Title

On the sparsity of fitness functions and implications for learning

Journal

Proceedings of the National Academy of Sciences of the United States of America, 119(1)

ISSN

0027-8424

Authors

Brookes, David H
Aghazadeh, Amirali
Listgarten, Jennifer

Publication Date

2022-01-05

DOI

10.1073/pnas.2109649118

Peer reviewed

On the sparsity of fitness functions and implications for learning

David H. Brookes[a], Amirali Aghazadeh[b], and Jennifer Listgarten[b,c,1]

[a]Biophysics Graduate Group, University of California, Berkeley, CA 94720; [b]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720; and [c]Center for Computational Biology, University of California, Berkeley, CA 94720

Edited by Günter Wagner, Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT; received May 25, 2021; accepted November 11, 2021

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 fitness datasets available to learn these functions are typically small relative to the large combinatorial space of sequences; characterizing how much data are 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 develop a framework to 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 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. We validate our framework by demonstrating that GNK models with parameters set according to structural considerations can be used to accurately approximate the number of samples required to recover two empirical protein fitness functions and an RNA fitness function. In addition, we show that these GNK models identify important higher-order epistatic interactions in the empirical fitness functions using only structural information.

fitness functions | compressed sensing | epistasis | protein structure

Advances in high-throughput experimental technologies now allow for the probing of the fitness of thousands, and sometimes even millions, of biological sequences. However, even in these high-throughput scenarios, the number of measurements generally represents only a tiny fraction of those required to comprehensively characterize a fitness function. Thus, methods to estimate fitness functions from incomplete measurements are necessary. Many such methods have been proposed, 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 (9–11) and are crucial components of in silico sequence design tools (12–15).

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 scope of an experimental probe such that the resulting data allow one to accurately estimate the function under study. Our central focus herein is to make progress on answering 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 (WH) basis, which encodes fitness functions in terms of all possible "epistatic" interactions (i.e., nonlinear contributions to fitness due to interacting sequence positions) (3, 16, 17). Further, this sparsity property has been exploited to improve estimators of such functions (18–20). 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 subfield 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

Significance

The properties of proteins and other biological molecules are encoded in large part in the sequence of amino acids or nucleotides that defines them. Increasingly, researchers estimate functions that map sequences to a particular property using machine learning and related statistical approaches. However, an important question remains unanswered: How many experimental measurements are needed in order to accurately learn these "fitness" functions? We leverage perspectives from the fields of biophysics, evolutionary biology, and signal processing to develop a theoretical framework that enables us to make progress on answering this question. We demonstrate that this framework can be used to make useful calculations on real-world data and suggest how these calculations may be used to guide experiments.

Author contributions: D.H.B. and A.A. designed research; D.H.B. performed research; and D.H.B., A.A., and J.L. wrote the paper.

Competing interest statement: J.L. is on the Scientific Advisory Board for Foresite Labs and Patch Biosciences.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: jennl@berkeley.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2109649118/-/DCSupplemental.

Published December 22, 2021.

fitness, which assign fitness values to sequences based on stochastic processes constructed to mimic the statistical properties of natural fitness functions (23, 24). We 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 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,* 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 random (23). NK models have been shown to model a number of properties of empirical fitness functions, including fitness correlation functions (26, 27) and adaptive walk statistics (25, 28, 29). The Generalized NK (GNK) model (30) extends the model by allowing neighborhoods to be of arbitrary size and content. We refer to simulated fitness functions sampled from the GNK model as "GNK fitness functions."

Buzas and Dinitz (30) 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 (31) 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 and to do so in the biologically relevant scenarios where sequences are made up of nonbinary elements (e.g., nucleotide or amino acid alphabets). To achieve this, we extend the results of refs. 30 and 31 to the case of nonbinary 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 measurements required to recover GNK fitness functions in the Fourier basis. This framework of using CS theory in tandem with the GNK model allows us to test the effects of sequence length, alphabet size, and interaction structure on the sample complexity of estimating GNK fitness functions.

We validate the practical utility of our framework by demonstrating that suitably parameterized GNK models can accurately approximate the sparsity of several empirical landscapes, and, thus, we can successfully leverage our sample complexity results to determine how many measurements are needed to accurately estimate these landscapes. In particular, we use GNK models that incorporate structural information to show this for two empirical protein landscapes and one "quasi-empirical" RNA landscape. Our analysis also demonstrates that structure-based GNK models correctly identify many of the important higher-order epistatic interactions in the corresponding empirical fitness functions, despite using only pairwise structural contact information. This insight bolsters a growing understanding of the importance of structural contacts in shaping fitness functions.

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

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 (32). 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 $q=20$ 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 be mapped to a common alphabet of integers. 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 *SI Appendix*, 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

$$\mathbf{f}=\mathbf{\Phi}\boldsymbol{\beta}, \qquad [1]$$

where $\mathbf{f}$ is the vector of all q^L fitness values (one for each unique 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 WH 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 $\mathcal{S}^{(L,q)}$.

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. Since $\mathbf{\Phi}$ is an orthogonal matrix, this is equivalent to the MSE between the true fitness values $\mathbf{f}$ and those predicted by using the estimated coefficients.

