Importance of higher-order epistasis in protein sequence-function relationships

Palash Sethi[a] and Juannan Zhou[a,1]

[a]Department of Biology, University of Florida, Gainesville, FL, 32611
[1]juannanzhou@ufl.edu

Abstract

Protein sequence–function relationships are inherently complex, as amino acids at different positions can interact in highly unpredictable ways. A key question for protein evolution and engineering is how often epistasis extends beyond pairwise interactions to involve three or more positions. Although experimental data has accumulated rapidly in recent years, addressing this question remains challenging, as the number of possible interactions is typically enormous even for proteins of moderate size. Here, we introduce an interpretable machine learning framework for studying higher-order epistasis scalable to full-length proteins. Our model builds on the transformer architecture, with key modifications allowing us to assess the importance of higher-order interactions by fitting a series of models with increasing complexity. Applying our method to 10 large protein sequence-function datasets, we found that while additive effects explain the majority of the variance, within the epistatic component, the contribution of higher-order epistasis ranges from negligible to up to 60%. We also found higher-order epistasis is particularly important for generalizing locally sampled fitness data to distant regions of sequence space and for modeling an additional multi-peak fitness landscape. Our findings suggest that higher-order epistasis can play important roles in protein sequence-function relationships, and thus should be properly considered in protein engineering and evolutionary data analysis.

Introduction

Understanding how amino acid sequence determines protein function is critical for biomedical engineering, synthetic biology, and evolutionary biology [1–4]. High-throughput techniques such as deep mutational scanning (DMS) have enabled the functional characterization of thousands to millions of protein variants in a massively parallel manner [5–24]. A general observation from these empirical studies is that while sequence–function relationships can often be reasonably approximated by the independent contributions of individual mutations, epistatic interactions between amino acids also play important roles and must be properly accounted for to accurately model protein function [25–29].

Although pairwise interactions between amino acids are both predicted by biophysical models and widely supported by empirical data [30–33], our understanding of how frequently epistasis extends beyond pairwise interactions to involve three or more positions remains incomplete. These higher-order interactions pose significant conceptual and practical challenges for deciphering protein sequence–function relationships. A thorough assessment of their prevalence is therefore essential for informing future experimental designs and the development of analytical methods [4, 34]. Current evidence on this question is mixed: some studies suggest that higher-order epistasis is widespread and plays a significant role in protein function [28, 30, 35–37], while others argue that protein sequence–function landscapes are simpler and epistasis beyond pairwise interactions is rare [31, 33].

For example, a recent paper [31] investigated the genetic architecture for 20 protein datasets with combinatorial mutagenesis by comparing the performance of pairwise and higher-order epistatic model, while simultaneously fitting a sigmoidal function to account for non-specific epistasis. In contrast to conclusions from some of the earlier studies, the authors observed small to negligible contribution of higher-order epistasis and argued that most protein fitness landscapes can be well characterized by only additive effects and pairwise interactions. However, a key limitation of this study is that it is based on relatively small-scale protein sequence-function relationships, involving either exhaustive amino acid substitutions at no more than four sites or simultaneous single amino acid substitutions at up to 20 sites.

To draw general conclusions about the role of higher-order epistasis, we must extend these tests to larger protein sequence-function relationships as interactions beyond pairwise epistasis become more probable due to simple combinatorics and biophysical principles. However, addressing higher-order epistasis in full-length proteins is challenging because naive approaches to modeling higher-order interactions using parametric regression methods lead to extremely high computational demands, as well as issues in overfitting and model interpretability due to the enormous number of model parameters. Although nonparametric methods represent a desirable, and highly scalable approach by modeling higher-order interactions without explicitly fitting regression coefficients [29, 38], generalizing it to account for the confounding effects of non-specific epistasis using a global nonlinearity [31, 39, 40] is not straightforward due to the resulting non-Gaussian likelihood [41].

Artificial neural networks (ANNs) have recently been widely applied to model sequence-function relationships [19, 42, 43]. Their flexibility makes them well-suited to capturing complex, higher-order epistatic interactions. However, naïvely applying ANNs provides limited insight into the prevalence or structure of higher-order epistasis. This is because there is no direct correspondence between the architecture of a neural network and the specific orders of epistasis it models. As a result, the standard approach of detecting higher-order interactions by comparing models with and without higher-order terms is not readily applicable to ANNs.

To address this challenge, we develop a modified transformer architecture. The key feature of our method is that it enables explicit control over the maximum order of epistasis the network fits by simply adjusting the number of attention layers. This design allows us to systematically assess the contribution of higher-order interactions by fitting a series of models with increasing epistatic complexity and evaluating their predictive performance. Crucially, unlike traditional regression-based approaches, our method captures higher-order interactions implicitly through learned neural network weights. As a result, model complexity does not grow exponentially with sequence length or interaction order. This enables us to analyze higher-order epistasis involving up to eight sites, while accurately accounting for nonspecific epistasis in sequence-function datasets for full-length protein, whereas previous methods can only achieve this for a limited number of mutagenized positions.

We begin by introducing the main ideas of our method and demonstrating its application to a simulated dataset. Next, we present results on the role of higher-order epistasis across 10 combinatorial protein DMS

datasets. We then provide a detailed analysis of how higher-order epistasis contributes to out-of-distribution predictions when the training data consist of locally sampled genotypes. Finally, we investigate how higher-order epistasis shapes a multi-peak fitness landscapes by jointly analyzing the combinatorial DMS data from four orthologous green fluorescent proteins.

Results

Epistatic transformer for fitting higher-order epistasis

We aim to quantify the contribution of higher-order specific epistatic interactions to the total phenotypic variance measured for empirical protein sequence-function relationships. Our overall strategy is to fit models of increasing complexity to these datasets and examine how inclusion of higher-order interactions improves model generalization to held-out test genotypes.

To account for epistasis resulting from interactions among specific mutations, as well as from nonlinear scaling on the observation scale—i.e., nonspecific or global epistasis[3, 39, 40]—we study models of the general form

$$f(x) = g \circ \phi(x). \tag{1}$$

Here $x = x_1, x_2, \cdots, x_L$ is the amino acid sequence of a protein of length L. $\phi(x)$ is a function that models the independent effects of individual amino acids as well as specific epistatic interactions, such that

$$\phi(x) = \sum_{i=1}^{L} e(x_i) + \sum_{1 \leq i < j \leq L} e(x_i, x_j) + \sum_{1 \leq i < j < k \leq L} e(x_i, x_j, x_k) + \cdots . \tag{2}$$

Specifically, $e(x_i)$ captures the independent effect of residue x_i at site i, $e(x_i, x_j)$ captures the interaction effect between residues x_i and x_j. $e(x_i, x_j, x_k)$ accounts for three-way interaction, etc. In our framework, ϕ may include epistatic interactions among up to $K \leq L$ residues. g is a nonlinear monotonic function that transforms the latent phenotype $\phi(x)$ to the measurement scale for modeling global epistasis [39, 43–46]. Importantly, Eq. 1 allows us to decompose the sequence-function relationship as a nonspecific component ϕ, and a specific component f, such that we may study the contribution of higher-order interactions by changing the complexity of the function ϕ with increasing values of K and examine the generalizability of the trained model to novel genotypes.

Fitting the function ϕ is challenging when higher-order interactions are considered, as the number of regression coefficients in Eq. 2 grows roughly exponentially with interaction order, such that for a protein of moderate length, e.g. 100 amino acids, there are $\binom{100}{3} \times 20^3 = 1.1$ billion three-way interactions and 511 billion four-way interactions. Thus for large protein sequence spaces, current methods that directly fit the regression coefficients in Eq. 2 cannot model interactions beyond pairs of sites [46].

Neural networks provide an attractive alternative for fitting epistasis interactions due to their ability at fitting complex nonlinear relationships. Here instead of directly fitting the the regression coefficients in Eq. 2, our strategy is to implicitly infer the function ϕ by fitting the weights of the neural network such that the network predictions $\phi(x)$, when decomposed as Eq. 2 contains epistatic interactions exactly up to a given order K. This is an attractive option as the complexity of the neural network need not scale exponentially with interaction order. Furthermore, many existing machine learning packages with GPU support may be employed to ensure fast model training.

