Optimizing Proteins via Generative Models through Hierarchical Data Selection*

Ue-Yu Pen*[1,†] and Vedant Sachdeva*[1,‡]

[1] *Unaffiliated*

Protein engineering through generative machine learning models has enabled the discovery of novel proteins to perform biological functions. Targeting specific levels of function, however, is

I. INTRODUCTION

Lian et al., [1] have shown that variational autoencoders (VAEs) trained by multiple sequence alignment (MSA) data of natural Src homology 3 (SH3) family are able to generate synthetic SH3 variants that express a wild-type $Sho1^{SH3}$ SH3 phenotype.

The SH3 is a small protein domain of about 60 amino acid residues. It binds to type II polyproline-containing peptides and mediates diverse signaling functions in cells. For example, the SH3 domain in the Sho1 transmembrane receptor ($Sho1^{SH3}$) in fungi mediates external osmotic stress response through binding to a polyproline ligand in the Pbs2 MAP kinase. As the Sho1 pathway has been conserved through speciation within the fungal kingdom, one can find a diverse ensemble of extant $Sho1^{SH3}$ ortholog sequences in the natural Sh3 MSA. On the other hand, gene duplication during evolution creates paralogous SH3 domains that have diverged to distinct and non-overlapping ligand specificities. In *S. cerevisiae*, the $Sho1^{SH3}$ is the only SH3 domain that mediates the osmosensing in the Sho1 pathway among 26 other paralogous domains in its genome [2].

[MM: THIS JUMP SEEMS SEVERE. FROM STUFF ABOUT SH3 AND FUNGI TO VAES. SOFTEN IT.] The VAE consists of an encoder $q_\phi(z|x)$ that compresses the information content of sequences in the MSA into low-dimensional latent space vectors z, and a decoder $p_\theta(x|z)$ that transforming latent vectors z back into protein sequences x (Fig. 1A). When trained by the natural protein sequence, the VAE model should learn the underlying design pattern of the input natural protein sequences and be able to generate novel sequences with similar properties of the natural input sequences. In addition, one should expect the VAE latent space to represent the phenotypic and/or phylogenetic relationship between sequences. Sequences generated from latent space coordinates not occupied by the natural train sequences should be novel synthetic sequences with similar phenotypic and/or phylogenetic properties to the surrounding sequences in the latent space.

[MM: A SUMMARY OF THE APPROACH AND CENTRAL RESULTS NEEDS TO BE PUT HERE.]

* These authors contributed equally
† ueyupen@gmail.com
‡ sachdved@gmail.com

II. RESULTS AND DISCUSSION

A. Global and local VAE

In Lian et al., the SH3 MSA used to train the VAEs contains natural SH3 paralogs and their orthologs. The sequences in the same paralogous are found to be grouped together in the VAE latent space. Here we are able to reproduce this observation in our VAE model [SI]. In this paper, we choose a two-dimensional VAE for better visualization while generality is found [MM: BE MORE PRECISE MAYBE? WHAT DOES GENERALITY IS FOUND FOR HIGHER D?] across VAEs with different numbers of dimensions [SI]. To test the phenotype of the sequences, Lian et al. choose the $Sho1^{SH3}$ activity as a target function which can be measured by a high-throughput quantitative select-seq assay [MN: REFERENCE TO THE ASSAY]. This assay utilizes a Sho1 deletion *S. cerevisiae* strain whose growth rate can be made to report the binding free energy between $Sho1^{SH3}$ domain and Pbs2 (Fig. 1B). The result shows the $Sho1^{SH3}$ activity of the natural sequences matches the $Sho1^{SH3}$ ortholog label in the MSA. By color labeling the $Sho1^{SH3}$ activity, the same as reported in [1], sequences with high $Sho1^{SH3}$ activity are grouped together in the VAE latent space. This localization is also found in the VAE-generated Sho1 orthologs (Fig. 1C).

