On the sparsity of fitness functions and implications for learning

David H. Brookes[1], Amirali Aghazadeh[2], and Jennifer Listgarten[2,3,*]

[1]Biophysics Graduate Group, University of California, Berkeley, Berkeley, CA, USA
[2]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA, USA
[3]Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA

Abstract

Fitness functions map biological sequences to a scalar property of interest. Accurate estimation of these functions yields biological insight and sets the foundation for model-based sequence design. However, the amount of fitness data available to learn these functions is typically small relative to the large combinatorial space of sequences; characterizing how much data is needed for accurate estimation remains an open problem. There is a growing body of evidence demonstrating that empirical fitness functions display substantial sparsity when represented in terms of epistatic interactions. Moreover, the theory of Compressed Sensing provides scaling laws for the number of samples required to exactly recover a sparse function. Motivated by these results, we study the sparsity of fitness functions sampled from a generalization of the NK model, a widely-used random field model of fitness functions. In particular, we present theoretical results that allow us to test the effect of the Generalized NK (GNK) model's interpretable parameters—sequence length, alphabet size, and assumed interactions between sequence positions—on the sparsity of fitness functions sampled from the model and, consequently, the number of measurements required to exactly recover these functions. Further, we show that GNK fitness functions with parameters set according to protein structural contacts can be used to accurately approximate the number of samples required to estimate two empirical protein fitness functions, and are able to identify important higher-order epistatic interactions in these functions using only structural information.

Introduction

Advances in high-throughput experimental technologies now allow for the probing of the fitness of thousands, and sometimes even millions, of biological sequences. However, these measurements generally represent only a tiny fraction of those required to comprehensively characterize a fitness function. It is therefore critical to develop methods that can estimate fitness functions from an incomplete set of measurements. Many methods have been proposed for this purpose, ranging from the fitting of regularized linear models [1], and parameterized biophysical models [2, 3] to nonparametric techniques [4, 5], and various nonlinear machine learning approaches [6], including deep neural networks [7, 8]. In addition to providing basic biological insight, such methods have been used to improve the efficiency and success rate of experimental protein engineering approaches [9, 10, 11] and are crucial components of *in-silico* sequence design tools [12, 13, 14, 15].

Despite these advances in fitness function estimation, the answer to one fundamental question remains elusive—namely how many experimental fitness measurements are required to accurately estimate a fitness function. We refer to this problem as that of determining the sample complexity of fitness function estimation. Insights on this topic can be used to inform researchers on which of a variety of experimental techniques should be used to probe a particular fitness function of interest, and on how to restrict the

*To whom correspondence should be addressed. Email: jennl@berkeley.edu

scope of an experimental probe such that the resulting data allows one to accurately estimate the function under study. Our central focus herein is to elucidate the open question of the sample complexity of fitness function estimation.

It has recently been observed that some empirical fitness functions—those for which experimental fitness measurements are available for all possible sequences—are sparse when represented in the Walsh-Hadamard basis, which represents fitness functions in terms of all possible "epistatic" interactions (i. e., nonlinear contributions to fitness due to interacting sequence positions) [16, 17, 3]. Further, this sparsity property has been exploited to improve estimators of such functions [18, 19, 20]. Indeed, it is well known in the field of signal processing that sparsity enables more statistically efficient estimation of functions. Additionally, results from Compressed Sensing (CS), a sub-field of signal processing, provide scaling laws for the number of measurements required to recover a function in terms of its sparsity [21, 22]. These results suggest that by studying the sparsity of fitness functions in more depth, we may be able to predict the sample complexity of fitness function estimation.

Although an increasing number of empirical fitness functions are available that could allow us to investigate sparsity in particular example systems, these data necessarily only report on short sequences in limited environments. A common approach in evolutionary biology to overcome the lack of sufficient empirical fitness functions is to instead study 'random field' models of fitness, which assign fitness values to sequences based on stochastic processes constructed to mimic the statistical properties of natural fitness functions [23, 24]. We will follow a similar line of reasoning and study the sparsity of fitness functions sampled from random field models, allowing us to probe properties of a much broader class of fitness functions than the available empirical data. We will make use of a particular random field model, namely a generalization of the widely-used NK model [25]. The NK model is known to represent a rich variety of realistic fitness functions despite requiring only two parameters to be defined: L, the sequence length,[1] and K, the maximum degree of epistatic interactions. In the NK model, each sequence position interacts with a "neighborhood" of $K-1$ other positions that either include directly adjacent positions or are chosen uniformly at randomly [23]. The Generalized NK (GNK) model [26], extends the model by allowing neighborhoods to be of arbitrary size and content. We will refer to simulated fitness functions sampled from the GNK model as 'GNK fitness functions'.

Buzas and Dinitz [26] calculated the sparsity of GNK fitness functions represented in the WH basis as a function of the sequence length and the composition of the neighborhoods. Nowak and Krug [27] expanded on this work by calculating the sparsity of GNK fitness functions with a few specific neighborhood schemes, as a function of only the size of the neighborhoods. Notably, these works consider only binary sequences, and use sparsity as a tool to understand the properties of adaptive walks on GNK landscapes, without connecting it to fitness function estimation. In contrast, our aim is to determine the sample complexity of estimating GNK fitness functions in the biologically-relevant scenarios where sequences are made up of non-binary elements (e. g., nucleotide or amino acid alphabets). In order to do so, we extend the results of refs. 26 and 27 to the case of non-binary alphabets by employing "Fourier" bases, which are generalizations of the WH basis that can be constructed for any alphabet size. We then leverage CS theory to determine the minimum number of samples required to recover GNK fitness functions in the Fourier basis.

Additionally, we describe a novel scheme to construct GNK neighborhoods based on contacts in a given atomistic protein structure, and show that the resulting GNK fitness functions can be used to accurately predict the sample complexity of estimating empirical fitness functions of two proteins: the mTagBFP2 fluorescent protein [18] and the protein coded by the His3 gene in *Saccharomyces cerevisiae*. Surprisingly, these GNK fitness functions correctly identify many of the important higher-order epistatic interactions in the empirical fitness functions while being constructed using only the second-order structural contact information. These results demonstrate the practical utility of our theory and further bolster a growing understanding of the importance of structural contacts in shaping protein fitness functions.

In the next sections, we summarize the relevant background material required for our main results.

[1] In the original definition of the model, N is used for the sequence length, but here we reserve N for the number of observed measurements.

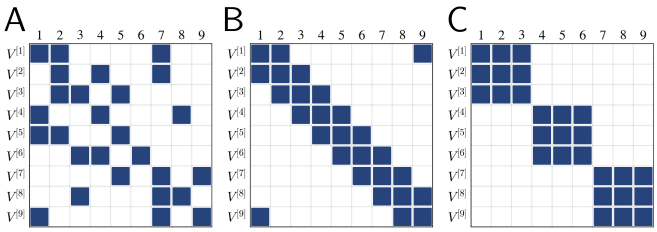

Fig. 1: Graphical depictions of GNK neighborhood schemes for $L = 9$ and $K = 3$. Rows on the vertical axis represent neighborhoods and columns on the horizontal axis represents sequence positions. A square in the $(i, j)^{\text{th}}$ position in the grid denotes that sequence position j is in the neighborhood $V^{[i]}$. (a) Random Neighborhoods (b) Adjacent Neighborhoods (c) Block Neighborhoods.

Fitness functions and estimation

A fitness function maps sequences to a scalar property of interest, such as catalytic efficiency [17], binding affinity [2], or fluorescent brightness [28]. In particular, let $\mathcal{S}^{(L,q)}$ be the set of all q^L possible sequences of length L whose elements are members of an alphabet of size q (e. g., $q = 4$ for nucleotides, and 20 for amino acids); then a fitness function is any function that maps the sequence space to scalar values, $f : \mathcal{S}^{(L,q)} \to \mathbb{R}$. In practice, sequences may contain different alphabets at different positions, but these can usually be mapped to a common alphabet. For instance, one position in a nucleotide sequence may be restricted to A or T, and another to G or C, but both of these can be mapped onto the binary alphabet {0,1}. In the SI we consider the case where the size of the alphabet may be different at each position.

Any fitness function of sequences of length L and alphabet size q can be represented exactly as

$$\mathbf{f} = \mathbf{\Phi}\boldsymbol{\beta}, \tag{1}$$

where $\mathbf{f}$ is the vector of all q^L fitness values, one for each possible sequence, $\mathbf{\Phi}$ is a $q^L \times q^L$ orthogonal basis, and $\boldsymbol{\beta}$ is the vector of q^L coefficients corresponding to the fitness function in that basis. Although any orthogonal basis may be used, here we restrict $\mathbf{\Phi}$ to refer to either the Walsh-Hadamard basis (when $q = 2$), or the Fourier basis (for $q > 2$), which will be defined shortly. Each row of $\mathbf{\Phi}$ represents an encoding of a particular sequence in $\mathcal{S}^{(L,q)}$. Now suppose we observe N fitness measurements, $\mathbf{y} \in \mathbb{R}^N$, for N different sequences, each corresponding to one of the rows of $\mathbf{\Phi}$. The goal of fitness function estimation is then to recover a good approximation to $\boldsymbol{\beta}$ using these N measurements, which correspond to only a subset of all possible sequences. In general, this is an underdetermined linear system that requires additional information to be solved, and many methods have been developed for this purpose. The extent to which a fitness function is recovered by such a method can be assessed by the mean squared error (MSE) between the estimated and true coefficients. Because $\mathbf{\Phi}$ is an orthogonal matrix this is equivalent to the MSE between the true fitness values $\mathbf{f}$ and those predicted using the estimated coefficients.

The field of Compressed Sensing is primarily concerned with studying algorithms that can solve underdetermined systems, and specifying the conditions under which recovery with a specified amount of error in the estimated coefficients can be guaranteed. Therefore, it stands to reason that CS may be helpful for characterizing fitness function estimation problems. The LASSO algorithm is among the most widely-used and well-studied for solving underdetermined systems, both in CS and also in machine learning [29]. The key determinant of success of algorithms such as LASSO in recovering a particular function is how sparse that function is when represented in a particular basis, or how well it can be approximated by a function that is sparse in that basis. Using the fitness function estimation problem as an example, a central result from CS [30] states that if $\boldsymbol{\beta}$ is an S-sparse vector (meaning that is has exactly S nonzero elements), then with high probability LASSO can recover $\boldsymbol{\beta}$ exactly with

$$N \geq C \cdot S \log q^L \tag{2}$$

noiseless fitness measurements, where C is an unknown constant. For this bound to hold, the N sequences with observed fitness measurements must be sampled uniformly from the space of sequences [30]. It has

also been shown that if $\boldsymbol{\beta}$ is only approximately sparse (i. e., it has many small, but nonzero, coefficients), or if there is noise in the measurements, then the error in the recovery can still be bounded (Materials and Methods).

Eq. 2 shows that if we are able to calculate the sparsity of a fitness function, and estimate a value for the constant C, then we can calculate the number of samples required to recover that fitness function with LASSO. Note that the "sparsity" of a fitness function is defined as the number of nonzero coefficients when the fitness function is expanded in a particular basis.[2] Sparsity is defined with respect to the particular orthonormal basis which must therefore be chosen carefully. In the next section, we discuss bases that can be used to represent fitness functions.

Fourier bases for fitness functions

The sparsity of a class of natural signals depends crucially on the basis with which they are represented. Many fitness functions have been shown to be sparse in the Walsh-Hadamard (WH) basis [16, 17, 3], which has also been used extensively in theoretical studies of fitness landscapes [31, 32, 33] and even to unify multiple definitions of epistasis [34]. The WH basis can be interpreted as encoding fitness functions in terms of epistatic interactions [35, 34]. In particular, when a fitness function of binary sequences of length L is represented in the form of Eq. 1 (with $\boldsymbol{\Phi}$ being the WH basis), then the sequence elements are encoded as $\{-1, 1\}$ and the fitness function evaluated on a sequence, $\mathbf{s} = [s_1, s_2, ..., s_L]$, has the form of an intuitive multi-linear polynomial [20],

$$f(\mathbf{s}) = \sum_{U \in \mathcal{U}} \beta_U \prod_{i \in U} s_i, \tag{3}$$

where $\mathcal{U} := \mathscr{P}(\{1, 2, ..., L\})$ is the power set of sequence position indices. Each of the 2^L elements of $\mathcal{U}$ is a set of indices that corresponds to a particular epistatic interaction, with the size of that set indicating the order of the interaction (e. g., if a $U \in \mathcal{U}$ is of size $|U| = r$, then it represents an rth order interaction). The coefficient β_U is an element of $\boldsymbol{\beta}$, indexed by its corresponding epistatic interaction set.

The WH basis can only be used to represent fitness functions of binary sequences, which poses a challenge in biological contexts where common alphabets include the nucleotide ($q = 4$) and amino acid ($q = 20$) alphabets. This issue is typically skirted by encoding elements of a larger alphabet as binary sequences (e. g., by using a "one-hot encoding"), and using the WH basis to represent fitness functions of these encoded sequences. However, doing so results in an inefficient representation, which has dramatic consequences on the calculation of sample complexities. To see this, consider the "one-hot" encoding scheme of amino acids, where each amino acid is encoded as a length 20 bit string. The number of amino acid sequences of length L is 20^L, while the one-hot encodings of these sequences are elements of a binary sequence space of size $2^{20L} = 1,048,576^L$. This latter number also corresponds to the number of WH coefficients required to represent the fitness function in the one-hot encoding, and is much too large to be of any practical use.

Although it is not widely recognized in the fitness function literature, it is possible to construct bases analogous to the WH basis for arbitrarily-sized alphabets, which we refer to as "Fourier" bases (Materials and Methods, 36, 37). The WH basis is the Fourier basis for $q = 2$. The Fourier basis for a larger alphabet shares much of the WH basis's intuition of encoding epistatic interactions between positions in a sequence. In particular, we have an analogous expression to Eq. 3 for the Fourier basis, in which the fitness function is represented as a sum of 2^L terms, each of which corresponds to an epistatic interaction. In the WH basis, an rth epistatic interaction U in a sequence s is encoded as the scalar $\left(\prod_{i \in U} s_i\right) \in \{-1, 1\}$, while in the Fourier basis, it is represented by a length $(q-1)^r$ vector, which we denote as $\boldsymbol{\phi}_U(\mathbf{s})$. Similarly, in the WH basis each epistatic interaction is associated with a single coefficient, while in the Fourier basis,

[2] In a quirk of common terminology, a signal is called "sparse" when it contains many zero coefficients, but the "sparsity" is formally defined as the number of nonzero coefficients. Thus, a 'sparse' signal has *low* "sparsity".

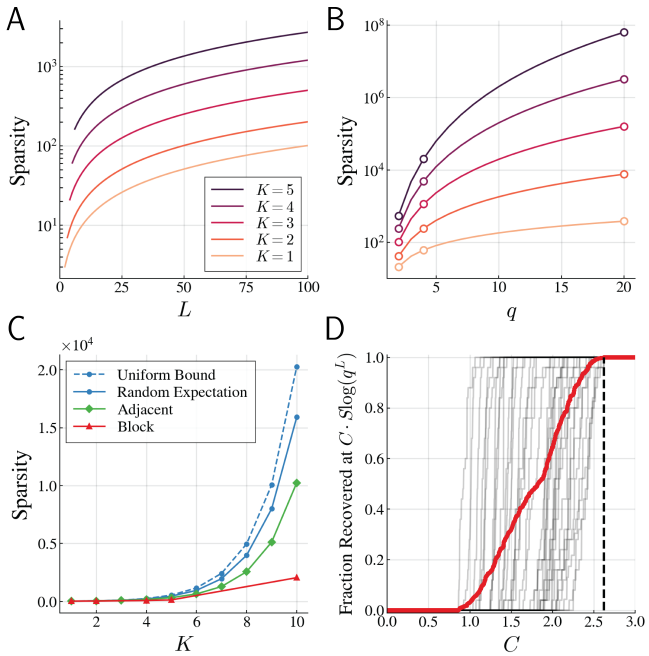

Fig. 2: The sparsity of GNK fitness functions. (A) Upper bound on sparsity of GNK fitness functions with constant neighborhood sizes for $q = 2$ and a range of settings of the L and K parameters (B) Upper bound for $L = 20$ and a range of settings of the alphabet size q and the K parameter (colors as in (A)). Alphabet sizes corresponding to binary ($q = 2$), nucleotide ($q = 4$), and amino acid ($q = 20$) alphabets are highlighted with open circles. (C) Sparsity of GNK fitness functions with neighborhoods constructed with each of the standard neighborhood schemes for $L = 20$ and $q = 2$, and and a range of settings of K, denoted by markers. (D) Fraction of sampled GNK fitness functions with Random Neighborhoods recovered at a range of settings of C. Each grey curve represents sampled fitness functions at a particular values of $L \in \{5, 6, ..., 13\}$, $q \in \{2, 3, 4\}$ and $K \in \{1, 2, 3, 4, 5\}$. The red curve averages over all 907 sampled functions. The value $C = 2.62$, which we chose to use for subsequent numerical experiments, is highlighted with a dashed line.

each epistatic interaction is associated with $(q-1)^r$ coefficients. All together, the evaluation of a fitness function represented in the Fourier basis on a sequence $\mathbf{s}$ is given by

$$f(\mathbf{s}) = \sum_{U \in \mathcal{U}} (\boldsymbol{\beta}_U)^T \, \boldsymbol{\phi}_U(\mathbf{s}). \tag{4}$$

As we will see, when GNK fitness functions are represented in the Fourier basis, then we have the intuitively pleasing result that all of the Fourier coefficients associated with a particular epistatic interaction are identically distributed, and thus the GNK model can be interpreted in terms of epistatic interactions.

The Generalized NK model

Sampling fitness functions from a random field model provides a means to simulate fitness functions of sequences of any length or alphabet size. A random field model specifies a stochastic process that assigns fitness values to all possible sequences. This process implicitly defines a joint probability distribution over the fitness values of all sequences, and another over all of the Fourier coefficients, $\boldsymbol{\beta}$.

We will focus on the Generalized NK (GNK) model [26]. In order to be defined, the GNK model requires the specification of the sequence length L, alphabet size q, and an interaction "neighborhood" for each position in the sequence. A neighborhood, $V^{[j]}$, for sequence position j is a set of position indices that contains j itself, and $K_j - 1$ other indices, where we define $K_j := |V^{[j]}|$ to be the size of the neighborhood.

Given L, q, and a neighborhood for each position, the GNK model assigns fitness to every sequence in the sequence space via a series of stochastic steps (Materials and Methods). In the GNK model, two sequences have correlated fitness values to the extent that they share subsequences corresponding to the positions in the neighborhoods. For example, consider a GNK model defined for nucleotide sequences of

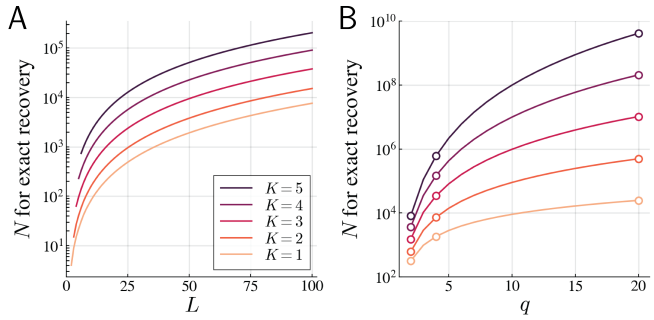

Fig. 3: Minimum number of measurements required to exactly recover GNK fitness functions with constant neighborhood sizes. (A) Upper bound on the minimum N required to recover GNK fitness functions with constant neighborhood sizes for $q = 2$ and a range of settings of the L and K parameters. (B) Upper bound for $L = 20$ and a range of settings of the alphabet size q and the K parameter (colors as in (A)). Alphabet sizes corresponding to binary ($q = 2$), nucleotide ($q = 4$), and amino acid ($q = 20$) alphabets are highlighted with open circles.

length 3, where the first neighborhood is $V^{[1]} = \{1, 3\}$. Then the sequences ACG and AAG will have partially correlated fitness values because they both contain the subsequence AG in positions 1 and 3. One of the key intuitions of the GNK model is that larger neighborhoods will produce more "rugged" fitness functions in which many fitness values are uncorrelated, because it is less likely for two sequences to share subsequences when the neighborhoods are large. Note that larger neighbourhoods also implies the presence higher order epistatic interactions.

The key choice in specifying the GNK model is in how the neighborhoods are constructed. We will characterize the sparsity induced by three "standard" schemes for constructing neighborhoods [27, 33]: the Random, Adjacent and Block Neighborhood schemes. These schemes all restrict every neighborhood to be the same size, K, which provides a basis for comparing how different interaction structures induce sparsity in fitness functions. Graphical depictions of these three schemes are shown in Fig. 1. We will additionally consider a novel scheme where neighborhoods are constructed based on contacts between residues in an atomistic protein structure, which is described in more detail in a later section.