The challenge in implementing this strategy is that existing neural network architectures do not allow us control the highest order of specific epistasis fit by the model, thus making it impossible to test if inclusion of pairwise or higher order epistasis improves model prediction. Instead, these models usually can only fit all orders of interactions simultaneously. Consider, for example, one of the simplest architecture, the multilayer perceptron (MLP). In an MLP, the value for a hidden unit is first calculated by taking a weighted sum of inputs across different positions. To model nonlinear relationships between the inputs, the sum is then passed through a nonlinear activation function. Translating this to the language of genotype-phenotype models, we see that the model performs global epistasis-like operations on the input sequence where the underlying phenotype is additive with respect to the positions [22, 43]. We can add more layers to perform similar operations, which allows one to model complex specific epistasis among the input positions. However, one

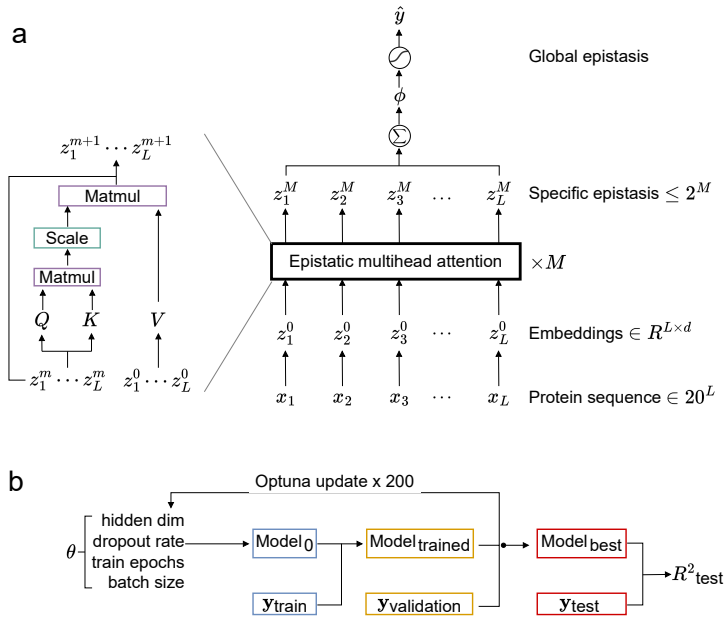

Figure 1: Epistatic transformer for jointly modeling fixed-order specific epistasis and nonspecific epistasis. a. Overall model architecture. The input amino acid token x_l is first used to generate position-specific amino acid embeddings z_l^0. The embeddings are then passed through M layers of modified multi-head attention (MHA) such that the output embeddings contain specific epistasis up to order 2^M. Left: Details for a single MHA layer. Hidden states z^m from the previous layer is passed through linear layers to generate the query (Q) and key (K) tensors, while the raw embeddings Z^0 are used directly as the value (V) tensor. Attention weights are calculated by taking scaled dot products between Q and K, which are used to generate the final output of this layer. This bypassing of the raw embedding Z^0, together with the removal of LayerNorm, softmax operation on the attention weights, and the feedforward layer, allow us to model only interactions among up to 2^M sites when using M MHA layers. b. Automatic hyperparameter search scheme using Optuna [48]

.

can easily see that the function by an MLP cannot be decomposed into a ϕ and a g component as in Eq. 1, due to of the lack of compartmentalization of global epsitasis and specific epistasis operations.

Here, we developed a novel neural network architecture, the 'epistatic transformer', for modeling specific epistatic interactions of fixed orders. The model allows one to fit Eq. 1, with ϕ containing specific epistasis among up to $K = 2^M$ sites with M attention layers, such that by setting $M = 1, 2, 3, \cdots$, we can fit pairwise, four-way, eight-way, or higher-order interaction models. Our model is based on a transformer architecture [47] typically employed for regression and classification tasks. The high-level model architecture is shown in Figure 1a. The input protein sequence $x_1, \cdots, x_L$ is first used to generate amino acid embeddings $z_1^0, \cdots, z_L^0$ of dimension d. The embeddings are then passed through M modified multi-head attention layers, such that the output embeddings $z_1^M, \cdots, z_L^M$ contain specific interactions exactly up to order 2^M. We then calculate a single weighted sum of the flattened embeddings, which is passed through a nonlinear function g for modeling global epistasis.

We model interaction among input positions through the multi-head attention (MHA) mechanism [47]. In a standard attention layer, three different linear transformations are applied to an input embedding vector to generate the query Q, key K, and value V vectors. Attention scores are computed by comparing the query of a focal position to the keys at all positions in the input sequence. A weighted sum of values across all positions with weights provided by the attention scores are then added to the input embedding to produce the position specific outputs. Importantly, pairwise interaction occurs during the computation of attention scores, where each query token interacts with all key tokens through a dot product operation followed by a softmax function. We can see that as one increases the number of attention layers M, the order of interactions occurring at a position increases roughly exponentially. This property makes the transformer architecture a potential framework for fitting a series of epistatic models with increasing complexity.

However, the original transformer architecture is not sufficient for fitting specific epistasis, as standard transformer models have the same problem of mixing global epistasis and specific epistasis operations. Here we made three key modifications to the MHA layer (Figure 1a, left panel; Methods) to allow us to precisely control the orders of specific epistasis in the intermediate model output and to better fit the complex epistatic

interactions in the data. First, we removed all non-linear operations in the hidden layers of the model, such that the ϕ function we infer only contains specific interactions. In particular, we remove the softmax operation on the attention weights of the MHA layer, the LayerNorm [49] operation on the embeddings across all positions, as well as the feedforward network applied to each position. Second, to ensure that the order of specific epistasis in our model scales exactly as 2^M with the number of layers M, we replace the value (V) input to each MHA layer with the raw amino acid embedding $(z_1^0, \cdots, z_L^0)$. Thus, by setting $M = 1, 2, 3$, we are able to fit Eq. 2 with up to pairwise, four-way, and eight-way interactions, respectively (see Supplemental Methods for proof). Third, we used a special positional encoding for the input amino acids. Without positional information, the same residues across different positions will interact in the same manner. To prevent this, typical transformer models combine continuous positional encoding with the input embedding. However, in our experiments, we found that this is insufficient to capture the range of interactions between residues in real protein datasets. Thus, in our model, we introduce a discrete positional encoding scheme, where the original amino acid token x_l is transformed to $x'_l = x_l + a \times (l-1)$, where a is the alphabet size (usually $a = 20$), thereby effectively allowing us to treat amino acids at different positions in completely different ways.

While our modified transformer architecture allows us to control the maximal order of specific epistasis incorporated into the model, we also experimented with different choices of nonlinear function g for mapping the sum of the flattened embeddings of the last transformer layer to a scalar value in order to model global epistasis. Specifically, we tested the performance of sums of various numbers of independent sigmoid functions[19, 46]. We found that a single scaled sigmoidal function with four parameters $\sigma(z) = \frac{a}{1+e^{-(z-\mu)/s}} + b$ performs equally well as more complex nonlinear functions for all datasets examined. For all subsequent analyses we therefore used this simple logistic link function to model nonspecific epistasis.

Our preliminary experiment showed that the performance of the trained models can be sensitive to choices of certain hyperparameters. Thus, to ensure that we fit the optimal model within a family of models with M attention layers, we used a hyperparameter optimization strategy. Specifically, we train our model iteratively on a training dataset while varying the values for hyperparameters including batch size, hidden dimensions, dropout rate, and number of epochs, while keeping the number of attention heads and learning rate constant to ensure a low dimensional search space. The trained model is evaluated on a validation dataset to inform the choice of hyperparameters of the next iteration. We used Optuna [48] to automate the hyperparameter search, with 200 updates and validation R^2 as the criterion (Figure 1b). The model trained under the optimal hyperparameters was then evaluated on the independent test dataset to produce the test R^2 value.

Epistatic transformer recapitulates the genetic architecture of a simulated fitness landscape

We first apply our model to a simulated dataset to characterize its behavior when the ground truth is known. We simulated a sequence-function map for a binary sequence space with 13 sites, with specific epistatic interactions up to order 8 (see Supplemental Methods). The choice of a small genotype space allows us to easily generate data for all possible genotypes and analytically characterize the formal structure of epistasis in the simulated data and the predictions in terms of variance components (proportions of total phenotypic variance explained by additive effects, pairwise, and higher-order interactions), which is not possible for larger sequence spaces.

Figure 2a summarizes the structure of specific epistasis in our simulated data in terms of the fraction of total phenotypic variance due to contributions of additive effects and epistatic interactions up a specified order (i.e., cumulative variance explained). We can see that additive effects only accounts for 10% of the total variance, whereas additive effects and pairwise interactions together explain 30% of variance. Further, 70% of the total phenotypic variance is due to interactions among up to four sites. Finally, considering interactions among up to 8 sites allow us to explain all variance in the data.