To further investigate the relation between the VAE latent space coordinate and the phenotype of SH3 sequences, we trained a VAE (the GlobalVAE) with all available sh3 (natural and synthetic) sequences to observe the organization of the latent space. Again we choose the 2-dimensional VAE for better visualization. As expected, the sequences with high r.e. are placed in the same latent space region (Figure 2A). However, by zooming into the locality that embeds the sequences with Sho1 activity (high r.e.), as shown in the left panel of figure. 2B, the model fails to separate functional, partial-rescue, and unfunctional sequences. That is, although being able to classify the orthologs in their latent space, the GlobalVAE cannot resolve the phenotype within an ortholog cluster. The ability to do so is important in both scientific and engineering aspects. A latent space organized by function provides predictability, interpolatability, interpretability, and can be used as a "function landscape" that guides the direction of protein design.

[MM: SOMETHING NEEDS TO COME BEFORE THIS SENTENCE. SOMETHING LIKE, OUR APPROACH BLAH BLAH. OTHERWISE, ITS NOT

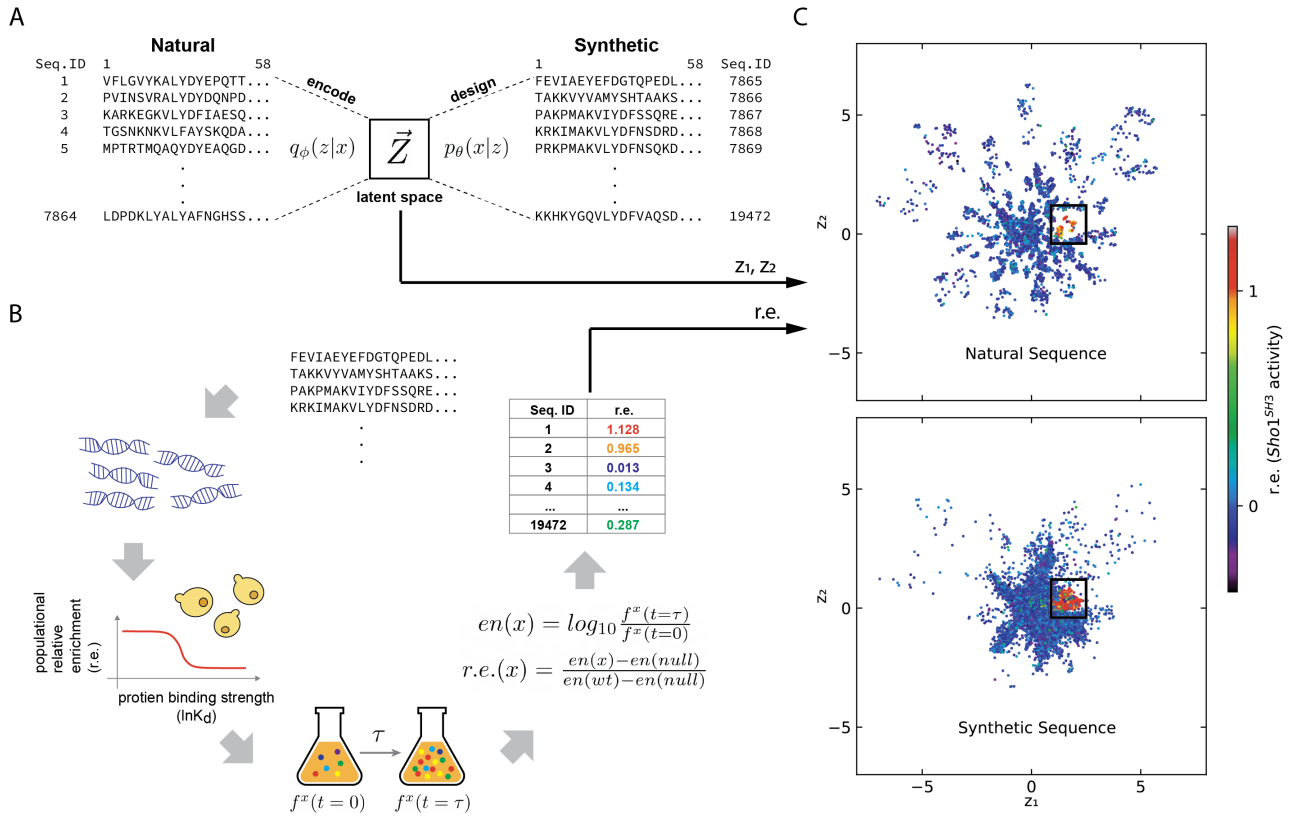

FIG. 1. **Evolutionary-based deep generative models of SH3 domain and high-throughput select-seq assay for $Sho1^{SH3}$ function in *S. cerevisiae*.** (A) Schematic of evolutionary-based generative models. The model consists of an encoder that projects a sequence alignment of natural SH3 homologs to a low-dimensional Gaussian latent space, a latent space where each sequence x is embedded and defined by a vector $\vec{z}$, and a decoder that converts the latent space coordinates to protein sequences. As the VAE model is designed to generate novel data similar to what it was trained on; the sequences decoded from unoccupied latent coordinates can be considered the synthetic member of SH3 protein family. (B) Workflow of characterizing Sho1 functionality of the sequences. Libraries of sequences of interest are constructed and transformed into Sho1-deficient *S. cerevisiae* cells. The variants were then grown under selective conditions for a period τ. Deep sequencing is performed to measure the frequency