The field of CS 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 Least Absolute Shrinkage and Selection Operator (LASSO) algorithm is among the most widely used and well-studied for solving underdetermined systems, both in CS and also in machine learning (33). 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 (34) states that if $\boldsymbol{\beta}$ is an S-sparse vector (meaning that it has exactly S nonzero elements), then with high probability, LASSO can recover $\boldsymbol{\beta}$ exactly with

$$N\geq C\cdot S\log q^L, \qquad [2]$$

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 at random from the space of sequences (34). 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

*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.

the fitness function is expanded in a particular basis.[†] Sparsity is defined with respect to a particular basis, which must therefore be chosen carefully. In *Fourier Bases for Fitness Functions*, 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 WH basis (3, 16, 17), which has also been used extensively in theoretical studies of fitness landscapes (27, 35–37) and even to unify multiple definitions of epistasis (38). The WH basis can be interpreted as encoding fitness functions in terms of epistatic interactions (38, 39). In particular, when a fitness function of binary sequences of length L is represented in the form of Eq. **1** (with $\mathbf{\Phi}$ being the WH basis), then the sequence elements are encoded as $\{-1, 1\}$, and the fitness function evaluated on a sequence, $\mathbf{s} = [s_1, s_2, \ldots, s_L]$, has the form of an intuitive multilinear polynomial (20),

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

where $\mathcal{U} := \mathscr{P}(\{1, 2, \ldots, L\})$ 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 r^{th}-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*; refs. 40 and 41). 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 r^{th}-order 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, 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

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

It is shown in *Results* that 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 GNK 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, $\boldsymbol{\beta}$.

Herein, we focus on the GNK random field model (30). 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 length three, 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 neighborhoods also imply the presence of higher-order epistatic interactions.

The key choice in specifying the GNK model is in how the neighborhoods are constructed. We will consider three "standard" schemes for constructing neighborhoods (31, 36): the Random, Adjacent, and Block Neighborhood schemes. These schemes all restrict every neighborhood to be the same size, K, which allows for a direct comparison of 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 scheme where neighborhoods are

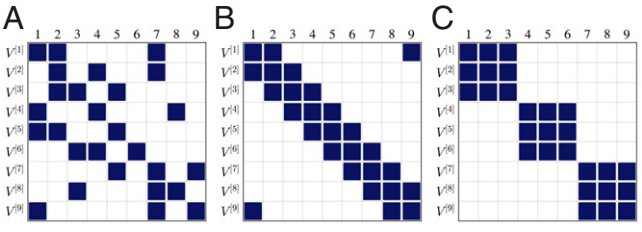

Fig. 1. Graphical depictions of GNK neighborhood schemes for $L = 9$ and $K = 3$. In each grid, rows represent neighborhoods, and columns represent sequence positions. A filled square in the $(i, j)^{\text{th}}$ position in the grid denotes that sequence position j is in the neighborhood $V^{[i]}$. (*A*) Random neighborhoods. (*B*) Adjacent neighborhoods. (*C*) Block neighborhoods.

[†]In a quirk of common terminology, a signal is considered 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.

constructed based on contacts between residues in an atomistic protein structure, which is described in more detail in *Results*.

Notably, the GNK model is an example of a spin glass, a popular model in statistical physics, with different neighborhood schemes corresponding to different types of spin glasses (42). Further, the recovery of sparse spin-glass Hamiltonians has been investigated in some depth (43).

In *Results*, we present 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 *SI Appendix*.

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 r^{th} order epistatic interaction, U, have equal variances given by

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

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 coefficient is 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 function f sampled from a GNK model is given by

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

where $\mathcal{T} := \bigcup_{j=1}^{L} \mathscr{P}(V^{[j]})$ is the union of the power sets 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 *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 Random neighborhoods. We also provide an upper bound on sparsity of GNK fitness functions with any neighborhood scheme with constant neighborhood size, K. In Fig. 2 *A* and *B*, we plot this upper bound for a variety of settings of L, q, and K. Further, in Fig. 2*C*, 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 needed 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 analyzed these estimates to determine 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. 2*D* 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. A more detailed analysis of these experiments is shown in *SI Appendix*, Fig. S3, which makes clear

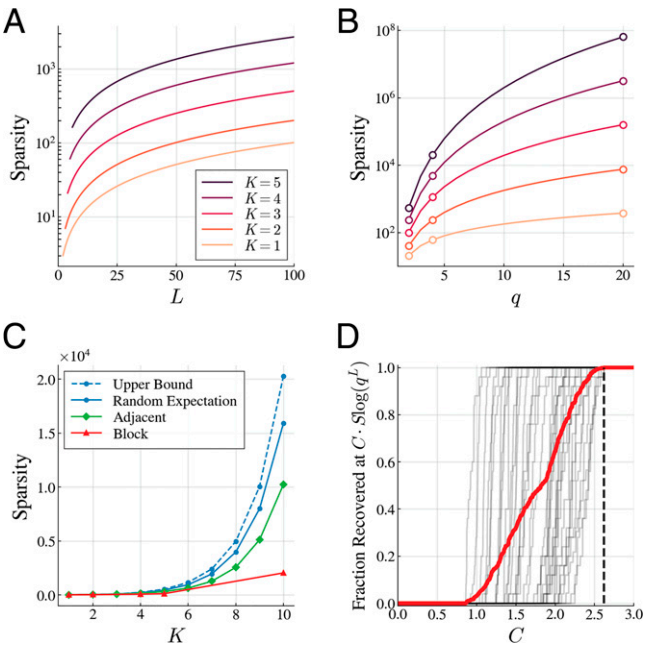

Fig. 2. The sparsity of GNK fitness functions. (*A*) Upper bound on sparsity of GNK fitness functions with constant neighborhood sizes for $q = 2$ and a range of settings of the L and K parameters. (*B*) Upper bound for $L = 20$ and a range of settings of the alphabet size q and the K parameter (colors 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 $L = 20$, $q = 2$, 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 gray curve represents sampled fitness functions at a particular value of $L \in \{5, 6, \ldots, 13\}$, $q \in \{2, 3, 4\}$, 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.

that the minimum possible setting of C is a function of L, q, and K and, therefore, that $C = 2.62$ may be a conservative setting for certain reasonable settings of these parameters.

