FROM MECHANISTIC INTERPRETABILITY TO MECHANISTIC BIOLOGY: TRAINING, EVALUATING, AND INTERPRETING SPARSE AUTOENCODERS ON PROTEIN LANGUAGE MODELS

Etowah Adams,* **Minji Lee, Yiyang Yu & Mohammed AlQuraishi**
Department of Systems Biology
Columbia University
New York, NY 15213, USA
{ea2901,ml5045,yy3448,ma4129}@columbia.edu

Liam Bai *
Ginkgo Bioworks
Boston, USA
lbai@ginkgobioworks.com

ABSTRACT

Protein language models (pLMs) are powerful predictors of protein structure and function, learning through unsupervised training on millions of protein sequences. pLMs are thought to capture common motifs in protein sequences, but the specifics of pLM features are not well understood. Identifying these features would not only shed light on how pLMs work, but potentially uncover novel protein biology—studying the model to study the biology. Motivated by this, we train sparse autoencoders (SAEs) on the residual stream of a pLM, ESM-2. By characterizing SAE features, we determine that pLMs use a combination of generic features and family-specific features to represent a protein. In addition, we demonstrate how known sequence determinants of properties such as thermostability and subcellular localization can be identified by linear probing of SAE features. For predictive features without known functional associations, we hypothesize their role in unknown mechanisms and provide visualization tools to aid their interpretation. Our study gives a better understanding of the limitations of pLMs, and demonstrates how SAE features can be used to help generate hypotheses for biological mechanisms. We release our code, model weights and feature visualizer.[1].

1 INTRODUCTION

Protein language models (pLMs) are language models trained on large datasets of protein sequences. In the process of minimizing loss on their pre-training task–usually masked token prediction–they learn representations of proteins useful for downstream tasks including protein structure (Lin et al., 2023) and function (Rao et al., 2019; Notin et al., 2022) prediction. For this reason, pLMs have become a growing component of the protein biologist's toolkit.

As these models are widely adopted, a number of studies have investigated their inner workings to understand their limitations. Previous work suggests pLMs do not learn the biophysics of proteins, but rather store common sequence motifs and contacts (Zhang et al., 2024). For many downstream tasks, pLM performance is driven by features learned in early layers and does not scale well with pre-training (Li et al., 2024). Yet exactly what pLMs have learned about proteins remains unknown.

*Equal contribution

[1]Code for training, evaluating, and visualizing: https://github.com/etowahadams/interprot; visualizer for exploring feature activations overlaid with structure: https://interprot.com; SAE weights: https://huggingface.co/liambai/InterProt-ESM2-SAEs

Recent work in mechanistic interpretability points to a path forward. Sparse autoencoders (SAEs) have successfully extracted interpretable features from large language models (LLMs) like GPT-4 and Claude (Gao et al., 2024; Templeton et al., 2024). SAEs decompose model activations into a sparse, high-dimensional representation where individual latent dimensions often have interpretable activation patterns. These features provide a window into the inner workings of these complex models. Training SAEs on pLMs can serve an additional purpose. Unlike natural languages, which we intuitively understand, the language of proteins is more cryptic. SAE features from pLMs could potentially unveil patterns in protein sequences that we have not yet discovered.

In this work, we train SAEs on the residual stream of ESM-2, a widely used pLM (Lin et al., 2022) (Figure 1a). To interpret the SAE latents, we develop InterProt, a tool that visualizes latent activations on protein sequences and structures. We find that some features are highly interpretable, corresponding to a range of concepts from secondary structure elements to entire domains, as reported contemporaneously by Simon & Zou (2024). Furthermore, we find that many features activate highly only on specific protein families, suggesting that pLMs rely on internal representations of protein families to perform their pre-training task. We introduce a method to categorize SAE latents based on their family specificity and activation pattern (Figure 1b).

After developing a framework for understanding features, we explore the impact of SAE hyperparameters and ESM layers. We find that increasing the sparsity imposed on SAE latents increases the number of features that are family-specific as does increasing the expansion factor. The number of family-specific features also varies substantially by ESM layer, with middle layers containing the most family-specific features.

With the goal of mapping SAE features to important protein properties, we train linear probes on SAE features and analyze the most predictive features. Our method identifies features corresponding to known sequence determinants for properties such as thermostability and subcellular localization. These results demonstrate that SAEs can help us not only understand pLMs, but also the data on which they are trained. By making pLM representations interpretable, we envision that SAE features may help generate hypotheses for unknown biological mechanisms (Figure 1c).