of each sequence x before (i.e., $f^x(t=0)$ and after (i.e., $f^x(t=\tau)$) selection, which are then used for calculating the enrichment of the sequence x $en(x)$. Finally, the relative enrichment (r.e.) of sequence x is derived by normalizing/subtracting $en(x)$ by the enrichment of wild type Sho1 allele $en(wt)$ and the enrichment of a non-growing "dead" allele $en(null)$ as shown in the equation. While the growing condition and the yeast strain were tested and verified to link the relative abundance (r.e) with protein's $Sho1^{SH3}$ activity (relative binding dissociation constant lnK_d of pbs2 MAPKK ligand), the measurement of r.e. is used as the proxy of the Sho1 protein activity. (C) The two-dimensional latent space for the natural SH3 MSA and synthetic sequence. Colored dots represent the embedding of each sequence in the SH3 MSA. Colors indicate the experimentally measured r.e. ($Sho1^{SH3}$ activity). The black box shows the locality in the latent space where the sequences with active $Sho1^{SH3}$ activity are embedded.

CLEAR THE READER IS ABOUT THE READ THE CENTRAL NOVEL PIECE OF THE STUDY.] To achieve this, we train another VAE (LocalVAE) with the same architecture by subsetted training data. We choose a reference sequence within the high r.e. locality and search in the MSA for sequences within 30 hamming distance as our training data. The latent space of the LocalVAE is shown in Figure. 2C. One can see that the LocalVAE separates the Sho1 activity levels better. This observation is quantitatively verified by the increased predictability of the latent coordinate on the Sho1 activity by using a random forest regressor. In addition, besides giving a better quantitative separation of function, the LocalVAE is able to organize its latent space in the way that the partial rescue sequences are placed between the functional and unfunctional sequences, showing a gradient of sho1 function in the latent space, so that the latent space can be viewed as the fitness landscape of the sho1 function.

B. Impacts of Subsetting

When subsetting the data set to construct our LocalVAE, we select sequences by Hamming distance cutoff from the top-performing sequence (sequence with highest