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. Fig. 3 shows 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 much smaller than both the total size of sequence space and the total number of possible interactions in the fitness functions. Consider, for instance, the $K = 5$ curve in Fig. 3*A* at $L = 50$; in this case, the size of sequence space is $2^{50} \approx 10^{15}$, and the total number of interactions is $\sum_{r=0}^{5} \binom{50}{r} \approx 2 \times 10^6$, while the number of measurements required to recover these fitness functions is about 5×10^4.

Additionally, comparing Fig. 3 *A* and *B* 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.

Analysis of Empirical Protein Fitness Functions. In order to validate our framework, we next tested the extent to which our results could be used to predict the sample complexity of estimating empirical protein fitness functions. To do so, we made use of a scheme for constructing GNK neighborhoods that uses information derived from the three dimensional structure of a given protein, which we call the Structural neighborhood scheme. In particular, Structural neighborhoods are constructed based on contacts between amino acid residues in a given atomistic protein structure, where, following refs. 4 and 44, we define two residues to be in contact if any two atoms in the residues are within 4.5 Å 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 r^{th}-order epistatic interaction has nonzero Fourier coefficients when $r - 1$ positions are in structural contact with a central position.

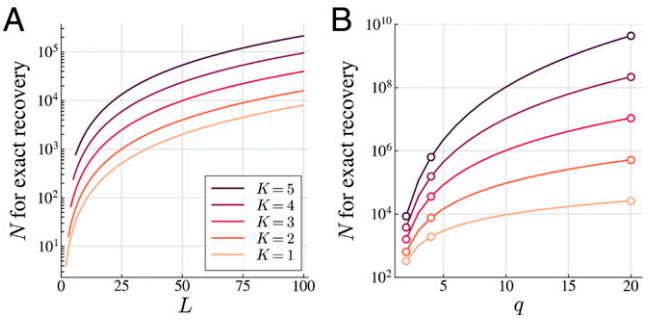

Fig. 3. Minimum number of measurements required to exactly recover GNK fitness functions with constant neighborhood sizes. (*A*) Upper bound on the minimum *N* required to recover GNK fitness functions with constant neighborhood sizes for $q = 2$ and a range of settings of the *L* and *K* parameters. (*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.

We instantiated GNK models with Structural neighborhoods for two proteins: the TagBFP fluorescent protein (45) and the protein encoded by the His3 gene in *Saccharomyces cerevisiae* (His3p). We then used the results described in *Exact Recovery of GNK Fitness Functions* to calculate the sparsity of GNK fitness functions with these Structural neighborhoods, the variance of each the functions' Fourier coefficients, and the sample complexity of estimating these functions.

Both TagBFP 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 TagBFP structure, the associated empirical fitness function contains functional observations (blue fluorescence brightness) of mutations to the mTagBFP2 protein (18), which is closely related to TagBFP, but has no available structure. This data contain measurements for all combinations of two possible amino acids in 13 positions, (i.e., $L = 13$ and $q = 2$), yielding $2^{13} = 8,192$ total fitness observations. A graphical depiction of the Structural neighborhoods associated with these 13 positions is shown in Fig. 4 *A*, *Top*. 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 nonzero Fourier coefficients and the epistatic interactions to which they corresponded.

For the case of His3p, we used a nearly combinatorial complete empirical fitness function that is embedded in the data of ref. 46. In particular, the data contain 2,030 out of the possible 2,048 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 (47) predicted structure of His3p (46) (Fig. 4 *A*, *Middle*), which resulted in sparsity $S(f) = 76$ for GNK fitness functions with these neighborhoods. We again computed the distribution of these coefficients and determined the corresponding epistatic interactions.

The comparisons of the mTagBFP2 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 5^{th} and 6^{th}-order epistatic interactions are shown in Fig. 4*B* for the TagBFP and His3p cases, respectively. Many of the epistatic interactions with the largest empirical coefficients also have nonzero coefficients in the GNK model with Structural neighborhoods. In *SI Appendix*, we quantify the overlap between the largest coefficients in the empirical and GNK fitness functions. Although the expected magnitudes of the GNK coefficients do not necessarily approximate the magnitudes of the empirical coefficients (*SI Appendix*, Fig. S9), the coefficients with nonzero variance in the GNK model have significantly higher ranks in the empirical coefficients than those that are deterministically zero in the GNK model (*SI Appendix*, Figs. S10–S13).

