MotifAE Reveals Functional Motifs from Protein Language Model: Unsupervised Discovery and Interpretability Analysis

Chao Hou[1,#], Di Liu[2], Yufeng Shen[1,2,3,#]

1 Department of Systems Biology, Columbia University Irving Medical Center, New York, NY 10032

2 Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY 10032

3 JP Sulzberger Columbia Genome Center, Columbia University, New York, NY 10032

Corresponding author: ch3849@cumc.columbia.edu (C. H.), ys2411@cumc.columbia.edu (Y. S.)

Abstract

Protein motifs are conserved elements that mediate processes such as folding, binding, catalysis, and post-translational modifications. While motif identification is critical for protein study, experimental methods are labor-intensive, only a few hundred motifs are cataloged in databases like ELM, and existing supervised models are typically limited to predicting motifs with a specific function. Here, we present MotifAE, an unsupervised framework for discovering functional motifs from the protein language model ESM2, which captures evolutionary-scale sequence regularities. MotifAE is based on the sparse autoencoder (SAE), an encoder-decoder architecture that projects ESM2 embeddings into a sparse latent space, with an additional local similarity loss that encourages coherent latent feature activations. When benchmarked against known ELM motifs, MotifAE achieves a median AUROC of 0.88, outperforming the standard SAE (0.80). We also calculated Position-specific scoring matrices (PSSMs) for MotifAE features and found that features with similar decoder weights share similar PSSMs. Furthermore, by aligning MotifAE features with experimental data through gated feature selection, we identified features associated with specific properties such as folding stability. Steering these features enabled designing proteins with enhanced stability, as evaluated *in silico*. Overall, MotifAE provides a general framework for systematic motif discovery and interpretation, with the potential to advance protein function analysis, mutation effect interpretation, and rational protein engineering.

Introduction

Proteins mediate essential biological processes, such as catalysis, immune defense, signal transduction, and molecular transport. Their functional regions, such as catalytic centers, binding interfaces, and post-translational modification sites, exhibit conserved sequence patterns, commonly referred to as motifs. Systematic discovery of functional motifs is crucial for advancing our understanding of protein function and has broad applications in gene annotation, mutation effect interpretation, and protein engineering.

Experimental Identification of functional motifs requires first locating protein regions that share the same function, then aligning these regions to get the motif sequence pattern, typically represented as a regular expression or a position-specific scoring matrix (PSSM)[1]. This process is labor-intensive and time-consuming. As of 2025, only 353 motifs have been curated in the Eukaryotic Linear Motif (ELM) database[2]. High-throughput screens have been developed to map functional regions[3–6], but these screens remain limited in scale and are tailored to specific and readily measurable functions, such as protein abundance or binding to a specific partner, making them unsuitable for systematically identifying the full spectrum of functional motifs. To complement experimental efforts, machine learning methods have also been developed[7–9], often trained on curated resources like ELM or datasets from high-throughput screens. However, these supervised models are typically tailored to specific functions and inherently biased toward their training datasets, limiting their generalizability.

In recent years, deep learning has driven remarkable advances in modeling protein sequence[10], structure[11], and function[12]. A major development is protein language models (pLMs), such as ESM2[10], which are self-supervised models trained on large-scale protein databases using masked or next-token prediction. Through this pre-training, pLMs implicitly learn conserved sequence patterns at the evolutionary scale. For example, Zhang et al.[13] demonstrated that ESM2 predicts protein structure by storing pairwise contact motifs, Vig et al.[14] found that pLMs' attentions capture target binding sites, Zhang et al.[15] showed that mutational constraints predicted by ESM2 are predictive of conserved sequence motifs. Moreover, numerous studies have leveraged pLM embeddings to predict functional regions such as signal peptides, transit peptides, and post-translational modification sites[8]. Collectively, these works demonstrate that pLMs encode rich information about functional motifs within their parameters. However, because pLMs operate as black boxes, the motif information they encode is not directly accessible.

Various attempts have been made to extract interpretable features from language models. Among them, sparse autoencoders (SAEs) have shown strong potential for interpretability[16]. The core assumption of SAEs is that the model embedding can be expressed as a linear combination of different features[17], with each feature ideally corresponding to an interpretable factor. To achieve this, SAE uses an encoder to project embeddings into a higher-dimensional, sparse latent space, and uses a decoder to reconstruct the embeddings from this space. Hereafter, we refer to each neuron in the SAE latent space as a feature, and neurons with positive values are activated. During SAE training, a reconstruction loss was used to ensure that the embeddings are accurately reconstructed, and a sparsity constraint was used to encourage sparse feature activations. Typically, only tens of features are active per residue, far fewer than the original embedding dimension, which often exceeds a thousand. SAEs have been successfully applied in large language models (LLMs) to find interpretable semantic features[16,17]. Adapting the training and interpretation strategies developed for LLM SAEs, similar approaches in biological sequence models have identified known functional elements, including DNA features such as exons, introns, and transcription factor binding sites[18], as well as protein features such as binding motifs and structural domains[19–21].

Here, we developed MotifAE, introducing methodological advances in both SAE model training and interpretation to make them better suited for biological sequences and compatible with experimental data. We incorporated an additional local similarity loss during training, encouraging MotifAE's latent features to capture the sequential nature of motifs. Compared to standard SAEs, MotifAE markedly improves the identification of functional motifs across diverse benchmarks. Building on this capability, we further align MotifAE features with experimental protein fitness data using a gated feature selection approach, which enables the identification of associated features and enhances performance in fitness prediction. Moreover, steering these selected features allows protein design with desired properties, as evaluated *in silico*. Overall, MotifAE provides a systematic framework for motif discovery from pLMs and annotation using experimental data, facilitating deeper insights into protein function, and offering potential applications in gene annotation, mutation effect interpretation, and protein engineering.

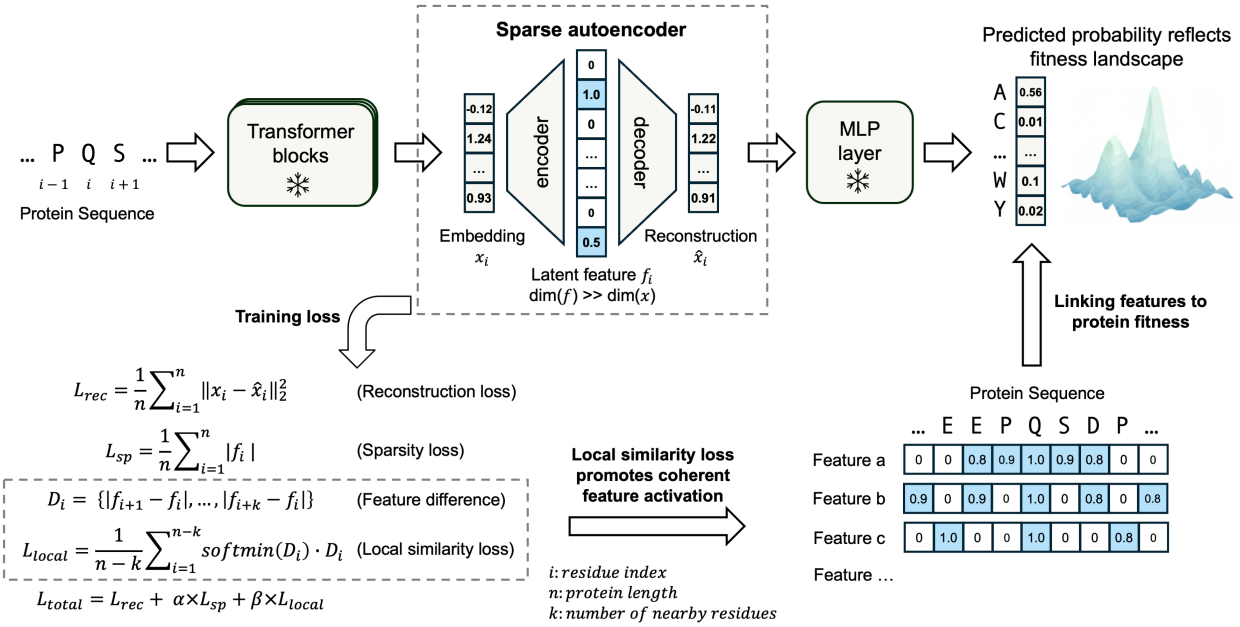

Figure 1. The MotifAE framework.

MotifAE uses the same architecture as the sparse autoencoder (SAE) but is trained with an additional local similarity loss. MotifAE projects ESM2 embeddings (dimension 1,280) into a high-dimensional latent space (dimension 40,960) using an encoder, and then reconstructs the embeddings through a decoder. The reconstructed embeddings can be mapped to amino acid probabilities via the fixed ESM2 MLP layer, enabling prediction of protein fitness landscape. MotifAE is trained with three objectives: a reconstruction loss, a sparsity loss, and a local similarity loss, the latter encourages locally coherent latent feature activations.

Results

MotifAE is a sparse autoencoder with coherent latent feature activation.

We developed MotifAE, an adaptation of the sparse autoencoder (SAE) for discovering functional motifs from pLMs. A standard SAE is trained with a reconstruction loss and a sparsity constraint (**Figure 1**), ensuring that the original embeddings are accurately reconstructed while enforcing sparse activation of latent features. To enhance biological relevance, we introduced a local similarity loss (see Methods) that encourages MotifAE features to activate coherently along the sequence. This design is motivated by the fact that protein motifs often involve local contiguous residues[2], and basic structural elements also exhibit local patterns, such as alternating contacts in β-strands and periodic interactions every three or four residues in α-helices. The local similarity loss requires that at least one neighboring residue within a defined window exhibits activation features similar to the current residue (**Figure 1**; see Methods), thereby promoting local coherence in feature activation. A window size of three residues was used in this study; since the local similarity loss is computed over the entire sequence, it can capture relationships extending beyond the three-residue window.