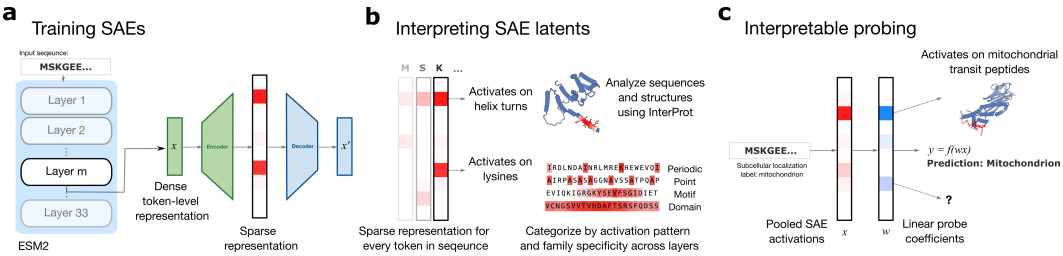

Figure 1: Overview of our paper. **a**. We train SAEs on the output of ESM intermediate layers. **b**. We interpret SAE latents using our latent visualizer InterProt and categorize features based on their activation pattern and family specificity. **c**. By interpreting the weights of linear models trained on SAE latents, we show how features that correspond to known sequence determinants can be identified.

2 Related Works

2.1 Interpreting protein language models

Previous evaluations of pLM representations on downstream tasks provide evidence of features that pLMs may be learning (Rao et al., 2019; Dallago et al., 2021; Detlefsen et al., 2022; Li et al., 2024). In particular, Li et al. 2024 performs a comprehensive analysis of ESM performance on a range of downstream tasks across different model sizes and layers. Of the tasks they tested, they find that only structure-related prediction tasks scales with model size, and that most tasks use "low-level features" learned early in pre-training. A key limitation of using downstream tasks to understand the features learned by pLMs is that features cannot be discovered without supervision. In contrast, SAEs can uncover features in a more unsupervised manner.

pLMs seem to use a notion of sequence homology and recurring protein motifs. Using influence functions, Gordon et al. 2024 found that pLM-derived sequence likelihoods are driven by homologous protein training data. Zhang et al. 2024 proposed that pLMs store common motifs in proteins and the pairwise contacts between them. They arrive at this proposition by analyzing what residue segments must be unmasked to recover a sequence contact (via the "categorical Jacobian").

Many studies have explored how pLMs work mechanistically as well. Analyses of attention matrices in pLMs have been shown to resemble pairwise contact maps (Vig et al., 2021; Rao et al., 2021; Lin et al., 2023), and contain information such as binding sites (Vig et al., 2021), allosteric sites (Kannan et al., 2024; Dong et al., 2024), and other known biological concepts (Chen et al., 2025).

Concurrent to our work, Simon & Zou (2024) trained SAEs on ESM-2 and presented methods to analyze their latents: visualizing them, evaluating them against Swiss-Prot annotations, and interpreting them with a language model. Our work builds on this foundation with different SAE training techniques, new feature categorizations, and new analysis methods with an emphasis on coevolution and scientific discovery. Mechanistic interpretability has also been applied to non-protein models in biology; we discuss those related works in subsection A.1.

3 Background and Methods

3.1 Sparse autoencoders (SAEs)

An SAE is an autoencoder designed to learn meaningful representations by enforcing sparsity constraints on encoder output. We use SAEs with TopK activation, as introduced in Gao et al. 2024. This activation function only allows the k largest latents to be non-zero. We choose TopK SAEs for their improved reconstruction at a given level of sparsity compared to other techniques, and because it allows us to directly set the L0-norm of the encoding using k. We note that improved reconstruction may come at the cost of increased feature absorption (Karvonen et al., 2024). We train TopK SAEs on 1 million randomly-selected sequences under 1022 residues from UniRef50 (Suzek et al., 2007).

Following Lieberum et al. 2024, we refer to the hidden dimensions of the SAE as "latents" to clearly distinguish them from the underlying conceptual features of the model. This is in contrast to previous works which use the word feature to refer to both Bricken et al. (2024).

3.2 Downstream tasks

Secondary structure. The secondary structure dataset from TAPE (Rao et al., 2019) contains residue-level labels in 3 classes: alpha helix, beta strand, or other. To prevent data leakage, sequences are divided into training and test sets based on a threshold of 25% sequence identity.

Subcellular localization. Subcellular localization is a protein-level label for where a protein is localized within a cell. The dataset from Almagro Armenteros et al. (2017) consists of UniProt-sourced eukaryotic proteins, with stringent homology-based train-test splits.

Thermostability. The thermostability dataset from the Melteome Atlas (Jarzab et al., 2020) measures the melting temperature of 48,000 proteins across 13 species. We use the train-test split from FLIP (Dallago et al., 2021) containing all sequences clustered at 25% sequence identity.