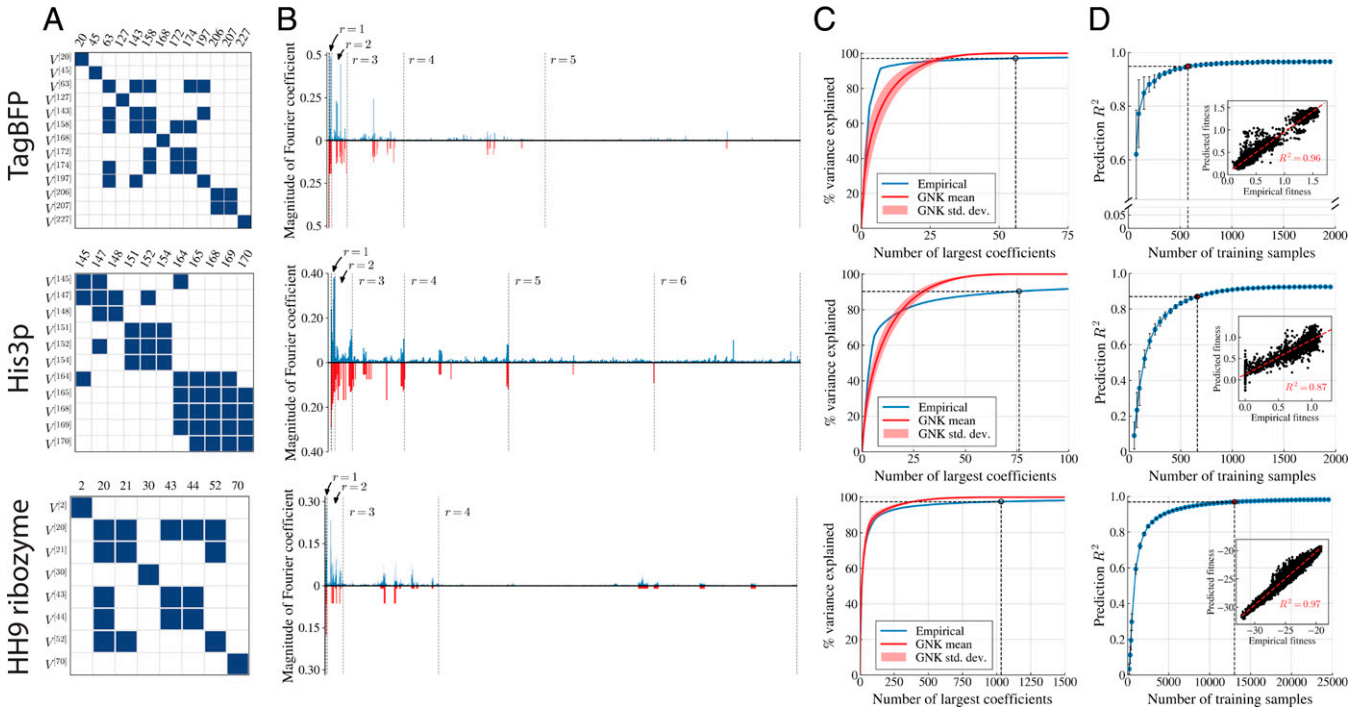

Fig. 4. Comparison of empirical fitness functions to GNK models with Structural neighborhoods. (*Top*) Comparison to mTagBFP2 fitness function from ref. 18. (*Middle*) Comparison to His3p fitness function from ref. (46). (*Bottom*) Comparison to quasi-empirical fitness function of the Hammerhead ribozyme HH9. (*A*) Structural neighborhoods derived from crystal structural of TagBFP (*Top*), I-TASSER predicted structure of His3p (*Middle*), and predicted secondary structures of the Hammerhead Ribozyme HH9 (*Bottom*). (*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 r^{th}-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 in *Top*, 76 in *Middle*, and 1,033 in *Bottom*, at which points 97.1%, 90.4%, and 97.5% of the empirical variances are explained, respectively. Std. dev., SD. (*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 were 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 *Top*, $N = 660$ in *Middle*, and $N = 13,036$ in *Bottom*) is highlighted with a red dot and dashed lines; at this number of samples, the mean prediction R^2 is 0.948 in *Top*, 0.870 in *Middle*, and 0.969 in *Bottom*. Error bars indicate the SD of R^2 over training replicates. *D*, *Insets* show paired plots between the estimated and predicted fitness function for one example training set of size $N = 575$ (*Top*), $N = 660$ (*Middle*), and $N = 13,036$ (*Bottom*).

Fig. 4 *B*, *Top* and *Middle* display a number of false positives (i.e., coefficients that have nonzero variance in the GNK model, but are very small in the empirical fitness function) and false negatives that deserve some comment. To explain these errors, it is first important to remember that the red bars in Fig. 4*B* represent the expected magnitudes of zero-mean GNK coefficients; even among fitness functions sampled directly from the GNK model, we would expect to see "false positives" where the sampled magnitudes were smaller than the expected magnitudes. The false negatives may be explained by three similar causes, all regarding the insufficiency of using a single crystal or predicted structure to construct Structural neighborhoods for proteins. First, the structures we used may simply be inaccurate: In one case, we used the TagBFP crystal structure, while the fitness function reports on mutations to mTagBFP2; in the His3p case, we used an I-TASSER predicted structure that may have inaccuracies. Secondly, static structures do not capture dynamical effects that may impact fitness; for instance, two residues may be in contact in a nonnative conformation of the protein that differs from the crystallized or predicted conformation. Finally, the crystal or predicted structures of wild-type proteins cannot capture the potential structural changes that may occur when the protein is mutated, as is done to collect fitness data. Additionally, we used a fixed contact threshold of 4.5 Å, but adjusting this threshold can moderately change the GNK Fourier coefficients (*SI Appendix*, Figs. S4–S7); most notably, the largest empirical $r = 6$ coefficient in the His3p fitness function has nonzero variance in the GNK model when the cutoff distance is increased to 7 Å.