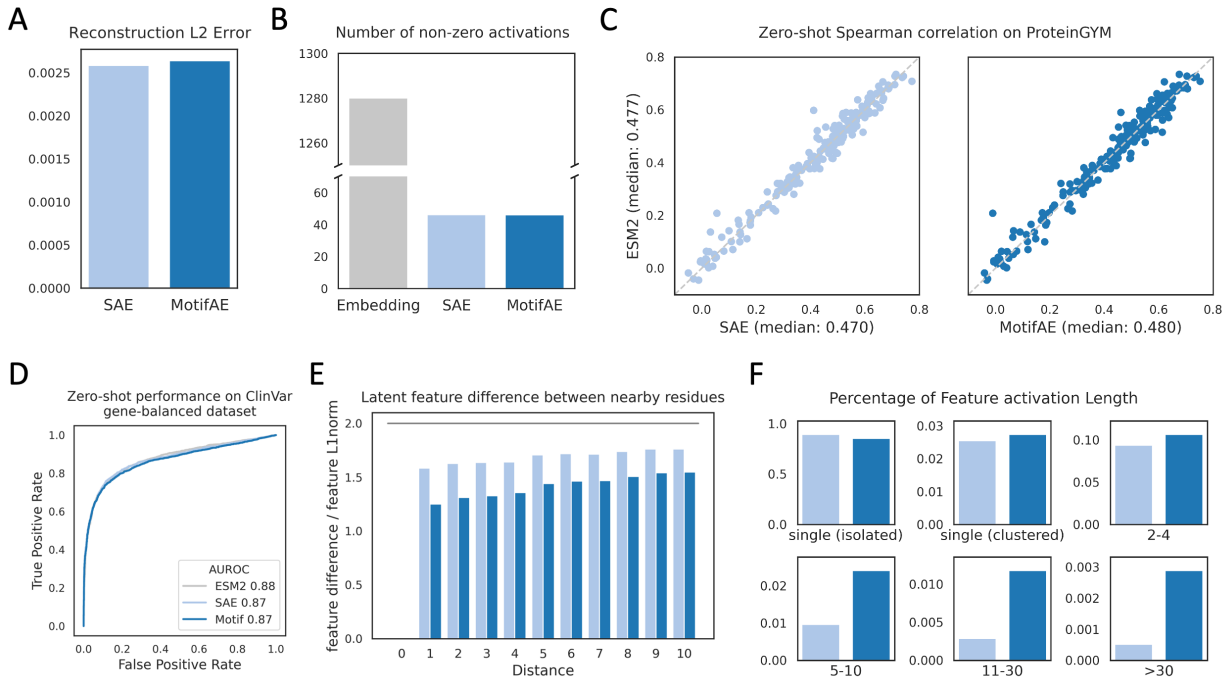

Figure 2. Comparison of MotifAE and SAE in terms of reconstruction error, sparsity, fitness prediction, and activation length distribution.

(**A**) Embedding reconstruction error on the evaluation set (see Methods). (**B**) Mean number of non-zero neurons per residue. The ESM2 embedding has 1280 non-zero neurons, while both SAE and MotifAE have ~46 non-zero (positive) latent features per residue. (**C**) Zero-shot performance (spearman correlation) on ProteinGYM DMS datasets, only single substitution mutations were evaluated. Each dot represents a DMS dataset. Wild-type marginal log-likelihood ratio (LLR) was used to estimate mutation effects. (**D**) Zero-shot performance on classifying pathogenic and benign missense mutations in the ClinVar gene-balanced dataset (see Methods). (**E**) Latent feature difference between nearby residues with different distances, the value of y-axis is 0 for identical activations and ~2 for random sparse activations. (**F**) Length distribution of activations from all features. For each feature, A single (clustered) activation is defined as an activated single residue with other activated residues within two amino acids along the sequence. In all plots, light blue denotes SAE, and dark blue denotes MotifAE.

We trained MotifAE on the final-layer embeddings from ESM2-650M, which has shown strong performance across diverse downstream tasks. ESM2-650M captures the protein fitness landscape by being trained on evolutionary-scale sequences and is widely used to predict mutation effects[22–24]. We used the last-layer embeddings because they are directly used to predict amino acid probability which reflects the protein fitness landscape[22,23], allowing us to investigate the relationship between latent features and fitness. To ensure that MotifAE captures representative protein features, training was performed on 2.3 million representative proteins obtained from structure-based clustering of the AlphaFold2 structure database[25]. The dimension of the latent space and weights for different losses were determined by jointly considering reconstruction error, latent sparsity, and the protein fitness prediction performance (**Figure S1**; see Methods). Based on these criteria, we selected a hidden dimension of 40,960, corresponding to a 32-fold expansion over the ESM2 embedding dimension. While only embeddings from unmasked sequences were used for model training, MotifAE can reconstruct embeddings of masked residues, and the reconstructed embeddings preserve essential information for predicting the masked amino acids when processed with the fixed ESM2 MLP layer (**Figure S1C**). We note that the reconstruction of ESM2 embeddings is highly fragile. Although the reconstruction

errors remain similar across different latent space dimensions (**Figure S1B**), the quality of the reconstructed embeddings varies substantially, as reflected in their performance on masked residue prediction (**Figure S1C**) and protein fitness prediction (**Figures S1D–E**).

We trained a standard SAE without the local similarity loss for comparison (see Methods). Both MotifAE and SAE achieve low reconstruction errors, with approximately 46 latent features activated per residue on average in both models (**Figure 2A–B**). To assess the quality of the reconstructed embeddings, we evaluated their performance in protein fitness prediction. We projected the reconstructed embeddings through the fixed ESM2 MLP layer to predict amino acid probability (**Figure 1**), calculated the wild-type marginal log-likelihood ratio (LLR; see Methods)[22,24], and used the LLR to estimate mutation effects. Across deep mutational scanning (DMS) experiments in ProteinGYM[23] and pathogenic/benign mutations in ClinVar[26] (see Methods), reconstructions from both MotifAE and SAE perform comparably to the original ESM2 embeddings (**Figure 2C-D**). These results demonstrate that both MotifAE and SAE preserve critical functional signals in ESM2 while requiring far fewer neurons to be activated in the latent space compared to the dimension of embedding.

The local similarity loss of MotifAE was introduced to encourage coherent latent feature activation. To evaluate this, we first analyzed feature activation similarity between nearby residues (measured using the absolute activation difference divided by the L1 norm of the activations: 0 for identical activations and ~2 for random sparse activations). MotifAE features were more similar among neighboring residues compared to SAE features (**Figure 2E**). We further examined the length distribution of activations and found that MotifAE features more frequently form clustered and contiguous activations than those of SAE, with a notably higher frequency of activations longer than four amino acids (**Figure 2F**). Together, these results showed that MotifAE preserves reconstruction quality and latent space sparsity, while the local similarity loss promotes locally coherent feature activations, with the potential to enable the identification of localized motifs and basic structural elements.

MotifAE captures known functional motifs.

To investigate how MotifAE latent features capture functional motifs, we compared their activations with experimentally validated motifs from the ELM[2] database, which covers six categories: ligand-binding, docking, post-translational modifications, targeting signals, degradation, and proteolytic cleavage sites. Each ELM motif has several experimentally verified functional regions, 270 motifs with at least 20 residues within verified regions were analyzed. We compared all MotifAE features against all ELM motifs, evaluating whether the feature activation scores could distinguish residues located inside versus outside the experimentally verified motif regions.

Considering the best-match feature for each motif, MotifAE demonstrates strong performance, achieving a median AUROC of 0.88 across the 270 motifs, significantly outperforming SAE (median AUROC 0.80) (**Figure 3A**). The best-match MotifAE features achieve AUROCs above 0.8 for 193 motifs and above 0.9 for 114 motifs out of 270 evaluated. One representative feature–motif pair is f13268 with the ELM motif MOD_TYR_CSK: the C-terminal phosphorylation motif in Src-family proteins targeted by the non-receptor tyrosine kinase Csk family[27]. Across 12 proteins with known MOD_TYR_CSK motif in ELM, f13268 predicts residues in motif regions with an AUROC of 0.97. Notably, the highest activation of f13268 was observed for the phosphorylated tyrosine residue (**Figure 3B**), indicating that this feature captures the biological characteristics of the motif. Additionally, MotifAE performs consistently across structural contexts, with median AUROCs of 0.88 in ordered regions and 0.84 in disordered regions, compared to 0.79 and 0.76, respectively, for SAE (**Figure 3C**, see Methods).

Furthermore, we evaluated the specificity of the relationship between ELM motifs and MotifAE features. We defined a matched motif–feature pair as one with AUROC > 0.9, resulting in 322 matched pairs involving 114 ELM motifs and 146 MotifAE features (**Figure S2**). Among these, 47 ELM motifs are matched to a single MotifAE feature, whereas 67 motifs are matched to multiple features (**Figure 3D**). On the feature side, 91 MotifAE features match only one ELM motif (**Figure 3E**), whereas 55 features match multiple motifs. Among the MotifAE features that match multiple ELM motifs, f27416 is one of the most prominent: it matches 17 ELM motifs, all of which locate at the C-terminus of protein. Across all ELM motifs, f27416 achieves a median AUROC of 0.97 for 25 motifs exclusively located at the C-terminus but show no predictive power for other motifs (**Figure 3F**). Importantly, f27416 is not universally activated at the C-terminus of all proteins: among 320 proteins with known C-terminal motifs in ELM, 96% show f27416 activation, compared with only 39% of randomly sampled proteins (**Figure 3G**). Overall, these results indicate that MotifAE latent features align well with known functional motifs and perform substantially better than the standard SAE. MotifAE features also exhibit granularity: some capture motifs associated with very specific functions, while others represent more general or broadly shared motifs.

