Separation of Scales in the Protein Universe[*]

The protein universe is dictated of rules that govern foldability, functionality, and phylogeny. These rules, however, are optimized over different timescales, and emerge at different strengths in the data. Here, we demonstrate that a model that explains one set of rules cannot explain another, due to the separation of scales. We then propose a method for building a model and making predictions at each of these scales.

I. TESTING ON SYNTHETIC DATA

A. Model

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. 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:

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

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, $\xi_{\text{short}}^{(i)}$ and $\xi_{\text{long}}^{(i)}$. Note that by construction, encoding $\xi_{\text{long}}^{(i)}$ into the landscape also encodes $-\xi_{\text{long}}^{(i)}$ 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 $x > 0.6N$. 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.

B. Impacts of Overcompression and Undercompression in VAEs

We begin by encoding 2 full sequence length orthogonal patterns into our landscape through $\xi_{\text{long}}^{(i)}$ into into our landscape and encoding 0 shorter sequence length motifs, and drawing 10000 samples from our landscape. We draw samples by initializing sequences randomly, and then Gibbs sampling for 100 steps at temperature $\beta = 4$.

We then train a VAE with 2 latent dimensions on this data and 4 latent dimensions. 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.

Each of the training data's points are embedded in the latent space of the two dimensional VAE (Fig. 1A). The dashed lines correspond to trajectories of single steps that transform a $\xi_{\text{long}}^{(i)}$ to $-\xi_{\text{long}}^{(i)}$. We observe that the trajectories for one pattern are orthogonal to the trajectories of the other pattern, showing that the latent dimensions preserves the intuitive expectation - the orthogonality of patterns in sequence space appears in the latent space, and the axes directly correspond to the pattern-antipattern directions. Further, we observe that the VAE organizes sequences that are near a long motif - and hence, near a ground state in the energy function - to be distant from the origin of latent space (Fig. 1B). The four-dimensional VAE, however, appears to break this observation. Pattern-antipattern paths seem to be randomly distributed throughout latent space and cannot be readily interpreted (Fig. 1C, D). However, by linearly transforming the axes to be defined $Z(\xi_{\text{long}}^{(i)}) - Z(-\xi_{\text{long}}^{(i)})$ and the null space thereof, we can recover the order observed in the two-dimensional latent space (Fig. 2A), with two additional auxiliary dimensions that are noisy (Fig. 2B). We can thus connect VAEs trained on the same dataset with different numbers of latent dimensions. This result has implications for models with supervised latent dimensions, as this supervision could be recast as a linear transformation of an unsupervised VAE.

The results we showed above explain how VAEs behave when properly compressing and undercompressing datasets. We will now consider the overcompression