None of the empirical Fourier coefficients are exactly zero; however, 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 mTagBFP2 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. 4*C* 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 SD, 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 TagBFP 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 $S = 76$ in the

TagBFP 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. 4*D*). Our framework predicts that 548 and 630 samples are minimally needed for exact recovery of the GNK fitness functions with TagBFP 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 mTagBFP2 and His3p fitness functions, respectively.

In *SI Appendix*, we show analogous results to those in Fig. 4 for another nearly complete subset of the His3p fitness data of ref. 46 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. All together, these results suggest that 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.

Analysis of a Quasi-empirical RNA Fitness Function. As further validation, we next tested the ability of our framework to predict the sample complexity of estimating a quasi-empirical RNA landscape. In particular, we studied the fitness function of all possible mutations to the *Erinaceus Europaes* Hammerhead ribozyme HH9 wild-type sequence (Rfam accession no. AANN01066007.1) at positions 2, 20, 21, 30, 43, 44, 52, and 70, where the fitness of each sequence in this $L = 8$, $q = 4$ sequence space is given by the Minimum Free Energy of the secondary structures associated with the sequence, as calculated by the ViennaRNA package (48). We follow ref. 49 in referring to this as a quasi-empirical fitness function, as it is constructed from an established physical model rather than direct experimental measurements. The magnitudes of the Fourier coefficients associated with this fitness function are shown as blue bars in Fig. 4 *B*, *Bottom.* This is a sparse landscape, with the largest 150 out of 65,536 possible coefficients explaining over 90% of the quasi-empirical variance.

We then used a GNK model with RNA-specific Structural neighborhoods to predict the sample complexity of estimating this quasi-empirical landscape. In order to construct these neighborhoods, we first used ViennaRNA to sample 10,000 secondary structures from the Boltzmann ensemble of structures associated with the wild-type sequence. We then built neighborhoods where a position j was included in the neighborhood of position k if 1) j and k were directly adjacent in the sequence or 2) j and k were paired in any of the sampled secondary structures (Fig. 4 *A*, *Bottom*). The expected magnitudes of the Fourier coefficients in the GNK model with these neighborhoods are shown as red bars in Fig. 4*B*. Once again, we see that the GNK model with Structural neighborhoods identifies many of the most important higher-order epistatic interactions in this fitness function.

As with the empirical protein fitness functions, we compared the sparsity of the GNK and quasi-empirical fitness functions (Fig. 4 *C*, *Bottom*) and tested the ability of our framework to predict the sample complexity of estimating the quasi-empirical fitness function with LASSO (Fig. 4 *D*, *Bottom*). These results demonstrate that a suitably parameterized GNK model can accurately model the sparsity of a realistic RNA fitness function, which bolsters our results on empirical protein fitness functions and further suggests that the GNK model can be a practical tool for estimating the sample complexity of fitness function estimation.

Discussion

By leveraging perspectives from the fields of CS and evolutionary biology, we developed a framework 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 1) the fitness functions are sampled from a specified GNK model; 2) fitness measurements are noiseless; 3) fitness measurements correspond to sequences sampled uniformly at random from the space of sequences; and 4) the fitness functions are represented in the Fourier basis. Under these assumptions, our results allow 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 (1) and (2) 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 a quasi-empirical RNA fitness function 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 framework to these fitness functions, which are neither exactly sparse nor noiseless (in the case of the protein fitness functions), 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 should be noted that assumptions (3) and (4) 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 guide experimental probes of fitness by suggesting how one might consider restricting the scope of mutagenesis such that the resulting data can likely be used to accurately estimate the fitness function under study. For example, one might restrict the number of mutated positions, informed by biophysical considerations (3, 50) or previous experimental results (51–54). Alternatively, one might restrict the alphabet of amino acids that are allowable at a position, for instance, by choosing only amino acids present in homologous sequences (17, 18, 46). Of course, one should take care not to minimize the sample-size requirements at the expense of probing important areas of the protein or nucleotide sequences under study.

Few attempts have been made at understanding how many measurements are required to estimate fitness functions, despite the practical importance of this question for experimental design. By making the connection between this question and the known sparsity of fitness functions in certain bases, we provide a much-needed framework for probing the sample complexity of estimating fitness functions. Further, we show that the GNK model, given protein and RNA structural information, can gauge the sparsity of empirical fitness functions enough to make useful statements about the sample complexity of estimating such

functions. As biotechnology progresses to reveal more complete and larger empirical landscapes, the tools and theoretical frameworks to analyze sample complexity may have to correspondingly progress; our work provides a solid foundation on which to do so.

Materials and Methods

Compressed Sensing. As described in *Fitness Functions and Estimation*, the fitness function estimation problem is to solve the underdetermined linear system $\mathbf{y} = \mathbf{X}\beta$ for an unknown β, where $\mathbf{y}$ is a vector of N fitness measurements, and $\mathbf{X}$ is a matrix containing the N corresponding rows of Φ that represent the sequences with fitness measurements. Here, 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 the following convex optimization program:

$$\min_{\hat{\beta}} \|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2 + \nu\|\hat{\beta}\|_1, \quad [7]$$

where ν is a hyperparameter that determines the strength of regularization. Candes and Plan (34) proved that when the rows of an orthogonal basis, such as Φ, are sampled uniformly at random, and the number of samples satisfies Eq. **2**, then the solution to the program in Eq. **7**, denoted β^*, satisfies with high probability

$$\|\beta - \beta^*\|_2 \leq C_1 \frac{\|\beta - \beta_S\|_1}{\sqrt{S}} + C_2\sigma, \quad [8]$$

where C_1 and C_2 are constants and β_S is the best S-sparse approximation to β, i.e., the vector that contains the S elements of β 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 β is itself S-sparse, then, in a noiseless setting, it can be recovered exactly with $\mathcal{O}(S \log q^L)$ measurements. Otherwise, if β is not exactly sparse, but is well approximated by a sparse vector, then it can be approximately recovered with error on the order of $\frac{1}{\sqrt{S}}\|\beta - \beta_S\|_1$, which is proportional to the sum of the magnitudes of the $q^L - S$ elements of β 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 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 $K(q)$ is given by the columns of the following Householder matrix:

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

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 one 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, 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 column 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 to which the row corresponds). More specifically, let $\tilde{\mathbf{P}}_q$ be the matrix containing the final $q - 1$ unnormalized columns of $\mathbf{P}_q$, such that $\mathbf{P}_q = \frac{1}{\sqrt{q}}\left[\mathbf{1}_q \mid \tilde{\mathbf{P}}_q\right]$, where $\mid$ denotes column-wise concatenation. Then, the i^{th} row of $\tilde{\mathbf{P}}_q$ encodes the i^{th} 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 the 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 $H(L, q)$ is given by the columns of following the $q^L \times q^L$ matrix (57):

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

where $\mathbf{P}_q$ is defined in Eq. **9**. In the basis defined in Eq. **10**, an epistatic interaction between positions in the set U is encoded by the length $(q-1)^{|U|}$ vector $\phi_U(\mathbf{s}) \coloneqq \frac{1}{\sqrt{q^L}} \bigotimes_{i \in U} \mathbf{p}_q(s_i)$. These encodings are used in the Fourier basis representation of fitness functions shown in Eq. **4**. The results of Eqs. **9** and **10** are proved in *SI Appendix*. Note that an equivalent form of this basis for $q = 4$ was given in ref. 40, and an alternative form for any alphabet size was given in ref. 41.

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 $j = 1, 2, \ldots, L$.
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. 30. In particular, in ref. 30, the authors allow subsequence fitness values to be sampled from any appropriate distribution, whereas 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 $j = 1, 2, \ldots, L$). In the Random scheme, each neighborhood $V^{[j]}$ contains j and $K - 1$ other position indices selected uniformly at random from $\{1, 2, \ldots, L\} \backslash j$. 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 (58, 59)] 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 *Results* and are all proved in *SI Appendix*. First, the sparsity of any GNK fitness with constant neighborhood sizes is bounded above by

$$S(f) \leq 1 + L(q-1) + L(q^K - Kq + K - 1). \quad [11]$$

All curves in Fig. 2 *A* and *B* 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. 2*C* with $L = 20$ and $q = 2$. Additionally, the sparsity of GNK fitness functions with Block neighborhoods can be calculated exactly as

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

Eq. **12** is plotted as the red curve in Fig. 2*C* with $L = 20$ and $q = 2$. Similarly, the sparsity of GNK fitness functions with Adjacent neighborhoods is given by

$$S(f) = 1 + Lq^{K-1}(q-1), \quad [13]$$

which is plotted as the green curve in Fig. 2*C* 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

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

where

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

and $\alpha(r) = \frac{(K-1)!}{(L-1)!}\frac{(K-r)!}{(L-r)!}$. Eq. **14** with $L = 20$ and $q = 2$ is shown as the solid blue curve in Fig. 2*C*. The results of Eqs. **11**–**14** are proved in *SI Appendix*.

Numerical Calculation of *C*. In order to determine an appropriate value of *C*, we ran experiments where we 1) sampled a fitness function from a GNK model, 2) subsampled *N* sequence-fitness pairs uniformly at random from the complete fitness function for a range of settings of *N*, 3) ran LASSO on each of the subsampled datasets, and 4) 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

$$\hat{C} = \frac{\hat{N}}{S(f)\log_{10}(q^L)}, \quad [15]$$

is the minimum value of *C* that satisfies Eq. **2**, where $S(f)$ is calculated with Eq. **6**. We ran multiple replicates of this experiment for neighborhoods sampled according to the Random neighborhood 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 five 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 five training sets explained $> 99.99\%$ of the variance in the fitness function's coefficients.

Equipped with an estimate of *C*, we calculated the minimum number of samples required to exactly recover a GNK fitness function by using Eq. **2** along with the sparsity calculations of Eq. **6**. Specifically,

$$N = \lceil C \cdot S(f)\log_{10}(q^L) \rceil, \quad [16]$$

is the minimum number of samples that guarantees exact recovery, where $\lceil\rceil$ represents the ceiling operator. Eq. **16** was used along with the bound in Eq. **11** to calculate the curves in Fig. 3.

Percent Variance Explained. In Fig. 4*C*, 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*. We refer to this as the "percent variance explained" by the largest *S* coefficients and calculate it as:

$$\%\ \text{variance explained}\ (S) := 100\% \cdot \left(1 - \frac{\|\beta_S - \beta\|_2^2}{\|\beta\|_2^2}\right). \quad [17]$$

Data Availability. The code for our analyses is available on GitHub, https://github.com/dhbrookes/FitnessSparsity. The mTagBFP2 fitness data used in this work is available in Supplementary Data 3 of ref. 18 (https://doi.org/10.1038/s41467-019-12130-8). The His3p fitness data used in this work is described in ref. 46 and is available in the NCBI Gene Expression Omnibus repository, accession no. GSE99990. All other data generated or analyzed in this study are included in the article and/or supporting information.