In the next section, we present theoretical results that enable us to calculate the sparsity of GNK fitness functions given the sequence length, alphabet size, and a set of neighborhoods. The proofs of these results are given in the SI.

Results

The sparsity of GNK fitness functions

A somewhat remarkable feature of the GNK model is that it can be shown that the Fourier coefficients of GNK fitness functions are independent normal random variables whose mean and variance can be calculated exactly given the sequence length, alphabet size, and neighborhoods. In particular, the Fourier coefficients of fitness functions sampled from the GNK model are distributed according to $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\lambda}\mathbf{I})$, where $\boldsymbol{\lambda}$ is a vector of variances corresponding to each element of $\boldsymbol{\beta}$ and $\mathbf{I}$ is the $q^L \times q^L$ identity matrix. Further, each of the $(q-1)^r$ Fourier coefficients corresponding to an rth order epistatic interaction, U, have equal variances given by

$$\lambda_U = \frac{1}{L} \sum_{j=1}^{L} q^{L-K_j} I(U \subseteq V^{[j]}), \tag{5}$$

where, with a slight abuse of notation, $I(U \subseteq V^{[j]})$ is an indicator function that is equal to one if U is a subset of the neighborhood $V^{[j]}$, and zero otherwise. Eq. 5 shows that the variance of a Fourier coefficient is roughly proportional to the number of neighborhoods that contain the corresponding epistatic interaction as a subset. Most importantly for our purposes, Eq. 5 implies that a Fourier coefficient only has nonzero

variance when the corresponding epistatic interaction is a subset of at least one neighborhood; otherwise the coefficients are deterministically zero. Consequently, we can use Eq. 5 to calculate the total number of Fourier coefficients that are not deterministically zero in a specified GNK model, which is equal to the sparsity of all fitness functions sampled from the model. In particular, the sparsity, $S(f)$, of a fitness functions f sampled from a GNK model is given by

$$S(f) = \sum_{U \in \mathcal{T}} (q-1)^{|U|}, \tag{6}$$

where $\mathcal{T} := \bigcup_{j=1}^{L} \mathscr{P}(V^{[j]})$ is the union of the powersets of the neighborhoods. Eq. 6 makes the connection between neighborhoods and epistatic interactions concrete: the GNK model assigns nonzero Fourier coefficients to any epistatic interactions whose positions are included in at least one of the neighborhoods. For example, if positions 3 and 4 in a sequence are both in some neighborhood $V^{[j]}$, then all elements of $\boldsymbol{\beta}_{\{3,4\}}$ are nonzero. Further, by the same reasoning, the coefficients corresponding to all subsets of positions $\{3, 4\}$ are also nonzero (i. e., the coefficients corresponding to the first order effects associated with positions 3 and 4).

Eq. 6 provides a general formula for the sparsity of GNK fitness functions as a function of L, q and the neighborhoods. We can use this formula to calculate the sparsity of GNK fitness functions with each of the standard neighborhood schemes—Random, Adjacent and Block—for a given neighborhood size, K. In the Materials and Methods, we provide exact results for the sparsity of GNK fitness functions with Adjacent and Block neighborhoods, and the expected sparsity of GNK fitness functions with for Random neighborhoods. We also provide an upper bound on sparsity of GNK fitness functions with any neighborhood scheme with constant neighborhood size, K. In Figs. 2A and 2B we plot this upper bound for a variety of settings of L, q and K. Further, in Fig. 2C we plot the upper bound along with the exact or expected sparsity of GNK fitness functions with each of the standard neighborhood schemes. We can see that even at the same setting of K, different neighborhood schemes result in striking differences in the sparsity of sampled fitness functions.

Exact recovery of GNK fitness functions

The sparsity result of Eq. 6 allows us to apply CS theory to determine the number of fitness measurements required to recover GNK fitness functions exactly. Specifically, we can use Eq. 2 to determine a minimal N such that exact recovery is guaranteed for an $S(f)$-sparse fitness function f when there is no measurement noise. However, to do so, we first need to determine an appropriate value for the constant C in Eq. 2, which we did via straightforward numerical experiments. In particular, we used LASSO to estimate GNK fitness functions using varying numbers of randomly sampled, noiseless fitness measurements and from these estimates, determined the minimum number of training samples required to exactly recover the fitness functions (allowing for a small amount of numerical error–see Materials and Methods for more details). We then determined the minimum value of C such that Eq. 2 holds in each tested case. Fig. 2D summarizes the experiments, showing that $C = 2.62$ is sufficiently large to ensure recovery of all of the over 900 tested fitness functions, and we use this value for all further calculations.

We next used this estimate of C, along with our results for the sparsity of GNK fitness functions, and the CS result of Eq. 2, to determine the minimum number of measurements required to exactly recover GNK fitness functions. Figs. 3A and 3B show examples of these calculations, where we used the bound on sparsity for GNK fitness functions with constant neighborhood sizes to calculate an upper bound on the minimum number of samples required to recover these fitness functions. A number of important insights can be derived from Fig. 3. First, the number of measurements required to perfectly estimate these fitness functions is many orders of magnitude smaller than the total size of sequence space. Consider, for instance the point in Fig. 3a where $L = 50$ and thus the size of sequence space is $2^{50} \approx 10^{15}$, 10 orders of magnitude greater than the largest plotted sample complexity. Additionally, comparing Figs. 3A and 3B clearly indicates that increasing the alphabet size within biologically relevant ranges increases the number of samples required to recover fitness functions at a faster rate than increasing the length of the sequence.

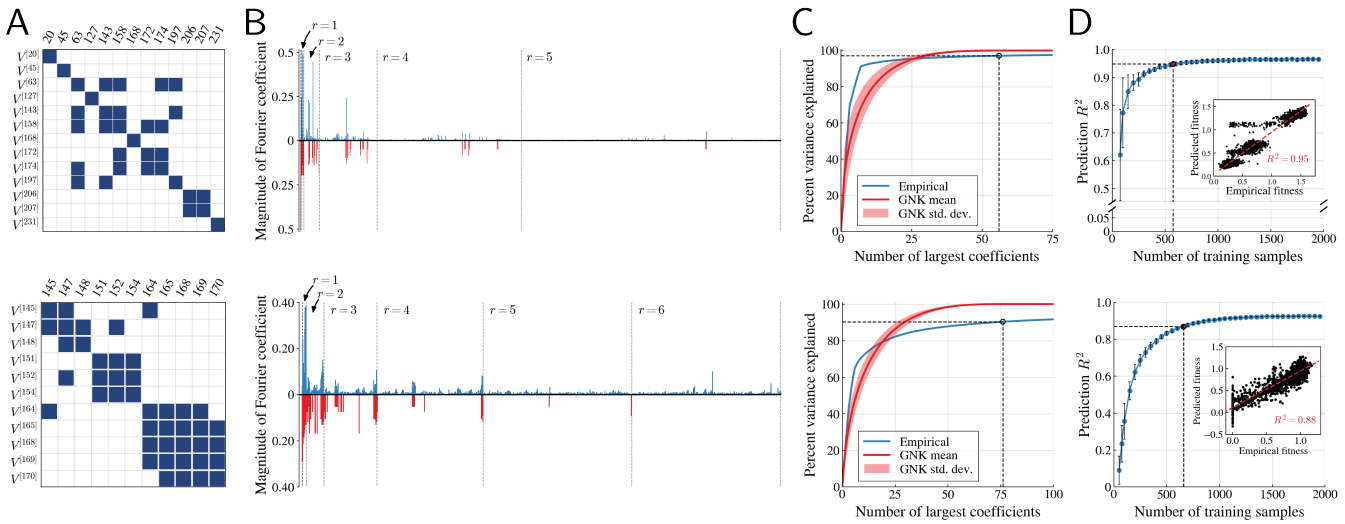

Fig. 4: Comparison of empirical fitness functions with the GNK fitness with Structural Neighborhoods. *First row*: comparison to mTagBFP2 fitness function from ref. 18. *Second row*: comparison to His3p fitness function from ref. 38. (A) Structural Neighborhoods derived from crystal structural of mTagBFP (first row) and I-TASSER predicted structure of His3p (second row). (B) Magnitude of empirical Fourier coefficients (upper plot, in blue) compared to expected magnitudes of coefficients in the GNK model (reverse plot, in red). Dashed lines separate orders of epistatic interactions, with each group of rth order interactions indicated. (C) Percent variance explained by the largest Fourier coefficients in the empirical fitness functions and in fitness functions sampled from the GNK model. The dotted line indicates the exact sparsity of the GNK fitness functions, which is 56 is in the first row and 76 in the second, at which points 97.1% and 90.4% of the empirical variances are explained, respectively. (D) Error of LASSO estimates of empirical fitness functions at a range of training set sizes. Each point on the horizontal axis represents the number of training samples, N, that are used to fit the LASSO estimate of the fitness function. Each point on the blue curve represents the R^2 between the estimated and empirical fitness functions, averaged over 50 randomly sampled training sets of size N. The point at the number of samples required to exactly recover the GNK model with Structural Neighborhoods ($N = 575$ in the first row, and $N = 660$ in the second) is highlighted with a red dot and dashed lines; at this number of samples, the mean prediction R^2 is 0.948 in the first row and 0.870 in the second. Error bars indicate the standard deviation of R^2 over training replicates. Insets show paired plots between the estimated and predicted fitness function for one example training set of size $N = 575$ (first row) and $N = 660$ (second row).

Application to protein fitness functions

We next tested the extent to which our results could be used to predict the sample complexity of estimating fitness functions of proteins. To do so, we made use of a novel scheme for constructing GNK neighborhoods, which we call the Structural neighborhood scheme, that uses information derived from 3D structure of a given protein. In particular, Structural neighborhoods are constructed based on contacts between amino acid residues in a given atomistic protein structure where, following refs. 39 and 4, we define two residues to be in contact if any two atoms in the residues are within 4.5Å of each other. Then the Structural neighborhood of a position j contains all positions that are in structural contact with it.

An interesting aspect of the Structural neighborhood scheme is how it encodes epistatic interactions through Eq. 6. In particular, in a GNK model with Structural neighborhoods, higher-order epistatic interactions arise from only pairwise structural contact information—that is, an rth order epistatic interaction has nonzero Fourier coefficients when $r-1$ positions are in structural contact with a central position; these $r-1$ positions do not necessarily have to be in structural contact with one another for the interaction to arise. We examine this hypothesis for the emergence of higher-order epistasis in protein fitness functions in the results below.

We instantiated GNK models with Structural neighbourhoods for two proteins: the mTagBFP fluorescent protein (PDB: 3M24 [40]) and the protein encoded by the His3 gene in yeast (His3p). We then used the results described in the previous section to calculate the sparsity of GNK fitness functions with these Structural neighborhoods, the sample complexity of estimating these functions, and the variance of each

the functions' Fourier coefficients.

Both mTagBFP and His3p are associated with empirical fitness functions with complete or nearly complete sets of experimental measurements. We calculated the Fourier coefficients associated with each of these empirical fitness functions using Ordinary Least Squares (or regression with a small amount of regularization when the measurements were only nearly complete), so as to be able to compare the resulting sparsity and magnitude of the empirical Fourier coefficients to those of the corresponding GNK fitness functions with Structural neighborhoods. Next, to assess whether the sample complexity of estimating GNK fitness functions with Structural neighborhoods can be used to inform the sample complexity of estimating real protein fitness functions, we fit LASSO estimates of the empirical fitness functions with varying numbers of randomly sampled empirical measurements, and determined how well each recovered the empirical fitness function.

In the case of the mTagBFP fluorescent protein, the associated empirical fitness function contains functional observations (blue fluorescence brightness) of mutations to the mTagBFP2 protein [18], for which no structure is available. This data contains measurements for all combinations of mutations in 13 positions, where each position is allowed to mutate to only one other amino acid (i. e., $L = 13$ and $q = 2$), yielding $2^{13} = 8192$ total fitness observations. A graphical depiction of the Structural neighborhoods associated with these 13 positions is shown in the first row of Fig. 4A. Using Eq. 6 for the GNK model with these Structural neighborhoods yielded a sparsity of $S(f) = 56$, while application of Eq. 5 enabled us to determine the distribution of these 56 non-zero Fourier coefficients and the epistatic interactions that they corresponded to.

For His3p, we used a nearly combinatorial complete empirical fitness function that is embedded in the data of ref. 38. In particular, the data contains 2030 out of the possible 2048 fitness observations for sequences corresponding to 11 positions in His3p, each taking on one of two amino acids (i. e., $L = 11$ and $q = 2$). We constructed Structural neighborhoods based on the I-TASSER [41] predicted structure of His3p [38] (Fig. 4A, second row), which resulted in sparsity $S(f) = 76$ for GNK fitness functions with these neighborhoods, and we again computed the distribution of these coefficients and determind the corresponding epistatic interactions.

The comparisons of the mTagBFP and His3p empirical fitness functions with the associated GNK models with Structural neighborhoods are summarized in Fig. 4. First, we examined the magnitudes of the Fourier coefficients of the empirical and GNK fitness functions. Since the Fourier coefficients in the GNK model are independent normal random variables, the expected magnitude of a coefficient with variance λ is $\sqrt{2\lambda/\pi}$. A comparison of all coefficients corresponding to up to 5th and 6th order epistatic interactions are shown in Fig. 4B for the mTagBFP and His3p cases, respectively. Many of the epsistatic interactions with the largest empirical coefficients also have nonzero coefficients in the GNK model with Structural neighborhoods, suggesting that these models are reasonable approximations to protein fitness functions.

The comparisons shown in Fig. 4B also support the hypothesis implicit in the GNK model with Structural neighborhoods that higher-order epistasis can be inferred from pairwise structural contacts. For example, in the mTagBFP empirical fitness function, the largest 5th order coefficient corresponds to an interaction between positions $\{63, 143, 158, 174, 197\}$, which is exactly the Structural neighborhood of position 63. Further, this neighborhood contain positions 143 and 174, which are not in contact with one another; this demonstrates that higher-order epistasis can arise between positions even when not all positions are in structural contact.

Although none of the empirical Fourier coefficients are exactly zero, these coefficients display substantial approximate sparsity. In particular, over 95% and 80% of the total variance in the coefficients can be explained by the 25 coefficients with the largest magnitude in the mTagBFP and His3p fitness functions, respectively. To more holistically assess whether GNK fitness functions with Structural neighborhoods approximate the sparsity of the empirical fitness functions well, we compared the percent variance explained by the S Fourier coefficients with the largest magnitudes in both the empirical and GNK fitness functions, for a range of settings of S. Fig. 4C shows the results of this comparison, with the blue curve showing

the percent variance explained by the largest empirical coefficients, and the red curve and red shaded region showing the mean and standard deviation, respectively, of the percent variance explained by the largest coefficients in 1,000 sampled GNK fitness functions. Considering that these plots show only the first few of all possible coefficients that could be included on the horizontal axis (75 out of the 8,192 for mTagBFP and 100 out of 2,048 for His3p), it is clear that the GNK model approximates the sparsity of the empirical fitness function qualitatively well. Of particular importance is the point at which all of the nonzero coefficients of the GNK fitness functions are included in the calculation (i. e., 100% of the variance is explained), which occurs at $S = 56$ and 76 in the mTagBFP and His3p cases, respectively; at this point, more than 90% of the empirical variance is explained in both cases.

These promising sparsity comparisons suggest that the sample complexity of estimating GNK fitness functions with Structural neighborhoods may be used to approximate the number of measurements required to effectively estimate protein fitness functions. We confirmed this by using LASSO to estimate the empirical fitness functions with varying number of training points and regularization parameter chosen by cross-validation (Fig. 4D). Our theory predicts that 548 and 630 samples are minimally needed for exact recovery of the GNK fitness functions with mTagBFP and His3p Structural neighborhoods, respectively. In both cases, we see these sample sizes produce effective estimates of the corresponding empirical fitness functions, with a mean R^2 of 0.95 and 0.87 for estimates of the mTagBFP and His3p fitness functions, respectively.

In the SI we show analogous results to those in Fig. 4 for another nearly complete subset of the His3p fitness data of ref. 38 that contains 48,219 out of 55,296 fitness measurements for the same 11 positions discussed above and alphabets that differ in size at each position. Altogether, these results suggest the GNK model with Structural neighborhoods can be used to approximate the sparsity of protein fitness functions, and the sample complexity of estimating such functions.

Discussion

By leveraging perspectives from the fields of Compressed Sensing and evolutionary biology, we developed a theory for calculating the sparsity of fitness functions and the number of fitness measurements required to exactly recover those functions with the LASSO algorithm (or another sparse recovery algorithm with CS guarantees) under a well-defined set of assumptions. These assumptions are that (i) the fitness functions are sampled from a specified GNK model, (ii) fitness measurements are noiseless, (iii) fitness measurements correspond to sequences sampled uniformly at random from the space of sequences, and (iv) the fitness functions are represented in the Fourier basis. Under these assumptions, our results allowed us to test the effect of sequence length, alphabet size, and positional interaction structure on the sparsity and sample complexity of fitness function estimation.

We have additionally demonstrated that in certain cases our results can be used to estimate the sample complexity of estimating protein fitness functions when assumptions (i) and (ii) may not be exactly satisfied. In particular, we showed that GNK models with Structural neighborhoods accurately approximate the sparsity of two empirical protein fitness functions, and can be used to estimate the number of measurements required to recover those empirical fitness functions with high accuracy. The success of applying our theory to these empirical fitness functions, which are neither exactly sparse nor noiseless, is at least partially due to the fact that sparse recovery algorithms such as LASSO are robust to approximate sparsity and noisy measurements (Materials and Methods, Eq. 8). It additionally suggests that GNK models with Structural neighborhoods can reliably simulate protein fitness functions, which may be a useful insight for a variety of applications that make use of simulated fitness landscapes.

It should be noted that assumptions (iii) and (iv) likely result in conservative estimates for the sample complexity of fitness function estimation. Uniform sampling of sequences is optimal when one has no *a priori* knowledge about the fitness function; however, if one knows which coefficients in a fitness function are likely to be nonzero, then it may be possible to construct alternative sampling schemes, or deterministic sets of sequences to measure, such that the fitness function can be recovered with many fewer measurements

than with uniform sampling. Additionally, it may be possible to construct a basis in which certain classes of fitness functions are more sparse than in the Fourier basis, and this will in turn result in fewer measurements being required to recover those fitness functions when they are represented in the alternative basis.

Our sample complexity predictions could be used to a certain extent to guide experimental probes of fitness by suggesting how one should restrict the scope of mutagenesis such that recovery of the resulting fitness function can be expected with a certain amount of data. Using protein mutagenesis experiments as an example, this could be done by limiting the number of positions that are mutated, perhaps based on biophysical considerations [42, 3] or previous experimental results [43, 44, 45, 46], or by allowing each position to mutate to only a restricted alphabets of amino acids, for instance by choosing only amino acids that are present in homologous sequences [18, 38, 17]. Of course, one should take care not to minimize the sample size requirements at the expense of probing important areas of the protein or nucleotides sequences under study.

Complementing our main contributions, we have also demonstrated that GNK models with Structural neighborhoods can predict the identity of many of the largest higher-order epistatic interactions in empirical fitness functions (Fig. 4B). This result bolsters a growing understanding in the field that structural contacts are a major factor in protein fitness. In particular, there is a growing body of evidence that evolutionary sequence data can be used to predict both fitness [47, 48, 49] and structural contacts [50, 51, 52], and that structural contacts can be recovered from the data of Deep Mutational Scanning experiments [53, 54], which select sequences based on their fitness. The success of these contact prediction methods suggests that protein fitness functions are intimately related to structural contacts. Our results provide further evidence for this conjecture by demonstrating that many higher-order epistatic interactions in protein fitness functions can be inferred from structural contacts.

Materials and Methods

Compressed Sensing

As described in the main text, the fitness function estimation problem is to solve the underdetermined linear system $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ for an unknown $\boldsymbol{\beta}$, where $\mathbf{y}$ are N fitness measurements, and $\mathbf{X}$ are the N corresponding rows of $\boldsymbol{\Phi}$ that represent the sequences with fitness measurements. In this section, we assume that each element of $\mathbf{y}$ is corrupted with independent Gaussian noise with variance σ^2. LASSO solves for an estimate of the Fourier coefficients by solving following convex optimization program:

$$\min_{\hat{\boldsymbol{\beta}}} \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2 + \nu\|\hat{\boldsymbol{\beta}}\|_1 \tag{7}$$

where ν is a hyperparameter that determines the strength of regularization. Candes and Plan [30] proved that when the rows of an orthogonal basis such as $\boldsymbol{\Phi}$ are sampled uniformly at random, and the number of samples satisfies Eq. 2, then the solution to the program in Eq. 7, denoted $\boldsymbol{\beta}^*$, satisfies with high probability

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq C_1 \frac{\|\boldsymbol{\beta} - \boldsymbol{\beta}_S\|_1}{\sqrt{S}} + C_2\sigma \tag{8}$$

where C_1, and C_2 are constants and $\boldsymbol{\beta}_S$ is the best S-sparse approximation to $\boldsymbol{\beta}$, i. e., the vector that contains the S elements of $\boldsymbol{\beta}$ with the largest magnitude and sets all others elements to zero. Eq. 8 has a number of important implications. First, it tells us that if $\boldsymbol{\beta}$ is itself S-sparse, then, in a noiseless setting, it can be recovered *exactly* with $\mathcal{O}(S \log q^L)$ measurements. Otherwise, if $\boldsymbol{\beta}$ is not exactly sparse but is well approximated by a sparse vector, then it can be approximately recovered with error on the order of $\frac{1}{\sqrt{S}}\|\boldsymbol{\beta} - \boldsymbol{\beta}_S\|_1$, which is proportional to the sum of the magnitudes of the $q^L - S$ elements of $\boldsymbol{\beta}$ with the smallest magnitudes.