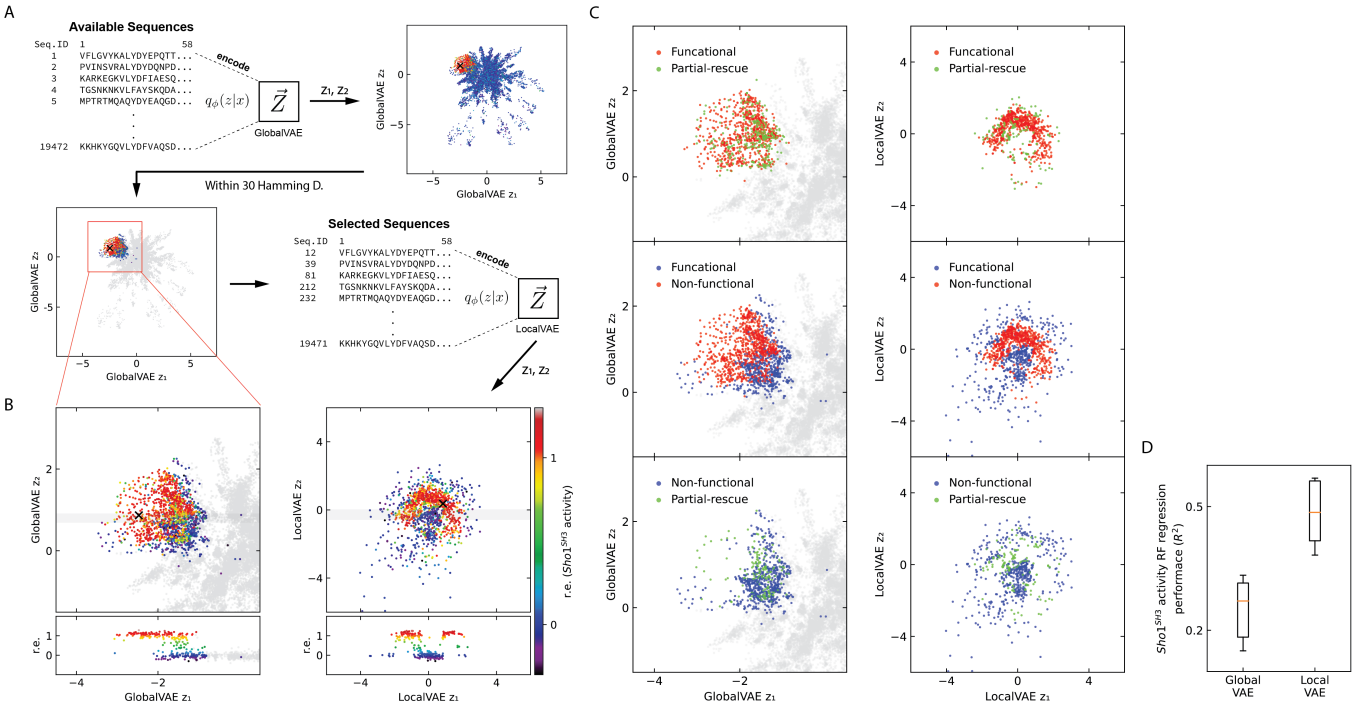

FIG. 2. **The VAE trained on a properly selected subset of MSA data gives better predictability of latent coordinates to Sho1 activity.** (A) Illustration of building the GlobalVAE and LocalVAE. The GlobalVAE is constructed by training a VAE model with the entire available MSA data, while the LocalVAE is trained by a selected subset of the MSA. The training data for the LocalVAE is selected for the sequences that are within 30 hamming distance to the natural $Sho1^{SH3}$. The $Sho1^{SH3}$ and the selected train data are marked as the black "X" and color dots in the GlobalVAE latent space, respectively. (B) Organization of the 2D latent space of the GlobalVAE and LocalVAE. The colored dots indicate sequences and their experimentally measured r.e. ($Sho1^{SH3}$ activity). The grey dots in the left panel represent the sequences whose hamming distance to the natural $Sho1^{SH3}$ sequence (the black "X") is larger than 30. The bottom panels show the relation between a latent coordinate z1 and the relative enrichment within a given area. This area is labeled as the grey stripes in both latent spaces. (C) Organization of GlobalVAE (left panel) and LocalVAE (right panel) latent space. The sequences' $Sho1^{SH3}$ activity is categorized into three classes: Functional ($r.e. \geq 0.8$, red), partial rescue ($0.8 > r.e. \geq 0.2$, green) and non-functional ($r.e. < 0.2$, blue). Only sequences within 30 hamming distance to the natural $Sho1^{SH3}$ are colored in these figures. Grey dots in the left panel indicate the sequences whose hamming distance is larger than 30 to the natural $Sho1^{SH3}$. Sequences with two of three activity classes are plotted in each plot for clarity. (D) Predictability of GlobalVAE and LocalVAE latent space coordinates to the relative enrichment of sequences by a random forest regressor (R^2). Each case is tested with 10 replicates. The box extends from the lower to upper quartile values of the data, while the orange lines represent the medians. The whiskers extend from the box to show the range of the data.