Mammalian cell expression. The dataset curated by Masson et al. (2022) contains expression data of 2,165 proteins from the human secretome in Chinese hamster ovary (CHO) cells, commonly used mammalian hosts for therapeutic protein production. We use this dataset as a binary classification task to determine whether a protein can be successfully expressed in CHO cells. We split proteins into training and test sets based on a threshold of 40% sequence identity.

3.3 Linear probes

For each task and for varying ESM layers, we train linear probes on both ESM and SAE embeddings. We mean-pool embeddings over the length of the sequence for tasks with protein-level labels: subcellular localization, thermostability, and mammalian cell expression. For all linear regression and logistic regression tasks, we perform grid search over a range of regularization strengths and pick the best model using a validation set. Table 1 summarizes probe implementation for each task.

4 Feature Analysis

4.1 SAEs uncover interpretable features

We train TopK SAEs on the residual stream of different layers of ESM-2 (650M) to explore its learned representations. To interpret a given SAE latent dimension (*i.e.*, a latent), we examine the top activating protein sequences, *i.e.*, those that produce the largest activation for that dimension. We developed a latent visualizer, InterProt, to streamline the process of identifying features. InterProt allows users to align the most-activating sequences and highlight their corresponding protein structure by activation strength. It is open source and can be accessed at `interprot.com`.

Through these visualizations, we gain insight into how ESM internally represents protein sequences (Figure 2). Many discovered features correspond to recognizable biological concepts, such as secondary structure elements (*e.g.*, alpha helices and beta strands), short conserved motifs, or entire functional domains. Some features capture biochemical concepts like cysteine-cysteine bonds or the orientation of residue side chains. In many ways, latents capture characteristics of a protein that a biologist might notice.

Notably, while many latents are intuitive (*e.g.*, focused on conserved motifs), others are unexpectedly specific. For instance, rather than a single, generic "alpha helix" latent, we observe multiple helix-related latents that activate under more context-dependent conditions (*e.g.*, at the beginning of a helix, only when the helix is buried, or when the helix meets a certain length criterion). Other examples include highly context specific amino acids, such as asparagine residues in disordered regions.

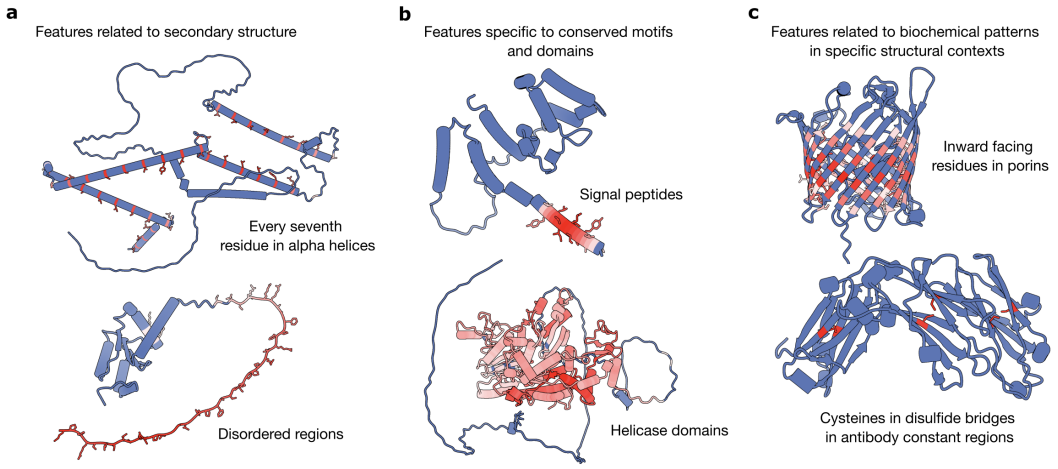

Figure 2: Examples of SAE features. We find features related to secondary structure (a), conserved motifs and domains (b), and biochemical patterns in specific structural contexts (c). The structures are colored according to activation (red: activation, blue: no activation).

4.2 SAE latents reveal a learned notion of protein families

Intriguingly, a large subset of latents appear to be protein family-specific. These latents activate weakly on most sequences but strongly on proteins belonging to a particular family. For example, the latent shown in Figure 3a appears to highly activates only for the central helix in beta-lactamases, but only exhibits marginal activations when a similar structural motif appears in proteins outside the beta-lactamase family.

By contrast, a non-family-specific latent such as the "disordered" latent in Figure 3b, responds broadly to intrinsically disordered regions in proteins across many families. These observations collectively suggest that pLMs learn internal representations that, in part, mirror known sequence-homology based groupings of proteins, akin to a model-internal notion of protein families.