[*] equal contribution

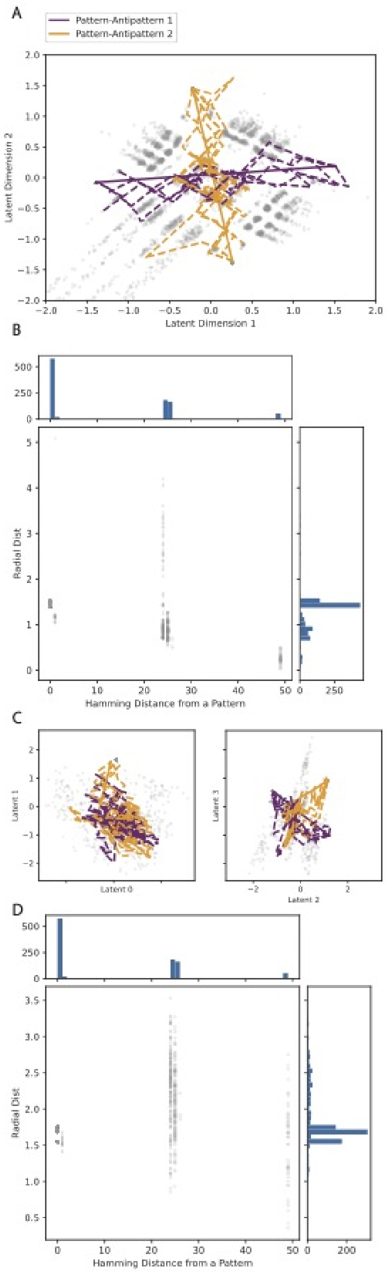

FIG. 1. **Results of a trained VAE with two latent dimensions (A, B) and four latent dimensions (C, D)on a dataset drawn from a landscape with two encoded patterns.** (A) With two encoded patterns, we observe that interpolating between a pattern and its degenerate antipattern results in traversing a straight line in between two latent coordinates. Further, the pattern-antipattern pairs become orthogonal from one another. (B) We find that radial distance becomes inversely correlated with hamming distance from one of the encoded patterns or antipatterns, with a correlation of -0.57. (C) In a VAE with four latent dimensions, we find an apparent breaking of our pattern-antipattern separation. (D) We now find that a correlation of 0.37 between the hamming distance from long motif and radial distance from the origin.

limit, by comparing a VAE model with four latent dimensions to a model with two latent dimensions, when trained on a dataset drawn from a distribution with four encoded patterns. We expand the dataset to be of size 20000, to ensure we still have approximately 5000 datapoints near each pattern. When we train a VAE with two latent dimensions, we observe that the VAE is capable of clustering and separating the sequences (Fig. 3A). However, unlike in the well-compressed limit, sequences distance from a pattern does not appear to correlate well with its radial distance. In this sense, no clear region in the VAE that corresponds to a ground state of our energy function (Fig. 3B). This reduces the interpretability of our model. However, with proper compression, we recover the properties we had observed earlier, such as orthoganality of pattern-anti-pattern dimensions(Fig. 3C). Notably, however, the latent space is not in a natural coordinate system. By linearly transforming the latent space into a coordinate system that corresponds to our encoded patterns, we further see that latent space is well-organized (Fig. 3D).

C. Semi-supervision in VAE models

We now progress to a model that includes two competing motifs - a short motif and a long motif. We achieve this by defining $\xi^{(i)}_{\text{long}}$ in the same way as before, but including a $\xi^{(i)}_{\text{short}}$ specific to each long motif. By defining the short motif in this way, we can tune the relevance of the short motif through its length relative to the long motif and its satisfiability, by controlling how much the long and short motif overlap. We will select that long motifs and short motifs are in tension, mirroring tradeoffs between phylogenetic history and performance. We will tune how unique a short motif is to any given long motif and present how results differ in each case.

D. Overlapping Short Motifs

Here, we impose that short motifs are shared across long motifs. We aim to develop a VAE that can infer the short motif overlap parameter of each sequence.

1. *Non-overlapping Short Motifs*

Here, we impose that the short motifs are fully unique to each long motif. In our model, we aim to model both long and short motif overlaps.

We draw 10000 datapoints from a model with two long and two short motifs and train an unsupervised model with two latent dimensions. We train a VAE, as before, on this dataset, and observe that the points are embedded in a similar manner to the way they were