$Sho1^{SH3}$ activity). By reducing the Hamming distance cutoff, we are also increasing the fraction of sequences we train on which are active (Fig. 3A). However, on the other hand, as Hamming distance cutoffs are made stricter, the LocalVAE is trained on fewer sequences. This results in a tradeoff between the relevance of our training dataset to our desired observable and the support for the learned model. Consequently, enhancement in predictive performance peaks at an intermediate value of Hamming distance cutoffs Fig. 3B.

This tradeoff emerges from the phylogenetic organization of the sequences. When constructing a maximum likelihood phylogenetic tree based on the Sh3 protein sequences, it is apparent that the active Sho1 sequences are more closely related to one another than they are to other sequences in the MSA Fig. 3C(i). Consequently, when selecting sequences based on a Hamming distance from the top performer, the effect is an increased focus on the phylogenetic branch corresponding to Sho1-binding domains. However, selecting a too stringent cutoff results in ignoring some key parts of the phylogenetic branch [MM: THIS STATEMENT NEEDS TO BE MADE WITH MORE DETAILS AND MORE CLEARLY]. This suggests an organizing principle behind which the hierarchical data selection can be cast, which is that the navigation between a focus on the global archetype of homologs, and the local archetype of a set of paralogs from a particular phylogenetic branch [MM: I ACTUALLY DONT UNDERSTAND THIS SENTENCE. IT IS PERHAPS THE MOST CRUCIAL STATEMENT IN THE PAPER, BUT SOMETHING ABOUT THE GRAMMAR OR ORDER OF WORDS IS PREVENT-

ING ME FROM GETTING ITS MEANING.]. We argue that both are necessary for generating and evaluating novel protein candidates.

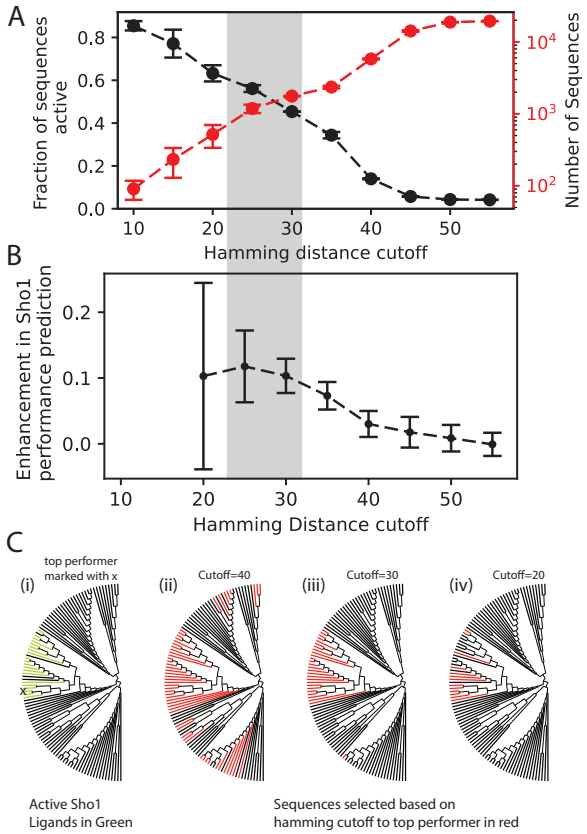

FIG. 3. **Sequence selection based on Hamming distance navigates a tension between relevance and dataset size.** (A) As we select sequences based on the hamming distance to the top performer in the training data set, we enrich our subset of sequences with active sequences, at the cost of the overall number of sequences we can train. (B) Enhancement in prediction performance depends on the tradeoff between the number of sequences in the subset with the relative enrichment of active sequences, resulting in an improvement maximizing at an intermediate value of Hamming distance cutoffs. (C) (i) By constructing a maximum likelihood phylogenetic tree on the SH3 sequences, and then coloring the sequences by $Sho1^{SH3}$ activity, we observe that the active sequences localize to a single branch of the phylogenetic tree. (ii-iv) By selecting different hamming distance cutoffs from the top performer, we observe that the effect focuses more or less on the phylogenetic branch of interest. Too strict a cutoff, however, causes the exclusion of parts of the phylogenetic branch of interest.