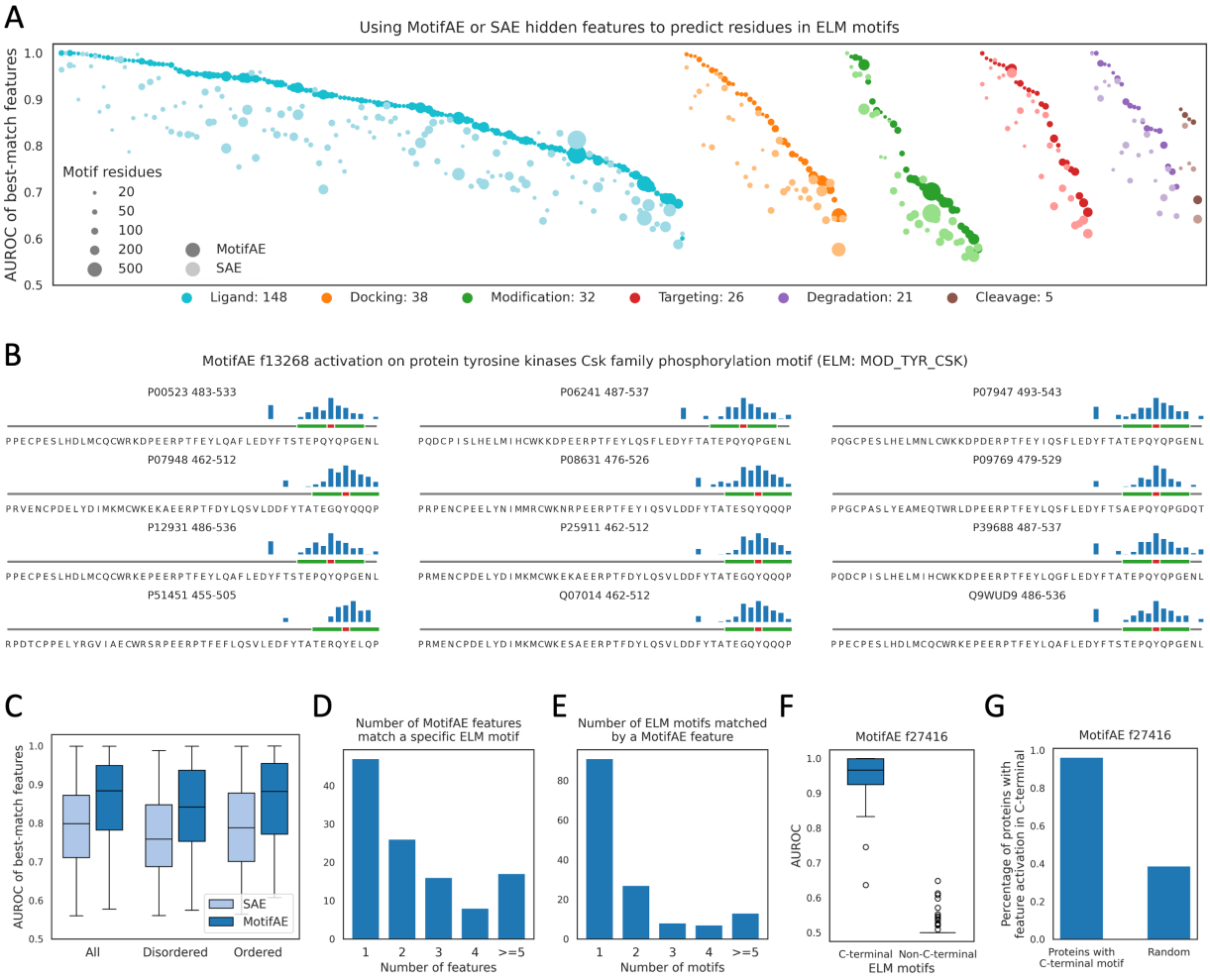

Figure 3. MotifAE captures known functional motifs.

(**A**) AUROC of the best-match feature for each ELM motif. Each column represents the performances of two models on an ELM motif; dots with dark colors indicate MotifAE, while dots with light colors indicate SAE. Dot size indicates the number of residues in motif regions, and colors represent different ELM categories. The six categories and the number of motifs in each are listed below the plot. (**B**) Feature activation along protein sequences. The height of the blue bars indicates the normalized feature activation values, the green region denotes the motif, and the red region marks the phosphorylated tyrosine residue. Only the C-terminal 50 residues of each protein are shown for clarity. (**C**) Distribution of AUROCs of best-match features for ELM motifs. Disordered and ordered regions were classified using IUPred3[28]. The box extends from the first quartile to the third quartile, with a line marking the median. Whiskers span to the most extreme points within 1.5× the interquartile range. (**D**) Number of features match each ELM motif, (**E**) Number of motifs matched by each MotifAE feature, a motif–feature match is defined by AUROC > 0.9. (**F**) AUROC of MotifAE feature f27416 on C-terminal versus non-C-terminal ELM motifs. Non-C-terminal ELM motifs are defined as those with no verified regions within 10 residues of the C-terminus. (**G**) Percentage of proteins with activation of f27416 at the C-terminus. Proteins with known C-terminal motifs were obtained from the ELM dataset, random proteins were sampled from the evaluation set.

MotifAE captures homodimerization interfaces.

Next, we evaluated MotifAE on three-dimensional functional sites, focusing on homodimer interfaces. A previous study has shown that ESM2 captures evolutionary signals of homo-oligomer symmetry directly from single sequences[29]. To construct the evaluation set, we retrieved homodimer structures from the PINDER[30] database and removed redundancy by clustering sequences at 30% identity (see Methods), yielding 1,565 non-redundant homodimers with at least five contact residues (contacts are defined by a heavy-atom distance cutoff of 5 Å). We used the activation value of each MotifAE feature to distinguish contact from non-contact residues within each protein. The best-match MotifAE feature for each homodimer achieves a median AUROC of 0.73 in identifying contact residues, outperforming SAE (0.69; **Figure 4A**). When stratifying homodimers by the number of contact residues (5–25, 25–50, >50), MotifAE performs particularly well on smaller homodimer interfaces, with a median AUROC of 0.82 for homodimers with 5–25

contact residues (SAE 0.76). Across homodimers with varying interface sizes, MotifAE consistently outperforms SAE. Representative homodimers with different number of contact residues are shown in **Figure 4B**, along with the activation values of their best-match MotifAE features.

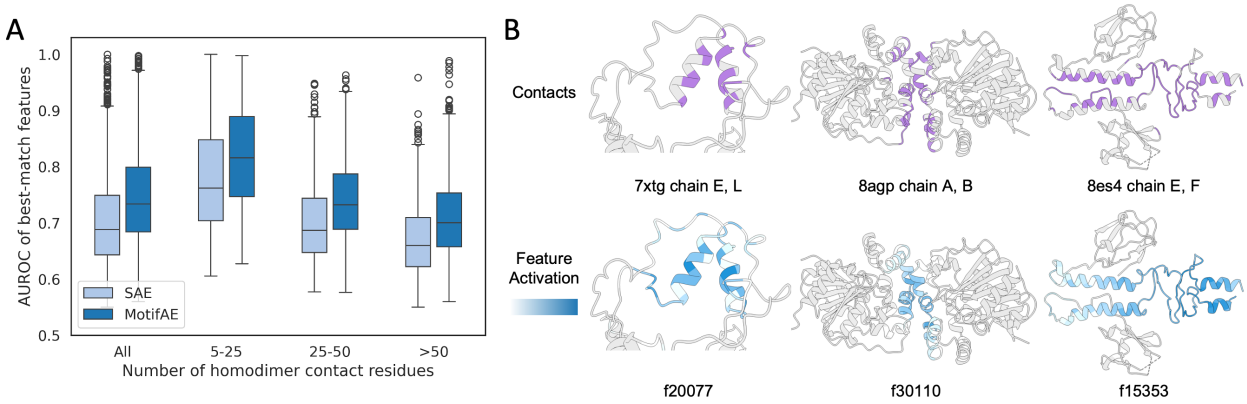

Figure 4. MotifAE captures homodimer interfaces.

(**A**) AUROC of the best-match MotifAE features for predicting each homodimer interfaces in the PINDER dataset. Homodimers are shown either collectively or grouped by the number of contact residues (5–25, 25–50, >50). (**B**) Representative homodimer structures. In the upper panel, contact residues are highlighted in purple. In the lower panel, the saturation of blue indicates the relative activation values of the best-match MotifAE feature.

The universe of sequence patterns of MotifAE features.

Next, we set out to characterize the sequence pattern of motifs captured by MotifAE features. Since functional motifs often exhibit distinct sequence preferences, we first analyzed the amino acid composition of activated peptides for each feature. Among the 40,960 features, 6,421 show coherent activations in the representative protein dataset[25]. By recording activated peptides for each feature (see Methods), we found that MotifAE features display significantly biased amino acid preferences compared with random peptides (**Figure 5A**). Notably, approximately 400 features are almost exclusively activated by a single amino acid (**Figure 5B**; with one amino acid type accounting for >99% of all activations). To describe the sequence patterns, we constructed position-specific scoring matrices (PSSMs) for each MotifAE feature: activated peptides of each feature were aligned using GibbsCluster[31], and the aligned cores were used to calculate the PSSMs (see Methods, PSSMs were set to a length of five residues for all features). The resulting PSSMs exhibited median KL divergence values (summed across the five residues) relative to the background amino acid distribution of 10.7 (**Figure 5C**). Among the PSSMs with well-defined sequence patterns, as indicated by higher KL divergence, some correspond to homopolymeric repeats (e.g., f13432 and f28775), while some exhibit multiple amino acid types (e.g., f2302, f36881, and f38747) (**Figure 5D**).