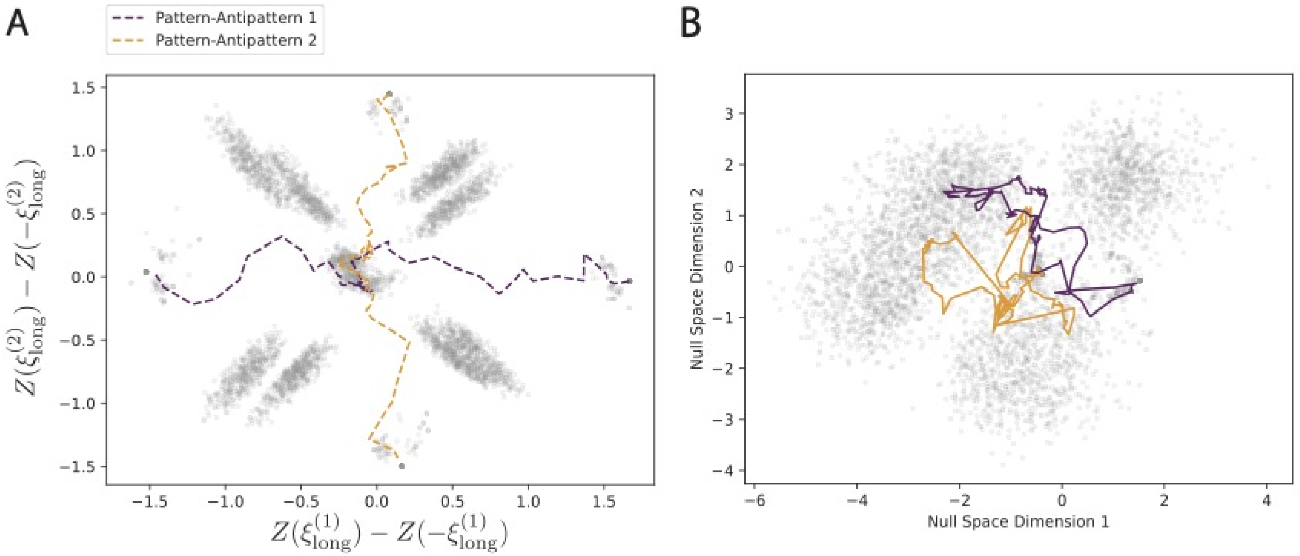

FIG. 2. **Results of a transformation of a four latent dimension VAE trained on a dataset with encoded patterns.** (A) By defining two of the latent axes to be along the pattern-antipattern vectors, we recover the structure of the latent space observed when a VAE is trained with just two latent dimensions. (B) The remaining two dimensions appear to be uninformative of any structure in the data.

in models with two long motifs and two latent dimensions, suggesting that most of the encoding corresponds to overlap with the long motif (Fig. 4A). However, there is a significantly larger spread within the region of latent space corresponding to a long motif. This corresponds to heightened diversity for sequences within the vicinity of a long motif. In order to assess if the VAE has effectively learned the short motif component of the sequence, we generate 10000 sequences with varying overlap to the short motifs for each pattern and embed them into the latent space (Fig. 4B). We observe that there does not appear to be a structure in the embedding of the sequences as a function of the short motif overlap. This suggests the VAE has not resolved the rules governing the short motif overlaps. However, a core engineering aim may, for example, be to optimize the short motif overlap for a given sequence. In order to achieve this, we attempt to regress the sequences, either one hot encoded, or latent space embedded, against the short motif overlap. We observe that regressing directly on the one hot encoded sequences onto short motif overlap (Fig. 4C) yields a coefficient of determination of 0.66 on a held-out set of data. Regression on the latent space embedding, on the other hand, yields a coefficient of determination of 0.57 (Fig. 4D). However, neither regression is performant on out-of-sample data, where we explicitly control short motif overlap. Both models yield a negative coefficient of determination (Fig. 4E,F). This suggests that though the data is recapitulating the detail of the sample dataset, it has not accurately resolved the rules corresponding to the short motif overlap.

We now compare these results to performance when incorporating regression directly into the VAE architecture. This explicitly enforces that one VAE dimension corresponds to each sequence's overlap with the short motif. The embedding is shown in Fig. 5A. Latent 1 is the supervised dimension, corresponding to short motif overlap. This latent dimension represents the overlap in the data set well, having a coefficient of determination of 0.99 (Fig. 5B). However, the VAE continues to poorly represent the out of sample sequences, having a coefficient of determination of -0.67 (Fig. 5C). This is a worse performance on out of sample data than linear regression directly on the one-hot encoded sequences. Notably, however, the trace of phenotypic prediction differs depending on which long-motif-short-motif pair is being predicted. This suggests that the model may fail to resolve the fact that the rules that explains short motif overlap because it is specific to each long motif, while the VAE itself is learning global organizational rules. Consequently, in order to properly predict short motif overlaps, a VAE focused on each clade is necessary.