To better understand the relationship between a family-specific latent and protein family, we investigate what happens when we "steer" ESM using the latent. Specifically, we set the activation of a family-specific latent to some multiplicative factor of the maximum activation of the sequence, and continue the forward pass of the model using the reconstructed activation. We find that the steered sequence retains awareness of family-level conservation, as it changes most minimally at evolutionarily conserved positions (Figure 5).

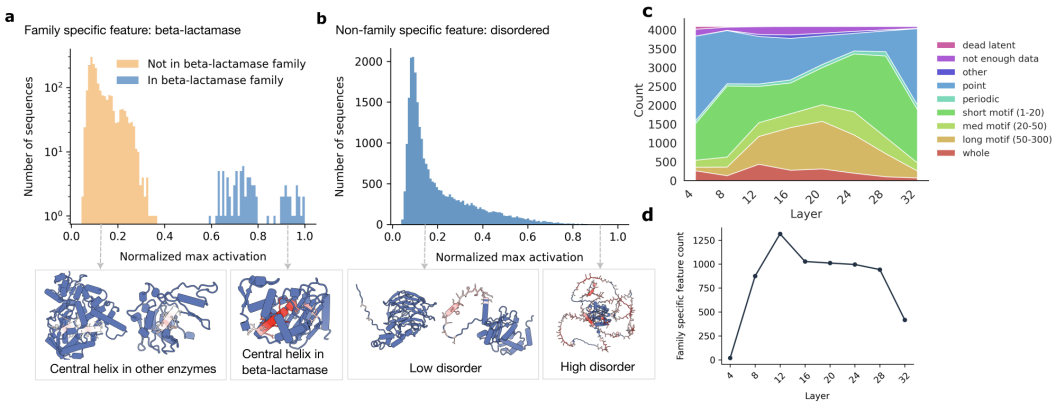

Figure 3: **a**. Example of a family specific feature, the central helix of a beta lactamase. The histogram shows the distribution of maximum activations in the sequences which activate the latent. **b**. Example of a non-family specific feature, corresponding to disordered sequences. **c**. Latents classified by activation pattern across layers. **d**. Number of family-specific latents across layers. The SAEs used in c and d have a hidden dimension of 4096 and k=64.

4.3 Classifying Latents

Motivated by these observations, we propose two classification schemes for the SAE latents.

Activation Pattern. Latents exibit distinct activation patterns across their top sequences (Figure 1b). To capture these differences, we use a rule-based scheme that classifies features into categories such as: point (activates on single residues at a time), periodic (activates in a regular interval, repeating every n residues), motif (activates in short, medium, or long contiguous intervals), and domain (activates over nearly the entire sequence). Refer to Table 2 for precise definitions.

Family Specificity. To quantify family specificity, we use each latent for binary classification: given a protein, can this latent's activations be used to predict membership in a particular family? We label a latent "family-specific" if it achieves an F1 score > 0.7 at some activation threshold (subsection A.4).

These classifications allow us to systematically group SAE features and draw broader conclusions about their roles in representing protein structure and function.

4.4 Influence of SAE Training Choices on Latent Classification

Prior literature on LLMs suggests that SAE-learned features depend on multiple factors, including model hyperparameters (*e.g.*, the sparsity level and latent dimension size) and training data (*e.g.*, which layer's activations are used). We therefore examine how these factors impact the types of features that emerge in pLMs.

Effect of k. In TopK SAEs, k controls the number of latents active for a given input, effectively setting the sparsity level. We find that, for a fixed hidden dimension, lower k (*i.e.*, higher sparsity) yields more family-specific features (Figure 6a). One plausible explanation is that family-specific features constitute the most salient signals used in reconstruction. Identifying family specific features may help us understand how *de novo* proteins are represented by pLMs.

Effect of expansion factor. We also experiment with increasing the latent dimension, called the expansion factor. While we initially hypothesized that a larger latent dimension might shift the distribution of feature activation patterns, we observed that the activation pattern classifications (point, motif, domain, etc.) remain largely consistent (Figure 6b).

Comparison across layers. Finally, we investigate layer-wise differences in the learned features. Previous studies have shown that downstream performance can vary substantially depending on the layer, suggesting that different layers capture different features. In our experiments, family-specific features tend to peak in certain early-to-mid layers, then decline in later layers (Figure 3d). Similarly, features which activate contiguously over long sections (motif/domain) are more common in earlier layers, whereas later layers exhibit features with shorter contiguous activations (Figure 3c). Given that later layers are thought to specialize for final logit computation (Gao et al., 2024), one possibility is that shorter, more specific activations better serve this goal.

5 Interpretable Probing

Representations learned by pLMs can be used in downstream tasks. Often, this is as simple as training a linear model on the pLM embeddings. Although this leads to a predictive model, it does so at the expense of interpretability, as the coefficients of the linear model do not correspond to interpretable features. In this section, we explore the ability of SAE representations to overcome this limitation. Since SAE latents are often interpretable, those that contribute significantly to prediction could reveal biological reasons that explain the prediction.