In the MotifAE decoder, each feature has a weight vector that is multiplied by its activation value to reconstruct the embedding. Thus, the decoder weights reflect the properties of features in the latent embedding space. Notably, features that have similar decoder weight vectors tend to have similar PSSMs (**Figure 5E**). For example, both f6643 and f20738 have an “H” residue centered between four “R” residues; both f2459 and f22016 capture an “LPLP” pattern; both f2182 and f23778 exhibit paired “Q” residues separated by a gap (**Figure 5D**). Together, by systematically characterizing the sequence patterns of MotifAE features, we found that they exhibit specific sequence preferences, we also found that these patterns are reflected in the decoder weight space.

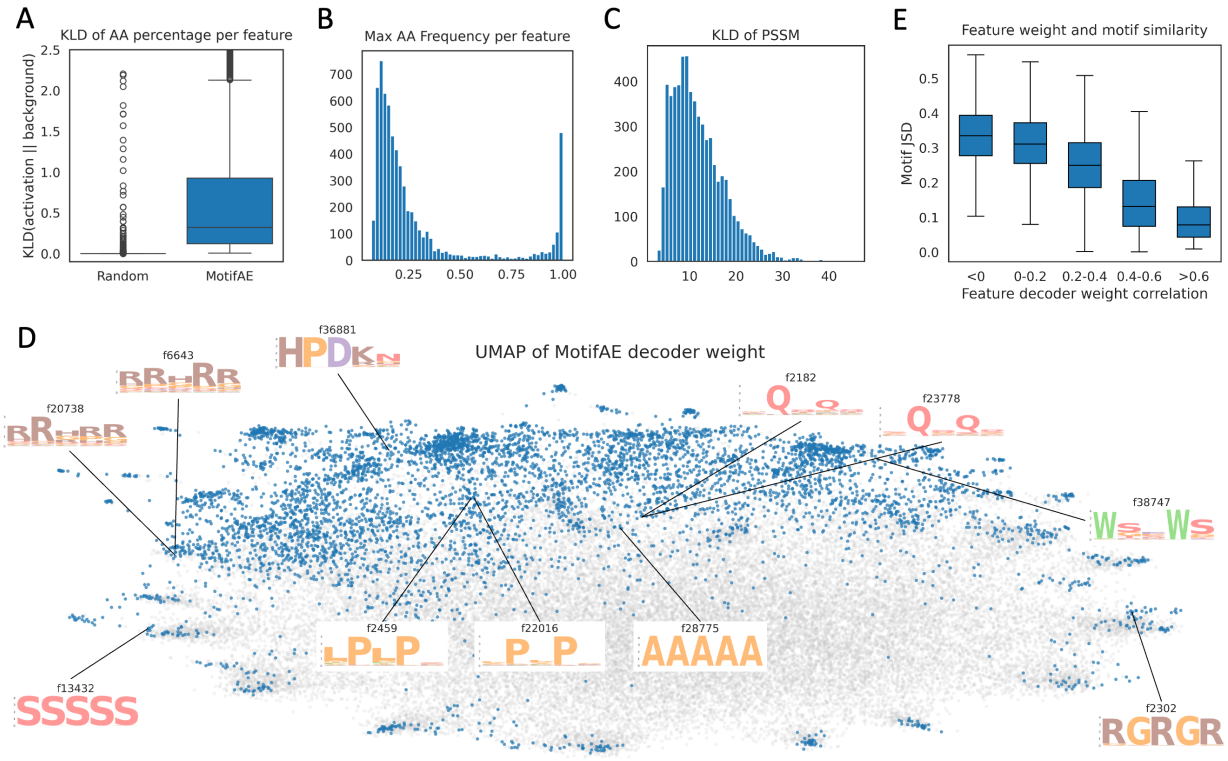

Figure 5. The universe of sequence patterns of MotifAE features.

(**A**) KL divergence between the amino acid composition of activated peptides of each feature and the background (amino acid composition in the 2.3M representative proteins). Random peptides matched in number and length distribution were used as a control. (**B**) Frequency of the most abundant amino acid in the activated peptides of each feature. (**C**) KL divergence of PSSMs calculated for each feature compared to the background amino acid frequency. KL divergence values were summed across the five aligned core positions. (**D**) UMAP visualization of MotifAE decoder weights. Each point represents a feature; gray points correspond to features without coherent activations, while blue points correspond to features with coherent activations. PSSM sequence logos are shown for some features. (**E**) Relationship between feature similarity in decoder weight space and similarity of their PSSMs. Decoder weight similarity was measured using Pearson correlation, while PSSM similarity was measured using Jensen–Shannon divergence (JSD), where smaller JSD values indicate higher similarity.

Aligning MotifAE with experimental data identifies features associated with the specific property.

Interpreting latent features remains a major challenge. While comparing MotifAE feature activations with known functional sites reveals the biological signal captured by latent features, this approach is limited by the scarcity of known functional sites. To overcome this, we developed a supervised approach to annotate MotifAE features using experimental data, which do not necessarily provide explicit site-level annotations. This approach involves two steps. First, a model is trained to predict experimental measurements from embeddings, many existing fine-tuning models can be directly reused; in this step, MotifAE reconstructions serve as a drop-in replacement for ESM2 embeddings. Second, a gating layer is introduced to selectively amplify or attenuate the activation magnitudes of MotifAE features. The gating layer is further trained on the same dataset, thereby modulating both the reconstructed embeddings and downstream predictions. After training, the gate weights reflect the association between features and the measured property, and the activated residues of associated features may underlie the property. We refer to this framework as MotifAE-G.

Here, we applied MotifAE-G to DMS experiments (**Figure 6A**). Because the ESM2 MLP layer's predicted amino acid probabilities correlate well with mutation effects (**Figure 2C-D**), we don't need to train an additional predictor (step one described above). A binary gate was used, meaning that each feature was either retained or completely suppressed. A differentiable Spearman correlation loss was used to update the gate parameters (see Methods). We used the mega-scale protein stability DMS dataset[32], which contains 379,495 single substitution mutations across 412 proteins, including both natural and designed proteins. On this dataset, MotifAE without gating performs comparably to ESM2 (**Figure S3A**). 412 proteins were grouped into 189 clusters based on 30% sequence identity. Among these, 285 proteins from 133 clusters were used for training and validation, while 127 proteins from the remaining 56 clusters were reserved for testing. Using the MotifAE-G framework, we identified 1,404 stability-

associated features from the training set. With these features, MotifAE-G achieves a median Spearman correlation of 0.61 on the training set, significantly outperforming ESM2 (0.43), with 97.5% of training proteins showing performance improvement (**Figure 6B**). On the test set, MotifAE-G achieves a median Spearman correlation of 0.61, also significantly higher than ESM2 (0.44), with 88% of test proteins showing improvement (**Figure 6C**). For both training and test sets, the improvement was more pronounced for proteins that were initially predicted poorly by ESM2 (**Figure 6B-C**).

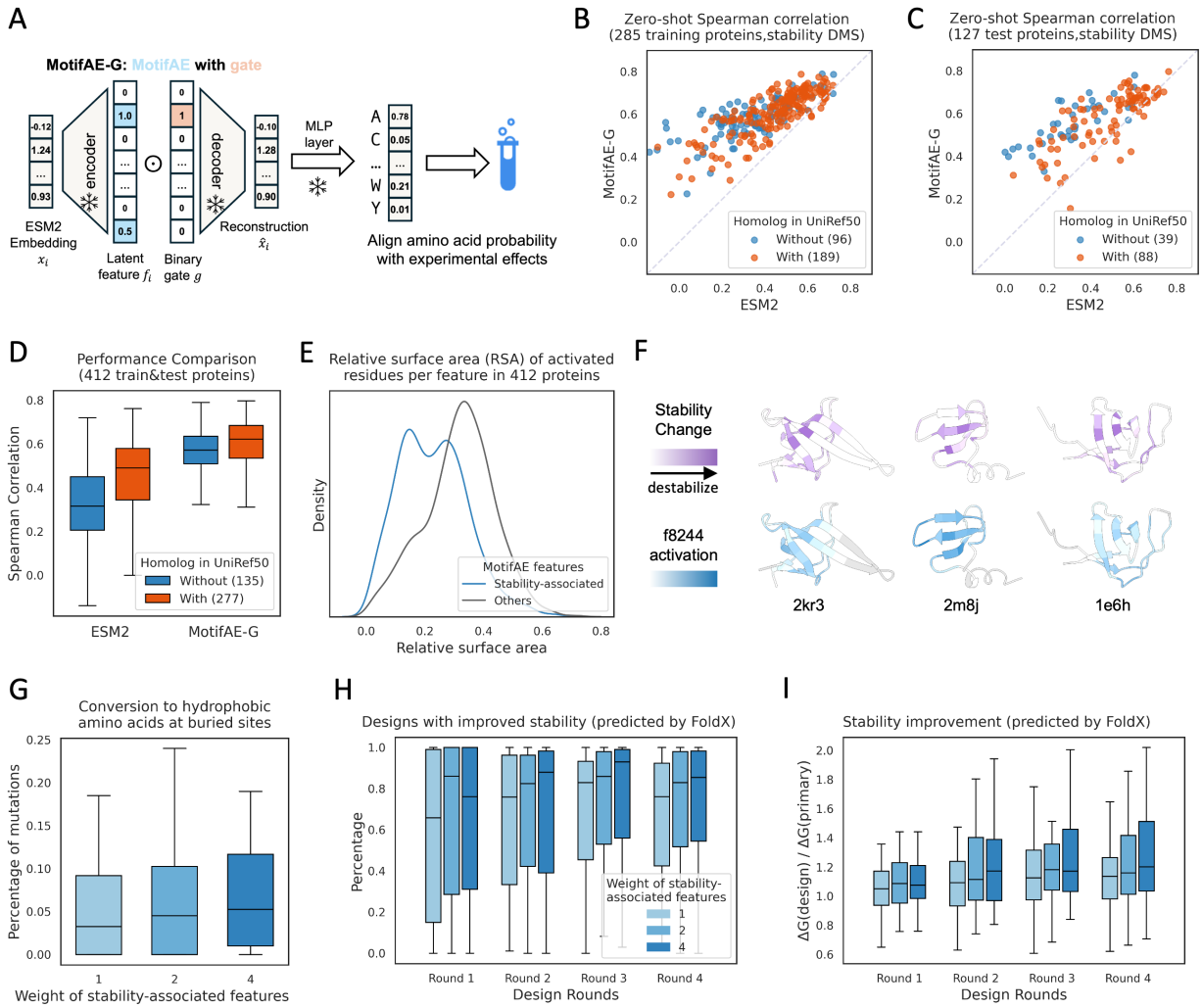

Figure 6. Aligning MotifAE with experimental data for feature annotation and protein design.