We begin by focusing on the set of sequences that have greater than an 80% overlap with one of the two patterns. This reduces the data to approximately 1500 sequences. Their location in the latent space of an unsupervised VAE is shown in dark red in Fig. 6A. We begin by training a semi-supervised VAE on this subset of the data. We observe that the performance of the regression continues to be poor on out of sample data, as seen in Fig. 6B. This is a result of many of the sites becoming constant, as we chose datapoints that cluster to the same position in latent space. Consequently, the VAE can achieve good reconstruction performance by focusing primarily on reconstructing those conserved positions. In order to enforce that the VAE learn how the sequences achieve

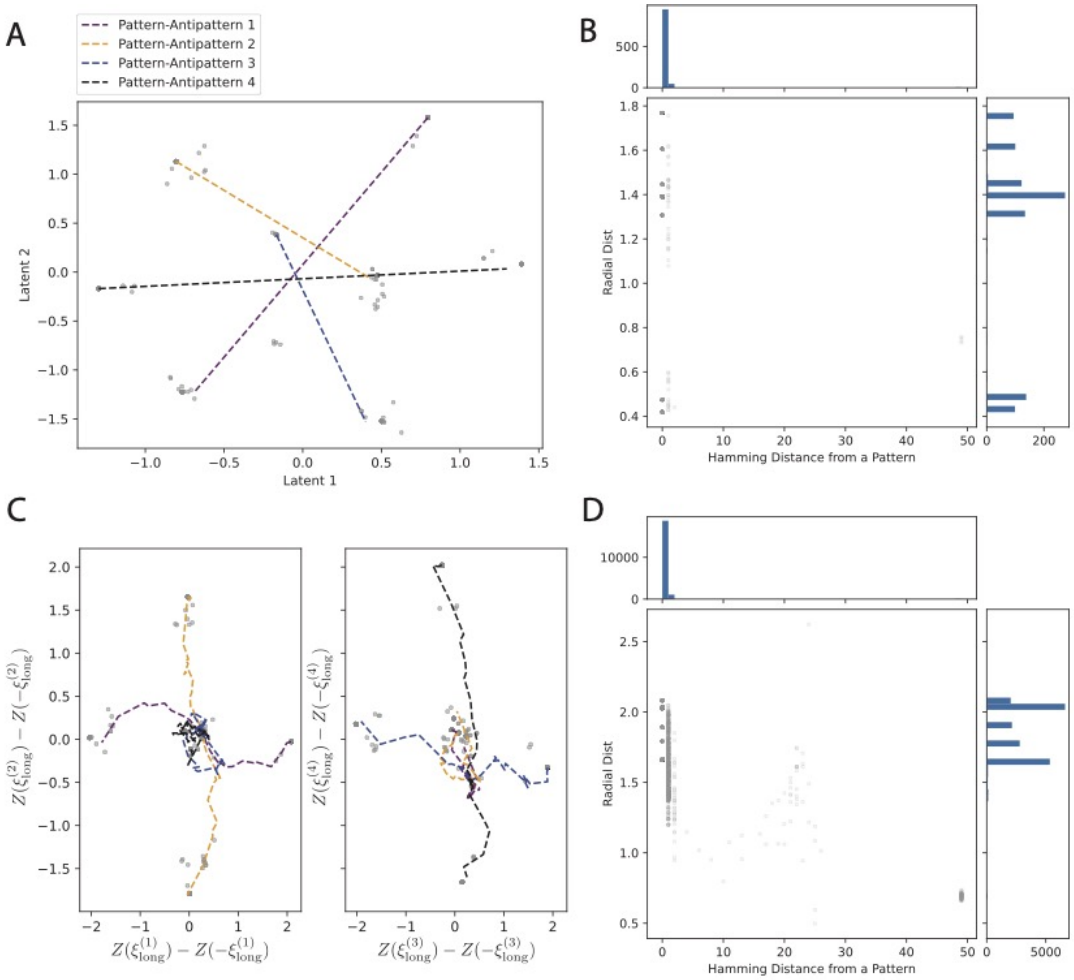

FIG. 3. **When overcompressing the the dataset, orthogonality of pattern-antipattern axes were lost, and hamming distance from pattern becomes decorrelated from radial distance.** (A) We embed the dataset with four encoded patterns into a VAE trained with two latent dimensions. We observe that latent space becomes patchy. Further, pattern-antipatterns cannot be orthogonalized, and interpolating between axes becomes impossible. (B) In this model, we find that the correlation between radial distance and hamming distance from a pattern is broken. (C) When encoding the dataset into a VAE trained with four latent dimensions, we find that the dimensions can be linearly transformed into a set of four orthogonal axes that each correspond to a pattern anti-pattern axis. This is not evident in the natural VAE latent axes, but appears after the linear transformation. (D) In this model, we recover the inverse correlation between hamming distance from a pattern and the radial distance.

differing performances on short-motif overlap, we truncate all the positions that feature no variation. We then find a significant improvement in our ability to regress short motif overlap on out of sample data, giving our model generalization capacity, as seen in Fig. 6. This demonstrates that the previously masked signal in the full dataset becomes apparent after having refined our data selection.