C. Parameter Regimes of Hierarchical Data Selection

Given that the benefits of hierarchical data selection emerge from a tradeoff between the local archetype of interest and global rules governing the protein family, the benefits likely depend on VAE model complexity, the amount of available data, and the balance of sequences illustrating the local rules to the sequences illustrating the global rules. We expect that increasing VAE model complexity will increase the likelihood that a VAE can resolve both local and global structures; however, this comes with the risk of increased overfitting. Further, the amount of available data controls the amount of support for the inferred global and local structure, and sufficient data is necessary to resolve it. Finally, the ratio of the amount of data displaying the local rules against the global rules matters, as heavily imbalanced ratios may obfuscate signals from the sequences local to our engineering target. There likely exists a limit where, with enough data, the difference between GlobalVAE and LocalVAE reduces. However, this likely requires a well-sampled dataset, which would be of size near L^{20} where L is the length of amino acid sequence, and 20 corresponds to the number of amino acids. Datasets of this size are intractably large for protein sequences of any meaningful length (Fig. 4A) [MM: GORGEOUS PARAGRAPH. TRULY DEEP.].

When increasing the number of latent dimensions, we observe that although the differences between the GlobalVAE and the LocalVAE's ability to resolve the function of interest decrease, we observe a steady increase in the redundancy of each additional latent dimension. This redundancy was measured through a random forest regression of the n^{th} dimension based on the first $n-1$ dimensions. This redundancy introduces potential pathologies when generating sequences, as only a subspace of the full latent space has actually been used for sequence representation. Consequently, some subspace of the latent space will be unstable for sequence generation (Fig. 4B).

In the regime where there is a benefit to training locally, we observe that the benefit depends on the ratio of the amount of data local to our phylogenetic branch of interest to the global sequence diversity. This manifests itself as the relative enrichment in focus on the phylogenetic branch. We vary this relative enrichment by either masking sequences from the phylogenetic branch from training, or from the global sequence diversity (Fig. 4C). We observe that the area of parameter space for which hierarchical data selection offers an improved predictive performance positively correlates with the relative enrichment of the local data by sequence selection. That is, when the data is imbalanced against the local phylogenetic branch of interest, hierarchical data selection becomes more important (Fig. 4D(i-iii)).

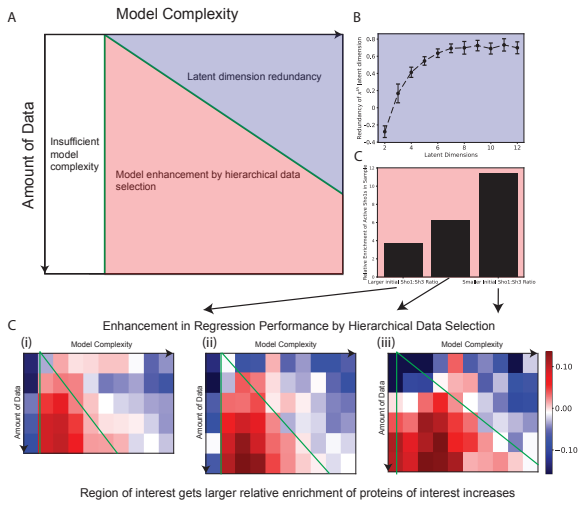

FIG. 4. **Benefits of hierarchical data selection depend on both model complexity and amount of available data.** (A) Hierarchical data selection offers the biggest benefit in regimes in which model complexity is relatively low, and the regime in which hierarchical data selection offers an advantage grows as the amount of data grows. We expect that at some point, with enough data, this regime will begin shrinking again; however, the number of samples needed to reach this threshold may be exceedingly large, as amino acid sequences exist in an extremely high dimensional space. (B) As model complexity grows, VAE latent dimensions become redundant, introducing pathologies to the generative process. (C) In regimes where the predictive capacity of models improves, the benefit depends on the relative enhancement in sequences of interest by hierarchical data selection dependence on the relative fraction of the sequences of interest to the overall number of sequences. (D) (i-iii) When the ratio is small - that is, the sequences of interest represent a large relative fraction of the total number of sequences - the benefit of hierarchical data selection is small, and in a limited regime. However, as this relative enhancement grows, the regime of interest becomes larger.