Next, we examine if we can infer the structure of specific epistasis in the data by fitting a series of epistatic transformer model with increasing complexity. Specifically, we fit the epistatic transformer model to 90% of randomly sampled data and examine its out-of-sample performance when the number of MHA layer is set to be 1, 2, or 3, corresponding to pairwise, 4th-order, and 8th-order interaction models (Figure 2a). To ensure that our epistatic transformer model can capture most epistatic interactions even when underparameterized, we constrain the hidden dimension to fewer than 8 and limit the number of attention heads to fewer than

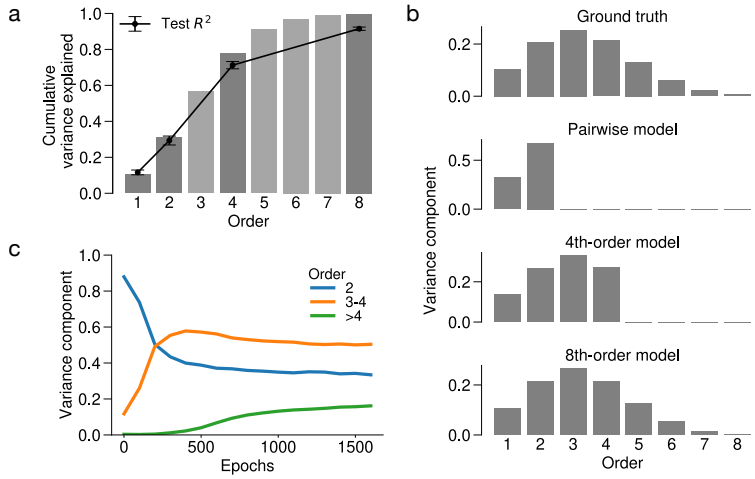

Figure 2: Epistatic transformer is able to capture the genetic architecture of simulated sequence-function relationships. Data was simulated for a binary fitness landscape with 13 sites, with specific epistasis among up to 8 sites. a. Test R^2 can be used to estimate cumulative variance components. Bar plot shows the cumulative variance explained by each order of interaction (proportion of variance due to epistasis up to a given order). Error plot shows the test R^2 for additive, pairwise, four-way, and eight-way interaction models using epistatic transformer to 90% of simulated data. Error bars correspond to one standard deviation with 5 random train-test splits. b. Epistatic transformer recapitulates true variance components. Bar plots correspond to variance components(proportion of variance explained by interactions of a given order) for the simulated landscape, the pairwise, 4-way, and 8-way models. Note that pairwise model only contains interaction up to second order, and 4th-order model only contains interaction among up to 4 positions, while 8th-order model model contains interaction among up to 8 positions with variance components aligning with the ground truth. c. Epistatic transformer model becomes increasingly complex with longer training. The three curves correspond to the variance component decomposition of the full landscape inferred by the 8th-order model at each training epoch. Blue: variance components for additive and pairwise interactions. Orange: variance components for three-way and four-way interactions. Green: variance components for interaction order higher than four.

2. Under these constraints, the best-performing epistatic transformer models with 1, 2, and 3 layers contain up to 1,926, 3,014, and 4,102 parameters, respectively—all smaller than the size of the training dataset. We see that the test R^2 increases with the interaction order of the model and remains slightly below—but close to—the ground truth cumulative variance explained, despite all models having fewer parameters than training data points. This results confirm that our model can be used to provide lower bounds for the contribution of pairwise and higher-order epistasis when fit to randomly sample training data.

We provided a mathematical proof (Supplemental Methods) that highest order of specific epistasis fit by our model with M MHA layers is exactly 2^M. To confirm this result and demonstrate that an all-order epistatic transformer model can capture arbitrary patterns of epistasis in the data, we directly compare the variance components (i.e., the fraction of variance explained by additive effects or epistasis of specific orders) between the data and the sequence-function relationships reconstructed by the low-order and all-order epistatic transformer models (Figure 2b). We see that for the low-order interaction models, the variance components are truncated at the highest order the model can capture, e.g., a 4th-order interaction model with 2 epistatic MHA layers only contains interactions among up to 4 sites. This results confirm that we can precisely control the order of specific epistasis in our model by varying the number of MHA layers. In contrast, the 3-layer model contains variance components up to order 8. Furthermore, comparing the top and bottom panel of Figure 2b, we can see that the variance components of the inferred landscape can faithfully recapitulate the ground-true variance components.

Lastly, we examine how our model fits higher-order interaction during training. In Figure 2c, we separately plotted the variance components of the 8th-order interaction model due to additive + pairwise interaction, three-way + four-way interaction, and interactions of order higher than 4 at different points along the training process. We see that when training begins, the model is predominantly fitting the additive and pairwise components. As training continues, the three-way and four-way interactions become increasingly more important, while the low-order contribution drops. Finally, we see an increase in the contribution of higher-order (> 4) interactions late in the training process (after epoch 500), leading to the convergence of of the variance components of the model landscape to the ground truth. This result shows an interesting behavior of our model which seems to first learn the simple features of the sequence-function relationship

(additive and pairwise), before preceding to model the more complex interactions.

Application to experimental protein sequence-function data

To use the epistatic transformer model to characterize the importance higher-order interactions in empirical protein fitness landscapes, we first curated a list of 10 deep mutational scanning datasets. Our criteria for choosing suitable datasets were (1) the sequence space should be large to allow higher-order epistasis to play significant roles. Specifically, the number of variable sites should be ideally larger than 15 for binary DMS datasets, and at least 5 when all amino acid mutations were constructed at every site. (2) the dataset should contain a large proportion of higher-order mutants, i.e., genotypes that are more than 3 mutations away from a reference wild type, to provide enough data for our model to detect higher-order interactions. We also applied a quality control step by fitting a simple linear regression to the data and only retain datasets where the test R^2 of the additive model is over 0.3 (Methods).

The resulting 10 datasets (Table 1) encompass a diverse set of proteins and phenotypes including protein abundance, binding, fluorescence, and cellular fitness. Importantly, the selected experiments were conducted in sequence spaces significantly larger than datasets used in previous studies on higher-order epistasis in proteins, with mutations spanning tens of amino acid sites, extending up to the entire length of the protein (e.g. the GFP datasets). The chosen datasets also exhibit a wide range of data distribution. In Supplemental Figure 1, we can see that while the measured genotypes are centered around the wild type sequence in some datasets (for example the GFP and His3 data), in other datasets the genotypes are concentrated at an intermediate distance to the WT (for example, GRB-1). Furthermore, the sparsity of the data also varies greatly. For instance, in the GRB-3-abundance and binding data, virtually all genotypes have been measured, whereas for large sequence spaces the proportion of measured genotype usually decays exponentially with Hamming distance, resulting in extremely low sampling of distant regions of the sequence space (e.g. data for various GFPs).

We applied the epistatic transformer model and a linear model with independent contributions from sites to these 10 datasets to examine how higher-order epistasis improves the ability to predict the phenotypes of novel sequences. For each dataset, we randomly assign random samples of all measured genotypes as training data. We used epistatic transformer models to fit pairwise, 4th-order, and 8th-order models in the form of Eq. 1, corresponding to models with 1, 2, and 3 layers of modified MHA attention. In addition, we also fit an additive model of the form in Eq. 1, which again includes composition with a sigmoid function in order to address any non-specific epistasis. To ensure that we fit the best model to the training data, we used the procedure in Figure 1b to identify the optimal hyperparameter combination using cross validation on a validation dataset, with 200 tuning steps (Methods).

We first observe that the additive model often provides a reasonably good fit to the data and explains the predominant proportion of total variance (test $R^2 > 0.55$ for all datasets, when trained on 80% of randomly sampled genotypes; Supplemental Figure 2. For some datasets, the additive model can explain as much as 90% of the variance in the test data (e.g. GRB-3-abundance, His-S5, and His-S12). We also found that incorporating global non-specific epistasis to the model with only additive effects leads to uniform improvement in performance (Supplemental Figure 2). The improvement varies across datasets and range

Table 1: Combinatorial protein mutagenesis datasets used in this paper.

Protein	**Dataset name**	**Genotype space**	**No. variants**	**Phenotype**	**Ref.**
GRB2-SH3 domain	GRB-1	2^{33}	129,320	Protein abundance	[23]
GRB2-SH3 domain	GRB-3-abundance	2^{15}	31,936	Protein abundance	[23]
GRB2-SH3 domain	GRB-3-binding	2^{15}	25,967	Protein binding	[23]
AAV2 capsid protein	AAV2-Capsid	20^{28}	42,328	DNA packaging	[21]
CreiLOV	CreiLOV	20^{15}	165,428	Fluorescence	[50]
cgreGFP	cgreGFP	20^{234}	26,165	Fluorescence	[20]
ppluGFP	ppluGFP	20^{221}	32,260	Fluorescence	[20]
His3 segment 2	His3-S2	20^{28}	116,935	Fitness	[22]
His3 segment 5	His3-S5	20^{31}	92,408	Fitness	[22]
His3 segment 12	His3-S12	20^{19}	62,305	Fitness	[22]