We primarily focus on cases where a fitness function is exactly sparse in the Fourier basis and we can calculate the sparsity. Although natural fitness functions are unlikely to be exactly sparse, they may be

well approximated by sparse vectors, and Eq. 8 tells us that the error of the estimator will be well controlled in this case. Similarly, measurement noise in experimental fitness data is unavoidable, but Eq. 8 shows that the error induced by this noise is dependent on the variance of the measurement noise, and not on the properties of the fitness function itself. Since here we are primarily concerned with understanding how assumed properties of fitness functions affect the sample complexity of estimating those functions, it is thus most appropriate to consider the noiseless setting and leave the estimation of error due to measurement noise to future work.

Fourier bases

Our generalization of the WH basis to larger alphabets is based on the theory of Graph Fourier bases. The Graph Fourier basis corresponding to a given graph is a complete set of orthogonal eigenvectors of the Graph Laplacian of the graph. Graph Fourier bases have many useful properties and have been used extensively for processing signals defined on graphs [55].

The WH basis is specifically the Graph Fourier basis corresponding to the Hamming graph $H(L, 2)$ [56]. The vertices of $H(L, 2)$ represent all unique binary sequences of length L; two sequences are adjacent in $H(L, 2)$ if they differ in exactly one position (i. e., the Hamming distance between the two sequences is equal to one). The Hamming graphs $H(L, q)$ are defined in the same way for sequences with alphabet size q. Thus, we can construct an analogous Graph Fourier basis to the WH basis to represent sequences with larger alphabets by calculating the eigenvectors of the Graph Laplacian of $H(L, q)$. Since we only consider functions defined on Hamming graphs, we refer to Graph Fourier bases corresponding to Hamming graphs simply as Fourier bases.

An important property of the Hamming graph $H(L, q)$ is that it can be constructed as the L-fold Graph Cartesian product of the "complete graph" of size q [56]. The complete graph of size q, denoted $K(q)$, has q vertices (which represent elements of the alphabet in our case) and edges between all pairs of vertices. Due to the spectral properties of graph products, the eigenvectors of the Hamming graph (i. e., the Fourier basis) can be calculated as a function of the eigenvectors of the complete graph. An orthonormal set of eigenvectors of the Graph Laplacian of the complete graph $K(q)$ are given by the columns of the following Householder matrix:

$$\mathbf{P}_q \coloneqq \mathbf{I}_q - \frac{2\mathbf{w}\mathbf{w}^T}{\|\mathbf{w}\|_2^2}, \tag{9}$$

where $\mathbf{w} \coloneqq \mathbf{1}_q - \sqrt{q}\mathbf{e}_1$, $\mathbf{1}_q$ is the vector of length q whose elements are all equal to one, $\mathbf{e}_1$ is the length q with the first element set to 1 and all others set to zero, and $\mathbf{I}_q$ is the $q \times q$ identity matrix.

The complete graph is equal to the Hamming graph $H(1, q)$, and thus Equation Eq. 9 constructs the Fourier basis for sequences of length one and alphabet size q. Each row of $\mathbf{P}_q$ corresponds to a sequence of length one; the first columns is constant for all rows while the remaining $q - 1$ columns encode the alphabet elements (i. e., the final $q - 1$ elements of a row uniquely identify the alphabet element that the row corresponds to). More specifically, let $\tilde{\mathbf{P}}_q$ be the matrix containing the final $q - 1$ unnormalized columns of $\mathbf{P}_q$, such that $\mathbf{P}_q = \frac{1}{\sqrt{q}}\left[\mathbf{1} \mid \tilde{\mathbf{P}}_q\right]$, where $\mid$ denotes column-wise concatenation. Then the ith row of $\tilde{\mathbf{P}}_q$ encodes the ith element of the alphabet; we denote each of these encodings as $\mathbf{p}_q(s)$, where s is an element of the alphabet (i. e., each $\mathbf{p}_q(s)$ is a row of $\tilde{\mathbf{P}}_q$).

Then, it can be shown that Fourier basis corresponding to the Hamming graph $H(L, q)$, which can be used to represent fitness functions of sequences of length L and alphabet size q, is given by the L-fold Kronecker product of the eigenvectors of the complete graph. More concretely, an orthonormal set of eigenvectors of the Graph Laplacian of the Hamming graph $H(L, q)$ are given by the columns of following the $q^L \times q^L$ matrix:

$$\mathbf{\Phi} = \bigotimes_{i=1}^{L} \mathbf{P}_q, \tag{10}$$

where $\mathbf{P}_q$ is defined in Eq. 9. In the basis defined in Eq. 10, an epistatic between positions in the set U

is encoded by the length $(q-1)^{|U|}$ vector $\boldsymbol{\phi}_U(\mathbf{s}) \coloneqq \frac{1}{\sqrt{q^L}} \bigotimes_{i \in U} \mathbf{p}_q(s_i)$. These encodings are used in the representation of fitness functions in the Fourier basis shown in Eq. 4. The results of Eq. 9 and Eq. 10 are proved in the SI.

GNK Model

Given sequence length, L, alphabet size, q, and set of neighborhoods $\mathcal{V} \coloneqq \{V^{[j]}\}_{j=1}^{L}$, a fitness function sampled from the GNK model assigns a fitness to every sequence $\mathbf{s} \in \mathcal{S}^{(L,q)}$ with the following two steps:

1. Let $\mathbf{s}^{[j]} \coloneqq (s_k)_{k \in V^{[j]}}$ be the subsequence of $\mathbf{s}$ corresponding to the indices in the neighborhood $V^{[j]}$. Assign a 'subsequence fitness', $f_j(\mathbf{s}^{[j]})$ to every possible subsequence, $\mathbf{s}^{[j]}$, by drawing a value from the normal distribution with mean equal to zero and variance equal to $1/L$. In other words, $f_j(\mathbf{s}^{[j]}) \sim \mathcal{N}(0, 1/L)$ for every $\mathbf{s}^{[j]} \in \mathcal{S}^{(K_j,q)}$, and for every $j = 1, 2, ..., L$.

2. For every $\mathbf{s} \in \mathcal{S}^{(L,q)}$, the subsequence fitness values are summed to produce the total fitness values $f(\mathbf{s}) = \sum_{j=1}^{L} f_j(\mathbf{s}^{[j]})$.

This definition of the GNK model is slightly more restrictive than that presented in ref. 26. In particular, in ref. 26 the authors allow subsequence fitness values to be sampled from any appropriate distribution. For simplicity, we consider only the case where subsequence fitness values are sampled from the scaled unit normal distribution.

Standard neighborhood schemes

We consider three standard neighborhood schemes: the Random, Adjacent and Block neighborhood schemes. In all of these, each neighborhood is of the same size, K (i. e., $K_j = K$ for all $j = 1, 2, ..., L$). In the Random scheme, each neighborhood $V^{[j]}$ contains j and $K-1$ other position indices selected uniformly at random from $\{1, 2, ..., L\} \backslash j$. In the Adjacent scheme when K is an odd number, each neighborhood $V^{[j]}$ contains the $\frac{K-1}{2}$ positions immediately clockwise and counterclockwise to j when the positions are arranged in a circle. When K is an even number, the neighborhood includes the $\frac{K-2}{2}$ counterclockwise positions and the $\frac{K}{2}$ clockwise positions. The Block scheme (also known as the Block Model [57, 58]), splits positions into $\frac{L}{K}$ blocks of size K and lets each block be "fully connected" in the sense that every neighborhood of a position in the block contains all other positions in the block, but no positions outside of the block. In order for Block neighborhoods to be defined, L must be a multiple of K.

Standard neighborhood sparsity calculations

The sparsity of GNK fitness functions with the standard neighborhood schemes can be calculated exactly as functions of L, q, and K. The following results are used in the main text and are all proved in the SI. First, the sparsity of any GNK fitness with uniform neighborhood sizes is bounded above by

$$S(f) \leq 1 + L(q-1) + L \sum_{r=2}^{K} \binom{K}{r} (q-1)^r. \tag{11}$$

All curves in Figs. 2A and 2B are calculated with this bound, and it is also used for the sample complexity calculations shown in Fig. 3. It is also plotted as the dashed blue curve in Fig. 2C with $L = 20$ and $q = 2$. Additionally, the sparsity of GNK fitness functions with Block neighborhoods can be calculated exactly and is given by

$$S(f) = \frac{L}{K}(q^K - 1) + 1. \tag{12}$$

Eq. 12 is plotted as the red curve in Fig. 2C. Similarly, the sparsity of GNK fitness functions with Adjacent neighborhoods is given by

$$S(f) = L \sum_{r=0}^{K} \binom{K-1}{r-1} (q-1)^r, \tag{13}$$

which is plotted as the green curve in Fig. 2C with $L = 20$ and $q = 2$. Finally, the expected sparsity of GNK fitness functions with Random neighborhoods, with the expectation taken over the randomly assigned neighborhoods, is given by

$$\mathbb{E}[S(f)] = \sum_{r=0}^{K} \binom{L}{r} p(r)(q-1)^r, \tag{14}$$

where

$$p(r) = 1 - (1-\alpha(r))^r \left(1 - \alpha(r)\frac{K-r}{L-r}\right)^{L-r}, \tag{15}$$

and $\alpha(r) = \frac{(K-1)!}{(L-1)!}\frac{(K-r)!}{(L-r)!}$. Eq. 14 is shown as the solid blue curve in Fig. 2C with $L = 20$ and $q = 2$.

Numerical calculation of C

In order to determine an appropriate value of C, we (i) sampled a fitness function from a GNK model, (ii) subsampled N sequence-fitness pairs uniformly at random from the complete fitness function for a range of settings of N, (iii) ran LASSO on each of the subsampled data sets and (iv) determined the smallest N such that the fitness function is exactly recovered by LASSO. Letting $\hat{N}$ be the minimum N for which exact recovery occurs, then

$$\hat{C} = \frac{\hat{N}}{S(f)\log_{10}(q^L)} \tag{16}$$

is the minimum value of C that satisfies Eq. 2, where $S(f)$ is calculated calculated with Eq. 6. We ran multiple replicates of this experiment for neighborhoods sampled according to the RN scheme, for different settings of L, q and K. This resulted in a test for 907 total fitness functions. For each of these fitness functions, we ran LASSO with 5 randomly sampled training sets for each size N, and a regularization parameter, ν, determined by cross-validation. We deemed the fitness function exactly recovered when the estimates resulting from all 5 training sets explained 99.99% of the variance in the fitness function's coefficients.

Equipped with an estimate of C, we can calculate the minimum number of samples required to exactly recover a GNK fitness function by using Eq. 2 along with the sparsity calculations discussed in the previous section. Specifically,

$$N = \lceil C_0 \cdot S(f)\log_{10}(q^L)\rceil \tag{17}$$

is the minimum number of samples that guarantees exact recovery, where $\lceil\rceil$ represents the ceiling operator. Eq. 17 was used along with the bound in Eq. 11 to calculate the curves in Fig. 3.

Percent variance explained

In Fig. 4C, we computed the percent of total variance in the Fourier coefficients explained by the S coefficients with the largest magnitudes, for a range of settings of S. The percent variance explained by the S largest elements of the vector of coefficients $\boldsymbol{\beta}$ is calculated as

$$\text{Percent variance explained}(S) \coloneqq 100\% \cdot \left(1 - \frac{\|\boldsymbol{\beta}_S - \boldsymbol{\beta}\|_2^2}{\|\boldsymbol{\beta}\|_2^2}\right). \tag{18}$$

References

[1] Jakub Otwinowski and Joshua B. Plotkin. Inferring fitness landscapes by regression produces biased estimates of epistasis. *Proc. Natl. Acad. Sci. U. S. A.* **111** (2014).

[2] Jakub Otwinowski. Biophysical inference of epistasis and the effects of mutations on protein stability and function. *Mol. Biol. Evol.* **35**, 2345–2354 (2018).

[3] Aditya Ballal, Caroline Laurendon, Melissa Salmon, Maria Vardakou, Jitender Cheema, Marianne Defernez, Paul E O'Maille, and Alexandre V Morozov. Sparse Epistatic Patterns in the Evolution of Terpene Synthases. *Mol. Biol. Evol.* **37**, 1907–1924 (2020).

[4] Philip A. Romero, Andreas Krause, and Frances H. Arnold. Navigating the protein fitness landscape with Gaussian processes. *Proc. Natl. Acad. Sci. U.S.A* **110**, E193–E201 (2013).

[5] Juannan Zhou and David M. McCandlish. Minimum epistasis interpolation for sequence-function relationships. *Nat. Commun.* **11** (2020).

[6] Kevin K. Yang, Zachary Wu, and Frances H. Arnold. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).

[7] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song, Evaluating protein transfer learning with tape in *Advances in Neural Information Processing Systems*, eds. H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. (Curran Associates, Inc.), Vol. 32, pp. 9689–9701 (2019).

[8] Surojit Biswas, Grigory Khimulya, Ethan C. Alley, Kevin M. Esvelt, and George M. Church. Low-N protein engineering with data-efficient deep learning. *Nat. Methods* **18**, 389–396 (2021).

[9] Richard J. Fox, S. Christopher Davis, Emily C. Mundorff, Lisa M. Newman, Vesna Gavrilovic, Steven K. Ma, Loleta M. Chung, Charlene Ching, Sarena Tam, Sheela Muley, John Grate, John Gruber, John C. Whitman, Roger A. Sheldon, and Gjalt W. Huisman. Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol.* **25**, 338–344 (2007).

[10] Claire N. Bedbrook, Kevin K. Yang, Austin J. Rice, Viviana Gradinaru, and Frances H. Arnold. Machine learning to design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization. *PLoS Comput. Biol.* **13** (2017).

[11] Zachary Wu, S. B. Jennifer Kan, Russell D. Lewis, Bruce J. Wittmann, and Frances H. Arnold. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 8852–8858 (2019).

[12] Anvita Gupta and James Zou. Feedback GAN for DNA optimizes protein functions. *Nat. Mach. Intell.* **1**, 105–111 (2019).

[13] David H. Brookes, Hahnbeom Park, and Jennifer Listgarten, Conditioning by adaptive sampling for robust design in *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, eds. Kamalika Chaudhuri and Ruslan Salakhutdinov. (PMLR, Long Beach, California, USA), Vol. 97, pp. 773–782 (2019).

[14] Christof Angermüller, David Dohan, David Belanger, Ramya Deshpande, Kevin Murphy, and Lucy Colwell, Model-based reinforcement learning for biological sequence design in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. (OpenReview.net), (2020).

[15] Clara Fannjiang and Jennifer Listgarten. Autofocused oracles for model-based design. *Adv. Neural Inf. Process. Syst.* **33** (2020).

[16] Zachary R. Sailer and Michael J. Harms. Detecting High-Order Epistasis in Nonlinear Genotype-Phenotype Maps. *Genetics* **205**, 1079–1088 (2017).

[17] Gloria Yang, Dave W. Anderson, Florian Baier, Elias Dohmen, Nansook Hong, Paul D. Carr, Shina Caroline Lynn Kamerlin, Colin J. Jackson, Erich Bornberg-Bauer, and Nobuhiko Tokuriki. Higher-order epistasis shapes the fitness landscape of a xenobiotic-degrading enzyme. *Nat. Chem. Biol.* **15**, 1120–1128 (2019).

[18] Frank J. Poelwijk, Michael Socolich, and Rama Ranganathan. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nat. Commun.* **10** (2019).

[19] Amirali Aghazadeh, Hunter Nisonoff, Orhan Ocal, Yijie Huang, O. Ozan Koyluoglu, Jennifer Listgarten, and Kannan Ramchandran. Sparse epistatic regularization of deep neural networks for inferring fitness functions. *bioRxiv* (2020).

[20] Amirali Aghazadeh, Orhan Ocal, and Kannan Ramchandran. CRISPRLand: Interpretable large-scale inference of DNA repair landscape based on a spectral approach. *Bioinformatics* **36**, i560–i568 (2020).

[21] Emmanuel J. Candes, Justin Romberg, and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Inf. Theory* **52**, 489–509 (2006).

[22] David L. Donoho. Compressed sensing. *IEEE Transactions on Inf. Theory* **52**, 1289–1306 (2006).

[23] Stuart Kauffman and Simon Levin. Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.* **128**, 11 – 45 (1987).

[24] Atish Agarwala and Daniel S. Fisher. Adaptive walks on high-dimensional fitness landscapes and seascapes with distance-dependent statistics. *Theor. Popul. Biol.* **130**, 13–49 (2019).

[25] Stuart A. Kauffman and Edward D. Weinberger. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *J. Theor. Biol.* **141**, 211 – 245 (1989).

[26] Jeffrey Buzas and Jeffrey Dinitz. An Analysis of NK landscapes: Interaction structure, statistical properties, and expected number of local optima. *IEEE Trans. Evol. Comput.* **18**, 807–818 (2014).

[27] Stefan Nowak and Joachim Krug. Analysis of adaptive walks on NK fitness landscapes with different interaction schemes. *J. Stat. Mech. Theory Exp.* **2015**, P06014 (2015).

[28] Karen S. Sarkisyan, Dmitry A. Bolotin, Margarita V. Meer, Dinara R. Usmanova, Alexander S. Mishin, George V. Sharonov, Dmitry N. Ivankov, Nina G. Bozhanova, Mikhail S Baranov, Onuralp Soylemez, Natalya S. Bogatyreva, Peter K. Vlasov, Evgeny S Egorov, Maria D. Logacheva, Alexey S. Kondrashov, Dmitry M. Chudakov, Ekaterina V. Putintseva, Ilgar Z. Mamedov, Dan S. Tawfik, Konstantin A. Lukyanov, and Fyodor A. Kondrashov. Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397 (2016).

[29] Trevor Hastie, Robert Tibshirani, and Martin Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations.* (Chapman & Hall/CRC), (2015).

[30] Emmanuel J. Candes and Yaniv Plan. A Probabilistic and RIPless Theory of Compressed Sensing. *IEEE Trans. Inf. Theory* **57**, 7235–7254 (2011).

[31] Robert B. Heckendorn and Darrell Whitley, A Walsh Analysis of NK-Landscapes in *Proceedings of the Seventh International Conference on Genetic Algorithms.* (Morgan Kaufmann), pp. 41–48 (1997).

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

[33] Sungmin Hwang, Benjamin Schmiegelt, Luca Ferretti, and Joachim Krug. Universality Classes of Interaction Structures for NK Fitness Landscapes. *J. Stat. Phys.* **172**, 226–278 (2018).

[34] Frank J. Poelwijk, Vinod Krishna, and Rama Ranganathan. The Context-Dependence of Mutations: A Linkage of Formalisms. *PLoS Comput. Biol.* **12** (2016).

[35] Daniel M. Weinreich, Yinghong Lan, C. Scott Wylie, and Robert B. Heckendorn. Should evolutionary geneticists worry about higher-order epistasis? *Curr. Opin. Genet. Dev.* **23**, 700–707 (2013).

[36] Gary D. Stormo. Maximally Efficient Modeling of DNA Sequence Motifs at All Levels of Complexity. *Genetics* **187**, 1219–1224 (2011).

[37] Peter F. Stadler, Rudi Seitz, and Günter P. Wagner. Population dependent fourier decomposition of fitness landscapes over recombination spaces: Evolvability of complex characters. *Bull. Math. Biol.* **62**, 399–428 (2000).

[38] Victoria O. Pokusaeva, Dinara R. Usmanova, Ekaterina V. Putintseva, Lorena Espinar, Karen S. Sarkisyan, Alexander S. Mishin, Natalya S. Bogatyreva, Dmitry N. Ivankov, Arseniy V. Akopyan, Sergey Ya. Avvakumov, Inna S. Povolotskaya, Guillaume J. Filion, Lucas B. Carey, and Fyodor A. Kondrashov. An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. *PLoS Genet.* **15**, 1–30 (2019).

[39] Christopher A. Voigt, Carlos Martinez, Zhen Gang Wang, Stephen L. Mayo, and Frances H. Arnold. Protein building blocks preserved by recombination. *Nat. Struct. Biol.* **9**, 553–558 (2002).

