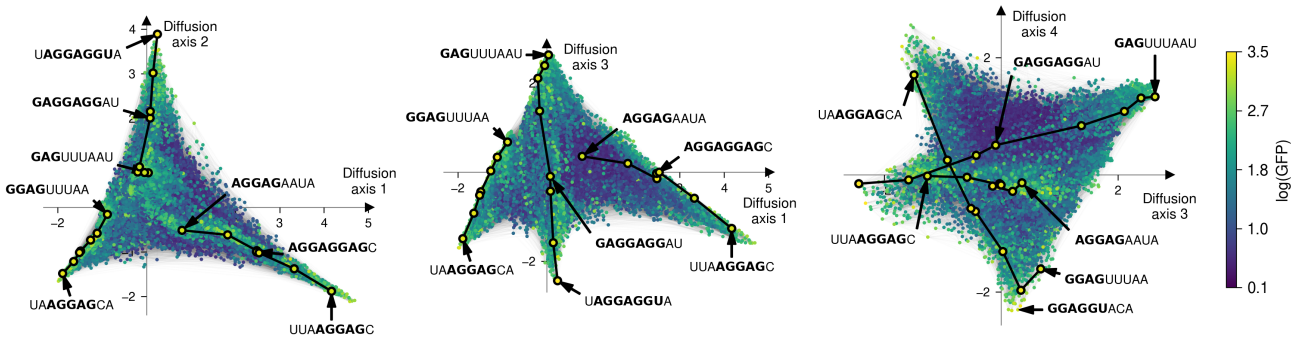

Fig. S6. Low dimensional representation of the Shine-Dalgarno genotype-phenotype map inferred with VC regression from MAVE data along Diffusion axes 1, 2, 3 and 4. Every dot represents one of the possible 4^9 possible sequences and is colored according to its inferred log(GFP) value. Sequences are laid 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.

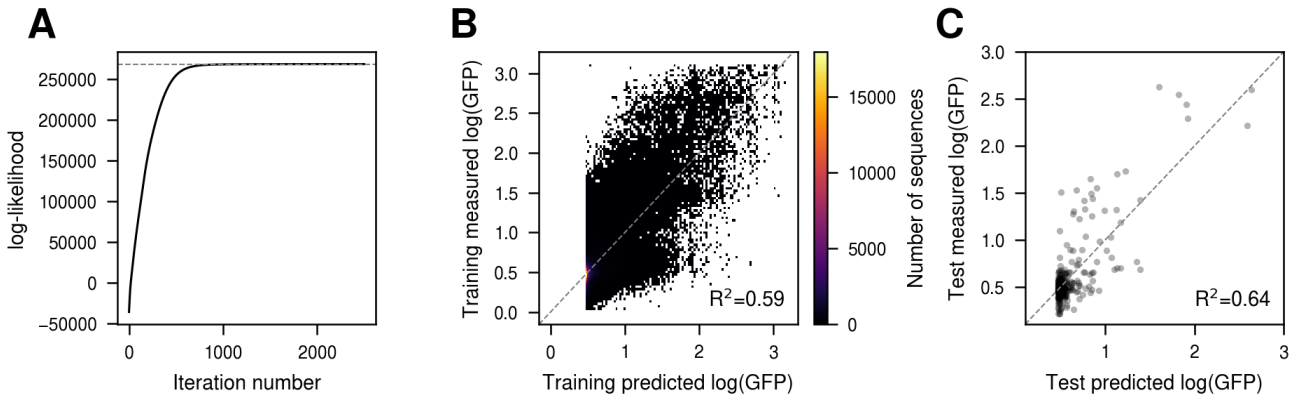

Fig. S7. Fitting a thermodynamic model to the Shine-Dalgarno genotype-phenotype map using MAVE data. (A) Training curve showing the evolution of the log-likelihood as a function of the number of iterations of the Adam optimizer. (B) Comparison of measured log(GFP) in the training data with the predicted 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.

Supplementary Information

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

$$L(i,j) = \begin{cases} -1 & \text{if } i \text{ and } j \text{ are neighbors} \\ \ell(\alpha-1) & \text{if } i = j \\ 0 & \text{Otherwise} \end{cases} \tag{14}$$

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 with memory requirements that scale only with the size of sequence space α^ℓ. 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 $\mathbf{B} \in \mathbb{R}^{(\alpha\times\alpha\times\ldots\times\alpha)}$ 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 $\mathbf{W}$, where $\mathbf{W}_{x_1,x_2,\ldots,x_\ell} = w_x$, can be easily computed in tensor form using broadcasting,

$$\mathbf{W} = \sum_i^\ell \mathbf{B}^{(i)} \tag{15}$$

where $\mathbf{B}^{(i)}_{x_1,\ldots,x_{i-1},*,x_{i+1},\ldots,x_\ell} = \sum_c^\alpha \mathbf{B}_{x_1,\ldots,x_{i-1},c,x_{i+1},\ldots,x_\ell}$, which can be efficiently computed by summing the entries of $\mathbf{B}$ over axis i. We can then use $\mathbf{B}$ and $\mathbf{W}$ to calculate $w = (\ell I + A)b$ as:

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

Properties of the P_U projection operators. Let f represent a genotype-phenotype map in the space of sequences with a single site and α different alleles. The function f can be projected into the constant subspace V_0 through the projection matrix $P_0 = b(b^T b)^{-1} b = \frac{1}{\alpha}\vec{1}\vec{1}^T$ and into the orthogonal subspace through the projection matrix $P_1 = I - P_0 = I - \frac{1}{\alpha}\vec{1}\vec{1}^T$. For any pair of sequences x, y, $P_0(x,y) = \frac{1}{\alpha}$ and

$$P_1(x,y) = \begin{cases} \frac{-1}{\alpha} & \text{if } x \neq y \\ \frac{\alpha-1}{\alpha} & \text{if } x = y. \end{cases}$$

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 $V_p = V_1$ for $p \in U$ and $V_p = V_0$ for $p \notin U$. Thus, the projection operator into the subspaces defined by U are obtained through Kronecker product $P_U = \bigotimes_p^\ell P_p$

$$P_U(x,y) = \alpha^{-\ell} \prod_{\substack{p\in U \\ x_p \neq y_p}} (-1) \prod_{\substack{p\in U_p \\ x_p = y_p}} (\alpha-1). \tag{16}$$

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 $(A\otimes B)(C\otimes D) = 0$ are orthogonal. If we consider two subspaces defined by different subsets of sites U and U', there will be at least one position at which the two site-specific elementary subspaces are orthogonal to each other $P_p P'_p = P_0 P_1 = 0$, such that $P_U P'_U = (\bigotimes_p^\ell P_p)(\bigotimes_p^\ell P'_p) = \bigotimes_p^\ell (P_p P'_p) = 0$. Next, we consider the subspaces defined by the direct sum of subspaces U containing exactly k sites $V_k = \bigoplus_{U:|U|=k} V_U$ and derive the corresponding projection operator P_k:

$$
\begin{aligned}
P_k(x,y) &= \sum_{U:|U|=k} P_U(x,y) = \sum_{U:|U|=k} \alpha^{-\ell} \prod_{\substack{p\in U\\ x_p\neq y_p}} (-1) \prod_{\substack{p\in U_p\\ x_p=y_p}} (\alpha-1) \\
&= \alpha^{-\ell} \sum_{U:|U|=k} \prod_{\substack{p\in U\\ x_p\neq y_p}} (-1) \prod_{\substack{p\in U_p\\ x_p=y_p}} (\alpha-1)
\end{aligned} \tag{17}
$$

We note that the elements in the sum can only be obtained by multiplying the factors $(\alpha-1)$ and (-1) k times. Therefore, these products can take only $k+1$ possible values $(\alpha-1)^0(-1)^l, (\alpha-1)^1(-1)^{\ell-1}, ..., (\alpha-1)^k(-1)^0$ or more generally $(-1)^q(\alpha-1)^{\ell-q}$, such that P_k can be expressed by:

$$
P_k(x,y) = \alpha^{-\ell} \sum_{q}^{k} (-1)^q (\alpha-1)^{k-q} n_q, \tag{18}
$$

where n_q is the number of times each unique value appears when summing through the corresponding P_U matrices. Whether we take $(\alpha-1)$ or (-1) depends on whether sequences x and y have the same alleles at the sites in U, but not on which alleles. Moreover, because we are summing over all possible U of the same size, n_q does not depends on the specific sites that are different, but only on the Hamming distance $d(x,y)$ between sequences x and y. Specifically n_q can be obtained by multiplying the number of ways in which we can select q sites within the set of different sites $d(x,y)$, given by $\binom{d(x,y)}{q}$, with the number of ways we can select the $k-q$ sites within the set of $\ell-d(x,y)$ sites with the same allele, given by $\binom{\ell-d(x,y)}{k-q}$.

$$
P_k(x,y) = \alpha^{-\ell} \sum_{q}^{k} (-1)^q (\alpha-1)^{k-q} \binom{d(x,y)}{q} \binom{\ell-d(x,y)}{k-q}. \tag{19}
$$

Note that this expression corresponds to the projection operator into the k-th order subspace (61), showing that the k-th order subspace can be decomposed into smaller 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 P_U orthogonality properties:

$$
P_k P_U = \sum_{U':|U'|=k} P_{U'} P_U = \begin{cases} P_U & \text{if } |U|=k \\ 0 & \text{if } |U|\neq k. \end{cases} \tag{20}
$$

Using the same argument, we can show that the columns of P_U span a non-zero subspace within the k-th order eigenspace of the graph Laplacian since $P_U P_k = P_U \sum_{U':|U'|=k} P_{U'}$ when $|U|=k$.