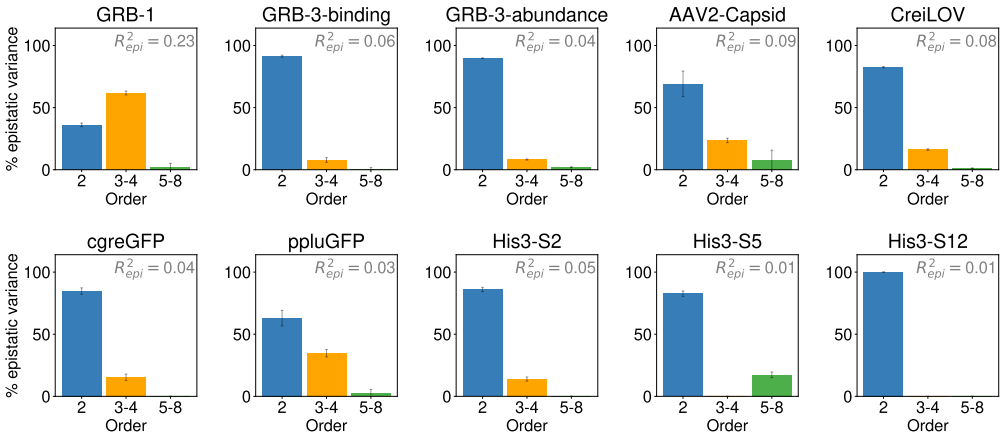

Figure 3: Importance of pairwise and higher-order specific epistasis in 10 experimental protein-sequence function datasets. For each dataset, pairwise, 4th-order and 8th-order models were fit using epistatic transformer with 1, 2, and 3 layers of epistatic multihead attention (MHA), along with an additive model. All models contain a final nonlinear activation function mapping a scalar value to the measurement scale for modeling non-specific epistasis. Models were fit to 80% of training data generated by randomly sampling all available data and evaluated on random test genotypes. In each panel, the number R^2_{epi} on the upper right corner is equal to the proportion of total variance in the test data explained by all orders of specific epistasis, equal to the difference between the R^2 of the 8-th order epistatic transformer model and the additive model. Importance of epistatic interactions of different orders is measured by percent epistatic variance, equal to the gain in R^2 by fitting an additional layer of MHA, normalized by R^2_{epi}. For example, the percent epistatic variance due to pairwise interactions is equal to the difference in R^2 between the pairwise and additive model divided by R^2_{epi}, and the percent epistatic variance due to 3-way and 4-way interactions is equal to the difference in R^2 by the 4-th order model and the pairwise divided by R^2_{epi}. Error bars represent 1 standard deviation calculated from 3 replicates.

from less than 10% improvement in R^2 in the three GRB datasets, to up to 30% in the cgreGFP dataset.

We next summarize the performance of our epistatic transformer models trained on 80% of randomly sampled genotypes (Figure 3). Here we use two metrics to gauge how incorporating epistasis and in particular, higher-order epistasis improves model prediction. The first metric, R^2_{epi} measures the proportion of total variance in the test data explained by all orders of specific epistasis, equal to the difference in test R^2 between the 8-th order epistatic transformer model with $M = 3$ (i.e., with 3 epistatic MHA layers, which fits up to 8-th order interactions) and the additive model. The second metric 'percent epistatic variance' measures the relative contribution of pairwise or higher-order interactions. This metric is calculated by first taking the difference in R^2 between an epistatic transformer model with given number of MHA layers and the preceding, less complex model. The percentage of epistatic variance is then acquired by dividing the resulting quantity by R^2_{epi} (for example, to quantify the importance of 3- and 4-th order epistasis, we calculate $(R^2_{M=2} - R^2_{M=1})/(R^2_{\text{epi}})$. Thus, the product of R^2_{epi} and percent epistatic variance returns the percentage of total variance in phenotype explained by certain orders of epistasis.

We first observe that the contribution of specific epistasis can vary significantly across datasets, with R^2_{epi} (top right corner of each panel in Figure 3) ranging from 0.01 (His3-S5 and His3-S12) to 0.23 (GRB-1). Second, we found that while pairwise interactions are typically the predominant form of epistasis, consistent with previous findings, higher-order epistasis can also play moderate to substantial roles. We found the strongest contribution of higher-order epistasis in the GRB-1 dataset, wherein three-way and four-way interactions together account for 62% of the epistatic variance, compared to 36% attributed to pairwise interactions (Figure 3). In contrast, we observed that interactions of order > 4 account for very small amount of epistatic variance, suggesting that three-way and four-way interactions are the predominant forms of epistasis in this dataset.

We also observed moderate contributions by higher-order epistasis in other datasets (e.g. AAV2-Capsid, ppluGFP, and CreiLOV), where epistasis of order 3-8 collectively accounts for roughly 1/5 to 1/3 of the total epistatic variance. Additionally, we found that the higher-order epistasis can make more substantial contributions when training data is more sparse, sometimes accounting for $> 80\%$ of the total epistatic variance (e.g. the AA2-Capisd and cgreGFP dataset, Supplemental Figure 3). Similar phenomenon has been observed previously, for instance, in non-parametric kernel methods [29, 38]. This suggests that higher-order models may, counterintuitively, provide a more parsimonious fit to sparse training data than the pairwise interaction model.

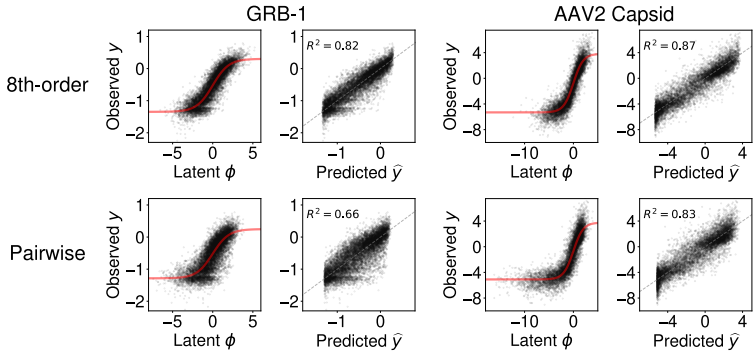

Figure 4: Improvement in prediction accuracy in higher-order models is due to specific epistasis. Scatter plots show the observed phenotypes (y) vs. latent model predictions (ϕ), or the final model predictions ($\widehat{y}$)) for the test genotypes, for the pairwise and 8th-order epistatic transformer models fit to the GRB-1 and AAV2-Capsid datasets. Models were fit to 80% of randomly sampled training data.

Improvement in prediction accuracy in higher-order models is due to specific epistasis

In order to understand why higher-order epistasis improves model prediction, we next examined the detailed model fit for the pairwise and higher-order models for the two datasets with the largest contribution of higher-order interactions, specifically the AAV2-Capsid and the GRB-1 datasets. In Figure 4, we show scatter plots comparing the actual experimental measurements for the test genotypes vs. the predicted latent phenotype (ϕ) before activation by the global epistasis function g (left panel) and the final predicted phenotype ($\widehat{y}$, containing both non-specific and specific epistasis, right panel). We first see that both datasets contain a prominent nonlinear relationship (red curve on the left panels), where genotypes with low predicted ϕ values, and genotypes with high predicted ϕ (to a lesser extent) are bounded by the sigmoid function. Next, we see that the final predictions for both models and datasets appear to have a linear correlation with the measurements. We also notice that the nonlinear functions for the pairwise and 8-th order models are highly similar, although the nonlinear function in these two models are fit separately.

Together, these results confirm that both the pairwise and higher-order epistatic transformer models have adequately captured the non-specific epistasis in the data, suggesting that the improvement by the higher-order order epistatic model over the pairwise model is due to the its ability to better fit the residual variance due to specific higher-order epistasis.

It is noteworthy that, for both the GRB-1 and AAV2 Capsid datasets, the 8th-order model offers better predictions across the entire range of measured phenotypes. This is evident from the reduced spread of test data points for different observed y values, both on the latent and predicted phenotype scales. Moreover, the improvement in prediction accuracy is most significant for genotypes with high or low observed y values (Supplemental Figure 4), indicating the higher-order model's ability to provide more parsimonious fits for genotypes at the extremes of the measurement scale.

Higher-order epistasis is important for predicting distant genotypes

Many of the datasets analyzed here consist of genotypes generated by introducing mutations to many positions of a wild type or reference genotype. This naturally leads to highly localized samples within the complete genotype space. A critical question then arises regarding whether models trained on these local data can be generalized to genotypes distant from the training samples. In this section, we provide a comprehensive characterization of our trained epistatic models in terms of their performance at predicting the phenotypes for genotypes at increasing distances to the training data. Specifically, we focus on the AAV2-Capsid and the cgreGFP dataset, where the data consist of measurements of genotypes in a dense local cloud concentrated near the wild-type and sparse higher-order mutants at greater mutational distances (Supplemental Figure 1). Furthermore, in both datasets we observed moderate improvement in predictive performance on randomly

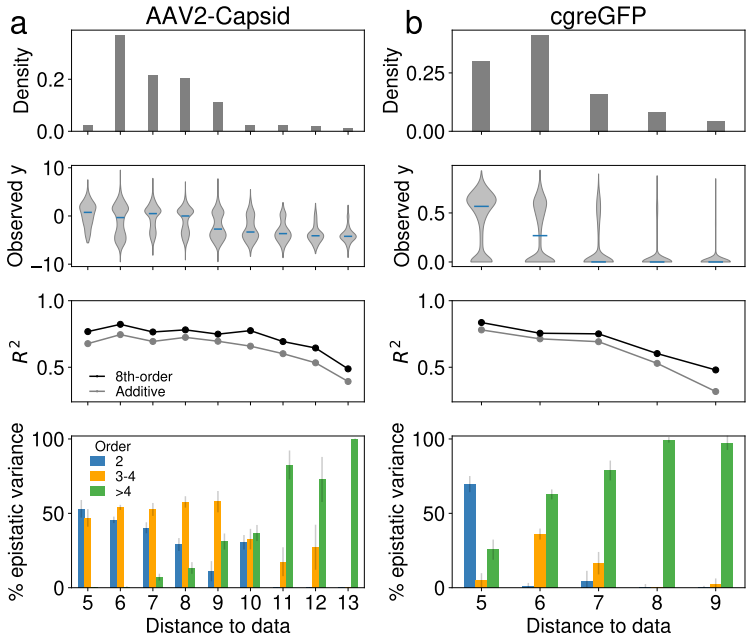

Figure 5: Importance of higher-order epistasis in predicting phenotypes for distant genotypes in the AAV2-Capsid (a) and the cgreGFP (b) datasets. For each dataset, genotypes are binned to discrete distance classes by their mean Hamming distances to the training data, which consist of 20% of randomly sampled genotypes. For both datasets, we retain only distance classes where the additive model has a test $R^2 > 0.3$. Top panel: Distribution of mean Hamming distance in the randomly sampled test data. Second panel: Distribution of the observed phenotypic values (y) for each distance class. Third panel: Test R^2 under the additive model and the 8-th order epistatic transformer for genotypes at different distance classes. The gap between the two curves is equal to R^2_{epi} for each distance class. Bottom panel: importance of specific pairwise and higher-order epistasis at different distance classes, measured by percent epistatic variance, equal to the gain in R^2 by fitting an additional layer of MHA, normalized by R^2_{epi}. All metrics were calculated for models fit to one training sample consisting of 20% of randomly sampled genotypes. Error bars represent 1 standard deviation calculated by bootstrapping the test data with 10 replicates.

sampled test genotypes by incorporating higher-order epistasis. Thus, we are interested in whether modeling higher-order epistasis leads to more pronounced performance gains in certain classes of genotypes.