[40] Oksana M. Subach, Vladimir N. Malashkevich, Wendy D. Zencheck, Kateryna S. Morozova, Kiryl D. Piatkevich, Steven C. Almo, and Vladislav V. Verkhusha. Structural characterization of acylimine-containing blue and red chromophores in mtagbfp and tagrfp fluorescent proteins. *Chem. & Biol.* **17**, 333–341 (2010).

[41] Jianyi Yang and Yang Zhang. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res.* **43**, W174–W181 (2015).

[42] Dave W. Anderson, Alesia N. McKeown, and Joseph W. Thornton. Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its dna binding sites. *eLife* **4**, e07864 (2015).

[43] Daniel M. Weinreich, Nigel F. Delaney, Mark A. DePristo, and Daniel L. Hartl. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114 (2006).

[44] Claudia Bank, Sebastian Matuszewski, Ryan T. Hietpas, and Jeffrey D. Jensen. On the (un)predictability of a large intragenic fitness landscape. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 14085–14090 (2016).

[45] 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**, e16965 (2016).

[46] 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. *Nat. Biotechnol.* (2021).

[47] 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. *Nat. Biotechnol.* **35**, 128–135 (2017).

[48] Adam J. Riesselman, John B. Ingraham, and Debora S. Marks. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).

[49] 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**, 440–445 (2020).

[50] 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. *Proc. Natl. Acad. Sci. U.S.A* **108**, E1293–E1301 (2011).

[51] Sergey Ovchinnikov, Hetunandan Kamisetty, and David Baker. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *eLife* **3**, e02030 (2014).

[52] Matteo Figliuzzi, Pierre Barrat-Charlaix, and Martin Weigt. How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins? *Mol. Biol. Evol.* **35**, 1018–1027 (2018).

[53] Nathan J. Rollins, Kelly P. Brock, Frank J. Poelwijk, Michael A. Stiffler, Nicholas P. Gauthier, Chris Sander, and Debora S. Marks. Inferring protein 3D structure from deep mutation scans. *Nat. Genet.* **51**, 1170–1176 (2019).

[54] Jörn M. Schmiedel and Ben Lehner. Determining protein structures using deep mutagenesis. *Nat. Genet.* **51**, 1177–1186 (2019).

[55] Benjamin Ricaud, Pierre Borgnat, Nicolas Tremblay, Paulo Gonçalves, and Pierre Vandergheynst. Fourier could be a data scientist: From graph Fourier transform to signal processing on graphs. *Comptes Rendus Phys.* **20**, 474–488 (2019).

[56] Peter F. Stadler, Towards a theory of landscapes in *Complex Systems and Binary Networks*, eds. Ramón López-Peña, Henri Waelbroeck, Riccardo Capovilla, Ricardo García-Pelayo, and Federico Zertuche. (Springer Berlin Heidelberg, Berlin, Heidelberg), pp. 78–163 (1995).

[57] Alan S. Perelson and Catherine A. Macken. Protein evolution on partially correlated landscapes. *Proc. Natl. Acad. Sci. U.S.A* **92**, 9657–9661 (1995).

[58] H. Allen Orr. The population genetics of adaptation on correlated fitness landscapes: The block model. *Evolution* **60**, 1113–1124 (2006).

[59] Yukio Shibata and Yosuke Kikuchi. Graph products based on the distance in graphs. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* **E83A** (2000).

[60] Richard Hammack, Wilfried Imrich, and Sandi Klavzar, *Handbook of Product Graphs, Second Edition.* (CRC Press, Inc., USA), 2nd edition, (2011).

[61] Kauê Cardoso. Principal eigenvector of the signless Laplacian matrix. *Comput. Appl. Math.* **40**, 50 (2021).

S1 Further details on mTagBFP2 fitness function

The data of ref. 18 reports on the fluorescence of every intermediate between the mTagBFP2 blue fluorescent and mKate2 red fluorescent proteins. These two proteins differ in only 13 positions, and the red and blue fluorescence of all 2^{13} possible combinations of the two sequences at these positions was tested. Table S1 shows the alphabet of each position in these empirical fitness functions. For our analysis of this data, we considered only the reported blue fluorescence of each tested sequence. Since there is no available structure for mTagBFP2, we calculated the Structural neighborhoods with the structure of mTagBFP, a blue fluorescent protein that differs in two positions from mTagBFP. No crystal structure is available for mKate2 or a closely related protein (e. g., mKate) that would have allowed to us to perform analogous analysis on the red fluorescent data.

Since the empirical fitness function in the data of ref. 18 is combinatorially complete, we can solve for the Fourier coefficients with Ordinary Least Squares regression. In particular, letting $\mathbf{y}$ be the empirical fitness values and $\mathbf{\Phi}$ the WH basis encoding all binary sequences of length 13, then the OLS estimate of the empirical Fourier coefficients is $\hat{\boldsymbol{\beta}} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{y} = \mathbf{\Phi}^T\mathbf{y}$, where the second equality is due to the orthogonality of $\mathbf{\Phi}$. The magnitudes of $\hat{\boldsymbol{\beta}}$ are plotted in Figure 4B (first row).

Position	20	45	63	127	143	158	168	172	174	197	206	207	231
mTagBFP2	D	V	L	T	F	N	S	A	A	Y	N	N	K
mKate2	N	A	M	P	S	A	G	C	L	R	D	K	R

Table S1: Alphabets at each mutated position in the empirical fitness function of ref. 18. The first row indicates the index of the position in the complete protein sequence, the second row and third rows indicate the amino acids present at these positions in the mTagBFP2 and mKate2 sequences, respectively.

S2 Further details and additional results on His3 fitness function

The complete data of ref. 38 reports the fitness of 875,151 unique amino acid sequences of the protein encoded by the His3 gene in yeast (which we refer to as His3p). In this case, the fitness was defined as the cellular growth rate when the the sequences were expressed in a strain of yeast. Embedded within this data, there exist a number of nearly combinatorially complete fitness functions. By 'embedded' we mean that if one only considers data reporting on a subset of the mutated positions, and a subset of the possible mutations at those positions, then nearly all combinations of the considered mutations at the considered positions have fitness values associated with them. Two such nearly complete fitness functions can be constructed by considering the 11 sequences positions $145, 147, 148, 151, 152, 154, 164, 165, 168, 169,$ and 170. If one considers only the two most frequently occurring amino acids at these positions, then 2,030 out of the $2^{11} = 2,048$ possible combinations of those amino acids at those positions have fitness data associated with them. The two most frequent amino acids at each of these positions are shown in Table S2. We will refer to the fitness function corresponding to these 11 positions and alphabets as the His3p(small) fitness function. The His3p(small) empirical fitness function is analyzed in the main text and compared to GNK fitnesss functions with Structural neighborhoods (Figure 4, second row). These Structural neighborhoods were constructed using the I-TASSER predicted structure of His3p that is analyzed in ref. 38. Determining the empirical Fourier coefficients for the His3p(small)) fitness function requires solving a very slightly underdetermined linear system (2,030 rows and 2,048 columns). In order to do so, we solved for a LASSO estimate using all 2,030 fitness measurements with a small amount of regularization ($\nu = 1 \times 10^{-12}$); the magnitudes of the resulting coefficients are shown in the second row of Figure 4B.

If one further considers larger alphabets at certain positions, then there is another nearly complete empirical fitness function associated with the 11 positions that are considered in the His3p(small) fitness

function. In particular, consider alphabets of size $\mathbf{q} = [2, 2, 3, 2, 2, 3, 3, 4, 2, 4]$ corresponding to each of the 11 positions; the particular alphabets corresponding to each position are shown in Table S3. We refer to the fitness function corresponding to these alphabets as the His3p(big) fitness function. The data of ref. 38 contains fitness measurements for 48,219 out of the $\prod_{i=1}^{11} q_i = 55,296$ sequences in the space corresponding to the alphabets in Table S3. We show how to construct Fourier bases for fitness functions of sequences with hybrid alphabet sizes in Section S6. The Fourier basis for $L = 11$ sequence positions and hybrid alphabet sizes $\mathbf{q}$ has 55,296 associated coefficients. We used LASSO with $\nu = 1 \times 10^{-9}$ to solve for the Fourier coefficients of the His3p(big) fitness functions. The magnitudes of these coefficients corresponding to up to 4th order epistatic interactions are shown in blue in Figure S1A. We additionally show in Section S6 how to extend our results on GNK fitness functions to the case of hybrid alphabet sizes. We used these results to calculate the sparsity of GNK fitness functions with $L = 11$, alphabet sizes $\mathbf{q}$ and Structural Neighborhoods as in Figure 4D in the main text, as well as the the number of measurements required to recover those fitness functions. Figures S1A and S1B and show the results of comparing the Fourier coefficients and sparsity of the His3p(big) empirical fitness function to the GNK fitness functions (these are calculated with the same methodology as Figures 4B and 4C). Figure S1C shows the result of estimating the His3p(big) fitness function with randomly sampled measurements (analogous to Figure 4D in the main text). We again see that GNK fitness functions with Structural Neighborhoods can approximate the sparsity of, higher-order epistatic interactions in, and number of samples required to estimate empirical protein fitness functions. These results also demonstrate that the Fourier bases produce sparse representations of fitness functions with non-binary alphabets.

145	147	148	151	152	154	164	165	168	169	170
L	F	R	K	I	Q	L	D	A	G	G
I	L	Q	M	V	E	M	H	S	E	S

Table S2: Alphabets at each mutated position in the His3p(small) fitness function. The first row indicates the sequence position. and the rows indicate amino acids that make up the alphabet at each position.

145	147	148	151	152	154	164	165	168	169	170
L	F	R	K	I	Q	L	D	A	G	G
I	L	Q	M	V	E	M	H	S	E	S
		K			H	I	E		R	A
					D		Q		Q	T

Table S3: Alphabets at each mutated position in the His3p(big) fitness function. The first row indicates the sequence position. and the rows indicate amino acids that make up the alphabet at each position.

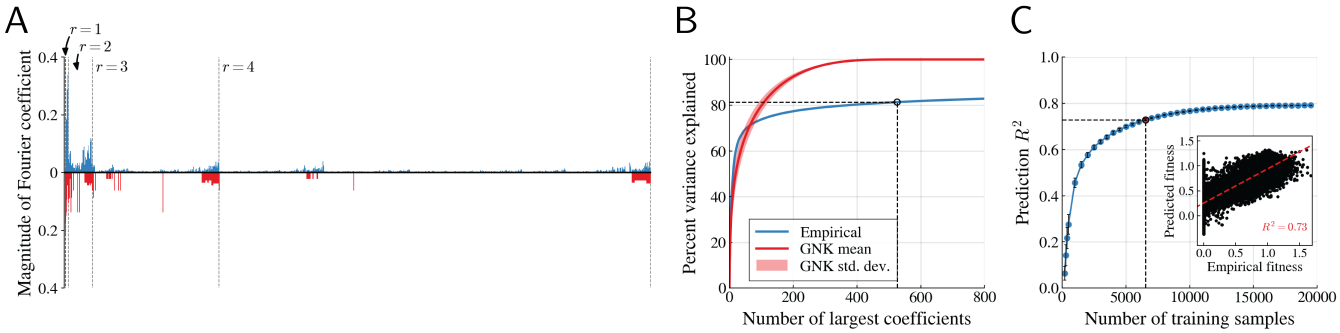

Fig. S1: Comparison of His3p(big) empirical fitness function to the GNK model with Structural Neighborhoods from Figure 4A (second row) in the main text. (A) Magnitude of empirical Fourier coefficients (upper plot, in blue) compared to the standard deviations of coefficients in the GNK model (reverse plot, in red). Dashed lines separate orders of epistatic interactions, with each group of rth order interactions indicated. (B) Percent of variance explained by the largest Fourier coefficients in the empirical fitness function and in fitness functions sampled from the GNK model. The dotted line indicates the exact sparsity of the GNK fitness functions (526) at which point 81.4% of the empirical variances is explained. (C) Error of LASSO estimates of empirical fitness functions at a range of training set sizes. Each point on the horizontal axis represents the number of training samples, N, that are used to fit the LASSO estimate of the fitness function. Each point on the blue curve represents the R^2 between the estimated and empirical fitness functions, averaged over 50 randomly sampled training sets of size N. The point at the number of samples required to exactly recover the GNK model with Structural Neighborhoods ($N = 6,537$) is highlighted with a red dot and dashed lines; at this number of samples, the mean prediction R^2 is 0.728. Error bars indicate the standard deviation of R^2 over training replicates. Inset shows paired plot between the estimated and predicted fitness function for one example training set of size $N = 6,537$.

S3 Effect of cutoff distance on sparsity of GNK fitness functions with Structural Neighborhoods

In the main text, we deem two positions of a protein sequence to be in structural contact when any pair of atoms in the residues at these positions are within 4.5 Åof one another in a given atomistic protein structure. We additionally constructed Structural neighborhoods for GNK fitness functions based on this definition of structural contacts. Here we test the effect of the 4.5 Å cutoff distance on the sparsity of the resulting GNK fitness functions. In order to do so, we constructed Structural Neighborhoods using structural contacts defined at different cutoff distances and calculated the sparsity of the GNK fitness functions with those Structural neighborhoods using Eq. 6. The results of this test for the two GNK models with Structural Neighborhoods considered in the main text, where the neighborhoods are derived from the mTagBFP crystal structure and the I-TASSER predicted structure of the protein encoded by the yeast His3 gene, are shown in Figure S2.

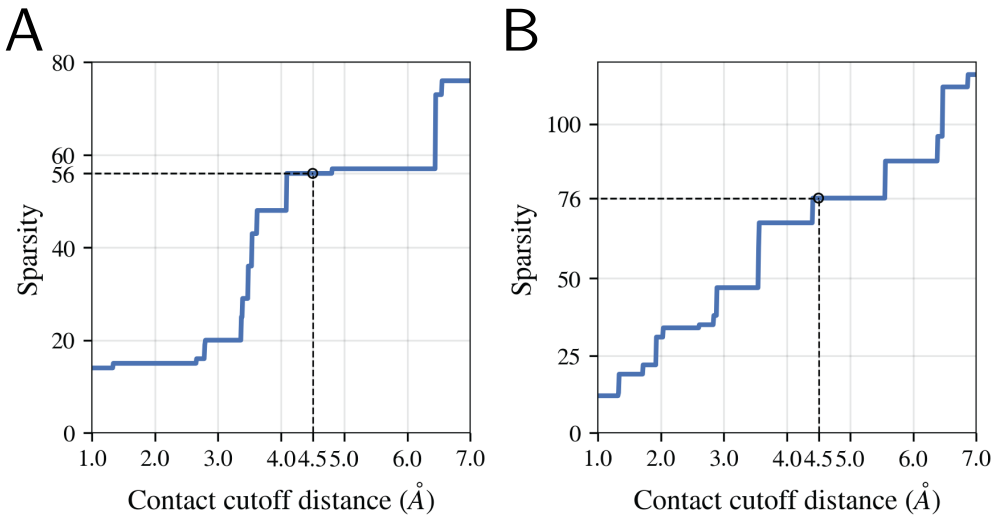

Fig. S2: Effect of structural contact cutoff distance on sparsity of GNK fitness functions with Structural Neighborhoods. Each plot corresponds to GNK models with a fixed L and q, and Structural Neighborhoods corresponding to a particular atomistic protein structure. The horizontal axis indicates the cutoff distance used to define structural contacts and thus used to construct the Structural Neighborhoods. The vertical axis indicates the sparsity of the GNK fitness functions when the Structural Neighborhoods are defined with each cutoff distance. The cutoff distance used in the main text (4.5 Å) and the corresponding sparsities are indicated with dashed lines. (A) GNK fitness functions with $L = 13$ and $q = 2$ and Structural Neighborhoods derived from the mTagBFP crystal structure. (B) GNK fitness functions with $L = 11$, $q = 2$ and Structural Neighborhoods derived from the I-TASSER predicted structure of the protein encoded by the His3 gene in yeast.

S4 Further details on numerical calculation of C

In order to determine an appropriate value of the C constant, we sampled fitness functions from the GNK model with Random Neighborhoods at different settings of L, q, and K, and determined the number of randomly sampled fitness measurements required to estimate these fitness functions. Figure 2D shows the results of these tests averaged over all settings of L, q, and K. In Figure S3, we display these results in more detail by showing the results for particular settings of L and q in separate plots. The curve in the plots of Figure S3 are also displayed as grey curves in 2D.

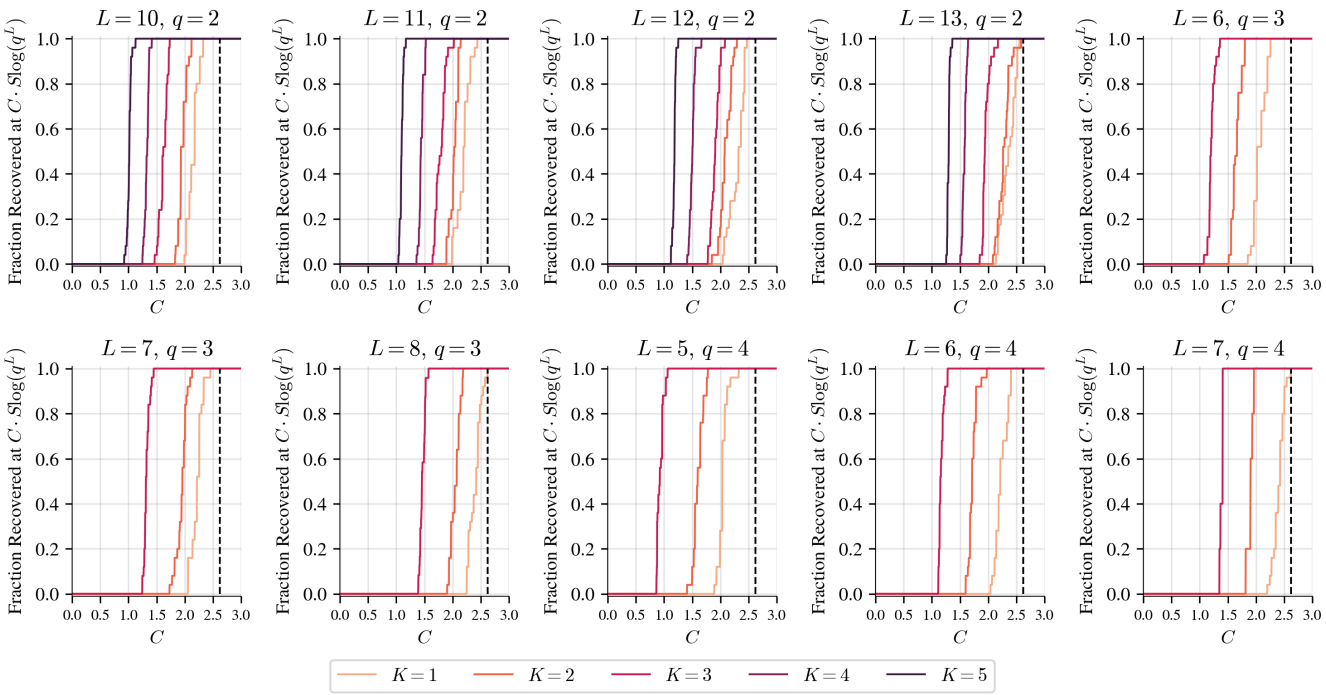

Fig. S3: Fraction of GNK fitness functions with Random Neighborhoods recovered at a range of settings of C. Each plot corresponds to fitness functions with the setting of L and q indicated above the plot. Colors indicate the value of K used when sampling Random Neighborhoods. The value $C = 2.62$ is highlighted with a dashed line in each plot.

S5 Proofs and additional theoretical results

In this section, we formally state the mathematical results from the main text and provide proofs.

S5.1 Graph theory preliminaries

Much of the following requires substantial graph-theoretic construction, so we first introduce the requisite notation and simple definitions. We will use the notation $V(G)$ and $E(G)$ to denote the vertex and edge sets of a graph G. The graph is then specified by $G = (V(G), E(G))$. The "degree" of a vertex v is the number of other vertices that are adjacent to v. A k-regular graph is a graph in which every vertex has degree equal to k. The Graph Laplacian of a graph G with vertices $V(G) = \{g_i\}_{i=1}^n$ is given by $\mathbf{L}(G) \coloneqq \mathbf{D}(G) - \mathbf{A}(G)$ where $\mathbf{D}(G)$ is an $n \times n$ diagonal matrix whose i^{th} diagonal element is equal to the degree of vertex i and $\mathbf{A}(G)$ is the $n \times n$ adjacency matrix of G with elements given by

$$\mathbf{A}_{ij}(G) = \begin{cases} 1 & \text{if } g_i \text{ is adjacent to } g_j \text{ in } G, \\ 0 & \text{otherwise.} \end{cases}$$

Graph Laplacians and adjacency matrices are real, symmetric matrices and thus have orthonormal sets of eigenvectors. In the case of a k-regular graph, $\mathbf{L}(G) = k\mathbf{I} - \mathbf{A}(G)$. Thus the Laplacian and adjacency matrices share eigenvectors, and the eigenvalues of the Laplacian are given by
$\lambda_j(\mathbf{L}) = k - \lambda_j(\mathbf{A})$ for $j = 1, ..., L$ where $\lambda_j(\mathbf{A})$ are the eigenvalues of the adjacency matrix.