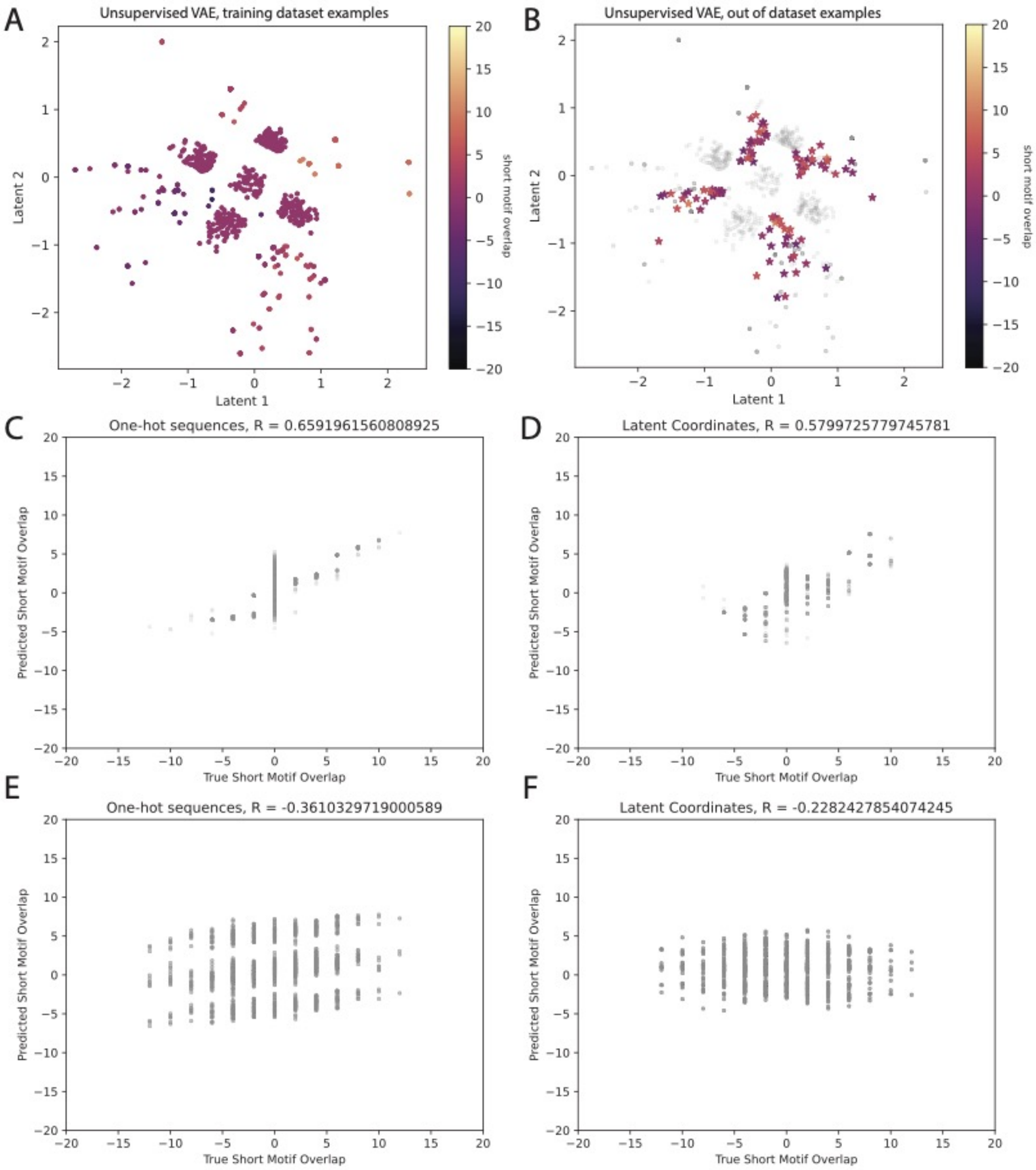

FIG. 4. **Attempting to infer the overlap of a sequence with a short motif using an unsupervised VAE fails when the short motif is specific to a particular long motif.** (A) We embed the dataset with two encoded patterns, each with a specific short motif to each pattern. We observe a greater spread in the regions of sequence space corresponding to a pattern. (B) We plot several out of sample sequences that span a range of overlaps with the relevant short motif. (C, E) We perform linear regression attempting to predict sequence overlap with the short motifs on a training dataset generated from a train-test split of the data on which the VAE was trained. Our linear regression performs reasonably well on the held out data. (D, F) We analyze the performance of our regression on out-of-sample data, and observe that the VAE's latent coordinates are less useful than direct regression on the one-hot encoded sequences.