Specifically, we used the mean Hamming distance as a measurement of the proximity of a genotype to the training data, such that for a test genotype x, its mean Hamming distance $\bar{d}(x) = \frac{1}{n}\sum_{x'} D(x, x')$, where $D(x, x')$ is the Hamming distance between x and a training genotype x'. We then grouped test genotypes into discrete bins by their mean Hamming distance and calculated the percent epistatic variance values for each bin using epistatic transformer models trained with a random training set consisting of 20% of the total data. Here we focused on models trained on 20% of randomly sampled data, instead of 80% as in Figure 3, to increase the number of test genotypes at higher distance classes from the training data, thereby enabling a better assessment of the model's generalizability.

For the AAV2-Capsid dataset, we first observed that the density of test genotypes decreases with mean distances, as a result of the localized data distribution (Figure 5a, top panel). This is accompanied by an overall decrease in the observed phenotypic scores (y) for genotypes in each class (Figure 5a, second panel). Furthermore, we also found that the model performance decreases monotonically with mean distance, such that the R^2 of the additive model starts out at around 0.8 for nearby genotypes and drops to 0.4 for distant genotypes (Figure 5a, third panel). The R^2 for the full epistatic transformer model with 3 MHA layers showed similar trend of decline with increasing distance, but maintained a relatively constant advantage over the additive model, with R^2_{epi} ranging from 0.05 to 0.14. Importantly, we found that the contribution of higher-order interactions measured in terms of percent epistatic variance increases substantially with distance to the data (Figure 5a, bottom panel). Specifically, for nearby genotypes, the pairwise and the 4-th order model explain similar proportions of variance. However, as we move to more distant genotypes, the proportion of variance explained by the pairwise model decreases, which is accompanied by a dramatic increase in the contribution from higher-order interactions, such that at distance > 10, interactions of order 3-8 virtually account for all variance due to epistasis, translating to $10 - 12\%$ of additional total variance explained in the test data.

One drawback of using the R^2 values within a distance class as a measure of model performance is that it may be sensitive to the overall phenotypic variance within a distance class. For example, if all genotypes within the class have the same score (e.g. are all nonfunctional), the models can predict the correct constant, but this accuracy can not be correctly characterized by the R^2 value. Thus, we repeat the same analysis above by comparing the mean squared errors (MSE) between the model predictions and the true phenotypes within each distance classes (Supplemental Figure 5). The new analysis with MSE confirms the results in Figure 5, in that the pairwise model reduces the prediction error of the additive model primarily for genotypes close to the data, and that the performance gain by higher-order models increases with more distant genotypes.

We found similar data and phenotype distribution in the cgreGFP dataset (Figure 5b, first three panels). The predictive accuracy of our models drop more rapidly than in the AAV2-Capsid dataset, such that the test R^2 of the additive model drops below 0.34 for genotypes at mean distance > 9. The performance of the 3-layer epistatic transformer model also decreases with distance, but at a slower rate than the additive model, with a $R^2 = 0.48$ at distance 9, equivalent to $R^2_{\text{epi}} = 0.14$, compared with $R^2_{\text{epi}} = 0.04$ distance 5. The pairwise interactions have a strong contribution (accounting for 61% of epistatic variance) at distance 5. However, for further distances, higher-order interactions become increasingly more important, such that at distance ≥ 8, almost all the epistatic variance is due to interactions among more than 4 sites, which explains an additional 10% of the total variance within the distance classes. Similar results were found when the analysis were repeated using MSE (Supplemental Figure 5).

In addition to the AAV2-Capsid and the cgreGFP dataset, we also apply the same analyses to four other datasets (His-S2, -S5, -S12, and ppluGFP) where data contains sufficient numbers of distance classes (Supplemental Figure 6). We observed that the higher-order epistasis plays moderate to important roles at further distances in three of the four datasets, with the exception being the His-S12, where little effects of epistasis was observed at all distances. We observed that higher-order epistasis plays a moderate to significant role at greater distances in three of the four datasets, with the exception of the His-S12 dataset, where epistatic effects were minimal or absent at all distances.

Together, our analyses using the AAV2-Capsid and the cgreGFP dataset showed that higher-order interactions can be more important for generalizing models trained on locally sampled data to distant regions of the sequence-function relationship. Furthermore, assessing the importance of higher-order epistasis using test data drawn from the same distribution as the training samples (e.g. in Figure 3) can be misleading as the signal of higher-order epistasis may be much more prominent for distant genotypes, but this can be easily obscured as distant genotypes typically consist of a small proportions of the all available data.

Higher-order epistasis for a multi-peak fitness landscape

Most of the datasets we have considered so far (with the exception of the GRB datasets and the CreiLOV dataset) consist of locally sampled genotypes generated by mutagenizing a WT sequence. These local fitness landscapes may not be ideal for detecting higher-order epistasis, which exerts its effects by modifying mutational effects or pairwise interactions across different genetic backgrounds ([38]). For locally sampled data, these effects may exhibit strong similarity due to lack of divergence between the genetic backgrounds, thereby making higher-order epistasis challenging to detect. To circumvent this limitation of locally sampled fitness landscape data, in this section, we examine the importance of higher-order epistasis by modeling a multi-peak landscape consisting of locally sampled data for four protein orthologs. In particular, the dataset we use is derived from the deep mutational scanning experiments for four full-length green fluorescent proteins, avGFP, amacGFP, cgreGFP, and ppluGFP [20] (note that cgreGFP and ppluGFP2 were studied individually in the previous section). A single dataset was generated by merging the four GFP datasets based on the multiple sequence alignment of the four WT sequences [20]. The four proteins exhibit moderate to very high sequence divergence ranging from 18% (between avGFP and amacGFP) to 83% (between ppluGFP2 and amacGFP) (Figure 6a).

This divergence between the genetic backgrounds among the four local GFP peaks can lead to higher variability in the local mutational and epistatic effects than when the local peaks are individually considered, which may allow us to better detect higher-order epistasis. A previous analysis of this combined dataset [20] showed that the four fitness peaks shared some gross similarities. For example, all GFPs show a threshold effect where the fluorescence level typically exhibits a sudden drop for variants accumulating mutations past a certain number. The authors also found that mutations in the chromophore are uniformly deleterious, while