(**A**) Schematic of applying MotifAE-G framework to DMS data. A learnable binary gate was used to select latent features to align predicted amino acid probability with experimental mutation effect data; during training, only the gate parameters were updated. (**B-C**) Comparison of MotifAE-G and ESM2 performance on stability mutation effects across training and test proteins; each point represents a protein; color represents whether the protein has homolog in UniRef50. (**D**) Comparison of MotifAE-G and ESM2 performance on protein with and without homolog in UniRef50. (**E**) Mean relative solvent accessibility (RSA) of activated residues for stability-associated versus other MotifAE features, only features with at least 50 activated residues in 412 proteins were analyzed. (**F**) AlphaFold2-predicted structures of representative proteins. Upper panels show experimentally measured stability changes, averaged over all mutations at each residue (purple indicates stronger destabilizing effects). In the lower panel, the saturation of blue indicates the relative activation values of the MotifAE feature. (**G–I**) Designing proteins with improved stability using iterative redesign sampling. Gate weights of 1, 2, and 4 were applied to stability-associated features (shown in progressively darker blue), while other features were fixed at a gate weight of 1. (**G**) Buried residues were defined as sites with RSA < 0.2. All mutations from four rounds of redesign were included in the analysis. (**H**) Percentage of designs showing improved stability compared to their primary (unmutated) sequences. (**I**) Relative FoldX ΔG compared to the primary sequence, 1 indicates no improvement. The median ΔG across all designs in each round for each protein is shown. **H-I** show results for 32 test proteins, only designs with ESMFold pLDDT > 0.8 and FoldX ΔG > 2 kcal/mol were included.

We previously observed that ESM2 models perform better on natural proteins but worse on designed proteins[33]. Among the 412 proteins in the dataset, 135 have no homologs in UniRef50[34] (the training set of ESM2), ESM2 performs much worse on these proteins (median Spearman correlation 0.32) compared to proteins with homologs (0.49; **Figure 6D**). MotifAE-G improves prediction performance for both, achieving median Spearman correlations of 0.57 for proteins without homologs and 0.62 for proteins with homologs (**Figure 6D**).

To further assess the biophysical relevance of the stability-associated features, we analyzed the relative solvent-accessible surface area of activated residues of each feature. We found that the stability-associated features tend to activate at more buried residues (**Figure 6E**), consistent with the fact that stability-disrupting mutations often occur at buried residues[24]. Furthermore, we compared the activation value of each feature with the mean mutation effect per residue within each protein. We found that stability-associated features tend to exhibit stronger correlations with mean mutation effects, indicating better alignment with experimentally observed stability changes (**Figure S3B**). One example feature is f8244, which shows strong correlations across multiple proteins, with representative cases visualized in **Figure 6F**: residues with high f8244 activation tend to harbor destabilizing mutations. Together, these results demonstrate that by aligning latent features with experimental data, MotifAE-G identifies meaningful features associated with specific functions or protein properties, which in turn enhance the prediction of mutational effects for both natural and designed proteins, highlighting its strong generalization.

Steering MotifAE-G enables designing protein with enhanced stability.

pLMs can be used for protein design by sampling from their predicted amino acid probabilities. However, directly generating sequences from pLMs does not necessarily yield proteins with the desired property. To address this, several approaches have been proposed to fine-tune pLMs to guide the design process toward specific functions or properties[35,36]. Here, we explored whether MotifAE-G can be used for property-specific protein design. Specifically, we steered the stability-associated features identified from the DMS dataset to design proteins with improved stability. Because ESM2 is not suitable for de novo sequence generation, we adopted an iterative redesign sampling strategy[37]. At each iteration, a single mutation was sampled according to the probabilities predicted by MotifAE-G, using higher gate weights to amplify the influence of stability-associated features and gate weights of one for other features (see Methods). This process was repeated over multiple rounds to progressively refine the sequence. The designed proteins were evaluated using FoldX[38], a physics-based force-field model that can estimate folding free energy (ΔG), applied to structures predicted by ESMFold[10] (see Method).

We applied the above strategy to the representative proteins from the 56 test protein clusters in the DMS dataset, among them, 32 proteins with ESMFold pLDDT > 0.8 and FoldX-predicted ΔG > 2 kcal/mol were selected for subsequent design. We applied weights of 1, 2, and 4 to the stability-associated features, with a weight of 1 as the baseline, equivalent to design using ESM2. For each weight, four rounds of iterative redesign were applied to each protein. We did not conduct more rounds because introducing a few new core interactions is typically sufficient to stabilize the protein, whereas excessive mutations may alter the native fold. The procedure was repeated 100 times per protein. By analyzing the introduced mutations, we found that increasing the weights of stability-associated features resulted in a higher proportion of mutations converting other amino acids into hydrophobic residues at buried sites (**Figure 6G**). This observation aligns with the fact that proteins are primarily stabilized by hydrophobic interactions within their cores. Furthermore, analysis of FoldX-predicted ΔG values showed that higher weights on stability-associated features generated more designs with improved stability relative to the original proteins (**Figure 6H**) and produced greater improvements in predicted ΔG on average across four design rounds (**Figure 6I**).

Discussion

In this study, we present MotifAE, an unsupervised framework for extracting interpretable functional motifs from pLMs. By incorporating a local similarity loss, MotifAE encourages coherent feature activation that reflects the sequential continuity of motifs and basic structural elements, thereby improving its ability to discover biologically meaningful motifs. Analysis of MotifAE feature sequence patterns reveals rich diversity, ranging from single amino acid specificity to multiple amino acids consensus motifs, highlighting its potential for large-scale motif discovery. Furthermore, MotifAE-G demonstrates how unsupervised MotifAE features can be aligned with experimental data, enabling the identification of features associated with specific functions or properties. Beyond interpretation, feature selection also allows MotifAE-G to enhance predictive performance on related tasks and to guide the rational design of proteins with desired characteristics.

While some SAEs have been trained for pLMs[19,21,39], they directly adopted the training and interpretation strategies developed for large language models. Our work introduces methodological advances in both model training and

interpretation, making it better suited for biological sequences and enabling integration with experimental data. For model training, we introduced the local similarity loss, defined as the L1 norm of latent feature differences between neighboring residues, sharing the same form as the L1 norm used for enforcing sparsity. Adding this loss not only preserves sparsity in the latent space and maintains the quality of reconstructed embeddings, but also promotes coherent activation of latent features, thereby facilitating more effective identification of functional motifs. Current local similarity loss operates at the sequence level, incorporating three-dimensional proximity information could enable MotifAE to capture more complex structural motifs. Interpreting latent features is the most challenging part. A common approach is to compare latent feature activation patterns with known functional sites, such as those annotated in UniProt (as in Simon et al.[39]) or ELM (as used here). However, this strategy is limited by the incompleteness of current annotations and cannot reveal novel motifs, as some latent features may capture functions absent from existing annotations. Although Simon et al.[39] employed large language models to further annotate latent features, this approach still relied on known protein functions described in literature. Our work provides two strategies for annotating and potentially discovering novel motifs. First, we calculated PSSMs for each MotifAE feature; features exhibiting well-defined sequence patterns (high KL divergence; **Figure 5C-D**) may correspond to novel motifs. Future studies are needed to systematically link these sequence patterns to biological functions or protein properties. Second, our MotifAE-G framework can be applied to experimental datasets that lack explicit site-level annotations, and the activated residues of selected features may reveal the mechanistic basis of the function or property measured in the experiment.

Looking forward, MotifAE could be further improved in several directions. First, we currently use the L1 norm for the sparsity loss because it shares the same form as the local similarity loss. Alternative sparsity-promoting strategies, such as TopK[40] and BatchTopK[41] approaches, have been shown to improve SAE training, and future work is needed to explore how these can be integrated with the local similarity constraint. Second, since deep learning models inherently bias to their training data, both the pre-training dataset of the underlying pLM and the dataset used to train MotifAE influence the biological signals captured by latent features. Investigating how dataset composition shapes latent features is important. Finally, applying MotifAE-G to other large-scale functional assays, such as enzymatic activity and protein degradation profiling, could link MotifAE features to other functions.

Overall, we established a framework for the unsupervised discovery and interpretation of functional motifs from pLMs, enabling the study of motifs at an evolutionary scale. By systematically uncovering the sequence determinants of protein function, MotifAE provides a versatile tool for protein annotation, mutational effect interpretation, and rational protein engineering.