In the context of LLMs, linear probes on SAE embeddings can achieve performance similar to, or in some cases better than, linear probes on their base model embeddings (Bricken et al., 2024; Kantamneni et al., 2024). To explore this dynamic in pLMs, we train linear probes on ESM and SAE embeddings across different layers. Following Li et al. (2024), we evaluate the probes on benchmarked downstream tasks: secondary structure, subcellular localization, thermostability, as well as a novel task, mammalian cell line expression prediction. Probes trained on SAE latents offer comparable performance to their ESM embedding counterparts across layers (subsection A.7).

5.1 SAE probes uncover interpretable latents corresponding to known mechanisms

To assess the interpretability of linear probes trained on SAE latents, we analyze the distribution of their coefficients (Figure 9) and manually inspect the latents with the largest coefficients in the InterProt visualizer [2]. We find that those latents often correspond to biological concepts with known relevance to each task (Figure 4). [3]

5.1.1 Secondary structure

Secondary structure prediction is a task that is known to utilize the features learned during pLM pre-training (Li et al., 2024). Our analysis of the highest-weighted SAE latents in our secondary structure probe confirms this and provides an explanation for why early layers perform worse.

Our SAE probes reveal that most layers contain features corresponding to secondary structure elements. The most positive coefficients for the alpha helix class are typically latents that activate exclusively on helices (Figure 4a). They range from specific helix residues, to entire helices, to helix-helix interactions. We observe similar results for the two other classes: beta strand and *other*, where *other* tends to correlate with features that activate on disordered regions.

We find that early layer performance is worse because they do not contain generic features for secondary structure. For example, the top latent for helix classification in layer 4 (L4/2624) detects proline (P) residues, which often appear at helix boundaries due to their disruptive effect on hydrogen

[2] Linear probe coefficients are not comparable across tasks because different tasks may apply different regularization strengths, optimized via grid search.

[3] In this section and in Figure 4, we discuss latents from a range of layers. The choice of layers is qualitative—primarily based on latents we find particularly interpretable or notable—and does not imply the lack of interesting latents in undiscussed layers.

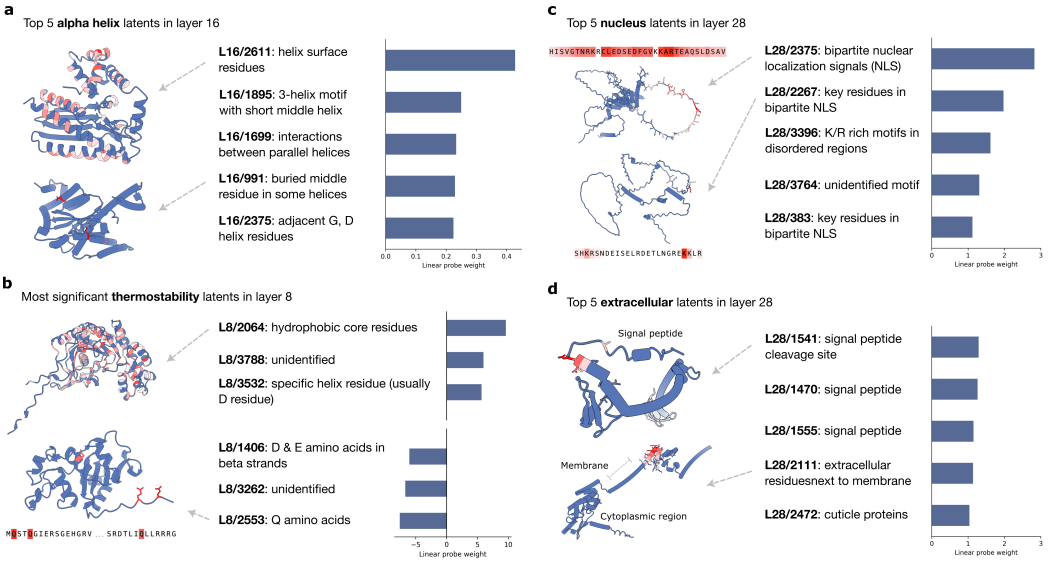

Figure 4: Linear probes identify predictive SAE latents for downstream tasks. We show the top 5 predictive latents for alpha helix prediction (**a**), the latents corresponding to the 3 most positive and 3 most negative coefficients from the linear probe on thermostability (**b**), top 5 nucleus and extracellular latents from the subcellular localization prediction task (**c-d**).

bonding. This reliance on a weak amino acid-level correlation highlights the absence of a more generalized helix feature. Similarly, the top latent for the disordered classifier activates on N-terminal residues, reflecting the spurious correlation that many proteins have disordered regions at their N-terminus, rather than capturing fundamental structural properties.