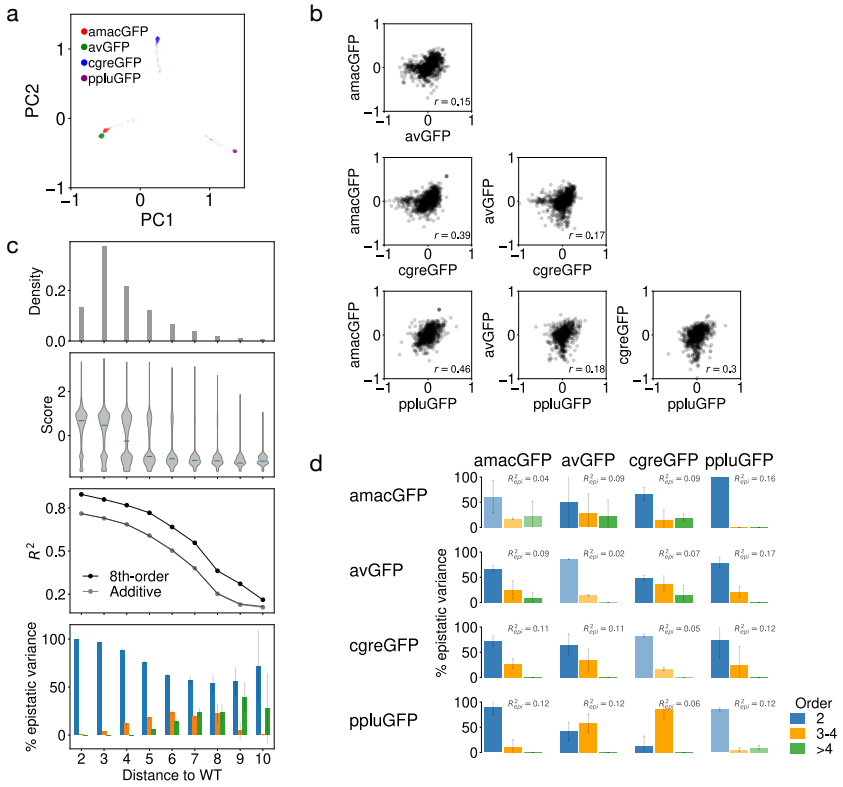

Figure 6: Higher-order epistasis in a multi-peak fitness landscape, consisted of four green fluorescent protein (GFP) orthologs (avGFP, amacGFP, ppluGFP2, cgreGFP). a. PCA coordinates for the one-hot embeddings of the all protein genotypes. Genotypes are extremely centered around the four wild types (WT), which exhibit varying degrees of sequence divergence. b. Scatter plots of shared mutational effects among the four GFPs, fit using separate additive model with a nonlinear sigmoid activation function. c. Higher-order epistasis allows better generalization to distant regions in sequence space. Models were fit use single and double mutant data for all GFPs. Models were tested in different distance classes, each containing all genotypes at given Hamming distance to their corresponding WT sequence. Error bars represent 1 standard deviation calculated by resampling random 90% of the test genotypes with 10 replicates. d. Higher-order epistasis allows better generalization across local peaks. Rows: GFP orthologs used to train the models. Columns: GFP orthologs used to test performance of the trained model. R^2_{epi} at top right corner of each panel measures the proportion of total variance explained by specific epistasis among up to 8 positions. Error bars correspond to 1 standard deviation across three model replicates with different training and test genotypes.

mutations on buried residues had stronger effect than surface sites. But the authors also noted substantial differences in the structure of these local landscapes, including the sharpness of the peaks and the importance of local epistasis [20]. To provide additional coarse characterization of the dissimilarity among the four local peaks, we first directly examine how the effects of mutations vary between GFP orthologs. Specifically, for each dataset we fit a model with additive mutational effects and a sigmoidal activation function. The parameters of the sigmoid function was shared among all four datasets to ensure the inferred mutational effects are on the same scale. In Figure 6b, we showed the scatter plots for the effects of shared mutations between pairs of GFPs. We found that mutational effects overall have low to moderate correlation between GFP orthologs, with Pearson r ranging between 0.15 to 0.46. Interestingly, we did not observe correlation between sequence divergence and mutational effect similarity. For example, the strongest correlation ($r = 0.46$) is between amacGFP and ppluGFP2, which have 83% sequence divergence.

With this basic understanding of the divergence between the local GFP fitness landscapes, we proceed to quantify the importance of higher-order epistasis in this multi-peak landscape. We first fit the epistatic transformer model to randomly sampled variants from all four GFP orthologs. In Supplemental Figure 7, we show the percent epistatic variance values for the pairwise, 4-th, and 8-th order epistatic transformer models trained on 80% of all data on randomly sampled test genotypes from all four local peaks. We first see that fitting epistasis of all orders leads to 12% increase in test R^2. In particular, we found that pairwise

interactions is the predominant form of epistasis, while higher-order epistasis accounts for 15% of the total test set epistatic variance.

This results suggest that pairwise interactions are capable of capturing most of the phenotypic variances among distant local fitness peaks, potentially by accounting for the divergence in local mutational effects (e.g. Figure 6b). Further, higher-order interactions do seem to play a role in determining the overall structure of this multi-peak landscape, although the role assessed by prediction for randomly sample genotypes seems moderate.

Next, we trained our models on two tasks where higher-order epistasis may play more critical roles. For the first task, we fit our models simultaneously to all four GFPs, but only used training data consisting of single and double mutants. We then assessed the performance of pairwise and higher-order epistatic transformer models on predicting phenotypes for distant, higher-order mutants. Our hypothesis is that although these local data only contain information for the effects of mutations and pairwise interactions, our models may nonetheless learn a coarse approximation of the actual logic of higher-order interactions based on how these local effects vary across the four divergent genetic background. This may lead to higher generalizability to distant genotypes, which likely are more affected by higher-order epistasis (e.g. Figure 5).

Our results (Figure 6c) show that the prediction accuracy of both the additive model and the full epistatic model with 3 epistatic MHA layers decreases for test genotypes farther from the local fitness peaks. Interestingly, although the total variance due to epistasis (R^2_{epi}) remains relatively constant, as evidenced by the gap between the two R^2 curves, contributions from higher-order interactions become increasingly important for genotypes more distant to their respective wild types, such that at Hamming distance of 7-9, roughly half of the epistatic variance is due to higher-order interactions.

This implies that higher-order interactions can be learned from assaying only the single and double mutants surrounding different fitness peaks, potentially by modeling how the these local effects vary across divergent genetic backgrounds, and that this can in turn help generalizing models trained on these highly localized samples to distant, higher-order mutants.

For the second task, we fit models with increasing complexity to random training genotypes for each GFP and examine how models fit to individual local peaks can extrapolate to other orthologous GFP landscapes. Our hypothesis is that the higher-order epistasis learned by our model when fit to one local fitness dataset consisting of single, double, and higher-order mutants will help predict how locally observed mutational effects and pairwise epistasis generalize to highly divergent, novel backgrounds, thus enabling better prediction. Specifically, for each GFP, we fit the models using 50% of randomly chosen genotypes and validation data consisting of the most distant 30% of the remaining genotypes. We then assess model generalizability by predicting fitness for random test genotypes for each of the three GFP orthologs. We first note that the the higher-order models explain low to moderate proportions of variance on test genotypes within the same local peak (Figure 6d, diagonal panels). In contrast, when focusing on predicting genotypes from orthologous fitness peaks (Figure 6d, off-diagonal panels), we notice that higher-order interactions frequently show more substantial contributions to total epistatic variance, sometimes on par with or exceeding the proportion of variance explained by the pairwise model (e.g. model trained on ppluGFP, evaluated on avGFP and cgreGFP). This result shows that higher-order epsitasis is important in determining the structure of this multi-peak landscape. Furthermore, by learning higher-order epistasis from data sampled from locally sampled data, our epistatic transformer model is able to achieve better generalization to distant fitness peaks.

Discussion

In this paper, we developed a novel neural network-based method capable of detecting higher-order epistasis in full-length protein sequence-function relationships. The study of epistasis in high-throughput protein datasets has been historically hindered by the combinatorial complexity of fitting higher-order interaction terms, which makes existing methods applicable to protein sequence-function datasets with at most 3-4 positions mutagenized to every other amino acid in a combinatorially complete manner. By contrast, our epistatic transformer model fits epistatic interactions implicitly through a neural network, and thus can be easily applied to study epistasis among hundreds of amino acid positions.

We applied our method to 10 protein sequence-function datasets. Our results showed that higher-order epistasis can have subtle to substantial effects in determining the overall structure of protein-sequence-

function relationships. The strongest effect of higher-order epistasis was observed in the abundance data for the SH3 domain of the protein GRB2, where higher-order epistasis explains 15% of the total variance in test genotypes, representing nearly two thirds of the total variance due to epistasis. This finding contrasts with a previous study [31] where Park et al. discovered no significant contribution by higher-order epistasis in 20 protein sequence-function relationship datasets. This discrepancy can be explained by the different scales of the study datasets. While Park et al. exclusively focused on combinatorially complete sequence-function data for small protein sequence space, we have predominantly used locally sampled sequence-function data embedded in much larger genotypic spaces, where higher-order epistasis likely plays more important roles due to combinatorial and biophysical reasons [51] . Our finding suggests that while some protein sequence-function relationship may be simple and explainable by additive effects, pairwise epistasis, and non-specific epistasis, higher-order specific epistasis can play a critical role in other cases and thus must be properly considered to fully understand the structure of protein sequence-function relationships.

It is also noteworthy that, in most of the other datasets we analyzed, the contribution of higher-order interactions is relatively modest when measured in terms of the total variance they explained. This observation aligns with the conclusions of Park et al. [31], who found that most protein sequence–function relationships can be well modeled using nonspecific and pairwise specific epistasis. One possible explanation for this limited apparent contribution of higher-order interactions is that these protein-sequence function relationships are truly devoid of higher-order interactions. Alternatively, higher-order epistasis may be present but obscured by experimental limitations, potentially due to the choice of mutagenized sites that do not interact strongly, or the limited scope of local sampling which does not ensure adequate statistical power to detect high-order interactions.

Importantly, even when datasets show little evidence of higher-order epistasis in summary statistics like R^2, such interactions can still play a crucial role in determining the phenotypes of a minority of genotypes with idiosyncratic behavior [52]. We demonstrated this by evaluating the importance of higher-order epistasis for making predictions for out-of-distribution genotypes. In particular, we showed that higher-order effects can be essential for predicting phenotypes in sparsely sampled regions of the sequence space, and for generalizing from data centered on one local GFP fitness peak to other, more distant orthologs. These findings suggest that coarse-grained metrics such as R^2 may indeed fail to fully capture the functional significance of higher-order epistasis. Relying on such summary statistics to guide protein engineering or interpret evolutionary patterns can therefore be misleading, particularly when extrapolating beyond the range of observed data into unexplored regions of the sequence space.