Methods

Sparse autoencoder architecture

Sparse autoencoder (SAE) projects language model embeddings into a sparse latent space. The model can be formulized as:

$$f = ReLU\left(W_{encoder}(x - b)\right) \quad (1)$$

$$\hat{x} = W_{decoder} \cdot f + b \quad (2)$$

where $W_{encoder}$ projects the original embedding x into the high dimensional latent space f and $W_{decoder}$ does the reverse to get the reconstruction $\hat{x}$. b is the bias term. ReLU was used as the activation function, setting negative values to zero and retaining positive values. Both the encoder and decoder consist of a single linear layer.

Here, the 650-million–parameter version of ESM2 was used, which produces embeddings with a dimensionality of 1,280. The SAE and MotifAE models were trained with varying latent space dimensions, and a dimension of 40,960 was selected (**Figure S1**).

MotifAE loss function

MotifAE employs three losses: the reconstruction loss (L_{rec}), the sparsity loss (L_{sp}), and the local similarity loss (L_{local}). The reconstruction and sparsity losses follow the SAE framework, preserving essential information from the input embeddings while enforcing sparse feature activations. The local similarity loss further constrains at least one nearby residue to have similar latent features, encouraging coherent activation patterns that capture the sequential continuity of protein motifs and basic structural elements. The loss functions are defined as follows:

$$L_{rec} = \frac{1}{n}\sum_{i=1}^{n}\|x_i - \hat{x}_i\|_2^2 \tag{3}$$

$$L_{sp} = \frac{1}{n}\sum_{i=1}^{n}|f_i| \tag{4}$$

$$D_i = \{|f_{i+1} - f_i|, \ldots, |f_{i+k} - f_i|\} \tag{5}$$

$$L_{local} = \frac{1}{n-k}\sum_{i=1}^{n-k} softmin(D_i) \cdot D_i \tag{6}$$

$$L_{total} = L_{rec} + \alpha \times L_{sp} + \beta \times L_{local} \tag{7}$$

Here, i denotes the residue index; n denotes the protein length; x and $\hat{x}$ represent the raw and reconstructed embeddings, respectively; and f denotes the latent feature activations. D_i quantifies the L1-norm difference in latent feature activations between residue i and its k neighboring residues. L_{local} encourages at least one nearby residue to have a similar latent feature activation. To achieve this, we apply a softmin over D_i to approximate the minimum difference. Only one direction along the sequence is considered for each residue, but since L_{local} is computed for the full sequence, the opposite direction is accounted for when evaluating neighboring residues. k is the local window size used for similarity calculation, which we set to three.

Model training

MotifAE and SAE were trained and evaluated on 2.3 million representative protein sequences clustered from the AlphaFold Protein Structure Database using Foldseek[42] (downloaded from https://afdb-cluster.steineggerlab.workers.dev/). Of these, 2.2 million sequences were used for training and the remainder for evaluation. Since ESM2 was trained with a maximum sequence length of 1,024, proteins exceeding this length were randomly truncated to a 1,024-residue region.

Models were implemented in PyTorch[43] (v2.2) and optimized using the Adam optimizer with a maximum learning rate of 0.001 and a 500-step linear warm-up. A batch size of 40 proteins was used. Residue order was preserved during training to retain positional information required by the local similarity loss. MotifAE was trained with both sparsity and local similarity loss weights set to 0.4, whereas the standard SAE was trained with a sparsity loss weight of 0.85 and no local similarity loss. The weights of the sparsity and local similarity losses were gradually annealed over 5,000 steps. Training was conducted for 80,000 optimization steps on a single NVIDIA L40S GPU, requiring approximately 10 hours per model.

Mutation effect prediction and data processing

The wild-type marginal method[22] was used to calculate log-likelihood ratio (LLR) to estimate mutation effects, which was calculated as:

$$LLR_i^{mt} = log\left(p(x_i^{mt}|\boldsymbol{x})\right) - log\left(p(x_i^{wt}|\boldsymbol{x})\right) \tag{8}$$

Where i represents residue index, mt and wt represent the mutant and wild-type amino acids, $\boldsymbol{x}$ is the full sequence without mask.

The ProteinGYM[23] database consists of over 200 protein deep mutational scanning (DMS) experiments, covering mutation effects across diverse protein functions. In this study, only single substitution mutations were evaluated.

The ClinVar[26] database provides expert-annotated mutations related to human diseases. Data preprocessing followed the procedure described in our previous work[44]. For this study, we considered only genes containing at least five pathogenic and five benign mutations, equal number of pathogenic and benign mutations were randomly sampled from each protein, resulting in a gene-balanced dataset of 2,272 pathogenic and 2,272 benign mutations in 207 genes. We used this dataset because ClinVar exhibits strong gene-level bias: the ratio of pathogenic to benign mutations varies substantially among genes[45].

For protein stability[32], the dataset "Tsuboyama2023_Dataset2_Dataset3_20230416.csv" and AlphaFold2-predicted protein structures were downloaded from https://zenodo.org/records/7992926. Proteins were defined using the "WT_name" column in the table. Only single substitution mutations with "ddG_ML" values were analyzed. The

"ddG_ML" values for the same sequence were averaged. Wild-type sequences of 412 proteins were clustered using MMseqs2[46] easy-cluster with the parameters: *--min-seq-id 0.3 -c 0.5 --cov-mode 1*. Homologs in the UniRef50 (downloaded in February 2025) were identified using MMseqs2 search with the parameters: *-s 7 -a 1*. Solvent accessibility was computed from predicted structures using mdtraj[47]. The raw solvent-accessible surface areas were normalized by the maximum accessible area of each amino acid residue, resulting in normalized values ranging from 0 to 1. For MotifAE-G training, 70% mutations in training proteins were used to select gates, while the remaining 30% were used as validation set to optimize hyperparameters.

Evaluation on ELM motifs and homodimer

ELM motifs were downloaded from the ELM[2] database in January 2025, and only instances annotated as *true positive* were analyzed. For proteins longer than 1,024 residues, a subsequence of length 1,024 was selected with the motif region positioned at the center. IUPred3[28] was used to predict disordered regions: residues with both long disordered region prediction > 0.5 and short disordered region prediction > 0.5 were classified as disordered, whereas all others were classified as ordered.

PINDER[30] is a database of PDB protein complexes. We filtered the dataset to include only structures released after September 30, 2021, and conducted data quality control following the PINDER supplementary material. From the resulting dataset, we got 3,303 homodimers and annotated contact residues as those with heavy atoms within 5 Å. We further removed redundancy by clustering the filtered dataset at a sequence identity threshold of 30% using MMseqs2 and retaining only the representative sequence from each cluster. Homodimers with less than five contacts were excluded. The final dataset comprised 1,565 dimers, with a median number of contacts of 44.

These two tasks were formulated as a binary classification problem, where each residue is labeled as either part of a motif/interface or not. In MotifAE and SAE, each residue is represented by a 40,960-dimensional feature vector. For each latent feature, its activation value was used as the residue-level prediction of motif/interface to calculate AUROC.

Activated peptides analysis and motif calculation

Activated peptides for each MotifAE feature were identified across the 2.3 million representative proteins. For each protein, only features with a summed activation value > 1 were analyzed. Coherent activations were recorded, and activations separated by a gap of one or two amino acids were merged. Only activated peptides with a mean feature activation value > 0.2 and lengths between 5 and 30 residues were retained for analysis.

To calculate position-specific scoring matrices (PSSMs), up to 1,000 activated peptides ranked by mean activation value were used for each feature. GibbsCluster 2.0[31] was employed to align the activated peptides and remove outliers with parameters: *-g 1 -l 5 -T -j 1 -I 1 -D 1*. The motif length was set to 5 amino acids. A trash cluster was used to remove sequences that did not align well with others, with the threshold set to 1. The alignment cores of length 5 generated by GibbsCluster were subsequently used to construct PSSMs, which were visualized using the Logomaker[48] Python package.

MotifAE-G architecture and training

The MotifAE-G model introduced a gating mechanism to amplify or attenuate feature signals during embedding reconstruction, which were subsequently used for downstream tasks. The gate was applied to each latent feature of MotifAE as:

$$f' = f \odot gate \tag{9}$$

where $\odot$ denotes elementwise multiplication.

For protein stability prediction, a learnable binary gate was introduced while keeping ESM2 and MotifAE parameters fixed. Each gate value could only take 0 or 1, and only features with a gate value of 1 were retained for reconstructing the embeddings. The gate was binarized using the Straight-Through Estimator[49,50]. In the forward pass, each gate was discretized as:

$$gate = \begin{cases} 1 & \sigma(g) > 0.5 \\ 0 & \sigma(g) \leq 0.5 \end{cases} \tag{10}$$

while in the backward pass, gradients were propagated directly through the sigmoid output $\sigma(g)$.

The reconstructed embeddings from the selected features were passed through the ESM2 MLP layer to predict logits for the 20 amino acids, which were then converted into LLRs between the mutant and wild-type amino acids (equation 8). The resulting LLRs were compared with experimentally measured mutation effects on protein stability "ddG_ML" to compute a soft Spearman correlation loss:

$$L = -\frac{Cov\big(r(ddG), r(LLR)\big)}{\sqrt{Var\big(r(ddG)\big)}\sqrt{Var\big(r(LLR)\big)}} + \lambda|g| \tag{11}$$

where $r(\cdot)$ denotes the differentiable rank function[51]. The first term maximizes the rank correlation between predicted and experimental mutation effects, while the second imposes an L1 regularization on the gate values to constrain the number of active features. The weighting coefficient λ was set to 1 based on performance on evaluation set.

Model training was implemented in PyTorch using the Adam optimizer (learning rate = 1×10^{-3}, default parameters). Each batch contained a single protein, and gradients were accumulated across all training proteins before updating the gate parameters, meaning the parameters were updated once per epoch. Models were trained for 60 epochs, and the checkpoint with the highest validation Spearman correlation after 50 epochs was selected for downstream analyses.