5.1.2 Subcellular localization

Transport mechanisms involving sequence motifs such as transit peptides and localization signals determine the location of proteins within the cell. We find that linear probes trained on SAE latents can uncover some of these known mechanisms.

Nuclear localization. The most highly weighted latents for nuclear localization often correspond to nuclear localization signals (NLS). For example, the top latents in the probe for layer 28 probe activates on K/R rich sequence motifs (Figure 4c) and on bipartite nuclear localization signals (L28/2375), which generally follow the pattern $\mathrm{R/K(X)}_{10-12}\mathrm{KRXK}$ (Lu et al., 2021). The latent activates strongly on the R/K motifs at each end and can flexibly recognize the variable-length linker sequence that connects them. These sequence motifs engage cellular mechanisms that enable the protein to enter the nucleus. As many NLS variants still remain unknown, our SAE-based approach has the potential to become a novel method to aid their discovery and characterization.

Extracellular localization. The most positive coefficients for the extracellular class relate to signal peptides (Figure 4d). Signal peptides are short sequences at the N-terminus that trigger the secretory pathway to transport proteins outside the cell. L28/1541 activates on the signal peptide cleavage site while L28/1470 and L28/1555 activate across the peptide. The top latents also include a latent that activates on the extracellular portion of membrane-bound proteins (L28/2111), demonstrating precise recognition of transmembrane regions, as well as a latent (L28/2472) that only activates on cuticle proteins, extracellular proteins that support the exoskeleton of insects.

For the other localization classes, we also find a variety of interpretable latents, including mitochondrial transit peptides (L28/3277) and transmembrane regions (L28/2818, L28/1298) (Figure 8c). Overall, this task uncovered many latents related to signal peptides, consistent with the observations from (Almagro Armenteros et al., 2017) that the attention maps of their transformer model trained on this task focus on sequence termini.

5.1.3 THERMOSTABILITY

Thermostability prediction is a task that does not scale with increased pre-training and larger models (Li et al., 2024). Our SAE probes reveal why—thermostability prediction relies most on simple amino acid statistics (Figure 4b).

Unlike the other tasks, the middle layers perform worse than early or late layers. We find that in the middle layers the most significant latents tend to recognize specific amino acids. For example, layer 24's most positive latents correspond to Arginine (R), Tyrosine (Y), and Leucine (L); its most negative latents recognize Glutamine (Q), Aspartic acid (D), and Threonine (T). Glutamine is known to correlate with instability (Farias & Bonato, 2003). In contrast, earlier layers seem to contain more relevant features. For instance, the most positive coefficient in layer 8 (L8/2064) corresponds to a latent that activates primarily on hydrophobic residues, which are important for stability (Figure 4b).

6 DISCUSSION

In this work, we train sparse autoencoders (SAEs) to extract interpretable features from a protein language model (pLM). We categorize SAE latents based on their activation patterns and specialization across protein families, systematically analyzing how SAE sparsity, width, and layer choice affect feature extraction. Our work offers a new perspective on previous efforts to understand pLMs, confirming known results such as the central role of sequence homology (Lin et al., 2022; Gordon et al., 2024), the use of motifs (Zhang et al., 2024), and the unexpected effectiveness of early layers on some downstream tasks (Li et al., 2024). To facilitate further exploration, we introduce InterProt, an open-source visualizer for interpreting SAE latents.

Mechanistic interpretability methods like SAEs hold promise not only for understanding how complex models function but also for advancing scientific discovery (Donhauser et al., 2024). However, existing approaches often rely on matching features to concepts already known to the scientific community, potentially overlooking novel features. To address this gap, we developed a probing strategy that associates latents with candidate functional roles by training linear probes on SAE latents. Through this process, we uncover features that align with known sequence determinants, such as nuclear localization signals, which our probe identified as strongly predictive of nuclear localization. Had the role of these signals been unknown, our probing results would have provided a strong hypothesis for subsequent validation. More broadly, probing SAE latents offers a means to explain how pLMs achieve their performance on downstream tasks. Notably, we find that thermostability prediction relies heavily on latents which provide simple amino acid composition statistics.

In our probing experiments, we observe that not all predictive latents are easily interpretable. For instance, a predictor for membrane localization, latent L28/3154, predominantly activates on poly-alanine motifs, whose function remains unclear. Such hard-to-interpret but predictive features could result from biases in training data, limitations of our SAE, or ESM. However, another intriguing possibility is that these features correspond to biological mechanisms that have yet to be discovered.