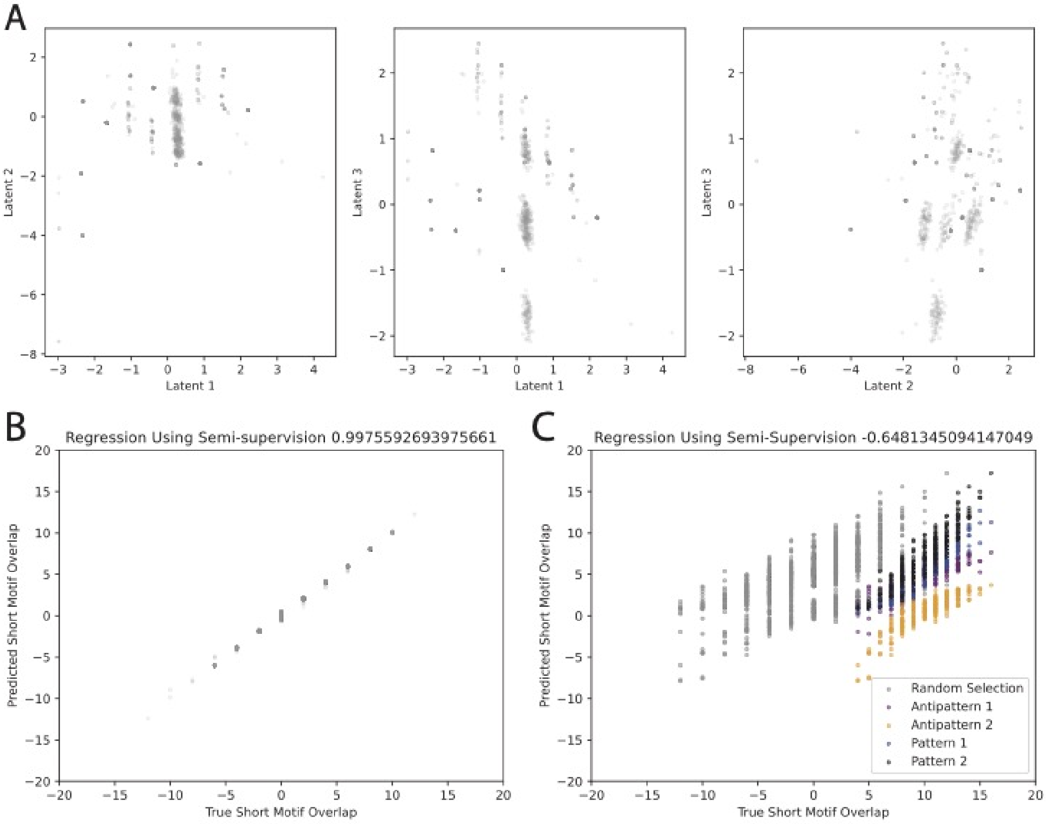

FIG. 5. **Explicitly regressing a single latent dimension against the short motif overlap value for each sequence yields improved performance in regression, but continues to fail on out of sample data.** (A) Here, Latent 1 corresponds to the dimension serving as a regression against the short motif overlap. (B) We embed the dataset with two encoded patterns, each with a specific short motif to each pattern. We observe a greater spread in the regions of sequence space corresponding to a pattern. (B) We plot several out of sampl e sequences that span a range of overlaps with the relevant short motif. (C, E) We perform linear regression attempting to predict sequence overlap with the short motifs on a training dataset generated from a train-test split of the data on which the VAE was trained. Our linear regression