D. A proposed mechanism for benefits of hierarchical data selection

1. Model

[MM;; CAN WE MAKE THE SAME PHASE DIAGRAM FOR THE MODEL AS WE DID FOR THE DATA IN FIGURE 4? CAN WE SHOW THE 3 PHASES?] Inspired by observations made on the *sh*3 protein family, we will now explore the capacity of a VAE to model both short and long-sequence motifs. Our proposal is that these long sequence motifs, or patterns, resemble the phylogenetic history of a member of a protein family, while short motifs can be thought of as corresponding to performance of a given protein. Further, we will propose a methodology for capturing both statistical structures when a single VAE fails. We begin by generating data from a Potts model with known long sequence motif and short sequence motif structure. We achieve this by writing down an energy function of the form:

$$\mathcal{H}(\mathbf{S}) = -\sum_{i=1}^{P} \alpha f(\mathbf{S}\cdot\xi^{(i)}_{\text{short}})^{\gamma} - (|\mathbf{S}\cdot\xi^{(i)}_{\text{long}}|^{\delta}) \qquad (1)$$

Here, sequences are represented by $\mathbf{S}$, a one-hot encoding of the sequence matrix of size $N \times q$. Here N represents sequence length and q represents the number of entries at any given site. There are P short and long sequence motifs indexed by i, $\xi^{(i)}_{\text{short}}$ and $\xi^{(i)}_{\text{long}}$. Note that by construction, encoding $\xi^{(i)}_{\text{long}}$ into the landscape also encodes $-\xi^{(i)}_{\text{long}}$ with the same energy depth. These can be thought of as anti-patterns. The relative importance of the short-sequence motifs to the long-sequence motifs is controlled by α, and by the function $f(x)$. Here, we take $f(x) = x$ if $x > 0.6N$ and is 0 otherwise. γ and δ control the range of interactions between the sites along the protein.

The probability of any given sequence being observed in a dataset is proportional to its Boltzmann factor, $P(\mathbf{S}) \propto \exp(-\beta\mathcal{H}(\mathbf{S}))$.

For this analysis, we take $\gamma = 1$, $\delta = 2$, and take $q = 2$. This results in the generative model behaving like a Hopfield landscape, where the patterns correspond to the long motifs, under a bias, corresponding to the short motifs. Notably, the short motifs' effect only emerges when the sequence is within 80% identity to one of the encoded pattern or anti-patterns.

2. Results of training a VAE on all sequence data

We begin by encoding 2 full sequence length orthogonal patterns into our landscape through $\xi^{(i)}_{\text{long}}$ into into our landscape and encoding short motifs of length 20% of the sequence, unique to each long motif. That is to say that $\xi^{(1)}_{\text{short}} \cdot \xi^{(2)}_{\text{short}} = 0$. We draw 10000 samples from our landscape. We draw samples by initializing sequences randomly, and then Gibbs sampling for 100 steps at inverse temperature $\beta = 2$.

We then train a VAE with 2 latent dimensions on this data. Our encoder and decoder architecture are 2 128 unit dense layers with scaled exponential linear activations. 2 latent dimensions are a natural choice, as there are only two patterns encoded into the generating distribution. Our loss function is based on a reconstruction term, given by the binary cross entropy, and a maximum mean discrepancy (MMD) loss, ensuring that the latent space remains informative of the input variables, while being constrained to be an uncorrelated univariate Gaussian. For comparison, we also present the results of the a 2 dimensional VAE trained on the Sh3 protein sequence family.

VAE representations of data generated from the phenomenological model show qualitative similarities to the

representation of data from Sh3. Sequences of high identity cluster together in both Sh3-trained VAEs and model-trained VAEs, and low identity is shown for inter-cluster sequences Fig. 5(A,B). Further, we observe that the locations along which high entropy exists differ between clusters in Sh3. By the construction of our short motifs, we similarly observe that positions of high entropy differ between clusters Fig. 5(C,D).