We will make use of the Cartesian product of graphs, defined below:

Definition 1 (Cartesian Product of Graphs)**.** The Cartesian product between two graphs $G = (V(G), E(G))$ and $H = (V(H), E(H))$ is defined as $G \square H = (V(G) \times V(H), E(G \square H))$, where two vertices (g, h) and (g', h') are adjacent in $G \square H$ if and only if either

1. $g = g'$ and h is adjacent to h' in H, or
2. $h = h'$ and g is adjacent to g' in G.

A direct consequence of Definition 1 is that the adjacency matrix of the Cartesian product can be constructed from the adjacency matrices of its components as [59]:

$$\mathbf{A}(G \square H) = \mathbf{A}(G) \otimes \mathbf{I}_m + \mathbf{I}_n \otimes \mathbf{A}(H), \tag{S1}$$

where $m = |V(H)|$ and $n = |V(G)|$ are the number of vertices in H and G, respectively.

We will additionally make use of the Lexicographic product of graphs [60].

Definition 2 (Lexicographic Product of Graphs)**.** The Lexicographic product between two graphs $G = (V(G), E(G))$ and $H = (V(H), E(H))$ is defined as $G \circ H = (V(G) \times V(H), E(G \circ H))$, where two vertices (g, h) and (g', h') are adjacent in $G \circ H$ if and only if either

1. g is adjacent to g' in G, or
2. $g = g'$ and h is adjacent to h' in H.

The adjacency matrix of a Lexicographic product of graphs is given by [59]:

$$\mathbf{A}(G \circ H) = \mathbf{A}(G) \otimes \mathbf{J}_m + \mathbf{I}_n \otimes \mathbf{A}(H), \tag{S2}$$

where $\mathbf{J}_m$ is the $m \times m$ matrix with every element equal to one.

The graphs described up until now have been 'simple' graphs, where each edge connects exactly two vertices and vertices are connected by at most one edge. We will also discuss 'hypergraphs', where 'edges' are sets that may contain more than two vertices (these are referred to as 'hyperedges'). Let $H = (V(H), E(H))$

be a hypergraph with n vertices, $V(H) = \{h_i\}_{i=1}^n$, and p hyperedges, $E(H) = \{e_j\}_{j=1}^p$. The *incidence* matrix of H, denoted $\mathbf{F}(H)$,9 is the $n \times p$ matrix with elements

$$\mathbf{F}_{ij}(H) = \begin{cases} 1 & \text{if } g_i \in e_i, \\ 0 & \text{otherwise.} \end{cases}$$

The degree of a vertex in a hypergraph is equal to the number of hyperedges that contain that vertex, and a k-regular hypergraph is one in which all vertices have degree equal to k. The *clique multigraph* corresponding to a hypergraph is the multigraph (another extension of simple graphs where two vertices can have multiple simple edges between them) with the same vertices as the hypergraph, and as many edges between two vertices as the number of times those vertices co-occur in a hyperedge of the hypergraph (i. e., if two vertices are both in two separate hyperedges of the hypergraph, then they will have two edges between them in the clique multigraph) [61]. The $(i, j)^{\text{th}}$ element of the adjacency matrix of a multigraph is equal to the number of edges that connect the ith and j^{th} vertices.

Below are two Lemmas regarding the spectrum of Cartesian and lexicographic graph products that will be useful.

Lemma 1. *Let G and H be regular graphs with n and m vertices, respectively. Let $\mathbf{A}(G) = \mathbf{P}\mathbf{\Lambda}_G\mathbf{P}^T$ and $\mathbf{A}(H) = \mathbf{Q}\mathbf{\Lambda}_H\mathbf{Q}^T$ be eigendecompositions of the adjacency matrices of G and H, respectively. Then the adjacency matrix of the Cartesian product $G\square H$ has the eigendecomposition given by:*

$$\mathbf{A}(G\square H) = \mathbf{R}\mathbf{\Lambda}_\square\mathbf{R}^T, \tag{S3}$$

where $\mathbf{R} = \mathbf{P} \otimes \mathbf{Q}$ and $\mathbf{\Lambda}_\square = \mathbf{\Lambda}_G \otimes \mathbf{I}_m + \mathbf{I}_n \otimes \mathbf{\Lambda}_H$.

Proof. We can use Equation S1 to prove the proposed Lemma directly:

$$\begin{aligned} \mathbf{R}\mathbf{\Lambda}_\square\mathbf{R}^T &= [\mathbf{P} \otimes \mathbf{Q}][\mathbf{\Lambda}_G \otimes \mathbf{I}_m + \mathbf{I}_n \otimes \mathbf{\Lambda}_H][\mathbf{P} \otimes \mathbf{Q}]^T \\ &= [\mathbf{P}\mathbf{\Lambda}_G \otimes \mathbf{Q} + \mathbf{P} \otimes \mathbf{Q}\mathbf{\Lambda}_H][\mathbf{P}^T \otimes \mathbf{Q}^T] \\ &= \mathbf{P}\mathbf{\Lambda}_G\mathbf{P}^T \otimes \mathbf{Q}\mathbf{Q}^T + \mathbf{P}\mathbf{P}^T \otimes \mathbf{Q}\mathbf{\Lambda}_H\mathbf{Q}^T \\ &= \mathbf{A}(G) \otimes \mathbf{I}_m + \mathbf{I}_n \otimes \mathbf{A}(H) \\ &= \mathbf{A}(G\square H), \end{aligned}$$

where in the second and third lines we have used the property of Kronecker products that $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AB} \otimes \mathbf{CD})$, in the second line we have also used the fact that the transpose is distributive over the Kronecker product, $(\mathbf{C} \otimes \mathbf{D})^T = \mathbf{C}^T \otimes \mathbf{D}^T$, and the final line is a result of Equation S1. □

Lemma 2. *Let G and H be regular graphs with n and m vertices, respectively. Let $\mathbf{A}(G) = \mathbf{P}\mathbf{\Lambda}_G\mathbf{P}^T$ and $\mathbf{A}(H) = \mathbf{Q}\mathbf{\Lambda}_H\mathbf{Q}^T$ be eigendecompositions of the adjacency matrices of G and H. Then the adjacency matrix of the lexicographic product $G \circ H$ has the eigendecomposition given by:*

$$\mathbf{A}(G \circ H) = \mathbf{R}\mathbf{\Lambda}_\circ\mathbf{R}^T, \tag{S4}$$

where $\mathbf{R} = \mathbf{P} \otimes \mathbf{Q}$, $\mathbf{\Lambda}_\circ = \mathbf{\Lambda}_G \otimes \mathbf{B} + \mathbf{I}_n \otimes \mathbf{\Lambda}_H$, and $\mathbf{B} = m\mathbf{e}_1\mathbf{e}_1^T$ (i. e., $B_{ij} = m$ if $i = j = 1$ and zero otherwise).

Proof. The adjacency matrix of any k-regular graph with m vertices has two eigenvalues, k and 0, with multiplicities 1 and $m-1$, respectively. The normalized eigenvector associated with the eigenvalue k is $\frac{1}{\sqrt{m}}\mathbf{1}_m$ and the normalized eigenvectors associated with the eigenvalue 0 are any set of length-m orthogonormal vectors that are orthogonal to $\mathbf{1}_m$ (i. e., vectors that sum to one). Since H is a regular graph, we then have that

$$(\mathbf{Q}^T\mathbf{J}_m)_{ij} = \begin{cases} \sqrt{m} & \text{if } i = 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$(\mathbf{Q}^T\mathbf{J}_m\mathbf{Q})_{ij} = \begin{cases} m & \text{if } i = j = 1 \\ 0 & \text{otherwise} \end{cases}$$

Therefore, $(\mathbf{Q}^T\mathbf{J}_m\mathbf{Q}) = \mathbf{B}$ and further $\mathbf{Q}\mathbf{B}\mathbf{Q}^T = \mathbf{J}_m$. Then we can use Equation Eq. S2 to show:

$$\begin{aligned}
\mathbf{R}\mathbf{\Lambda}_\circ\mathbf{R}^T &= [\mathbf{P}\otimes\mathbf{Q}][\mathbf{\Lambda}_G\otimes\mathbf{B} + \mathbf{I}_n\otimes\mathbf{\Lambda}_H][\mathbf{P}\otimes\mathbf{Q}]^T \\
&= \mathbf{P}\mathbf{\Lambda}_G\mathbf{P}^T\otimes\mathbf{Q}\mathbf{B}\mathbf{Q}^T + \mathbf{P}\mathbf{P}^T\otimes\mathbf{Q}\mathbf{\Lambda}_H\mathbf{Q}^T \\
&= \mathbf{A}(G)\otimes\mathbf{J}_m + \mathbf{I}_n\otimes\mathbf{A}(H) \\
&= \mathbf{A}(G\circ H)
\end{aligned}$$

□

S5.2 Graph Fourier basis results

Here we will prove results related to our construction of Graph Fourier bases. First, we prove a result regarding the eigenvectors of the complete graph, which is presented in the main text as Equation Eq. 9.

Proposition 1. *An orthonormal set of eigenvectors of the Graph Laplacian of the complete graph $K(q)$ are given by the columns of the $q\times q$ matrix:*

$$\mathbf{P}_q \coloneqq \mathbf{I}_q - \frac{2\mathbf{w}\mathbf{w}^T}{\|\mathbf{w}\|_2^2}, \tag{S5}$$

where $\mathbf{w} \coloneqq \mathbf{1}_q - \sqrt{q}\mathbf{e}_1$, $\mathbf{1}_q$ is the vector of length q whose elements are all equal to one, $\mathbf{e}_1$ is the length q with the first element set to 1 and all others set to zero, and $\mathbf{I}_q$ is the $q\times q$ identity matrix

Proof of Proposition 1. The Graph Laplacian of the complete graph $K(q)$ is given by $\mathbf{L}(K(q)) = q\mathbf{I}_q - \mathbf{J}_q$ (i. e., the $q\times q$ matrix with q on the diagonal and all other elements equal to -1). This matrix has two eigenvalues, 0 and q, with multiplicities 1 and $q-1$, respectively. The normalized eigenvector corresponding to the zero eigenvalue is $\frac{1}{\sqrt{q}}\mathbf{1}_q$. Since the graph Laplacian is symmetric, the eigenvectors are orthogonal and therefore the remaining eigenvectors are any set of $n-1$ orthogonal vectors that are orthogonal to $\mathbf{1}_q$ (i. e., vectors that sum to zero) and each other. In order to show that the columns of the Householder matrix given in Equation Eq. S5 are orthonormal eigenvectors of the complete graph, we will prove (i) that the first column of $\mathbf{P}_q$ is equal to $\frac{1}{\sqrt{q}}\mathbf{1}_q$ and (ii) that $\mathbf{P}_q$ is an orthogonal matrix:

(i) We can more explicitly write the $\mathbf{w} = [w_1, w_2, ..., w_q]$ vector as $\mathbf{w} = \mathbf{1}_q - \sqrt{q}\mathbf{e}_1 = [1-\sqrt{q}, 1, 1, ..., 1]^T$. Therefore, $\|\mathbf{w}\|_2^2 = (1-\sqrt{q})^2 + (q-1) = 2(q-\sqrt{q})$. Let α_i for $i = 1, 2, ..q$ be the elements of the first column of $\mathbf{P}_q$, respectively. Then,

$$\begin{aligned}
\alpha_1 &= 1 - \frac{2w_1w_1}{\|w\|_2^2} \\
&= 1 - \frac{(1-\sqrt{q})^2}{q-\sqrt{q}} \\
&= 1 - \left(1 - \frac{1}{\sqrt{q}}\right) = \frac{1}{\sqrt{q}},
\end{aligned}$$

and for $j = 2, 3, ..., q$, we have

$$\begin{aligned}
\alpha_j &= \frac{2w_jw_1}{\|w\|_2^2} \\
&= \frac{1-\sqrt{q}}{q-\sqrt{q}} \\
&= \frac{1}{\sqrt{q}}.
\end{aligned}$$

Therefore, all elements of the first columns of $\mathbf{P}_q$ are equal to $\frac{1}{\sqrt{q}}$.

(ii) The orthogonality of $\mathbf{P}_q$ follows directly from the definition:

$$\begin{aligned}(\mathbf{P}_q)^T\mathbf{P}_q &= \left(\mathbf{I}_q - \frac{2\mathbf{w}\mathbf{w}^T}{\|\mathbf{w}\|_2^2}\right)^T \left(\mathbf{I}_q - \frac{2\mathbf{w}\mathbf{w}^T}{\|\mathbf{w}\|_2^2}\right) \\ &= \mathbf{I}_q - \frac{4\mathbf{w}\mathbf{w}^T}{\|\mathbf{w}\|_2^2} + \frac{4\mathbf{w}\mathbf{w}^T\mathbf{w}\mathbf{w}^T}{\|\mathbf{w}\|_2^4} \\ &= \mathbf{I}_q.\end{aligned}$$

Thus, $\mathbf{P}_q$ is an orthogonal matrix whose first column is $\frac{1}{\sqrt{q}}\mathbf{1}_q$, and further, the columns are $\mathbf{P}_q$ are an orthonormal set of eigenvectors of $K(q)$. □

We additionally have the following result showing how to construct the Graph Fourier basis corresponding to the Hamming graph using the eigenvectors of the complete graph, which is presented in the main text as Equation Eq. 10.

Proposition 2. *An orthonormal set of eigenvectors of the Graph Laplacian of the Hamming graph $H(L, q)$ are given by the columns of the $q^L \times q^L$ matrix*

$$\mathbf{\Phi} = \bigotimes_{i=1}^{L} \mathbf{P}_q, \tag{S6}$$

where $\mathbf{P}_q$ is defined in Eq. S5.

In order to prove Proposition 2, we need a preliminary result. First remember that the Hamming graph $H(L, q)$ is the L-fold Cartesian product of the complete graph $K(q)$. We have the following result regarding the eigenvectors and eigenvalues of the adjacency matrices of Cartesian products of regular graphs.

Lemma 3. *Let G and H be regular graphs and let $\mathbf{P}$ and $\mathbf{Q}$ be matrices whose columns are eigenvectors of the Graph Laplacians of G and H, respectively. Then the columns of $\mathbf{P} \otimes \mathbf{Q}$ are eigenvectors of the Graph Laplacian of the Cartesian product $G \square H$.*

Proof. For a regular graph, the degree matrix is a constant multiplied by the identity matrix. Thus, in this case, the Graph Laplacian and adjacency matrices differ only by a constant added to the diagonal (and a constant multiplicative factor of -1). The Graph Laplacian and adjacency matrices of a regular graph therefore have the same eigenvectors. The result then follows from Lemma 1. □

Proof of Proposition 2. The Hamming graph $H(L, q)$ is defined as the L-fold Cartesian product of the complete graph $K(q)$ [56]:

$$H(L, q) = \square_{i=1}^{L} K(q). \tag{S7}$$

Thus, by Lemma 3, the eigenvectors of the Graph Laplacian of $H(L, q)$ are the L-fold Kronecker product of the eigenvectors of the Graph Laplacian of $K(q)$, as given in Equation Eq. S6. □

S5.3 Distribution of GNK Fourier coefficients

Here we prove our result regarding the distribution of the Fourier coefficients of fitness functions sampled from the GNK model (Equation Eq. 5). To begin, let $\mathrm{GNK}(L, q, \mathcal{V})$ be probability distribution over fitness functions induced by the GNK model for sequence length L, alphabet size q, and a set of neighborhoods corresponding to each position, $\mathcal{V} := \{V^{[j]}\}_{j=1}^{L}$. We now formally restate the result of Equation Eq. 5.

Theorem 1. *Let* $\mathbf{f} = (f(\mathbf{s}))_{\mathbf{s}\in\mathcal{S}^{(L,q)}}$ *be the complete vector of evaluations of a fitness function* $f \sim GNK(L, q, \mathcal{V})$*. Then the Fourier coefficients of* f*, given by* $\boldsymbol{\beta} = \boldsymbol{\Phi}^T\mathbf{f}$*, are distributed according to* $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\lambda}\mathbf{I})$ *(i. e., normally distributed with zero mean and diagonal covariance). Let* $\boldsymbol{\beta}_U$ *be the length* $(q-1)^r$ *sub-vector of* $\boldsymbol{\beta}$ *representing the epistatic interaction* U*. Then the variance of every element of* $\boldsymbol{\beta}_U$ *is given by*

$$\lambda_U = \frac{1}{L}\sum_{j=1}^{L} q^{L-K_j} I\left(U \subseteq V^{[j]}\right) \tag{S8}$$

where $I\left(U \subseteq V^{[j]}\right)$ *is an indicator function that is equal to one if* U *is a subset of or equal to* $V^{[j]}$ *and zero otherwise.*

The proof of Theorem 1 is quite involved and requires a number of lemmas; the proofs of the lemmas are shown after the proof of the main result.

In order prove Theorem 1, we will first provide an alternative definition of the GNK model in terms of hypergraphs. To start, we'll now assign an index to every sequence in the space of sequences, so $\mathcal{S}^{(L,q)} = \{\mathbf{s}_i\}_{i=1}^{q^L}$. As in the main text, $\mathbf{s}_i^{[j]}$ refers to the subsequence of $\mathbf{s}_i$ corresponding to the indices in the neighborhood $V^{[j]}$. Each neighborhood in the GNK model induces a hypergraph over sequence space, where the vertices represent all sequences in $\mathcal{S}^{(L,q)}$ and edges contain sequences that share subsequences corresponding to the indices in the neighborhood. We formally define this hypergraph and related quantities below.

Definition 3 (GNK hypergraph)**.** *Let* $G(V) = (\mathcal{S}^{(L,q)}, E(V))$ *be a 'GNK hypergraph' corresponding to a neighborhood* V *for a GNK model defined for sequences of length* L *and alphabet size* q*. The edge set,* $E(V)$*, corresponds to every possible subsequence of length* $|V|$*, and two sequences co-occur in an edge if and only if they share the subsequence corresponding to the positions in* V*. Additionally, let* $\mathbf{F}(V) := \mathbf{F}(G(V))$ *be the incidence matrix of* $G(V)$*,* $C(V)$ *be the clique multigraph of* $G(V)$ *and* $\mathbf{A}(V) := \mathbf{A}(C(V))$ *be the adjacency matrix of* $C(V)$*. Finally, when it is appropriate to consider the indexed neighborhoods* $V^{[j]}$*, then we will use this indexing for all of the GNK hypergraph quantities. Specifically, define* $G^{[j]} := G(V^{[j]})$*,* $\mathbf{F}^{[j]} := \mathbf{F}(V^{[j]})$*,* $C^{[j]} := C(V^{[j]})$*, and* $\mathbf{A}^{[j]} := \mathbf{A}(V^{[j]})$*.*

The following Lemma gives an immediate useful result of this definition.

Lemma 4. *Every GNK hypergraph,* $G(V)$*, is a 1-regular hypergraph.*

We will use the GNK hypergraphs to provide an alternative definition of the GNK model, which is shown in the following result. Note that this definition is equivalent to the matrix definition of the GNK model of ref. 26.

Lemma 5. *Define the matrix* $\mathbf{F}$ *as the column-wise concatenation of the incidence matrices* $\mathbf{F}^{[j]}$ *for* $j = 1, 2, ..., L$*:*

$$\mathbf{F} := \left[\,\mathbf{F}^{[1]} \,\middle|\, \mathbf{F}^{[2]} \,\middle|\, \ldots \,\middle|\, \mathbf{F}^{[L]}\,\right] \tag{S9}$$

Additionally, let $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \frac{1}{L}\mathbf{I})$ *be a length* $\sum_{j=1}^{L} q^{K_j}$ *normally distributed random vector. Then,* $\mathbf{f} = \mathbf{F}\mathbf{w}$ *contains all fitness evaluations of a fitness function* f *that is distributed according to* $GNK(L, q, \mathcal{V})$ *(i. e.,* $\mathbf{f} = (f(\mathbf{s}))_{\mathbf{s}\in\mathcal{S}^{(L,q)}}$ *and* $f \sim GNK(L, q, \mathcal{V})$*).*

The following conclusions regarding the statistics of the Fourier coefficients in the GNK model are immediate from this definition of the model. In what follows, we use angled brackets, $\langle\rangle$ to indicate expectation over the random field.