performs reasonably well on the held out data. (D, F) We analyze the performance of our regression on out-of-sample data, and observe that the VAE's latent coordinates are less useful than direct regression on the one-hot encoded sequences.

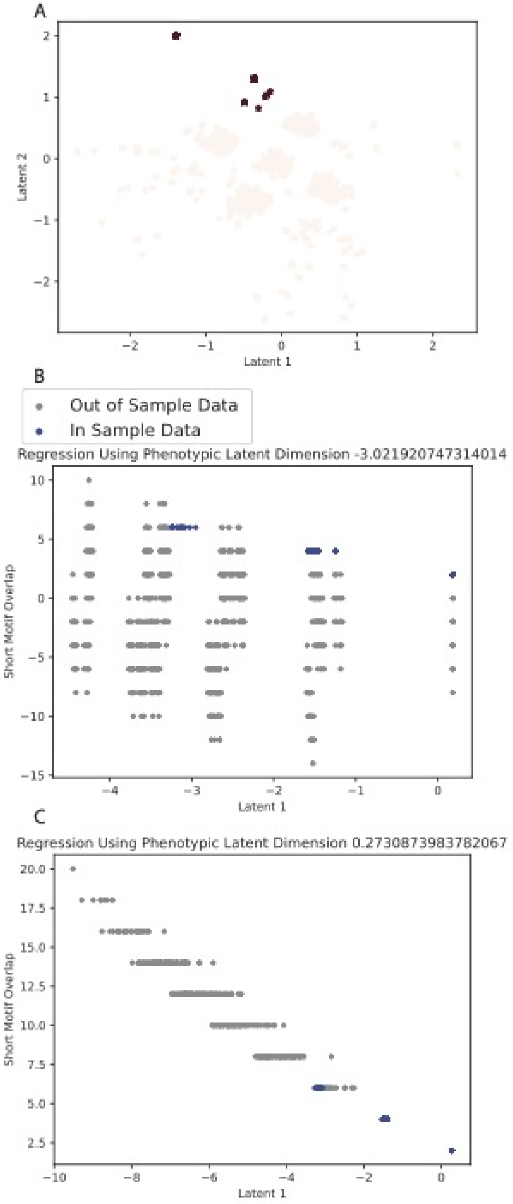

FIG. 6. **Training a second VAE on a subset of the data improves out of sample generalization if conserved residues are truncated.** (A) We select sequences in a small part of latent space. These sequences correspond to one of the long motifs, with diversity in the short motif overlap. (B) When training a semi-supervised VAE on the full sequence length, we observe poor performance on out-of-sample data. (C) When training a semi-supervised VAE on sequences with positions corresponding to complete conservation truncated, we are capable of generalizing to out of sample data well.