We now treat short motif overlap as a target variable for regression, with the features being the latent space representation of the sequence data. We represent short motif overlap in the latent space (6A) and observe that in some clusters, a gradient corresponding to the short motif overlap emerges, even without supervision (6B). This mirrors results observed in[frances arnold paper on gmp]. However, in other clusters, the local gradient does not emerge(6C). This results in our ability to predict short motif overlap being contingent on the region of latent space to which a sequence embeds. This illustrates that the GlobalVAE does not represent all phylogenetic branches equivalently. However, by training on individual phylogenetic branches, which, in this case, means clusters of sequences, we are able to recover well-organized latent spaces in LocalVAE (Fig. 6D,E). As a part of the training, we remove positions that have low entropy, as they do not contribute the predictions on short motif overlap. We now observe that the embedding of the VAE just on sequences in a given cluster shows local gradients that can be used for prediction. This suggests that while the VAE trained on all sequence data could not resolve all signals simultaneously, the VAE trained on just one cluster of sequences can resolve the signal, providing a potential mechanism for hierarchical data selection.

[1] R. R. A. F. Xinran Liana, Niksa Praljakb, Deep learning-enabled design of synthetic orthologs of a signaling protein, Biophysical Journal **122**, 311a (2023).

[2] W. A. L. A. Zarrinpar, S.-H. Park, Optimization of specificity in a cellular protein interaction network by negative selection, Nature **426**, 676 (2003).

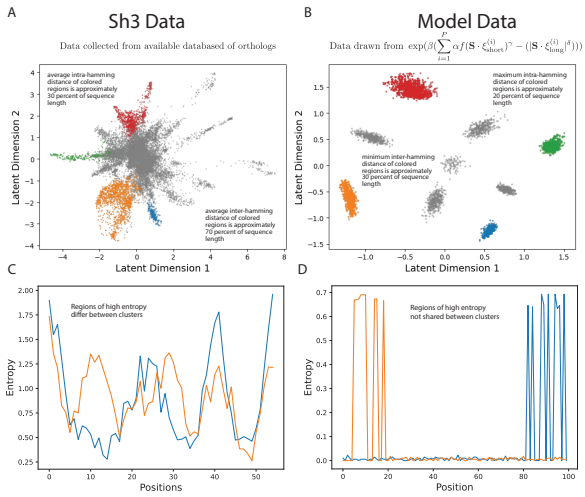

FIG. 5. **A simple statistical physics model captures the phenomenology of proteins and can be used to explore mechanisms for how hierarchical data selection works.** (A) In the Sh3 domains, sequences organize into clusters by sequence identity. Intra-cluster identity is typically quite high, while inter-cluster identity is quite low. (B) VAEs organize data drawn from our statistical physics model in a similar way, showing high intra-cluster identities with low inter-cluster identities. (C) Further, in Sh3 data, the distinct VAE clusters of sequences have high entropy at different sites, suggesting differences in the modes of variation within each cluster. (D) Similarly, the VAE organizes sequence data drawn from our statistical physics model into groups where sites of high entropy differ.

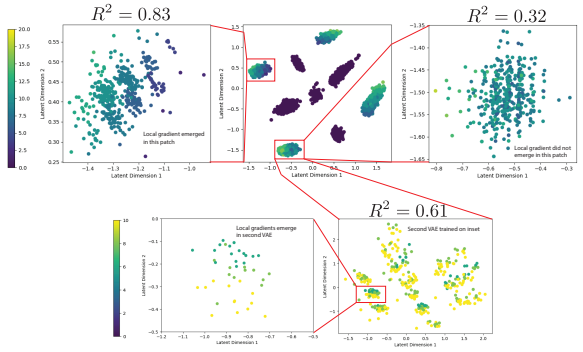

FIG. 6. **Hierarchical data selection improves local gradient organization in cases where not all signals can be resolved simultaneously** (A) Here, we present the embeddings of the sequence data on a VAE trained on all available sequence data in a training set, and color the embeddings by the corresponding sequence activity. (B) We observe that some clusters organize into local gradients of activity (C) while others do not. (D) However, by training a second VAE on the sequence data that did not organize by phenotypic activity, the VAE only needs to resolve the local gradients of the single cluster of sequences. (E) This yields improved organization of the sequence data and enables heightened predictive capacity of the model.