Our results may present a challenge to researchers in protein engineering and evolution as they suggest that higher-order epistasis cannot be altogether ignored when studying protein-sequence function relationships. We want to point out that the extent to which researchers should 'worry' [28] about higher-order epistasis depends on the particular application. For example, based on the evidence of the Park et al. study, higher-order epistasis is likely to be less prevalent when only a small number of sites are mutagenized. Similarly, higher-order epistasis may also play a less important role in determining the structure of local sequence-function relationships. However, if the researcher is interested in how large numbers of mutations across the whole protein combine to determine function, then higher-order epistasis should be properly incorporated to enable an faithful construction of the sequence-function relationship. Reassuringly, we found that higher-order epistasis can be effectively modeled based on moderate size data to allow high overall prediction accuracy.

Methods

Epistatic transformer architecture

In this section, we provide a detailed description of the architecture of the epistatic transformer model in Figure 1. Our input x consists of tokenized protein sequences, i.e. $x = x_1, x_2, \cdots, x_L$, where x_l is an integer between 0 and $a-1$ (a is the alphabet size, which depends on the study dataset). We used a special positional encoding to transform the raw input sequence such that for the new sequence x', $x'_l = x_l + (l-1) \times a$. That is, we essentially treat amino acids from different positions completely differently. This is different from typical protein language modeling where a continuous positional encoding is added to the amino acid embeddings, such that the identities of the amino acids are still largely preserved to allow the model to learn general rules of interaction among residues while informed by residue positions when trained on large number of natural protein sequences. In contrast, here we aim to have the maximum flexibility to allow the same amino acid combinations to have different interactions at different positions, in order to learn a model that well approximates the parameterization in Eq. 2.

The new tokenized sequence x' is passed through an embedding layer to generate position-wise continuous representations $Z^0 = (z_1^0, \cdots, z_n^0)$, where each position consists of d hidden dimensions. The embeddings are then passed through M layers of modified multi-head attention (MHA) to model epistatic interactions among sites. Our epistatic MHA layer is built on the standard MHA layer, but with some key differences. Specifically, the input $Z_m \in \mathbb{R}^{d \times L}$ from MHA layer m is first used to generate the query and key tensors for H attention heads in layer $m+1$, such that for head h, the query and key tensors are

$$Q_h = W_h^Q Z_m, \quad W_h^Q \in \mathbb{R}^{d/H, d/H} \tag{3}$$

$$K_h = W_h^K Z_m, \quad W_h^K \in \mathbb{R}^{d/H, d/H}. \tag{4}$$

Importantly, the value tensor V is directly generated from the raw embedding Z_0, instead of Z_m

$$V_h = W_h^Q Z_0. \tag{5}$$

In the Supplement, we show how this modification is essential to ensure that M layers of MHA leads to a models that strictly contain specific epistasis of order up to 2^M. Next, for each head, we perform a modified scaled dot attention (Figure 1)

$$\text{Attention}(Q_h, K_h, V_h) = \left(\frac{Q_h^T K_h}{\sqrt{d}} \right) V_h. \tag{6}$$

The key difference between this equation and the standard scaled dot attention is the lack of a softmax normalization, which is used to convert the raw attention scores obtained from the dot product of the query and key vectors into a probability distribution over the input positions. Since in the standard scaled dot attention the softmax operation would be applied across positions for each embedding dimension, it acts similarly to a global epistasis nonlinearity. This would cause the hidden nodes in the model to contain interactions of all orders, which is unsuitable for our purpose of only fitting specific epistasis up to a particular order. Using our modified scaled dot attention, the output from each head is then concatenated and passed through a linear layer to generate the output Z_{m+1}

$$Z_{m+1} = \text{MultiHead}(Q_h, K_h, V_h) = W^O \text{Concat}(\text{head}_1, \ldots, \text{head}_h), \quad W^O \in \mathbb{R}^{d \times d}. \tag{7}$$

The output of the final Mth MHA layer is then flattened and converted to a scalar ϕ

$$\phi = \text{vec}(Z_M)^T w + b. \tag{8}$$

The scalar ϕ can be interpreted as a hidden phenotype that only contains additive effects and effects due to specific interactions. Note that our construction of the model architecture makes sure that ϕ only contains specific interactions of order up to 2^M. The scalar ϕ is then passed through a sigmoid function mapping the hidden phenotype to the measurement scale to account for any non-specific epistasis

$$\hat{y} = \frac{a}{1 + e^{-(\phi - \mu)/s}} + b, \tag{9}$$

where a, μ, s, and b are trainable model parameters, which allow us to shift and scale the standard sigmoid function. We also experimented with more complex nonlinear activation functions parameterized by a sum of independent sigmoid functions. We did not observe improvement with this method over Eq. 9. Thus we trained all models using Eq. 9 as the global epistasis function. Another key difference between our model and standard transformer models is the lack of LayerNorm operation which is typically used in transformer models to normalize the activations of neurons within a layer. Similar to the softmax function, LayerNorm is a nonlinear function over all positions, which will introduce undesirable global epistasis effect in the hidden layers.

Model training

For all datasets studied in the main text, we train the epistatic transformer model with $1, 2$, and 3 layers of attention, corresponding to models with specific epistasis among up to $2, 4$ and 8 sites. In addition to the epistatic models, we also trained an additive model to benchmark the importance of specific epistasis. The additive model uses the same sigmoid activation function in Eq. 9 to model nonspecific epistasis. However, its hidden phenotype ϕ was calculated as a simple weighted sum of the flattened one-hot embeddings of the input sequence, which was used to model the sum of additive effects of all mutations.

As mentioned in the main text, we use Optuna [48] to optimize hyperparameters including hidden dimension (d), number of train epochs, batch size, and dropout rate. Optuna uses a Bayesian optimization algorithm called Tree-structured Parzen Estimator to model the history record of trials to determine which hyperparameter values to use for the next iteration. We use the R^2 value on a validation dataset to guide 200 Optuna optimization steps for all models. In our preliminary study, we found that learning rate and the number of attention heads have little effect on the final model performance, we therefore used 8 attention heads and the Adam optimizer with fixed learning rate (0.001) to train all models. All code was written in PyTorch v2.2.0. Individual models were trained on one NVIDIA Ampere A100 GPU with 80GB memory.

Selection of experimental datasets

We used the following criteria for selecting datasets for detecting higher-order epistasis by fitting our epistatic transformer model. First, the dataset must contain high-throughput measurements for proteins, and ideally embedded in a large sequence space to allow potentially non-significant contribution by higher-order epistasis. Second, the dataset must contain substantial numbers of higher-order mutants, i.e. genotypes with at least three mutations relative to the reference (WT) sequence, to allow higher-order epistasis to play a role. We found that combinatorial mutagenesis datasets satisfying these criteria are rare compared with the large body of studies that measure single mutant and/or double mutant effects. Our research result in 7 publications. However, initial fitting of our models to datasets from two publications [53, 54] resulted in very low model performance ($R^2 < .1$ for all models), which were subsequently discarded. The final datasets are listed in Table 1, containing data from five publications.

Three publications contain measurements for multiple proteins or functions. We used all three datasets from the 'GRB2-SH3 domain' study by Faure et al. [23]. The His3 publication [22] contains fitness measurements for combinatorial variants in 12 non-overlapping segments tiling the His3 protein. We chose segment 2, 5, and 12 for our study as these three segments showed the lowest R^2 when we fit an additive + sigmoid model, thus allowing for more room of improvement by the epistatic models. The GFP publication [20] contains fluorescence data for three GFPs (amacGFP, ppluGFP, and cgreGFP), complemented by a fourth GFP (avGFP) [43] from an earlier study. For examining our model performance on random training data, we only used the cgreGFP and the ppluGFP datasets. The aligned dataset containing all four GFPs were used to detect the importance of higher-order epistasis in the multi-peak sequence-function relationship formed by these orthologs.

Data Availability

Computer code to replicate all analyses can be found at `https://github.com/juannanzhou/EpistaticTransformer`.

Acknowledgement

We thank David McCandlish for thoughtful comments and suggestions. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM154908, and University of Florida College of Liberal Arts & Sciences.

Author contributions

J.Z planned research; P.S and J.Z performed research; and J.Z wrote the paper.

References

[1] Philip A Romero and Frances H Arnold. "Exploring protein fitness landscapes by directed evolution". In: *Nature reviews Molecular cell biology* 10.12 (2009), pp. 866–876.

[2] Charlotte M Miton and Nobuhiko Tokuriki. "How mutational epistasis impairs predictability in protein evolution and design". In: *Protein Science* 25.7 (2016), pp. 1260–1272.

[3] Tyler N Starr and Joseph W Thornton. "Epistasis in protein evolution". In: *Protein science* 25.7 (2016), pp. 1204–1218.

[4] Rosalie Lipsh-Sokolik and Sarel J Fleishman. "Addressing epistasis in the design of protein function". In: *Proceedings of the National Academy of Sciences* 121.34 (2024), e2314999121.