ACKNOWLEDGMENTS. We thank Clara Wong-Fannjiang and Nilesh Tripuraneni for enlightening discussions; and Akosua Busia and Chloe Hsu for helpful comments on the manuscript. D.H.B and J.L. were supported by the Chan Zuckerberg Investigator Program. A.A. was supported by Army Research Office Grant W911NF2110117.

1. J. Otwinowski, J. B. Plotkin, Inferring fitness landscapes by regression produces biased estimates of epistasis. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E2301–E2309 (2014).
2. J. Otwinowski, Biophysical inference of epistasis and the effects of mutations on protein stability and function. *Mol. Biol. Evol.* **35**, 2345–2354 (2018).
3. A. Ballal *et al.*, Sparse epistatic patterns in the evolution of terpene synthases. *Mol. Biol. Evol.* **37**, 1907–1924 (2020).
4. P. A. Romero, A. Krause, F. H. Arnold, Navigating the protein fitness landscape with Gaussian processes. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E193–E201 (2013).
5. J. Zhou, D. M. McCandlish, Minimum epistasis interpolation for sequence-function relationships. *Nat. Commun.* **11**, 1782 (2020).
6. K. K. Yang, Z. Wu, F. H. Arnold, Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).
7. R. Rao *et al.*, "Evaluating protein transfer learning with tape" in *Advances in Neural Information Processing Systems*, H. Wallach *et al.*, Eds. (Curran Associates, Inc., Red Hook, NY, 2019), **vol. 32**, pp. 9689–9701.
8. S. Biswas, G. Khimulya, E. C. Alley, K. M. Esvelt, G. M. Church, Low-N protein engineering with data-efficient deep learning. *Nat. Methods* **18**, 389–396 (2021).
9. R. J. Fox *et al.*, Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol.* **25**, 338–344 (2007).
10. C. N. Bedbrook, K. K. Yang, A. J. Rice, V. Gradinaru, F. H. Arnold, Machine learning to design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization. *PLOS Comput. Biol.* **13**, e1005786 (2017).
11. Z. Wu, S. B. J. Kan, R. D. Lewis, B. J. Wittmann, F. H. Arnold, Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 8852–8858 (2019).
12. A. Gupta, J. Zou, Feedback GAN for DNA optimizes protein functions. *Nat. Mach. Intell.* **1**, 105–111 (2019).
13. D. H. Brookes, H. Park, J. Listgarten, "Conditioning by adaptive sampling for robust design" in *Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research*, K. Chaudhuri, R. Salakhutdinov, Eds. (Proceedings of Machine Learning Research, PMLR, Long Beach, CA), **vol. 97**, pp. 773–782 (2019).
14. C. Angermüller *et al.*, "Model-based reinforcement learning for biological sequence design" in *8th International Conference on Learning Representations, ICLR 2020* (OpenReview.net, 2020). https://openreview.net/forum?id=HklxbgBKvr. Accessed 17 December 2021.
15. C. Fannjiang, J. Listgarten, "Autofocused oracles for model-based design" in *Advances in Neural Information Processing Systems 33*, H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, H.-T. Lin, Eds. (NeurIPS, 2020).
16. Z. R. Sailer, M. J. Harms, Detecting high-order epistasis in nonlinear genotype-phenotype maps. *Genetics* **205**, 1079–1088 (2017).
17. G. Yang *et al.*, Higher-order epistasis shapes the fitness landscape of a xenobiotic-degrading enzyme. *Nat. Chem. Biol.* **15**, 1120–1128 (2019).
18. F. J. Poelwijk, M. Socolich, R. Ranganathan, Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nat. Commun.* **10**, 4213 (2019).
19. A. Aghazadeh *et al.*, Epistatic Net allows the sparse spectral regularization of deep neural networks for inferring fitness functions. *Nat. Commun.* **12**, 5225 (2021).
20. A. Aghazadeh, O. Ocal, K. Ramchandran, CRISPRL and: Interpretable large-scale inference of DNA repair landscape based on a spectral approach. *Bioinformatics* **36** (suppl. 1), i560–i568 (2020).
21. E. J. Candes, J. Romberg, T. Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory* **52**, 489–509 (2006).
22. D. L. Donoho, Compressed sensing. *IEEE Trans. Inf. Theory* **52**, 1289–1306 (2006).
23. S. Kauffman, S. Levin, Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.* **128**, 11–45 (1987).
24. A. Agarwala, D. S. Fisher, Adaptive walks on high-dimensional fitness landscapes and seascapes with distance-dependent statistics. *Theor. Popul. Biol.* **130**, 13–49 (2019).
25. S. A. Kauffman, E. 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. W. Rowe *et al.*, Analysis of a complete DNA-protein affinity landscape. *J. R. Soc. Interface* **7**, 397–408 (2010).
27. J. Neidhart, I. G. Szendro, J. Krug, Exact results for amplitude spectra of fitness landscapes. *J. Theor. Biol.* **332**, 218–227 (2013).
28. Y. Hayashi *et al.*, Experimental rugged fitness landscape in protein sequence space. *PLoS One* **1**, e96 (2006).
29. T. Aita *et al.*, Extracting characteristic properties of fitness landscape from in vitro molecular evolution: A case study on infectivity of fd phage to *E. coli*. *J. Theor. Biol.* **246**, 538–550 (2007).
30. J. Buzas, J. Dinitz, An analysis of NK landscapes: Interaction structure, statistical properties, and expected number of local optima. *IEEE Trans. Evol. Comput.* **18**, 807–818 (2014).
31. S Nowak, J Krug, Analysis of adaptive walks on NK fitness landscapes with different interaction schemes. *J. Stat. Mech. Theory Exp.* **2015**, P06014 (2015).
32. K. S. Sarkisyan *et al.*, Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).
33. T. Hastie, R. Tibshirani, M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations* (Chapman & Hall/CRC, Boca Raton, FL, 2015).
34. E. J. Candes, Y. Plan, A probabilistic and RIPless theory of compressed sensing. *IEEE Trans. Inf. Theory* **57**, 7235–7254 (2011).