MotifAE-G protein design and stability prediction

For each representative test protein, we performed four rounds of iterative design. In each round, the probabilities of all possible single substitutions were predicted using ESM2 and MotifAE-G, with different weights applied to stability-associated features. The wild-type sequence was used without masking, following the same procedure as in mutation effect prediction. All possible mutations were ranked by their relative probability compared to the wild type: $p(x_i^{mt}|\boldsymbol{x})/p(x_i^{wt}|\boldsymbol{x})$. A top-k sampling strategy[37] (with k=10 in our experiments) was applied to select mutations, where the sampling probability of each mutation was proportional to its relative probability.

To evaluate the stability of the designed sequences, we predicted their structures using ESMFold. Global stability was then estimated using FoldX (version 20251231) based on the predicted structures. Specifically, we first ran the command "RepairPDB", followed by "Stability" to compute energy contributions, and used the sum (total energy) as an absolute stability metric. We reversed the sign of this total energy to represent ΔG. Only designs with ESMFold pLDDT > 0.8 and FoldX-predicted ΔG > 2 kcal/mol were analyzed.

Data availability

All data used in this work are publicly available. 2.3 million representative proteins were downloaded from https://afdb-cluster.steineggerlab.workers.dev/. ProteinGYM DMS data were downloaded from https://proteingym.org/download. ELM motifs were downloaded from http://elm.eu.org/downloads.html. For protein stability, the DMS data and AlphaFold2-predicted protein structures were downloaded from https://zenodo.org/records/7992926.

Code availability

Codes are available at GitHub: https://github.com/CHAOHOU-97/MotifAE.git.

Acknowledgments

This work was supported by NIH grants R35GM149527 and Simons Foundation SFARI #1019623.

Contributions

Y.S. and C.H. conceived the study. C.H. designed and implemented the experiments. D.L. assisted with analyses of homodimers and mutations. All authors evaluated and interpreted the results and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Gribskov, M., McLachlan, A. D. & Eisenberg, D. Profile analysis: detection of distantly related proteins. *Proceedings of the National Academy of Sciences* **84**, 4355–4358 (1987).
2. Kumar, M. *et al.* ELM-the Eukaryotic Linear Motif resource-2024 update. *Nucleic Acids Res* **52**, D442–D455 (2024).
3. Aditham, A. K., Markin, C. J., Mokhtari, D. A., DelRosso, N. & Fordyce, P. M. High-Throughput Affinity Measurements of Transcription Factor and DNA Mutations Reveal Affinity and Specificity Determinants. *Cell Systems* **12**, 112-127.e11 (2021).
4. Örd, M. *et al.* High-throughput investigation of cyclin docking interactions reveals the complexity of motif binding determinants. *Nat Commun* **16**, 7622 (2025).
5. Koren, I. *et al.* The Eukaryotic Proteome Is Shaped by E3 Ubiquitin Ligases Targeting C-Terminal Degrons. *Cell* **173**, 1622-1635 e14 (2018).
6. Zhang, Z. *et al.* Elucidation of E3 ubiquitin ligase specificity through proteome-wide internal degron mapping. *Molecular Cell* **83**, 3377-3392.e6 (2023).
7. Tokheim, C. *et al.* Systematic characterization of mutations altering protein degradation in human cancers. *Mol Cell* **81**, 1292-1308 e11 (2021).
8. Savojardo, C., Martelli, P. L. & Casadio, R. Finding functional motifs in protein sequences with deep learning and natural language models. *Current Opinion in Structural Biology* **81**, 102641 (2023).
9. Hou, C., Li, Y., Wang, M., Wu, H. & Li, T. Systematic prediction of degrons and E3 ubiquitin ligase binding via deep learning. *BMC Biol* **20**, 162 (2022).
10. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
11. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* https://doi.org/10.1038/s41586-021-03819-2 (2021) doi:10.1038/s41586-021-03819-2.
12. Kulmanov, M. *et al.* Protein function prediction as approximate semantic entailment. *Nature Machine Intelligence* **6**, 220–228 (2024).
13. Zhang, Z. *et al.* Protein language models learn evolutionary statistics of interacting sequence motifs. *Proc Natl Acad Sci U S A* **121**, e2406285121 (2024).
14. Vig, J. *et al.* Bertology meets biology: Interpreting attention in protein language models. (2020).
15. Zhang, Y., Zheng, J. & Zhang, B. Protein Language Model Identifies Disordered, Conserved Motifs Implicated in Phase Separation. *eLife* **14**, (2025).
16. Cunningham, H., Ewart, A., Riggs, L., Huben, R. & Sharkey, L. Sparse Autoencoders Find Highly Interpretable Features in Language Models. Preprint at https://doi.org/10.48550/arXiv.2309.08600 (2023).
17. Yun, Z., Chen, Y., Olshausen, B. A. & LeCun, Y. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. Preprint at https://doi.org/10.48550/arXiv.2103.15949 (2023).
18. Brixi, G. *et al.* Genome modeling and design across all domains of life with Evo 2. 2025.02.18.638918 Preprint at https://doi.org/10.1101/2025.02.18.638918 (2025).
19. Adams, E., Bai, L., Lee, M., Yu, Y. & AlQuraishi, M. From Mechanistic Interpretability to Mechanistic Biology: Training, Evaluating, and Interpreting Sparse Autoencoders on Protein Language Models. 2025.02.06.636901 Preprint at https://doi.org/10.1101/2025.02.06.636901 (2025).
20. Simon, E. & Zou, J. InterPLM: discovering interpretable features in protein language models via sparse autoencoders. *Nat Methods* **22**, 2107–2117 (2025).
21. Gujral, O., Bafna, M., Alm, E. & Berger, B. Sparse autoencoders uncover biologically interpretable features in protein language model representations. *Proceedings of the National Academy of Sciences* **122**, e2506316122 (2025).
22. Brandes, N., Goldman, G., Wang, C. H., Ye, C. J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat Genet* **55**, 1512–1522 (2023).

23. Notin, P. *et al.* ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. *Advances in Neural Information Processing Systems* **36**, 64331–64379 (2023).
24. Hou, C., Liu, D., Zafar, A. & Shen, Y. Understanding Language Model Scaling on Protein Fitness Prediction. 2025.04.25.650688 Preprint at https://doi.org/10.1101/2025.04.25.650688 (2025).
25. Barrio-Hernandez, I. *et al.* Clustering predicted structures at the scale of the known protein universe. *Nature* **622**, 637–645 (2023).
26. Landrum, M. J. *et al.* ClinVar: updates to support classifications of both germline and somatic variants. *Nucleic Acids Res* **53**, D1313–D1321 (2025).
27. Okada, M. Regulation of the Src Family Kinases by Csk. *Int J Biol Sci* **8**, 1385–1397 (2012).
28. Erdos, G., Pajkos, M. & Dosztanyi, Z. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. *Nucleic acids research* **49**, W297–W303 (2021).
29. Kshirsagar, M. *et al.* Rapid and accurate prediction of protein homo-oligomer symmetry using Seq2Symm. *Nat Commun* **16**, 2017 (2025).
30. Kovtun, D. *et al.* PINDER: The protein interaction dataset and evaluation resource. 2024.07.17.603980 Preprint at https://doi.org/10.1101/2024.07.17.603980 (2024).
31. Andreatta, M., Alvarez, B. & Nielsen, M. GibbsCluster: unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Res* **45**, W458–W463 (2017).
32. Tsuboyama, K. *et al.* Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620**, 434–444 (2023).
33. Hou, C., Zhao, H. & Shen, Y. Learning Biophysical Dynamics with Protein Language Models. *bioRxiv* 2024.10.11.617911 (2025) doi:10.1101/2024.10.11.617911.
34. UniProt, C. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506–D515 (2019).
35. Dieckhaus, H., Brocidiacono, M., Randolph, N. Z. & Kuhlman, B. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proceedings of the National Academy of Sciences* **121**, e2314853121 (2024).
36. Madani, A. *et al.* Large language models generate functional protein sequences across diverse families. *Nat Biotechnol* **41**, 1099–1106 (2023).
37. Darmawan, J. T., Gal, Y. & Notin, P. Sampling Protein Language Models for Functional Protein Design. in (2025).
38. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Res* **33**, W382-8 (2005).
39. Simon, E. & Zou, J. InterPLM: discovering interpretable features in protein language models via sparse autoencoders. *Nat Methods* **22**, 2107–2117 (2025).
40. Gao, L. *et al.* Scaling and evaluating sparse autoencoders. in (2024).
41. Bussmann, B., Leask, P. & Nanda, N. BatchTopK Sparse Autoencoders. Preprint at https://doi.org/10.48550/arXiv.2412.06410 (2024).
42. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* https://doi.org/10.1038/s41587-023-01773-0 (2023) doi:10.1038/s41587-023-01773-0.
43. Adam Paszke *et al.* PyTorch: an imperative style, high-performance deep learning library. in *Proceedings of the 33rd International Conference on Neural Information Processing Systems* Article 721 (Curran Associates Inc., 2019).
44. Zhong, G., Zhao, Y., Zhuang, D., Chung, W. K. & Shen, Y. PreMode predicts mode-of-action of missense variants by deep graph representation learning of protein sequence and structural context. *Nat Commun* **16**, 7189 (2025).
45. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* https://doi.org/10.1126/science.adg7492 (2023) doi:10.1126/science.adg7492.
46. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017).
47. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys J* **109**, 1528–32 (2015).
48. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (2020).
49. Huh, M., Cheung, B., Agrawal, P. & Isola, P. Straightening Out the Straight-Through Estimator: Overcoming Optimization Challenges in Vector Quantized Networks. Preprint at https://doi.org/10.48550/arXiv.2305.08842 (2023).
50. Yin, P. *et al.* Understanding Straight-Through Estimator in Training Activation Quantized Neural Nets. Preprint at https://doi.org/10.48550/arXiv.1903.05662 (2019).