Limitations. Our study has several limitations. First, we focus exclusively on the 650M-parameter version of ESM-2, and extending our analysis to other model sizes, checkpoints, and architectures could yield richer insights into how features vary with pre-training scale and model architecture. Second, the features uncovered by our SAEs are sensitive to the training data distribution; although we use a diverse set of sequences from UniRef50, different datasets may produce a distinct feature set. Third, our probing approach relies on mean-pooling for protein-level tasks, which has known limitations—alternative pooling methods may better capture functional associations (NaderiAlizadeh & Singh, 2024). Finally, our interpretations are constrained by current biological knowledge, meaning some latents with complex activation patterns may be misinterpreted or remain unexplained.

Future directions. The success of our SAE linear probes in uncovering known biological mechanisms suggests their potential for discovering new ones. A key next step is to formulate and experimentally test hypotheses around unexplained predictive features. Since SAE-derived features depend on the dataset (Kissane et al., 2024), investigating how data diversity affects feature emergence could be valuable–training on task-relevant sequences may yield more function-specific latents. Feature steering is another promising direction: while Simon & Zou (2024) provides a proof of concept, further research is needed to determine whether SAEs can be used as a practical tool for protein engineering.

MEANINGFULNESS STATEMENT

A meaningful representation of life is one that facilitates scientific discovery. In this work, we leverage sparse autoencoders (SAEs) to extract interpretable features from protein language models (pLMs). By probing SAE latents, we recover known sequence determinants for functions such as nuclear localization, and identify predictive but unexplained features that could correspond to novel biological mechanisms. Our approach contributes to meaningful biological representations by bridging learned representations with testable hypotheses, aiding both interpretability and discovery in protein science.

REFERENCES

Adam Green. Through a glass darkly: Mechanistic interpretability as the bridge to end-to-end biology, 2024. URL https://www.markov.bio/research/mech-interp-path-to-e2e-biology.

José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, 2017.

Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler, Marie-Claude Blatter, Anne Estreicher, Elisabeth Gasteiger, Maria J Martin, Karine Michoud, Claire O'Donovan, Isabelle Phan, et al. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic acids research*, 31(1):365–370, 2003.

Trenton Bricken, Jonathan Marcus, Siddharth Mishra-Sharma, Meg Tong, Ethan Perez, Mrinank Sharma, Kelley Rivoire, and Thomas Henighan. Using dictionary learning features as classifiers, 2024. https://transformer-circuits.pub/2024/features-as-classifiers/index.html.

Hedi Chen, Boxue Tian, Jingrui Zhong, Xiaochun Zhang, Jingke Chen, Lin Guo, Xiaoliang Xiong, Xiaonan Zhang, Xiangyu Liu, and Bailong Xiao. Automatically defining protein words for diverse functional predictions based on attention analysis of a protein language model. *bioRxiv*, pp. 2025–01, 2025.

Christian Dallago, Jody Mou, Kadina E Johnston, Bruce J Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K Yang. Flip: Benchmark tasks in fitness landscape inference for proteins. *bioRxiv*, pp. 2021–11, 2021.

Nicki Skafte Detlefsen, Søren Hauberg, and Wouter Boomsma. Learning meaningful representations of protein sequences. *Nature communications*, 13(1):1914, 2022.

Tianze Dong, Christopher Kan, Kapil Devkota, and Rohit Singh. Allo-allo: Data-efficient prediction of allosteric sites. *bioRxiv*, pp. 2024–09, 2024.

Konstantin Donhauser, Kristina Ulicna, Gemma Elyse Moran, Aditya Ravuri, Kian Kenyon-Dean, Cian Eastwood, and Jason Hartford. Towards scientific discovery with dictionary learning: Extracting biological concepts from microscopy foundation models, 2024. URL https://arxiv.org/abs/2412.16247.

Sávio T Farias and MC Bonato. Preferred amino acids and thermostability. *Genet Mol Res*, 2(4): 383–393, 2003.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Cade Gordon, Amy X Lu, and Pieter Abbeel. Protein language model fitness is a matter of preference. *bioRxiv*, pp. 2024–10, 2024.

Anna Jarzab, Nils Kurzawa, Thomas Hopf, Matthias Moerch, Jana Zecha, Niels Leijten, Yangyang Bian, Eva Musiol, Melanie Maschberger, Gabriele Stoehr, et al. Meltome atlas—thermal proteome stability across the tree of life. *Nature methods*, 17(5):495–503, 2020.

Gokul R Kannan, Brian Hie, and Peter S Kim. Single-sequence, structure free allosteric residue prediction with protein language models. *bioRxiv*, pp. 2024–10, 2024.