Linear operator for ℓ-Kronecker products. In the previous section, 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 ($P_U = \bigotimes_p^\ell P_p$). Specifically, for any matrix A for any matrix obtained through an ℓ-Kronecker product $A = \bigotimes_p^\ell A_p$, we can compute the matrix vector product Ab without explicitly constructing A as follows:

$$
\begin{aligned}
Ab &= (\bigotimes_{p=1}^{\ell} A_p)b = (A_1 \bigotimes_{p=2}^{\ell} A_p)b \\
&= \begin{bmatrix} (A_1)_{11}(\bigotimes_{p=2}^{\ell} A_p)b_1 + (A_1)_{12}(\bigotimes_{p=2}^{\ell} A_p)b_2 + \cdots + (A_1)_{1\alpha}(\bigotimes_{p=2}^{\ell} A_p)b_\ell \\ (A_1)_{21}(\bigotimes_{p=2}^{\ell} A_p)b_1 + (A_1)_{22}(\bigotimes_{p=2}^{\ell} A_p)b_2 + \cdots + (A_1)_{2\alpha}(\bigotimes_{p=2}^{\ell} A_p)b_\ell \\ \vdots \\ (A_1)_{\alpha 1}(\bigotimes_{p=2}^{\ell} A_p)b_1 + (A_1)_{\alpha 2}(\bigotimes_{p=2}^{\ell} A_p)b_2 + \cdots + (A_1)_{\alpha\alpha}(\bigotimes_{p=2}^{\ell} A_p)b_\ell \end{bmatrix} \\
&= \begin{bmatrix} \sum_c^\alpha (A_1)_{1c}(\bigotimes_{p=2}^{\ell} A_p)b_c \\ \sum_c^\alpha (A_1)_{2c}(\bigotimes_{p=2}^{\ell} A_p)b_c \\ \vdots \\ \sum_c^\alpha (A_1)_{\alpha c}(\bigotimes_{p=2}^{\ell} A_p)b_c \end{bmatrix} = \begin{bmatrix} (\bigotimes_{p=2}^{\ell} A_p)\sum_c^\alpha (A_1)_{1c}b_c \\ (\bigotimes_{p=2}^{\ell} A_p)\sum_c^\alpha (A_1)_{2c}b_c \\ \vdots \\ (\bigotimes_{p=2}^{\ell} A_p)\sum_c^\alpha (A_1)_{\alpha c}b_c \end{bmatrix} = \begin{bmatrix} (\bigotimes_{p=2}^{\ell} A_p)u_{11} \\ (\bigotimes_{p=2}^{\ell} A_p)u_{12} \\ \vdots \\ (\bigotimes_{p=2}^{\ell} A_p)u_{1\alpha} \end{bmatrix},
\end{aligned}
$$

where we let $u_{1i} = \sum_c^\alpha (A_1)_{ic} b_c$ and note that these vectors can be computed simultaneously by multiplying an $\alpha \times \alpha$ matrix with an $\alpha \times \alpha^{\ell-1}$ matrix

$$\begin{bmatrix} u_{11}^T \\ u_{12}^T \\ \vdots \\ u_{1\alpha}^T \end{bmatrix} = \begin{bmatrix} \sum_c^\alpha (A_1)_{1c} b_c^T \\ \sum_c^\alpha (A_1)_{2c} b_c^T \\ \vdots \\ \sum_c^\alpha (A_1)_{\alpha c} b_c^T \end{bmatrix} = A_1 \begin{bmatrix} b_1^T \\ b_2^T \\ \vdots \\ b_\alpha^T \end{bmatrix}.$$

Once these vectors are computed, we can use the same strategy to compute the $(\bigotimes_{p=2}^{\ell} A_p) u_{1j}$. Recursively repeating this calculation, we find that we can compute Ab using only ℓ different $\alpha \times \alpha$ by $\alpha \times \alpha^{\ell-1}$ matrix multiplication operations, avoiding storage and computation of the $\alpha^\ell \times \alpha^\ell$ dense matrix A.

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 am $n \times n$ 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

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

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:

$$\log p(f|y) \propto -\frac{1}{2} f^T C f - \frac{1}{2}(y - X^T f)^T D^{-1} (y - X^T f). \tag{21}$$

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

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

We next take the gradient with respect to f

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

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

$$\hat{f} = (C + X D^{-1} X^T)^{-1} X D^{-1} y. \tag{22}$$

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

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

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

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

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

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

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:

$$
\Sigma = \left(\nabla \nabla_f log(f|y) \right)^{-1} = \left(C + XD^{-1}X^T \right)^{-1}, \tag{23}
$$

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:

$$
\begin{aligned}
\Sigma_{zz} &= Z^T \left(C + XD^{-1}X^T \right)^{-1} Z \\
&= Z^T \left(C^{-1} - C^{-1} \left(X^T C^{-1} X + D \right)^{-1} C^{-1} \right) Z \\
&= Z^T \left(K - K \left(X^T KX + D \right)^{-1} K \right) Z \\
&= Z^T KZ - Z^T K \left(X^T KX + D \right)^{-1} KZ \\
&= K_{zz} - K_{zx} \left(K_{xx} + D \right)^{-1} K_{xz},
\end{aligned} \tag{24}
$$

as desired.