51. Blondel, M., Teboul, O., Berthet, Q. & Djolonga, J. Fast Differentiable Sorting and Ranking. Preprint at https://doi.org/10.48550/arXiv.2002.08871 (2020).

Supplementary information

Unsupervised Discovery of Functional Motifs from Protein Language Model with MotifAE

Figure S1-3

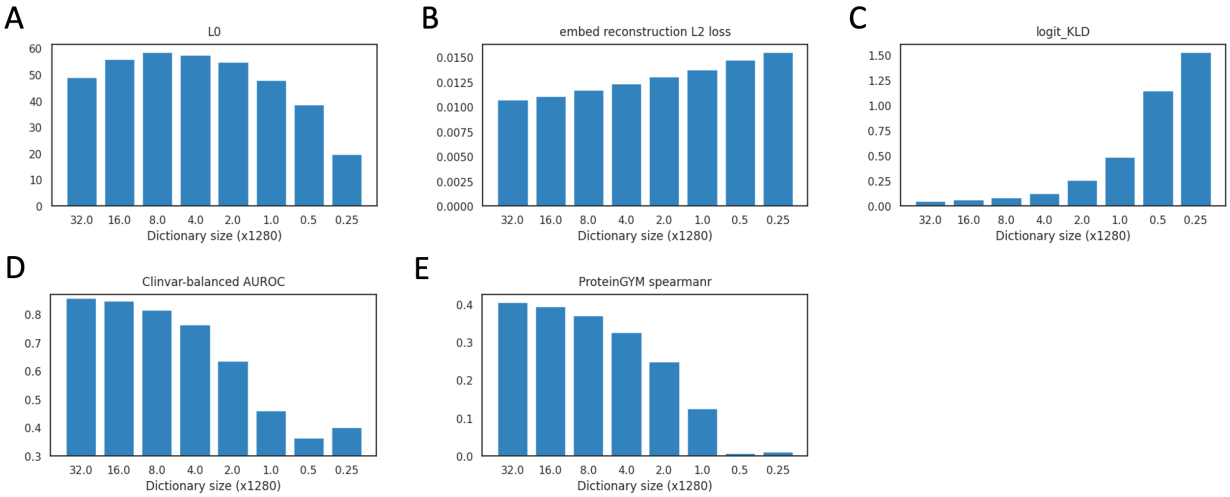

Figure S1. Comparison of different latent feature dimensions in MotifAE.

The values on the x-axis represent the expansion ratio relative to the embedding dimension (1,280). We note that the results shown here are obtained from MotifAE trained on normalized ESM2 embeddings in our preliminary experiment, whereas the results in the main figures are based on embeddings without normalization. Therefore, the absolute values in some figures may differ, but the overall conclusions should remain consistent. (**A**) Average number of non-zero latent feature activations per residue on the evaluation set. (**B**) Reconstruction loss on the evaluation set. (**C**) One residue in each evaluation protein was masked; amino acid probabilities were predicted using the MotifAE reconstruction with the fixed ESM2 MLP head and compared with predictions from ESM2 raw embeddings. A KLD value of zero indicates identical distributions. (**D–E**) Performance on mutation prediction using the wild-type log-likelihood ratio as the predictor. Mean Spearman correlation on ProteinGYM dataset was reported.

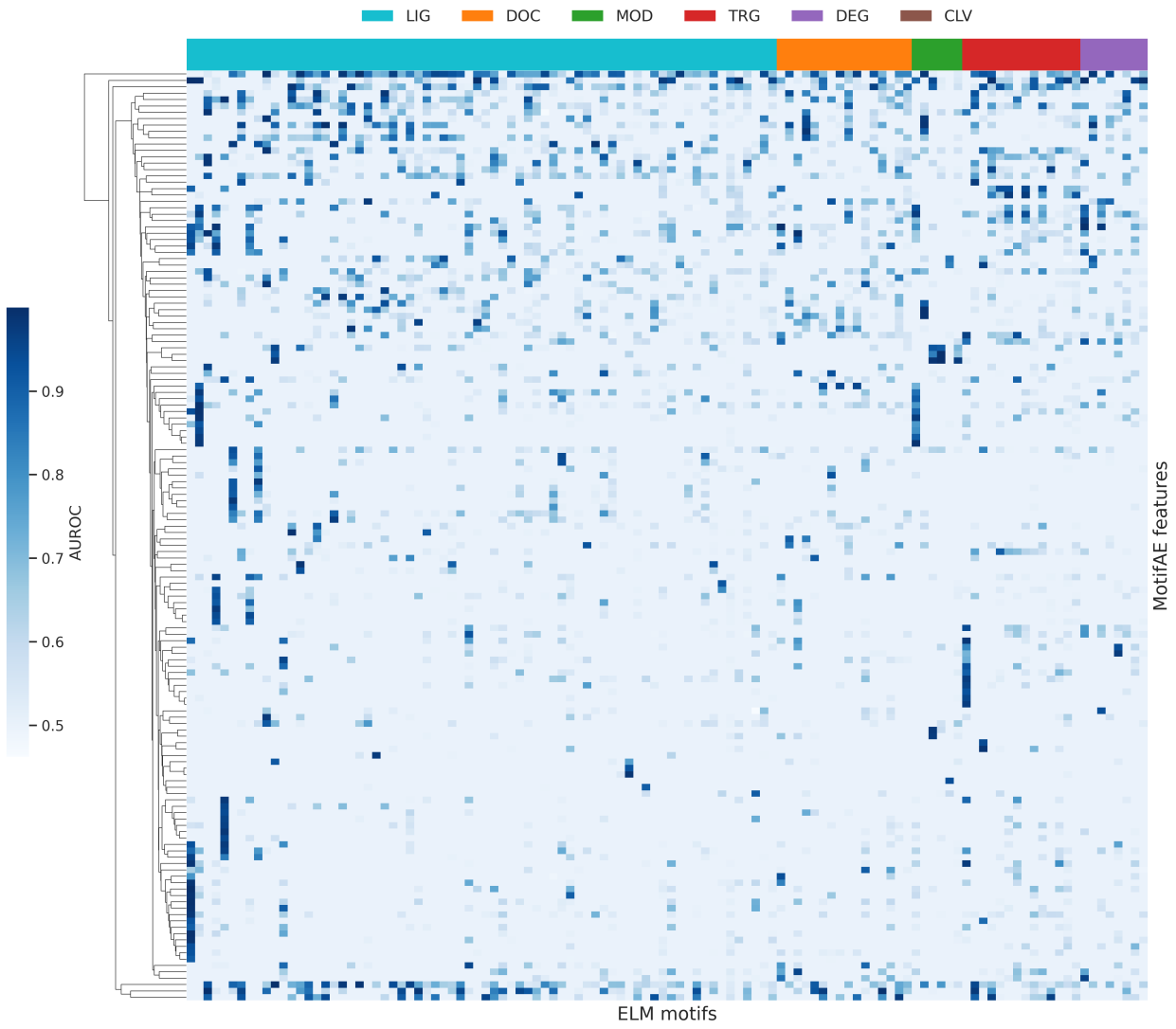

Figure S2. MotifAE feature performance on ELM motifs.

Matched motif–feature pairs were defined as those with AUROC > 0.9, resulting in 322 matched pairs involving 114 ELM motifs and 146 MotifAE features. The heatmap shows the performance (AUROC, indicated by blue intensity) of these 146 features across all 114 motifs. The six motif categories on the x-axis are the same as those shown in Figure 3A.

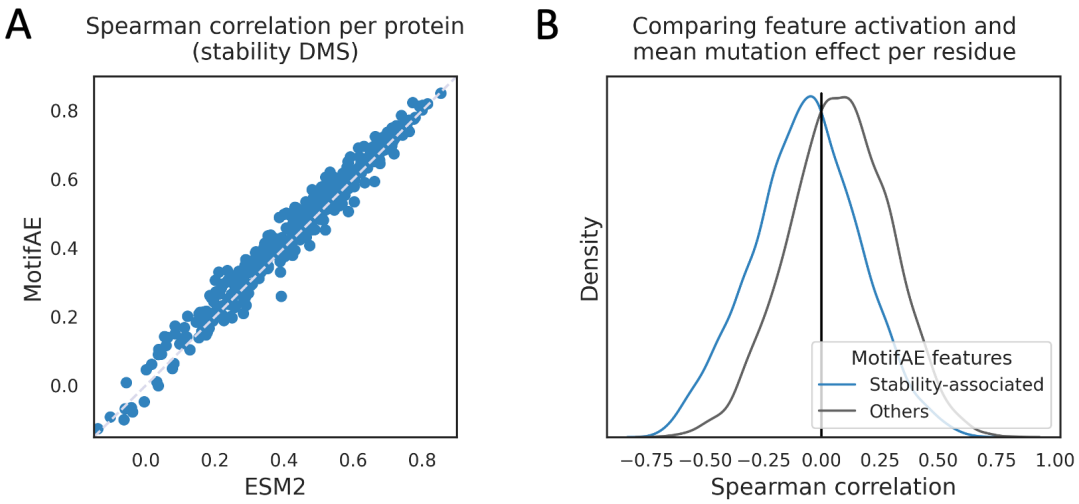

Figure S3.

(**A**) Zero-shot performance (Spearman correlation) on 412 protein stability DMS datasets. (**B**) Distribution of Spearman correlations between feature activation values and mean mutation effects per residue for each protein. Only features with at least ten activated residues in a protein were considered. Because destabilizing mutations have negative ΔΔG values, a negative Spearman correlation indicates that the feature is associated with protein stability.