35. R. B. Heckendorn, D. Whitley, "A Walsh analysis of NK-landscapes" in *Proceedings of the Seventh International Conference on Genetic Algorithms*, T. Bäck, Ed. (Morgan Kaufmann, Burlington, MA), pp. 41–48 (1997).
36. S. Hwang, B. Schmiegelt, L. Ferretti, J. Krug, Universality classes of interaction structures for NK fitness landscapes. *J. Stat. Phys.* **172**, 226–278 (2018).
37. D. M. Weinreich, Y. Lan, J. Jaffe, R. B. Heckendorn, The influence of higher-order epistasis on biological fitness landscape topography. *J. Stat. Phys.* **172**, 208–225 (2018).
38. F. J. Poelwijk, V. Krishna, R. Ranganathan, The context-dependence of mutations: A linkage of formalisms. *PLOS Comput. Biol.* **12**, e1004771 (2016).
39. D. M. Weinreich, Y. Lan, C. S. Wylie, R. B. Heckendorn, Should evolutionary geneticists worry about higher-order epistasis? *Curr. Opin. Genet. Dev.* **23**, 700–707 (2013).
40. G. D. Stormo, Maximally efficient modeling of DNA sequence motifs at all levels of complexity. *Genetics* **187**, 1219–1224 (2011).
41. P. F. Stadler, R. Seitz, G. P. Wagner, Population dependent Fourier decomposition of fitness landscapes over recombination spaces: Evolvability of complex characters. *Bull. Math. Biol.* **62**, 399–428 (2000).
42. E. D. Weinberger, Local properties of Kauffman's N-k model: A tunably rugged energy landscape. *Phys. Rev. A* **44**, 6399–6413 (1991).
43. P. Ravikumar, M. J. Wainwright, J. D. Lafferty, High-dimensional Ising model selection using l1-regularized logistic regression. *Ann. Stat.* **38**, 1287–1319 (2010).
44. C. A. Voigt, C. Martinez, Z. G. Wang, S. L. Mayo, F. H. Arnold, Protein building blocks preserved by recombination. *Nat. Struct. Biol.* **9**, 553–558 (2002).
45. O. M. Subach *et al.*, Structural characterization of acylimine-containing blue and red chromophores in mTagBFP and TagRFP fluorescent proteins. *Chem. Biol.* **17**, 333–341 (2010).
46. V. O. Pokusaeva *et al.*, An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. *PLoS Genet.* **15**, e1008079 (2019).
47. J. Yang, Y. Zhang, I-TASSER server: New development for protein structure and function predictions. *Nucleic Acids Res.* **43**, W174–W181 (2015).
48. R. Lorenz *et al.*, ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
49. L. du Plessis, G. E. Leventhal, S. Bonhoeffer, How good are statistical models at approximating complex fitness landscapes? *Mol. Biol. Evol.* **33**, 2454–2468 (2016).
50. D. W. Anderson, A. N. McKeown, J. W. Thornton, Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. *eLife* **4**, e07864 (2015).
51. D. M. Weinreich, N. F. Delaney, M. A. Depristo, D. L. Hartl, Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114 (2006).
52. C. Bank, S. Matuszewski, R. T. Hietpas, J. D. Jensen, On the (un)predictability of a large intragenic fitness landscape. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 14085–14090 (2016).
53. N. C. Wu, L. Dai, C. A. Olson, J. O. Lloyd-Smith, R. Sun, Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife* **5**, e16965 (2016).
54. D. H. Bryant *et al.*, Deep diversification of an AAV capsid protein by machine learning. *Nat. Biotechnol.* **39**, 691–696 (2021).
55. B. Ricaud, P. Borgnat, N. Tremblay, P. Gonçalves, P. Vandergheynst, Fourier could be a data scientist: From graph Fourier transform to signal processing on graphs. *C. R. Phys.* **20**, 474–488 (2019).
56. P. F. Stadler, "Towards a theory of landscapes" in *Complex Systems and Binary Networks*, R. López-Peña, H. Waelbroeck, R. Capovilla, R. García-Pelayo, F. Zertuche, Eds. (Lecture Notes in Physics, Springer, Berlin, 1995), vol. 461, pp. 78–163.
57. R. Hammack, W. Imrich, S. Klavzar, *Handbook of Product Graphs* (CRC Press, Inc., Boca Raton, FL, ed. 2, 2011).
58. A. S. Perelson, C. A. Macken, Protein evolution on partially correlated landscapes. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 9657–9661 (1995).
59. H. A. Orr, The population genetics of adaptation on correlated fitness landscapes: The block model. *Evolution* **60**, 1113–1124 (2006).