[5] Douglas M Fowler et al. "High-resolution mapping of protein sequence-function relationships". In: *Nat. Methods* 7.9 (2010), pp. 741–746.

[6] Lea M Starita et al. "Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis". In: *Proc. Natl. Acad. Sci. U.S.A.* 110.14 (2013), E1263–E1272.

[7] Daniel Melamed et al. "Deep mutational scanning of an RRM domain of the *Saccharomyces cerevisiae* poly (A)-binding protein". In: *RNA* 19.11 (2013), pp. 1537–1551.

[8] C Anders Olson, Nicholas C Wu, and Ren Sun. "A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain". In: *Curr. Biol.* 24.22 (2014), pp. 2643–2651.

[9] Michael B Doud, Orr Ashenberg, and Jesse D Bloom. "Site-specific amino acid preferences are mostly conserved in two closely related protein homologs". In: *Mol. Biol. Evol.* 32.11 (2015), pp. 2944–2960.

[10] Anna I Podgornaia and Michael T Laub. "Pervasive degeneracy and epistasis in a protein-protein interface". In: *Science* 347.6222 (2015), pp. 673–677.

[11] Karen S Sarkisyan et al. "Local fitness landscape of the green fluorescent protein". In: *Nature* 533.7603 (2016), p. 397.

[12] Barrett Steinberg and Marc Ostermeier. "Shifting fitness and epistatic landscapes reflect trade-offs along an evolutionary pathway". In: *J. Mol. Biol.* 428.13 (2016), pp. 2730–2743.

[13] Claudia Bank et al. "On the (un)predictability of a large intragenic fitness landscape". In: *Proc. Natl. Acad. Sci. U.S.A.* 113.49 (2016), pp. 14085–14090.

[14] Tyler N Starr, Lora K Picton, and Joseph W Thornton. "Alternative evolutionary histories in the sequence space of an ancient protein." In: *Nature* 549.7672 (Sept. 2017), pp. 409–413.

[15] Victoria O Pokusaeva et al. "An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape". In: *PLos Genet.* 15.4 (2019), e1008079.

[16] Calin Plesa et al. "Multiplexed gene synthesis in emulsions for exploring protein functional landscapes". In: *Science* 359.6373 (2018), pp. 343–347.

[17] Drew S Tack et al. "The genotype-phenotype landscape of an allosteric protein". In: *Molecular systems biology* 17.3 (2021), e10179.

[18] Tyler N Starr et al. "Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding". In: *Cell* 182.5 (2020), pp. 1295–1310.

[19] Louisa Gonzalez Somermeyer et al. "Heterogeneity of the GFP fitness landscape and data-driven protein design". In: *eLife* 11 (May 2022). Ed. by Daniel J Kliebenstein and Naama Barkai, e75842. ISSN: 2050-084X. DOI: `10.7554/eLife.75842`. URL: `https://doi.org/10.7554/eLife.75842`.

[20] Louisa Gonzalez Somermeyer et al. "Heterogeneity of the GFP fitness landscape and data-driven protein design". In: *Elife* 11 (2022), e75842.

[21] Drew H Bryant et al. "Deep diversification of an AAV capsid protein by machine learning". In: *Nature Biotechnology* 39.6 (2021), pp. 691–696.

[22] Victoria O Pokusaeva et al. "An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape". In: *PLoS genetics* 15.4 (2019), e1008079.

[23] Andre J Faure et al. "The genetic architecture of protein stability". In: *bioRxiv* (2023), pp. 2023–10.

[24] Justin B Kinney and David M McCandlish. "Massively parallel assays and quantitative sequence–function relationships". In: *Annual review of genomics and human genetics* 20 (2019), pp. 99–127.

[25] Patrick C Phillips. "Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems". In: *Nat. Rev. Genet.* 9.11 (2008), pp. 855–867.

[26] Dmitry A Kondrashov and Fyodor A Kondrashov. "Topological features of rugged fitness landscapes in sequence space". In: *Trends Genet.* 31.1 (2015), pp. 24–33.

[27] Júlia Domingo, Pablo Baeza-Centurion, and Ben Lehner. "The Causes and Consequences of Genetic Interactions (Epistasis)". In: *Annu. Rev. Genomics Hum. Genet.* 20 (2019).

[28] Daniel M Weinreich et al. "Should evolutionary geneticists worry about higher-order epistasis?" In: *Curr. Opin. Genet. Dev.* 23.6 (2013), pp. 700–707.

[29] Juannan Zhou and David M McCandlish. "Minimum epistasis interpolation for sequence-function relationships". In: *Nature Communications* 11.1 (2020), pp. 1–14.

[30] Frank J Poelwijk, Michael Socolich, and Rama Ranganathan. "Learning the pattern of epistasis linking genotype and phenotype in a protein". In: *Nat. Commun.* 10.1 (2019), pp. 1–11.

[31] Yeonwoo Park, Brian PH Metzger, and Joseph W Thornton. "The simplicity of protein sequence-function relationships". In: *bioRxiv* (2023).

[32] C Anders Olson, Nicholas C Wu, and Ren Sun. "A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain". In: *Current biology* 24.22 (2014), pp. 2643–2651.

[33] Brian PH Metzger et al. "Epistasis facilitates functional evolution in an ancient transcription factor". In: *Elife* 12 (2024), RP88737.

[34] Daniel M Weinreich et al. "The influence of higher-order epistasis on biological fitness landscape topography". In: *Journal of Statistical Physics* 172.1 (2018), pp. 208–225.

[35] Nicholas C. Wu, C. Anders Olson, and Ren Sun. "High-throughput identification of protein mutant stability computed from a double mutant fitness landscape". In: *Protein Sci.* 25.2 (2016), pp. 530–539. ISSN: 1469896X. DOI: 10.1002/pro.2840.

[36] Daniel M. Weinreich et al. "The Influence of Higher-Order Epistasis on Biological Fitness Landscape Topography". In: *J. Stat. Phys.* 172.1 (2018), pp. 208–225. ISSN: 00224715. DOI: 10.1007/s10955-018-1975-3. URL: https://doi.org/10.1007/s10955-018-1975-3.

[37] Gloria Yang et al. "Higher-order epistasis shapes the fitness landscape of a xenobiotic-degrading enzyme". In: *Nature Chemical Biology* 15.11 (2019), pp. 1120–1128.

[38] Juannan Zhou et al. "Higher-order epistasis and phenotypic prediction". In: *Proceedings of the National Academy of Sciences* 119.39 (2022), e2204233119.

[39] Jakub Otwinowski, David Martin McCandlish, and Joshua B Plotkin. "Inferring the shape of global epistasis". In: *Proceedings of the National Academy of Sciences* 115.32 (2018), E7550–E7558.

[40] Ammar Tareen et al. "MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect". In: *Genome biology* 23.1 (2022), p. 98.

[41] Carl Edward Rasmussen and Christopher K I Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

[42] Amirali Aghazadeh et al. "Epistatic Net allows the sparse spectral regularization of deep neural networks for inferring fitness functions". In: *Nature Communications* 12.1 (2021), pp. 1–10.

[43] Karen S Sarkisyan et al. "Local fitness landscape of the green fluorescent protein". In: *Nature* 533.7603 (2016), pp. 397–401.

[44] Zachary R Sailer and Michael J Harms. "Detecting High-Order Epistasis in Nonlinear Genotype-Phenotype Maps". In: *Genetics* 205.3 (2017), pp. 1079–1088.

[45] Peter D Tonner, Abe Pressman, and David Ross. "Interpretable modeling of genotype–phenotype landscapes with state-of-the-art predictive power". In: *Proceedings of the National Academy of Sciences* 119.26 (2022), e2114021119.

[46] Ammar Tareen et al. "MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect". In: *Genome Biology* 23.1 (2022), pp. 1–27.

[47] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[48] Takuya Akiba et al. "Optuna: A next-generation hyperparameter optimization framework". In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, pp. 2623–2631.

[49] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. "Layer normalization". In: *arXiv preprint arXiv:1607.06450* (2016).

[50] Yongcan Chen et al. "Deep Mutational Scanning of an Oxygen-Independent Fluorescent Protein CreiLOV for Comprehensive Profiling of Mutational and Epistatic Effects". In: *ACS Synthetic Biology* 12.5 (2023), pp. 1461–1473.

[51] Eric A Ortlund et al. "Crystal structure of an ancient protein: evolution by conformational epistasis". In: *Science* 317.5844 (2007), pp. 1544–1548.

[52] Thomas Dupic, Angela M Phillips, and Michael M Desai. "Protein sequence landscapes are not so simple: on reference-free versus reference-based inference". In: *bioRxiv* (2024), pp. 2024–01.

[53] Philipp Koch et al. "Optimization of the antimicrobial peptide Bac7 by deep mutational scanning". In: *BMC biology* 20.1 (2022), p. 114.

[54] Remkes A Scheele et al. "Droplet-based screening of phosphate transfer catalysis reveals how epistasis shapes MAP kinase interactions with substrates". In: *Nature Communications* 13.1 (2022), p. 844.