Lemma 6. *Let* $\boldsymbol{\beta}$ *be the Fourier coefficients of a fitness functions distributed according to* $GNK(L, q, \mathcal{V})$*. Then* $\boldsymbol{\beta}$ *is normally distributed with* $\langle\boldsymbol{\beta}\rangle = 0$ *and covariance matrix*

$$\langle\boldsymbol{\beta}\boldsymbol{\beta}^T\rangle = \frac{1}{L}\boldsymbol{\Phi}^T\mathbf{F}\mathbf{F}^T\boldsymbol{\Phi}, \tag{S10}$$

where $\mathbf{\Phi}$ is the Fourier basis defined in Equation Eq. S6 *and $\mathbf{F}$ is the column-wise concatenation of the incidence matrices of GNK hypergraphs defined in Equation* Eq. S9*.*

Lemma 6 provides a straightforward path towards proving Theorem 1. We now need to show that (i) $\mathbf{\Phi}$ diagonalizes $\mathbf{F}\mathbf{F}^T$ (i. e., $\mathbf{\Phi}$ is a basis of eigenvectors for $\mathbf{F}\mathbf{F}^T$) and (ii) that the eigenvalues of $\mathbf{F}\mathbf{F}^T$ are given by Equation Eq. S8. The problem can be further simplified by first noting the following simple result, which follows straightforwardly from the multiplication of block matrices.

Lemma 7. $\mathbf{F}\mathbf{F}^T = \sum_{j=1}^{L} \mathbf{F}^{[j]}(\mathbf{F}^{[j]})^T$.

This result tells us that if possible, it is sufficient to prove that $\mathbf{\Phi}$ diagonalizes each $\mathbf{F}^{[j]}(\mathbf{F}^{[j]})^T$ in order to prove that $\mathbf{\Phi}$ diagonalizes $\mathbf{F}\mathbf{F}^T$. Then the eigenvalues of $\mathbf{F}\mathbf{F}^T$ will simply be given by the sum of the eigenvalues of the $\mathbf{F}^{[j]}(\mathbf{F}^{[j]})^T$. We are further assisted by the following result regarding the outer product of incidence matrices of regular hypergraphs, due to [61].

Lemma 8. *Let C be the clique multigraph of a k-regular hypergraph H with incidence matrix $\mathbf{F}(H)$. Then $\mathbf{F}(H)\mathbf{F}(H)^T = \mathbf{A}(C) + k\mathbf{I}$, where $\mathbf{A}(C)$ is the adjacency matrix of C.*

Lemma 8 tells us if we can determine the spectrum of the adjacency matrices $\mathbf{A}^{[j]}$ of the clique multigraphs $C^{[j]}$, then it is straightforward to calculate the spectrum of $\mathbf{F}^{[j]}(\mathbf{F}^{[j]})^T$. In order to begin to calculate the spectrum of $\mathbf{A}^{[j]}$ we recognize the following simple fact regarding these clique multigraphs (remember that $G^{[j]}$ is a 1-regular hypergraph by Lemma 4).

Lemma 9. *The clique multigraph of a 1-regular hypergraph is a simple graph.*

We thus need to determine the spectrum of the simple graphs, $C^{[j]}$. $C^{[j]}$ contains edges between any two sequences that share a subsequence corresponding to the indices in the jth neighborhood $V^{[j]}$. In order to calculate the spectrum of $C^{[j]}$, we will first show how these clique multigraphs can be constructed recursively. In the next few Lemmas, we will provide results for clique graphs associated with a generic neighborhood V, and then return to considering the indexed neighborhoods $V^{[j]}$ when necessary.

Lemma 10. *Let $V \subseteq \{1, 2, ..., L\}$ be a GNK neighborhood. Additionally, let $O(q)$ be the empty graph of size q (i. e., the graph containing q vertices and no edges) and define the graphs $B_l(V)$, via the recursion relation:*

$$B_{l+1}(V) = \begin{cases} B_l(V) \square O(q) & \text{if } i+1 \in V \\ B_l(V) \circ K(q) & \text{otherwise,} \end{cases} \tag{S11}$$

for $i = 1, 2, ..., L-1$, where

$$B_1(V) = \begin{cases} O(q) & \text{if } 1 \in V \\ K(q) & \text{otherwise.} \end{cases} \tag{S12}$$

Then the vertices of $B_l(V)$ represent all sequences in $\mathcal{S}^{(l,q)}$ and two sequences $\mathbf{s}_i = [s_{i,1}, s_{i,2}, ..., s_{i,l}] \in \mathcal{S}^{(l,q)}$ and $\mathbf{s}_j = [s_{j,1}, s_{j,2}, ..., s_{j,l}] \in \mathcal{S}^{(l,q)}$ are adjacent in $B_l(V)$ if and only if $s_{i,k} = s_{j,k}$ for every $k \in V_{(l)}$ where we define $V_{(l)} := \{m \in V : m \leq l\}$ to be the l smallest elements of V.

The simple corollary of Lemma 10 is that the clique graphs $C(V)$ are the final results of the recursion in Equation Eq. S11.

Lemma 11. *Let $C(V)$ be the clique multigraph of a GNK hypergraph $G(V)$ and $B_L(V)$ be the graph defined by Equations* Eq. S11 *and* Eq. S12*. Then $C(V) = B_L(V)$.*

We have thus given a recursive definition of $C(V)$, which will allow us to calculate the spectrum of the adjacency matrix $\mathbf{A}(V)$ using the spectral properties of graph products presented in Lemmas 1 and 2. In particular, we have the following result regarding the eigenvectors of $\mathbf{A}(V)$.

Lemma 12. *The columns of the Fourier basis $\mathbf{\Phi}$ are a complete set of orthonormal eigenvectors of the adjacency matrix $\mathbf{A}(V) \coloneqq \mathbf{A}(C(V))$ of the clique multigraph $C(V)$.*

We could similarly use Lemmas 1 and 2 to calculate the eigenvalues of $\mathbf{A}(V)$; however, this would not allow us to connect the eigenvalues to epistatic interactions, as is required to prove Theorem 1. We will instead proceed by showing in Lemma 13 that the columns of suitably defined matrix are eigenvectors of the adjacency matrix $\mathbf{A}(V)$ with eigenvalues equal to a summand of Equation Eq. S8 up to additive constant. Then, in Lemma 14, we will show that this matrix is indeed equal to the columns of the Fourier basis corresponding to the epistatic interaction U. For the following results, recall from the main text that $\tilde{\mathbf{P}}_q$ is the matrix containing the final $q-1$ unnormalized columns of $\mathbf{P}_q$, such that $\mathbf{P}_q = \frac{1}{\sqrt{q}}\left[\mathbf{1} \mid \tilde{\mathbf{P}}_q\right]$, where $\mid$ denotes column-wise concatenation.

Lemma 13. *Let $U \subseteq \{1, 2, ..., L\}$ be a set of position indices representing an epistatic interaction and $V \subseteq \{1, 2, ..., L\}$ be a GNK neighborhood. Define the matrix $\mathbf{Z}_l(U)$ with the recursion relation:*

$$\mathbf{Z}_{l+1}(U) = \begin{cases} \mathbf{Z}_l(U) \otimes \tilde{\mathbf{P}}_q & \text{if} \quad l+1 \in U \\ \mathbf{Z}_l(U) \otimes \mathbf{1}_q & \text{otherwise} \end{cases} \tag{S13}$$

for $l = 1, 2, ..., L-1$, where

$$\mathbf{Z}_1(U) = \begin{cases} \tilde{\mathbf{P}}_q & \text{if} \quad 1 \in U \\ \mathbf{1}_q & \text{otherwise} \end{cases} \tag{S14}$$

Then the columns of $\mathbf{Z}_L(U)$ are eigenvectors of the adjacency matrix $\mathbf{A}(V)$, all associated with the eigenvalue given by

$$\mu(U, V) = q^{L-|V|} I(U \subseteq V) - 1. \tag{S15}$$

Lemma 14. *Define $\mathbf{\Phi}_U$ as the matrix of $(q-1)^{|U|}$ columns of the Fourier basis $\mathbf{\Phi}$ corresponding to the epistatic interaction $U \subseteq \{1, 2, ..., L\}$:*

$$\mathbf{\Phi}_U \coloneqq \begin{bmatrix} — & \boldsymbol{\phi}_U(\mathbf{s}_1)^T & — \\ — & \boldsymbol{\phi}_U(\mathbf{s}_2)^T & — \\ & \vdots & \\ — & \boldsymbol{\phi}_U(\mathbf{s}_{q^L})^T & — \end{bmatrix}, \tag{S16}$$

where $\boldsymbol{\phi}_U(\mathbf{s}_i) \coloneqq \frac{1}{\sqrt{q^L}} \bigotimes_{j \in U} \mathbf{p}_q(s_{i,j})$ is the encoding of sequence $\mathbf{s}_i$ in terms of the epistatic interaction U in the Fourier basis. Then, $\frac{1}{\sqrt{q^L}} \mathbf{Z}_L(U) = \mathbf{\Phi}_U$, where $\mathbf{Z}_L(U)$ is defined by Equations Eq. S13 *and* Eq. S14.

Equipped with these results, we are finally prepared to prove Theorem 1.

Proof of Theorem 1. In order to prove this theorem, we need to show (i) that the Fourier coefficients are normally distributed with zero mean and diagonal covariance and (ii) that the variance of the coefficients corresponding to a particular epistatic interaction are given by Equation Eq. 5.

First, Lemma 6 proves that the Fourier coefficients are normally distributed with zero mean. Next, Lemma 12 proves that the Fourier basis $\mathbf{\Phi}$ diagonalizes the adjacency matrix $\mathbf{A}^{[j]}$ of the clique multigraph of the GNK hypergraph $G^{[j]}$. Recalling that $G^{[j]}$ is a 1-regular hypergraph, then by Lemma 8, $\mathbf{F}^{[j]}(\mathbf{F}^{[j]})^T = \mathbf{A}^{[j]} + \mathbf{I}$. Therefore, the Fourier basis diagonalizes $\mathbf{F}^{[j]}(\mathbf{F}^{[j]})^T$ for all $j = 1, 2, ..., L$, and thus also diagonalizes $\mathbf{F}\mathbf{F}^T$ due to Lemma 7. Then, the covariance matrix of the Fourier coefficients, which is shown in Lemma 6 to be $\langle \boldsymbol{\beta}\boldsymbol{\beta}^T \rangle = \frac{1}{L}\mathbf{\Phi}^T \mathbf{F}\mathbf{F}^T \mathbf{\Phi}$, is diagonal. The eigenvalues of $\mathbf{F}\mathbf{F}^T$ are then equal to the variances of the Fourier coefficients.

Lemma 13 shows that the columns of the matrix $\mathbf{Z}_L(U)$ defined by Equations Eq. S13 and Eq. S14 are eigenvectors of $\mathbf{A}^{[j]}$. Further, Lemma 14 shows that this matrix is equal to the columns of the Fourier

basis corresponding to the epistatic interaction U, $\mathbf{\Phi}_U$. Thus, Lemma 13 shows that the eigenvalue of $\mathbf{A}^{[j]}$ associated with the columns $\mathbf{\Phi}_U$ is given by:

$$\mu(U, V^{[j]}) = q^{L-K_j} I(U \subseteq V^{[j]}) - 1.$$

By Lemma 8, the eigenvalues of $\mathbf{F}^{[j]}(\mathbf{F}^{[j]})^T$ are simply one plus those calculated with Eq. S15. Since the Fourier basis diagonalizes all $\mathbf{F}^{[j]}(\mathbf{F}^{[j]})^T$, the eigenvalues of $\mathbf{F}\mathbf{F}^T$ are simply the sum of those of the $\mathbf{F}^{[j]}(\mathbf{F}^{[j]})^T$. The eigenvalues of $\mathbf{F}\mathbf{F}^T$ associated with the eigenvectors given by the columns of $\mathbf{\Phi}_U$ are the variances of the Fourier coefficients corresponding to the epistatic interaction U. All of this together, we have:

$$\begin{aligned}
\langle \boldsymbol{\beta}_U \boldsymbol{\beta}_U^T \rangle &= \frac{1}{L} \left(\mathbf{\Phi}_U\right)^T \mathbf{F}\mathbf{F}^T \mathbf{\Phi}_U \\
&= \frac{1}{L} \sum_{j=1}^{L} \left(\mathbf{\Phi}_U\right)^T \mathbf{F}^{[j]} (\mathbf{F}^{[j]})^T \mathbf{\Phi}_U \\
&= \frac{1}{L} \sum_{j=1}^{L} (\mu(U, V^{[j]}) + 1) \mathbf{I} \\
&= \frac{1}{L} \sum_{j=1}^{L} \left(q^{L-K_j} I(U \subseteq V^{[j]}) \right) \mathbf{I},
\end{aligned}$$

which is the desired result for the variances of the Fourier coefficients. □

Proof of Lemma 4. Every sequence (i. e., vertex of $G(V)$) contains exactly one subsequence corresponding to the position indices in V. Therefore, each vertex is contained in exactly one edge of $G(V)$. □

Proof of Lemma 5. In order to prove this, we need to show (i) that the above formulation results in L unit normally distributed subsequence fitness values being assigned to each sequence, where each subsequence corresponds to the position indices in a neighborhood $V^{[j]}$ (ii) that sequences share subsequence fitness values when they share the corresponding subsequence, and (iii) that the L subsequence fitness values are summed to produce the total fitness value assigned to each sequence.

A direct result of Definition S5.3 is that $\mathbf{F}^{[j]}$ has elements given by:

$$\mathbf{F}_{ik}^{[j]} = \begin{cases} 1 & \text{if } \mathbf{s}_i^{[j]} = \tilde{\mathbf{s}}_k \\ 0 & \text{otherwise,} \end{cases} \tag{S17}$$

where we define $\tilde{\mathbf{s}}_k$ as the kth possible subsequence of length K_j (i. e., the kth element in $\mathcal{S}^{(K_j,q)}$). Since each hyperedge in $G^{[j]}$ represents a subsequence of length K_j, each hyperedge contains vertices that represent sequences that share subsequence fitness values in the GNK model. Therefore, letting $\mathbf{w}^{[j]} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be a length q^{K_j} normally distributed random vector representing the subsequence fitness values randomly assigned to each subsequence, then

$$\mathbf{f}^{[j]} = \mathbf{F}^{[j]} \mathbf{w}^{[j]},$$

where $\mathbf{f}^{[j]} = [f_j(\mathbf{s}_1), f_j(\mathbf{s}_2), ..., f_j(\mathbf{s}_{q^L})]^T$ is the vector of subsequence fitness values corresponding to neighborhood j that are assigned to each sequence in $\mathcal{S}^{(L,q)}$. Since $G^{[j]}$ is a 1-regular hypergraph (Lemma 4), each row of $\mathbf{F}^{[j]}$ contains exactly one nonzero element and therefore the subsequence fitness values of each sequence are distributed as $\mathcal{N}(0, 1)$, as in the original definition of the GNK model given in the main text. Additionally, the structure of the incidence matrix shown in Equation Eq. S17 ensures that two sequences that share a subsequence corresponding to the position indices in $V^{[j]}$ also share a subsequence fitness value in $\mathbf{f}^{[j]}$, as required by the GNK model.

Now, let $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be the random vector that is the concatenation of the $\mathbf{w}^{[j]}$ random vectors containing subsequence fitness values. Then we have:

$$\begin{aligned}\mathbf{f} &= \mathbf{F}\mathbf{w} \\ &= \sum_{j=1}^{L} \mathbf{F}^{[j]}\mathbf{w}^{[j]} \\ &= \sum_{j=1}^{L} \mathbf{f}^{[j]}.\end{aligned}$$

Therefore, the elements of $\mathbf{f}$ are simply the sums of the L subsequence fitness values corresponding to each sequence, which is the final step in definition of the GNK model given in the main text. □

Proof of Lemma 6. This result follows immediately from recognizing that $\mathbf{f} = \mathbf{F}\mathbf{w} = \mathbf{\Phi}\boldsymbol{\beta}$, and therefore $\boldsymbol{\beta} = \mathbf{\Phi}^T\mathbf{F}\mathbf{w}$. The Fourier coefficients are thus a linear transformation of a normally distributed random vector, $\mathbf{w}$, and are therefore normally distributed with mean $\langle\boldsymbol{\beta}\rangle = \mathbf{\Phi}^T\mathbf{F}\langle\mathbf{w}\rangle = \mathbf{0}$ and covariance matrix $\langle\boldsymbol{\beta}\boldsymbol{\beta}^T\rangle = \mathbf{\Phi}^T\mathbf{F}\langle\mathbf{w}\mathbf{w}^T\rangle\mathbf{F}^T\mathbf{\Phi} = \mathbf{\Phi}^T\mathbf{F}\mathbf{F}^T\mathbf{\Phi}$. □

Proof of Lemma 8. Each element of $\mathbf{F}(H)\mathbf{F}(H)^T$ is the inner product of two rows in $\mathbf{F}(H)$. Since each element in row i of $\mathbf{F}(H)$ indicates whether vertex i is in a particular edge, the inner product of row i and row j ($i \neq j$) counts the number of edges that contain both vertex i and vertex j. Of course, this is also the number of edges connecting vertex i and vertex j in the clique multigraph, and thus the off-diagonal elements of $\mathbf{F}(H)\mathbf{F}(H)^T$ are equal to the elements of $\mathbf{A}(C)$. The diagonal elements of $\mathbf{F}(H)\mathbf{F}(H)^T$ are equal to the total number of edges containing vertex i, which is L for every vertex. □

Proof of Lemma 9. Each vertex in a 1-regular hypergraph is in exactly one hyperedge, and thus the clique multigraph has at most one edge between any two vertices. □

Proof of Lemma 10. First, both the lexicographic and Cartesian products result in graphs whose vertex sets are the (set) Cartesian product of the vertex sets of the multiples. Since the vertex sets of both $O(q)$ and $K(q)$ represent elements of the alphabet of size q, an l-fold graph product of these graphs will result in each vertex representing a sequence of length l.

We will prove the adjacency property of these product graphs with induction. For ease of notation, we drop the dependence of $B_l(V)$ on V and let $B_l \leftarrow B_l(V)$. Assume that two sequences $\mathbf{s}_i = [s_{i,1}, s_{i,2}, ..., s_{i,l}] \in \mathcal{S}^{(l,q)}$ and $\mathbf{s}_j = [s_{j,1}, s_{j,2}, ..., s_{j,l}] \in \mathcal{S}^{(l,q)}$ are adjacent in B_l if and only if the adjacency condition, $s_{i,k} = s_{j,k}$ for every $k \in V_{(l)}$, is satisfied. We will show that these adjacency conditions remain true for $l+1$. There are two cases to consider: (i) $l+1 \in V$ and (ii) $l+1 \notin V$.

(i) ($l+1 \in V$). Let $\mathbf{s}_i = [\tilde{\mathbf{s}}_i | s_{i,l+1}] \in \mathcal{S}^{(l+1,q)}$ and $\mathbf{s}_j = [\tilde{\mathbf{s}}_j | s_{j,l+1}] \in \mathcal{S}^{(l+1,q)}$ be sequences of length $l+1$, where $\tilde{\mathbf{s}}_i$ contains the first l elements of $\mathbf{s}_i$. Since in this case $l+1 \in V_{(l+1)}$, we must prove that $\mathbf{s}_i$ and $\mathbf{s}_j$ are adjacent in B_{l+1} if and only if the adjacency condition is satisfied for $\tilde{\mathbf{s}}_i$ and $\tilde{\mathbf{s}}_j$ in B_l (which is true by inductive assumption) and $s_{i,l+1} = s_{j,l+1}$. By Equation S11, $B_{l+1} = B_l \square O(q)$ in this case. Note that the vertices in B_l represent the $\tilde{\mathbf{s}}_i$ sequences of length l and the vertices in $O(q)$ represent the new elements of the sequence, $s_{i,l+1}$. According to the definition of the graph Cartesian product (Definition 1), $\mathbf{s}_i$ and $\mathbf{s}_j$ are adjacent in B_{l+1} if and only if $\tilde{\mathbf{s}}_i$ and $\tilde{\mathbf{s}}_j$ are adjacent in B_l and $s_{i,l+1} = s_{j,l+1}$. Thus, in this case, the adjacency condition remains true for $l+1$ under the inductive assumption.

(ii) ($l+1 \notin V$). In this case, $l+1 \notin V_{(l+1)}$ and therefore we need to prove that $\mathbf{s}_i$ and $\mathbf{s}_j$ are adjacent in B_{l+1} if and only if $\tilde{\mathbf{s}}_i$ and $\tilde{\mathbf{s}}_j$ are adjacent in B_l or $\tilde{\mathbf{s}}_i = \tilde{\mathbf{s}}_j$. In this case, $B_{l+1} = B_l \circ K(q)$. Due to the definition of the lexicographic product (Definition 2), $\mathbf{s}_i$ and $\mathbf{s}_j$ are adjacent in B_{l+1} if and only if (1) $\tilde{\mathbf{s}}_i$ and $\tilde{\mathbf{s}}_j$ are adjacent in B_l or (2) $\tilde{\mathbf{s}}_i = \tilde{\mathbf{s}}_j$ and $s_{i,l+1}$ is adjacent to $s_{j,l+1}$ in $K(q)$. Since all

vertices in $K(q)$ are adjacent to one another, condition (2) simply results in $\mathbf{s}_i$ and $\mathbf{s}_j$ being adjacent in B_{l+1} if $\tilde{\mathbf{s}}_i = \tilde{\mathbf{s}}_j$. Thus, in this case, the required adjacency condition remains true for $l+1$ under the inductive assumption.

The base case of this induction is $l = 1$. If $1 \in V$, then $V_{(1)} = \{1\}$. Since $s_{i,1} \neq s_{j,1}$ for all $\mathbf{s}_i, \mathbf{s}_j \in \mathcal{S}^{(1,q)}$ with $i \neq j$ (i. e., since each length-one sequence represents an element of the alphabet, none of these sequences are equal to one another), the graph B_1 should contain no edges; this is indeed the case because $B_1 = O(q)$ in the case $1 \in V^{[j]}$ due to Equation Eq. S12. Similarly, if $i \notin V^{[j]}$, then $V^{[j]}_{(1)} = \emptyset$ and all vertices of B_1 should be adjacent to one another; this is indeed the case because $B_1 = K(q)$ when $1 \notin V^{[j]}$ due to Equation Eq. S12. □

Proof of Lemma 11. The vertex sets of both $C(V)$ and $B_L(V)$ are given by the space of sequences $\mathcal{S}^{(L,q)}$. By definition, two sequences co-occur in an edge of the hypergraph $G(V)$ if and only if they share the subsequence corresponding to the indices in V. Therefore two sequences are adjacent in the clique multigraph $C(V)$ if and only if they share the subsequence corresponding to the indices in V. Additionally, by Lemma 9, $C(V)$ is a simple graph. By Lemma 10, $B_L(V)$ is a simple graph in which two sequences are adjacent if and only if they share the subsequence corresponding to the indices in V. Thus, the vertex and edge sets of $C(V)$ and $B_L(V)$ are equivalent, and the graphs are equivalent. □

Proof of Lemma 12. Recall from Equation Eq. S6 that $\mathbf{\Phi} = \bigotimes_{i=1}^{L} \mathbf{P}_q$, where $\mathbf{P}_q$ is a complete set of orthonormal eigenvectors of the complete graph $K(q)$. Additionally, recognize that the adjacency matrix of the empty graph, $\mathbf{A}(O(q))$ has every element equal to zero and therefore any nonzero vector is an eigenvector of $\mathbf{A}(O(q))$; for our purposes we will use $\mathbf{P}_q$ as the eigenvectors of $\mathbf{A}(O(q))$. Let the columns of $\mathbf{\Theta}_l$ be orthonormal eigenvectors of the graph $B_l(V)$, which is defined in Lemma 10. Then we have

$$\mathbf{\Theta}_1 = \begin{cases} \mathbf{P}_q & \text{if } 1 \in V \\ \mathbf{P}_q & \text{otherwise,} \end{cases}$$

where the first and second lines on the RHS are due to $\mathbf{P}_q$ being a set of orthonormal eigenvectors of $O(q)$ and $K(q)$, respectively. We additionally have the recursive relation:

$$\mathbf{\Theta}_{l+1} = \begin{cases} \mathbf{\Theta}_l \otimes \mathbf{P}_q & \text{if } i+1 \in V^{[j]} \\ \mathbf{\Theta}_l \otimes \mathbf{P}_q & \text{otherwise,} \end{cases}$$

for $l = 1, 2, ...L-1$, where the first line on the RHS is due to Lemma 1 and the second is due to Lemma 2. Therefore, $\mathbf{\Theta}_L = \bigotimes_{i=1}^{L} \mathbf{P}_q = \mathbf{\Phi}$. The result follows from recognizing that $C(V) = B_L(V)$ by Lemma 11, and therefore $\mathbf{\Theta}_L = \mathbf{\Phi}$ are a set of orthonormal eigenvectors of $\mathbf{A}(V)$. □

Proof of Lemma 13. We will prove this by induction. For ease of notation, we will drop the dependence of the $\mathbf{Z}_l(U)$ matrices on U, and let $\mathbf{Z}_l \leftarrow \mathbf{Z}_l(U)$. Define $\mathbf{A}_l := \mathbf{A}(B_l(V))$ as the adjacency of the graph $B_l(V)$, which is the graph defined by Equations Eq. S11 and Eq. S12. Additionally let $V_{(l)} := \{m \in V : m \leq l\}$ and $U_{(l)} := \{m \in U : m \leq l\}$ be the l smallest elements of V and U, respectively. Our inductive assumption will be that

$$\mathbf{A}_l \mathbf{Z}_l = \mu_l \mathbf{Z}_l,$$

where we define $\mu_l := q^{l-|V_{(l)}|} I(U_{(l)} \subseteq V_{(l)}) - 1$. In other words, we will assume that the columns of $\mathbf{Z}_l$ are eigenvectors of $\mathbf{A}_l$ associated with the eigenvalue μ_l, and then will show that $\mathbf{A}_{l+1}\mathbf{Z}_{l+1} = \mu_{l+1}\mathbf{Z}_{l+1}$. There are four cases to consider: (i) $l+1 \in U$ and $l+1 \in V$, (ii) $l+1 \in U$ and $l+1 \notin V$, (iii) $l+1 \notin U$ and $l+1 \in V$, and (iv) $l+1 \notin U$ and $l+1 \notin V$

(i) ($l+1 \in U$ and $l+1 \in V$). In this case, $V_{(l+1)}$ and $U_{(l+1)}$ add the element $l+1$ to $V_{(l)}$ and $U_{(l)}$, respectively. Therefore, if $U_{(l)} \subseteq V_{(l)}$, it will be true that $U_{(l+1)} \subseteq V_{(l+1)}$, and if $U_{(l)} \nsubseteq V_{(l)}$, then $U_{(l+1)} \nsubseteq V_{(l+1)}$. Additionally, in this case, $|V_{(l+1)}| = |V_{(l)}| + 1$, so we have

$$\begin{aligned}\mu_{l+1} &= q^{l+1-|V_{(l+1)}|} I(U_{(l+1)} \subseteq V_{(l+1)}) - 1 \\ &= q^{l-|V_{(l)}|} I(U_{(l)} \subseteq V_{(l)}) - 1 \\ &= \mu_l\end{aligned}$$

Therefore, we must show that μ_l is the eigenvalue of $\mathbf{A}_{l+1}$ associated with the columns of $\mathbf{Z}_{l+1}$. In this case, $\mathbf{Z}_{l+1} = \mathbf{Z}_l \otimes \tilde{\mathbf{P}}_q$. Also in this case, $B_{l+1} = B_l \square O(q)$ (by Equation S11), so by Equation Eq. S1, $\mathbf{A}_{l+1} = \mathbf{A}_l \otimes \mathbf{I}_q$. Then we have

$$\begin{aligned}\mathbf{A}_{l+1}\mathbf{Z}_{l+1} &= (\mathbf{A}_l \otimes \mathbf{I}_q)(\mathbf{Z}_l \otimes \tilde{\mathbf{P}}_q) \\ &= \mathbf{A}_l\mathbf{Z}_l \otimes \tilde{\mathbf{P}}_q \\ &= \mu_l \mathbf{Z}_l \otimes \tilde{\mathbf{P}}_q \\ &= \mu_l \mathbf{Z}_{l+1},\end{aligned}$$

where the third line results from the inductive assumption. Thus, $\mu_l = \mu_{l+1}$ is the eigenvalue of $\mathbf{A}_{l+1}$ associated with the columns of $\mathbf{Z}_{l+1}$.

(ii) ($l+1 \in U$ and $l+1 \notin V$). In this case, $U_{(l+1)} \nsubseteq V_{(l+1)}$ because the element $l+1$ is in $U_{(l+1)}$ but not $V_{(l+1)}$. Therefore, in this case $\mu_{l+1} = -1$, and we must prove the -1 is the eigenvalue of $\mathbf{A}_{l+1}$ associated with the columns of $\mathbf{Z}_{l+1}$. In this case, $\mathbf{Z}_{l+1} = \mathbf{Z}_l \otimes \tilde{\mathbf{P}}_q$. Also, in this case, $B_{l+1} = B_l \circ K(q)$, so by Equation Eq. S2,

$$\begin{aligned}\mathbf{A}_{l+1} &= \mathbf{A}_l \otimes \mathbf{J}_q + \mathbf{I} \otimes \mathbf{A}(K(q)) \\ &= \mathbf{A}_l \otimes \mathbf{J}_q + \mathbf{I} \otimes (\mathbf{J}_q - \mathbf{I}_q).\end{aligned}$$

Then we have,

$$\begin{aligned}\mathbf{A}_{l+1}\mathbf{Z}_{l+1} &= (\mathbf{A}_l \otimes \mathbf{J}_q + \mathbf{I} \otimes (\mathbf{J}_q - \mathbf{I}_q))(\mathbf{Z}_l \otimes \tilde{\mathbf{P}}_q) \\ &= \mathbf{A}_l\mathbf{Z}_l \otimes \mathbf{J}_q\tilde{\mathbf{P}}_q + \mathbf{Z}_l \otimes \mathbf{J}_q\tilde{\mathbf{P}}_q - \mathbf{Z}_l \otimes \tilde{\mathbf{P}}_q \\ &= \mathbf{A}_l\mathbf{Z}_l \otimes \mathbf{0}_q + \mathbf{Z}_l \otimes \mathbf{0}_q - \mathbf{Z}_l \otimes \tilde{\mathbf{P}}_q \\ &= -\mathbf{Z}_l \otimes \tilde{\mathbf{P}}_q \\ &= -\mathbf{Z}_{l+1},\end{aligned}$$

where the third line results from recognizing that each column of $\tilde{\mathbf{P}}_q$ sums to zero, so $\mathbf{J}_q\tilde{\mathbf{P}}_q = \mathbf{0}_q$, where $\mathbf{0}_q$ is the $q \times q$ matrix of all zeros. Thus, $\mu_{l+1} = -1$ is the eigenvalue of $\mathbf{A}_{l+1}$ associated with the columns of $\mathbf{Z}_{l+1}$.

(iii) ($l+1 \notin U$ and $l+1 \in V$). In this case, the element $l+1$ is in $V_{(l+1)}$ but not $U_{(l+1)}$. Therefore, if $U_{(l)} \subseteq V_{(l)}$, it will be true that $U_{(l+1)} \subseteq V_{(l+1)}$, and if $U_{(l)} \nsubseteq V_{(l)}$, then $U_{(l+1)} \nsubseteq V_{(l+1)}$. Additionally, in this case, $|V_{(l+1)}| = |V_{(l)}| + 1$, so, as in case (i), we have $\mu_{l+1} = \mu_l$, and we must prove the μ_l is the eigenvalue of $\mathbf{A}_{l+1}$ associated with the columns of $\mathbf{Z}_{l+1}$. In this case, $\mathbf{Z}_{l+1} = \mathbf{Z}_l \otimes \mathbf{1}_q$ and $\mathbf{A}_{l+1} = \mathbf{A}_l \otimes \mathbf{I}_q$ (as in case (i)). Then,

$$\begin{aligned}\mathbf{A}_{l+1}\mathbf{Z}_{l+1} &= (\mathbf{A}_l \otimes \mathbf{I}_q)(\mathbf{Z}_l \otimes \mathbf{1}_q) \\ &= \mathbf{A}_l\mathbf{Z}_l \otimes \mathbf{1}_q \\ &= \mu_l \mathbf{Z}_l \otimes \mathbf{1}_q \\ &= \mu_l \mathbf{Z}_{l+1},\end{aligned}$$

where the third line results from the inductive assumption. Thus, $\mu_l = \mu_{l+1}$ is the eigenvalue of $\mathbf{A}_{l+1}$ associated with the columns of $\mathbf{Z}_{l+1}$.

(iv) ($l+1 \notin U$ and $l+1 \notin V$). In this case, $V_{(l+1)} = V_{(l)}$ and $U_{(l+1)} = U_{(l)}$, so $I(U_{(l+1)} \subset V_{(l+1)}) = I(U_{(l)} \subseteq V_{(l)})$ and $|V_{(l+1)}| = |V_{(l)}|$. Therefore,

$$\begin{aligned}
\mu_{l+1} &= q^{l+1-|V_{(l+1)}|} I(U_{(l+1)} \subseteq V_{(l+1)}) - 1 \\
&= q^{l+1-|V_{(l)}|} I(U_{(l)} \subseteq V_{(l)}) - 1 \\
&= q\left(q^{l-|V_{(l)}|} I(U_{(l)} \subseteq V_{(l)})\right) - 1 \\
&= q\left(q^{l-|V_{(l)}|} I(U_{(l)} \subseteq V_{(l)}) - 1\right) + q - 1 \\
&= q\mu_l + q - 1.
\end{aligned}$$

Thus, we must prove that $q\mu_l + q - 1$ is the eigenvalue of $\mathbf{A}_{l+1}$ associated with the columns of $\mathbf{Z}_{l+1}$. In this case, $\mathbf{Z}_{l+1} = \mathbf{Z}_l \otimes \mathbf{1}_q$ (as in case (iii)) and $\mathbf{A}_{l+1} = \mathbf{A}_l \otimes \mathbf{J}_q + \mathbf{I} \otimes (\mathbf{J}_q - \mathbf{I}_q)$ (as in case (ii)). Then, we have

$$\begin{aligned}
\mathbf{A}_{l+1}\mathbf{Z}_{l+1} &= (\mathbf{A}_l \otimes \mathbf{J}_q + \mathbf{I} \otimes (\mathbf{J}_q - \mathbf{I}_q))(\mathbf{Z}_l \otimes \mathbf{1}_q) \\
&= \mathbf{A}_l\mathbf{Z}_l \otimes \mathbf{J}_q\mathbf{1}_q + \mathbf{Z}_l \otimes \mathbf{J}_q\mathbf{1}_q - \mathbf{Z}_l \otimes \mathbf{1}_q \\
&= \mu_l\mathbf{Z}_l \otimes q\mathbf{1}_q + \mathbf{Z}_l \otimes q\mathbf{1}_q - \mathbf{Z}_l \otimes \mathbf{1}_q \\
&= (q\mu_l + q - 1)\,\mathbf{Z}_l \otimes \mathbf{1}_q \\
&= (q\mu_l + q - 1)\,\mathbf{Z}_{l+1},
\end{aligned}$$

where the third line results from the inductive assumption and recognizing that $\mathbf{J}_q\mathbf{1}_q = q\mathbf{1}_q$. Thus, $\mu_{l+1} = (q\mu_l + q - 1)$ is the eigenvalue of $\mathbf{A}_{l+1}$ associated with the columns of $\mathbf{Z}_{l+1}$.

We additionally have four analogous base cases for the induction: (i) $1 \in U$ and $1 \in V$, (ii) $1 \in U$ and $1 \notin V$, (iii) $1 \notin U$ and $1 \in V$, and (iv) $1 \notin U$ and $1 \notin V$:

(i) ($1 \in U$ and $1 \in V$). In this case, $U_{(1)} = V_{(1)} = \{1\}$, so $I(U_{(1)} \subseteq V_{(1)}) = 1$, $|V_{(1)}| = 1$, and therefore $\mu_1 = 0$. Additionally, $\mathbf{A}_1 = \mathbf{A}(O(q)) = \mathbf{0}_q$, so $\mathbf{A}_1\mathbf{Z}_1 = \mu_1\mathbf{Z}_1 = \mathbf{0}_q$.

(ii) ($1 \in U$ and $1 \notin V$). In this case, $V_{(1)} = \emptyset$, so $U_{(1)} \not\subseteq V_{(1)}$, $|V| = 0$, and $\mu_1 = -1$. Additionally, $\mathbf{Z}_1 = \tilde{\mathbf{P}}_q$ and $\mathbf{A}_1 = \mathbf{A}(K(q)) = \mathbf{J}_q - \mathbf{I}_q$, so we have $\mathbf{A}_1\mathbf{Z}_1 = \mathbf{J}_q\tilde{\mathbf{P}}_q - \tilde{\mathbf{P}}_q = -\tilde{\mathbf{P}}_q = -\mathbf{Z}_1$.

(iii) ($1 \notin U$ and $1 \in V$). In this case $U = \emptyset$ and $V = \{1\}$, so $U_{(1)} \subseteq V_{(1)}$, $|V| = 1$, and $\mu_1 = 0$. Since $\mathbf{A}_1 = \mathbf{A}(O(q)) = \mathbf{0}_q$, then $\mathbf{A}_1\mathbf{Z}_1 = \mu_1\mathbf{Z}_1 = \mathbf{0}$.

(iv) ($1 \notin U$ and $1 \notin V$). In this case, $U_{(1)} = V_{(1)} = \emptyset$, so $U_{(1)} \subseteq V_{(1)}$, $|V_{(1)}| = 0$ and thus $\mu_1 = q - 1$. Additionally, $\mathbf{Z}_1 = \mathbf{1}_q$ and $\mathbf{A}_1 = \mathbf{A}(K(q)) = \mathbf{J}_q - \mathbf{I}_q$, so we have

$$\begin{aligned}
\mathbf{A}_1\mathbf{Z}_1 &= \mathbf{J}_q\mathbf{1}_q - \mathbf{1}_q \\
&= (q-1)\mathbf{1}_q \\
&= (q-1)\mathbf{Z}_1.
\end{aligned}$$

Thus, μ_1 is the eigenvalue of $\mathbf{A}_1$ associated with the columns $\mathbf{Z}_1$ in each of these base cases.

By this induction, we have proved that μ_L is the eigenvalue of $\mathbf{A}_L$ associated with the columns of $\mathbf{Z}_L$. It is clear to see that $\mu_L = \mu(U, V)$. The result then follows from recognizing that, due to Lemma 11, $B_L(V) = C(V)$ and therefore $\mathbf{A}_L = \mathbf{A}(V)$. Thus, from the induction, the columns of $\mathbf{Z}_L(U)$ are eigenvectors $A(V)$ associated with the eigenvalue $\mu_L = \mu(U, V)$. □

Proof of Lemma 14. For this proof, recall that the ith row of $\tilde{\mathbf{P}}_q$ encodes the ith element of the alphabet. We will denote each of these encodings as $\mathbf{p}_q(s)$, where s is an element of the alphabet (i. e., each $\mathbf{p}_q(s)$ is

a row of $\tilde{\mathbf{P}}_q$). Let $\hat{\boldsymbol{\phi}}_U^{(L)}(\mathbf{s}_i) \coloneqq \sqrt{q^L}\boldsymbol{\phi}_U(\mathbf{s}_i)$ for $i = 1, 2, ...q^L$ be the unnormalized rows of $\boldsymbol{\Phi}_U$. These can be defined recursively. In particular, we have

$$\hat{\boldsymbol{\phi}}_U^{(l+1)}(\mathbf{s}_i) = \begin{cases} \hat{\boldsymbol{\phi}}_U^{(l)}(\tilde{\mathbf{s}}_i) \otimes \mathbf{p}_q(s_{i,l+1}) & \text{if } l+1 \in U \\ \hat{\boldsymbol{\phi}}_U^{(l)}(\tilde{\mathbf{s}}_i) & \text{otherwise.} \end{cases} \tag{S18}$$

for $l = 1, 2, ..., L$, where $\tilde{\mathbf{s}}_i$ are the first l positions of $\mathbf{s}_i$,

$$\hat{\boldsymbol{\phi}}_U^{(1)}(s_{i,1}) = \begin{cases} \mathbf{p}_q(s_{i,1}) & \text{if } 1 \in U \\ 1 & \text{otherwise,} \end{cases} \tag{S19}$$

and $\boldsymbol{\phi}_U^{(L)}(\mathbf{s}_i) = \boldsymbol{\phi}_U(\mathbf{s}_i)$. We can then recursively define the $\boldsymbol{\Phi}_U$ matrix, by letting

$$\boldsymbol{\Phi}_U^{(l)} \coloneqq \begin{bmatrix} — \ \boldsymbol{\phi}_U^{(l)}(\mathbf{s}_1)^T \ — \\ — \ \boldsymbol{\phi}_U^{(l)}(\mathbf{s}_2)^T \ — \\ \vdots \\ — \ \boldsymbol{\phi}_U^{(l)}(\mathbf{s}_{q^L})^T \ — \end{bmatrix}, \tag{S20}$$

For a given $\tilde{\mathbf{s}} \in \mathcal{S}^{(l,q)}$, there are q sequences in $\mathcal{S}^{(l+1,q)}$ whose first l positions are $\tilde{\mathbf{s}}$. Further, each of these sequences has a unique element in the final position. Thus,

$$\begin{bmatrix} — \ \hat{\boldsymbol{\phi}}_U^{(l+1)}([\tilde{\mathbf{s}}, 1])^T \ — \\ — \ \hat{\boldsymbol{\phi}}_U^{(l+1)}([\tilde{\mathbf{s}}, 2])^T \ — \\ \vdots \\ — \ \hat{\boldsymbol{\phi}}_U^{(l+1)}([\tilde{\mathbf{s}}, q])^T \ — \end{bmatrix} = \begin{cases} \hat{\boldsymbol{\phi}}_U^{(l)}(\tilde{\mathbf{s}}) \otimes \tilde{\mathbf{P}}_q & \text{if } l+1 \in U \\ \hat{\boldsymbol{\phi}}_U^{(l)}(\tilde{\mathbf{s}}) \otimes \mathbf{1}_q & \text{otherwise.} \end{cases} \tag{S21}$$

Now let, $\hat{\boldsymbol{\Phi}}_U^{(l)} \coloneqq \sqrt{q^l}\boldsymbol{\Phi}_U^{(l)}$. Applying Equation Eq. S21 to each row in $\hat{\boldsymbol{\Phi}}_U^{(l,q)}$ results in:

$$\hat{\boldsymbol{\Phi}}_U^{(l+1)} = \begin{cases} \hat{\boldsymbol{\Phi}}_U^{(l)} \otimes \tilde{\mathbf{P}}_q & \text{if } l+1 \in U \\ \hat{\boldsymbol{\Phi}}_U^{(l)} \otimes \mathbf{1}_q & \text{otherwise.} \end{cases} \tag{S22}$$

which is equivalent to the recursion in Equation Eq. S13 that defines $\mathbf{Z}_l(U)$. Additionally, repeated application of Equation Eq. S19 to each element in the alphabet results in the equivalent base case to Equation Eq. S14. Carrying out the recursion of Equation Eq. S21 to $l = L$ then gives
$\mathbf{Z}_L(U) = \hat{\boldsymbol{\Phi}}_U^{(L)} = \sqrt{q^L}\boldsymbol{\Phi}_U^{(L)} = \sqrt{q^L}\boldsymbol{\Phi}_U$. □

S5.4 The sparsity of GNK fitness functions

Here we prove our main result regarding the sparsity of the Fourier coefficients of fitness functions sampled from the GNK model, which is summarized in Equation Eq. 6. First we re-state this result formally.

Theorem 2. *Let $S(f) \coloneqq \#\text{supp}(\boldsymbol{\beta})$ be the sparsity of a fitness function f of sequences of length L and alphabet size q with Fourier coefficients $\boldsymbol{\beta}$, where $\text{supp}(\boldsymbol{\beta})$ is the set of nonzero elements of $\boldsymbol{\beta}$ and $\#$ is the counting measure. Then for any $f \sim GNK(L, q, \mathcal{V})$,*

$$S(f) = \sum_{U \in \mathcal{T}} (q-1)^{|U|} \tag{S23}$$

almost surely, where $\mathcal{T} \coloneqq \bigcup_{j=1}^{L} \mathscr{P}(V^{[j]})$ is the union of the powerset of each neighborhoods.

Proof of Theorem 2. Theorem 1 shows that all the Fourier coefficients associated with an epistatic interaction U are deterministically zero if $U \not\subseteq V^{[j]}$ for $j = 1, 2, ..., L$, which can be alternatively stated as $U \notin \mathscr{P}(V^{[j]})$ for $j = 1, 2, ..., L$. Recalling that $\mathcal{U} \coloneqq \mathscr{P}(\{1, 2, ..., L\})$, the epistatic interactions with nonzero Fourier coefficients are the $U \in \mathcal{U} \coloneqq \mathscr{P}(\{1, ..., L\})$ for which there exists a $j \in 1, 2, ..., L$ such that $U \in \mathscr{P}(V^{[j]})$. Since $\mathscr{P}(V^{[j]}) \subseteq \mathcal{U}$, we have

$$\{U \in \mathcal{U} : U \in \mathscr{P}(V^{[j]})\} = \mathscr{P}(V^{[j]})$$

and further,

$$\begin{aligned}\{U \in \mathcal{U} : \exists j \in \{1, 2, ..., L\} \text{ such that } U \in \mathscr{P}(V^{[j]})\} &= \bigcup_{j=1}^{L} \{U \in \mathcal{U} : U \in \mathscr{P}(V^{[j]})\} \\ &= \bigcup_{j=1}^{L} \mathscr{P}(V^{[j]}).\end{aligned}$$

There are $(q-1)^{|U|}$ Fourier coefficients associated with each $U \in \mathcal{U}$, so letting $\mathcal{T} \coloneqq \bigcup_{j=1}^{L} \mathscr{P}(V^{[j]})$, we have

$$\#\text{supp}(\boldsymbol{\beta}) \geq \sum_{U \in \mathcal{T}} (q-1)^{|U|},$$

where the bound results from recognizing that the RHS sums overs all Fourier coefficients that are *deterministically* zero, but the coefficients with nonzero variances may still equal zero. However, recognizing that each $\beta \in \boldsymbol{\beta}$ with nonzero variance is a normal random variable that can equal zero with zero probability, we have

$$\#\text{supp}(\boldsymbol{\beta}) = \sum_{U \in \mathcal{T}} (q-1)^{|U|}$$

almost surely. □

S5.5 The sparsity of GNK fitness functions with standard neighborhood schemes

Here we prove our results regarding the sparsity of GNK fitness functions with standard neighborhood schemes. In particular, we prove (i) an upper bound on the sparsity of any GNK fitness function with constant neighborhood sizes (Equation Eq. 11 in the main text), (ii) the sparsity of GNK fitness functions with Block Neighborhoods (Equation Eq. 12), (iii) the sparsity of GNK fitness functions with Adjacent Neighborhoods (Equation Eq. 13) and (iv) the expected sparsity of GNK fitness functions with Random Neighborhoods. Below we restate each of these results formally, and provide proofs. We start with the bound of Equation Eq. 11.

Proposition 3. *Let $\mathcal{V}_K$ be a set of neighborhoods where $K_j = K$ for $j = 1, 2, ..., L$ and $1 \leq K \leq L$. Then, the sparsity of any $f \sim GNK(L, q, \mathcal{V}_K)$ is bounded above by:*

$$S(f) \leq 1 + L(q-1) + L \sum_{r=2}^{K} \binom{K}{r} (q-1)^r \tag{S24}$$

Proof. Let $\mathcal{W}_r^{[j]} \coloneqq \{W \in \mathscr{P}(V^{[j]}) : |W| = r\}$ be the number of elements of the powerset of neighborhood j with cardinality r. Additionally define

$$\begin{aligned} n(r) &\coloneqq \left| \bigcup_{j=1}^{L} \mathcal{W}_r^{[j]} \right| \\ &= \#\{W \in \mathcal{T} : |W| = r\} \end{aligned}$$

as the number of elements in the union of powersets with cardinality r. For any set of neighborhoods, we have $n(0) = 1$ and $n(1) = L$. Additionally, for any $\mathcal{V}_K$ with constant neighborhood size K, $n(r) = 0$ for $r > K$. Then, for $r = 2, 3..., K$, we have

$$\begin{aligned} n(r) &= \left| \bigcup_{j=1}^{L} \mathcal{W}_r^{[j]} \right| \\ &\leq \sum_{j=1}^{L} |\mathcal{W}_r^{[j]}| \\ &= \sum_{j=1}^{L} \binom{K}{r} \\ &= L \binom{K}{r}, \end{aligned}$$

where the second line results from the union bound and the third from recognizing that there are $\binom{K}{r}$ sets of cardinality r in the powerset of a set with K elements. Using this within Theorem 2 we then have

$$\begin{aligned} S(f) &= \sum_{U \in \mathcal{T}} (q-1)^{|U|} \\ &= \sum_{r=0}^{L} n(r)(q-1)^r \\ &\leq 1 + L(q-1) + L \sum_{r=0}^{K} \binom{K}{r} (q-1)^r. \end{aligned}$$

□

The next result calculates the sparsity of GNK fitness functions with Block Neighborhoods.

Proposition 4. *Given an L, q, and K satisfying $L \bmod K = 0$ (i. e., K must be set such that L is a multiple of K), define a Block Neighborhood as*

$$V_{BN}^{[j]} = \left\{ j, K \left\lfloor \frac{j-1}{K} \right\rfloor + 1, K \left\lfloor \frac{j-1}{K} \right\rfloor + 2, ..., K \left\lfloor \frac{j-1}{K} \right\rfloor + K \right\}, \tag{S25}$$

where $\lfloor\rfloor$ is the floor operator, and we assume, without loss of generality, that the positions in each block are adjacent. Further let $\mathcal{V}_{BN} = \{V_{BN}^{[j]}\}_{j=1}^{L}$ be a set of L Block Neighborhoods. Then, the sparsity of a fitness function f sampled from $GNK(L, q, \mathcal{V}_{BN})$ is given by:

$$S(f) = \frac{L}{K}(q^K - 1) + 1 \tag{S26}$$

Proof. There are $\frac{L}{K}$ blocks. The blocks are fully connected, so all $\sum_{r=0}^{K} \binom{K}{r}(q-1)^r = q^K$ Fourier coefficients corresponding to intra-block epistatic interactions are nonzero. The only epistatic interaction shared by the blocks is the zeroth order interaction, so each block contributes $(q^K - 1)$ unique nonzero Fourier coefficients, and the total number of nonzero Fourier coefficients is given by Equation Eq. S26, where the final addition of one is due to the shared zeroth order interaction. □

Similarly, the sparsity of GNK fitness with Adjacent Neighborhoods is shown in the following proposition.

Proposition 5. *Given an L, q, and $K \leq \frac{L}{2}$, define an Adjacent Neighborhood as*

$$V_{AN}^{[j]} = \{j, a_j \bmod L+1, (a_j+1) \bmod L+1, ..., (a_j+K) \bmod L+1\}, \tag{S27}$$

where we define $a_j := j - \lfloor \frac{K-1}{2} \rfloor - 1$. Further let $\mathcal{V}_{AN} = \{V_{AN}^{[j]}\}_{j=1}^{L}$ be a set of L Adjacent Neighborhoods. Then, the sparsity of any $f \sim GNK(L, q, \mathcal{V}_{AN})$ is given by:

$$S(f) = 1 + L \sum_{r=1}^{K} \binom{K-1}{r-1} (q-1)^r \tag{S28}$$

Proof. Define $\mathcal{W}_r^{[j]} := \{W \in \mathscr{P}(V^{[j]}) : |W| = r\}$ and

$$n_l(r) := \left| \bigcup_{j=1}^{l} \mathcal{W}_r^{[j]} \right|$$

for $l = 1, 2, ..., L$. For $l \leq L - K + 1$, and $r = 1, 2, ..., K$, we have

$$n_l(r) = l \binom{K}{r} - (l-1) \binom{K-1}{r}. \tag{S29}$$

This can be shown by induction. In particular, assume Equation Eq. S29 is correct for $l < L - K + 1$ and then we find:

$$\begin{aligned}
n_{l+1}(r) &= \left| \bigcup_{j=1}^{l+1} \mathcal{W}_r^{[j]} \right| \\
&= \left| \left(\bigcup_{j=1}^{l} \mathcal{W}_r^{[j]} \right) \cup \mathcal{W}_r^{[l+1]} \right| \\
&= \left| \bigcup_{j=1}^{l} \mathcal{W}_r^{[j]} \right| + \left| \mathcal{W}_r^{[l+1]} \right| - \left| \left(\bigcup_{j=1}^{l} \mathcal{W}_r^{[j]} \right) \cap \mathcal{W}_r^{[l+1]} \right| \\
&= n_l(r) + \binom{K}{r} - \binom{K-1}{r} \\
&= (l+1) \binom{K}{r} - l \binom{K-1}{r},
\end{aligned}$$

where the fourth line results from recognizing that $V^{[l+1]}$ when $l + 1 \leq L - K + 1$ contains exactly one position that is not in $\bigcup_{j=1}^{l} V^{[j]}$; there are then $\binom{K-1}{r-1}$ sets in $\mathcal{W}_r^{[l+1]}$ that contain this element and are thus unique to $\mathcal{W}_r^{[l+1]}$, which leads to

$$\left| \left(\bigcup_{j=1}^{l} \mathcal{W}_r^{[j]} \right) \cap \mathcal{W}_r^{[l+1]} \right| = \binom{K}{r} - \binom{K-1}{r-1} = \binom{K-1}{r}.$$

It is clear that $n_1(r) = \binom{K}{r}$, and thus Equation Eq. S29 is proved by induction for $l \leq L - K + 1$.

Equation Eq. S29 accounts for redundancies in $\mathcal{W}_r^{[j]}$ that result from overlapping positions in the neighborhoods, without considering periodicity. There are additional redundancies that occur when $l > L - K + 1$ due to the periodicity of the neighborhoods. In particular, for $l = L - K + 2, ..., L$, due to periodicity $V^{[l]}$ contains $(l + k) \bmod L - 1$ additional positions that are already in $\bigcup_{j=1}^{l} V^{[j]}$ (outside

of those that are already $\bigcup_{j=1}^{l} V^{[j]}$ due to non-periodic overlap). Therefore $\mathcal{W}_r^{[l]}$ contains $\binom{(l+k) \bmod L-1}{r}$ additional sets that are already in $\bigcup_{j=1}^{l-1} \mathcal{W}_r^{[j]}$ due to periodicity. Then we have, for $l = L-K+2, ..., L$:

$$n_l(r) = l\binom{K}{r} - (l-1)\binom{K-1}{r} - \binom{(l+k) \bmod L-1}{r}.$$

At $l = L$, we then have

$$\begin{aligned} n_L(r) &= L\binom{K}{r} - (L-1)\binom{K-1}{r} - \binom{(L+k) \bmod L-1}{r} \\ &= L\binom{K}{r} - (L-1)\binom{K-1}{r} - \binom{K-1}{r} \\ &= L\left(\binom{K}{r} - \binom{K-1}{r}\right) \\ &= L\binom{K-1}{r-1}. \end{aligned}$$

The result follows from recognizing that $n_L(0) = 1$, and therefore

$$\begin{aligned} S(f) &= \sum_{U \in \mathcal{T}} (q-1)^{|U|} \\ &= \sum_{r=0}^{L} n_L(r)(q-1)^r \\ &= 1 + L\sum_{r=1}^{K} \binom{K-1}{r-1}(q-1)^r. \end{aligned}$$

□

Additionally, we are able to calculate the expected sparsity of GNK fitness functions with Random Neighborhoods, which is shown in the following result. The proof of this follows the analogous calculations of ref. 27, and we correct a mistake in their calculations.

Proposition 6. *Let $V_{RN}^{[j]}$ be a set of cardinality K, where the first $K-1$ elements are selected uniformly at random from $\{1, 2, ..., L\}$ without replacement, and the final element is j. Let $\mathcal{V}_{RN} = \{V_{RN}^{[j]}\}_{j=1}^{L}$ be a collection of such sets. Then, the expected sparsity of a fitness function f sampled from $GNK(L, q, \mathcal{V}_{RN})$, with the expectation taken over the possible realizations of $\mathcal{V}_{RN}$, is given by:*

$$\mathbb{E}_{\mathcal{V}_{RN}}[S(f)] = \sum_{r=0}^{K} \binom{L}{r} p(r)(q-1)^r, \tag{S30}$$

where

$$p(r) = 1 - (1-\alpha(r))^r \left(1 - \alpha(r)\frac{K-r}{L-r}\right)^{L-r}, \tag{S31}$$

and $\alpha(r) = \frac{(K-1)!}{(L-1)!}\frac{(K-r)!}{(L-r)!}$.

Proof. Consider a set $W \subseteq \{1, 2, ..., L\}$ of cardinality r. Define $\alpha(r)$ as the the probability that W is a subset of the random neighborhood $V^{[j]}$ given that $j \in W$, which is given by

$$\begin{aligned} \alpha(r) &\coloneqq \Pr(W \subseteq V^{[j]} | j \in W) \\ &= \frac{\binom{L-r}{K-r}}{\binom{L-1}{K-1}} \\ &= \frac{(L-r)!}{(K-r)!}\frac{(K-1)!}{(L-1)!}, \end{aligned}$$

where $\binom{L-1}{K-1}$ is the total number of ways to construct $V^{[j]}$ and $\binom{L-r}{K-r}$ is the number of ways to construct $V^{[j]}$ such that every element of W is in $V^{[j]}$. The probability that W is a subset of the random neighborhood $V^{[j]}$ given that $j \notin W$ is similarly given by:

$$\begin{aligned}\Pr(W \subseteq V^{[j]} | j \notin W) &= \frac{\binom{L-r-1}{K-r-1}}{\binom{L-1}{K-1}} \\ &= \frac{(L-r-1)!}{(K-r-1)!}\frac{(K-1)!}{(L-1)!} \\ &= \alpha(r)\frac{K-r}{L-r}\end{aligned}$$

There are r neighborhoods $V^{[j]}$ for which $j \in W$, and $L-r$ neighborhoods for which $j \notin W$. Define $p(r)$ as the probability that W is a subset of at least one, which is then:

$$\begin{aligned}p(r) &\coloneqq \Pr(\exists j : W \subseteq V^{[j]}) \\ &= 1 - \Pr(\nexists j : W \subseteq V^{[j]}) \\ &= 1 - \Pr(\nexists j \in W : W \subseteq V^{[j]})\Pr(\nexists j \notin W : W \subseteq V^{[j]}) \\ &= 1 - (1-\alpha(r))^r \left(1 - \alpha(r)\frac{K-r}{L-r}\right)^{L-r}\end{aligned}$$

There are $\binom{L}{r}$ sets of cardinality r; in expectation $\binom{L}{r}p(r)$ will be subsets of at least one neighborhood, and will therefore represent epistatic interactions or order r corresponding to $(q-1)^r$ nonzero Fourier coefficients. Equation Eq. S30 follows from summing over all possible cardinalities r. □

S6 Extension to non-constant alphabet sizes

It is common that alphabet sizes are not constant at every position. Here we generalize our formal results to the case of "hybrid alphabets", where alphabet sizes may differ at each position. Consider the case where the alphabet size at each position is given by the length L vector $\mathbf{q} = [q_1, q_2, ..., q_L]$ and let $\mathcal{S}^{(L,\mathbf{q})}$ be the space of all sequences corresponding to the alphabet sizes in $\mathbf{q}$. Denote as $H(L, \mathbf{q})$ the Generalized Hamming graph whose vertex set is $\mathcal{S}^{(L,\mathbf{q})}$ and whose edges connect sequences that differ in exactly one position [56]. The Generalized Hamming graph can be constructed as an L-fold graph Cartesian product:

$$H(L, \mathbf{q}) = \square_{i=1}^{L} K(q_i) \tag{S32}$$

We then have the following result for the Fourier basis corresponding to these Generalized Hamming graphs, which follows straightforwardly applying Lemma 3 to Equation Eq. S32.

Proposition 7. *An orthonormal set of eigenvectors of the Graph Laplacian of the Generalized Hamming graph $H(L, \mathbf{q})$ is given by*

$$\mathbf{\Phi}^{(\mathbf{q})} \coloneqq \bigotimes_{i=1}^{L} \mathbf{P}_{q_i}, \tag{S33}$$

where $\mathbf{P}_{q_i}$ is defined in Equation Eq. S5.

In this basis, each epistatic interaction U is represented by $\prod_{k \in U}(q_k - 1)$ columns of $\mathbf{\Phi}^{(\mathbf{q})}$.

The Fourier basis of Equation Eq. S33 can be used to represent fitness functions of sequences with non-constant alphabet sizes, $\mathbf{q}$. The GNK model can also be defined analogously to the definition given in the Materials and Methods section for the case of non-constant alphabet sizes; in particular let $\text{GNK}(L, \mathbf{q}, \mathcal{V})$ be the distribution over fitness functions of sequences of length L with non-constant alphabet sizes given

by $\mathbf{q}$ and neighborhood set $\mathcal{V}$. We then have the following results regarding the distribution and support of the Fourier coefficients of fitness functions sampled from this distribution. We present these results without proof, though it is straightforward to see how to adapt the proof of Theorems 1 and 2 to prove these.

Theorem 3. *Let* $\mathbf{f} = (f(\mathbf{s}))_{\mathbf{s}\in\mathcal{S}^{(L,\mathbf{q})}}$ *be the complete vector of evaluations of a fitness function* $f \sim GNK(L, \mathbf{q}, \mathcal{V})$*. Then the Fourier coefficients of* f*,* $\boldsymbol{\beta} = (\boldsymbol{\Phi}^{(\mathbf{q})})^T\mathbf{f}$*, are distributed according to* $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\lambda}\mathbf{I})$*. Let* $\boldsymbol{\beta}_U$ *be the length* $\prod_{k\in U}(q_k - 1)$ *sub-vector of* $\boldsymbol{\beta}$ *representing the epistatic interaction* U*. Then the variance of every element of* $\boldsymbol{\beta}_U$ *is given by:*

$$\lambda_U = \left(\prod_{i=1}^{L} q_i\right) \sum_{j=1}^{L} \left(\prod_{k\in V^{[j]}} \frac{1}{q_k}\right) I\left(U \subseteq V^{[j]}\right), \tag{S34}$$

where $I\left(U \subseteq V^{[j]}\right)$ *is an indicator function that is equal to one if* U *is a subset of or equal to* $V^{[j]}$ *and zero otherwise.*

Theorem 4. *The sparsity of any* $f \sim GNK(L, \mathbf{q}, \mathcal{V})$ *is given by*

$$S(f) = \sum_{U\in\mathcal{T}} \prod_{k\in U} (q_k - 1), \tag{S35}$$

almost surely.