Subhash Kantamneni, Josh Engel, Senthooran Rajamanoharan, and Neel Nanda. Sae probing: What is it good for? absolutely something, 2024. https://www.lesswrong.com/posts/NMLq8yoTecAF44KX9/sae-probing-what-is-it-good-for-absolutely-something.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Arthur Conmy, Callum McDougall, Kola Ayonrinde, Matthew Wearden, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders, December 2024. URL https://www.neuronpedia.org/sae-bench/info.

Connor Kissane, Robert Krzyzanowski, Neel Nanda, and Arthur Conmy. Saes are highly dataset dependent: A case study on the refusal direction. Alignment Forum, 2024. URL https://www.alignmentforum.org/posts/rtp6n7Z23uJpEH7od/saes-are-highly-dataset-dependent-a-case-study-on-the.

Francesca-Zhoufan Li, Ava P Amini, Yisong Yue, Kevin K Yang, and Alex X Lu. Feature reuse and scaling: Understanding transfer learning with protein language models. *bioRxiv*, pp. 2024–02, 2024.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, 2024. URL https://arxiv.org/abs/2408.05147.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Juane Lu, Tao Wu, Biao Zhang, Suke Liu, Wenjun Song, Jianjun Qiao, and Haihua Ruan. Types of nuclear localization signals and mechanisms of protein import into the nucleus. *Cell communication and signaling*, 19(1):60, 2021.

Helen O Masson, Chih-Chung Kuo, Magdalena Malm, Magnus Lundqvist, Åsa Sievertsson, Anna Berling, Hanna Tegel, Sophia Hober, Mathias Uhlén, Luigi Grassi, et al. Deciphering the determinants of recombinant protein yield across the human secretome. *bioRxiv*, pp. 2022–12, 2022.

Navid NaderiAlizadeh and Rohit Singh. Aggregating residue-level protein language model embeddings with optimal transport. *bioRxiv*, pp. 2024–01, 2024.

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan N Gomez, Debora Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *International Conference on Machine Learning*, pp. 16990–17017. PMLR, 2022.

Typhaine Paysan-Lafosse, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Peer Bork, Alan Bridge, Lucy Colwell, et al. Interpro in 2022. *Nucleic acids research*, 51(D1):D418–D427, 2023.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=fylclEqgvgd.

Viktoria Schuster. Can sparse autoencoders make sense of latent representations?, 2024. URL https://arxiv.org/abs/2410.11468.

Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular systems biology*, 7(1):539, 2011.

Elana Simon and James Zou. Interplm: Discovering interpretable features in protein language models via sparse autoencoders. *bioRxiv*, pp. 2024–11, 2024.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Baris E Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H Wu. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288, 2007.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. Bertology meets biology: Interpreting attention in protein language models, 2021. URL https://arxiv.org/abs/2006.15222.

Jian Ye, Scott McGinnis, and Thomas L Madden. Blast: improvements for better sequence analysis. *Nucleic acids research*, 34(suppl_2):W6–W9, 2006.

Zhidian Zhang, Hannah K Wayment-Steele, Garyk Brixi, Haobo Wang, Dorothee Kern, and Sergey Ovchinnikov. Protein language models learn evolutionary statistics of interacting sequence motifs. *Proceedings of the National Academy of Sciences*, 121(45):e2406285121, 2024.

A Appendix

A.1 Dictionary learning for feature extraction from models trained on biological data

Following the success of extracting interpretable features from LLMs, a number of works have applied sparse dictionary learning techniques to models trained on scientific data. Donhauser et al. 2024 evaluates the possibility of extracting features from models trained on microscopy data. Interpretable features have also been uncovered from single cell foundation models (Schuster, 2024; Adam Green, 2024). Associating the extracted features with functional labels via probing, however, has not yet been explored in this context.

A.2 Probe details

Table 1 summarizes the linear probes used in each downstream task.

Table 1: Summary of downstream tasks for interpretable probing

Dataset	Task type	Label level	Evaluation metric	Probe implementation
Secondary structure	Classification	Residue	Accuracy	PyTorch linear classifier
Subcellular location	Classification	Protein	Accuracy	Sklearn logistic regression
Thermostability	Regression	Protein	Spearman's ρ	Sklearn ridge regression
CHO cell expression	Classification	Protein	Accuracy	Sklearn logistic regression

A.3 Steering family-specific latents

We steer each family-specific latent and measure and analyze the resulting sequence changes. Steering refers to clamping of the activation of a specific latent to a fixed level–either increasing it to amplify the feature or decreasing it to suppress it. For each family-specific latent, corresponding to an InterPro family, we identify the top-activating sequences, conduct protein-protein BLAST (Ye et al., 2006), and align sequences within the same InterPro category and obtain consensus sequence using Clustal Omega (Sievers et al., 2011). Finally, we determine the maximum activation of all sequences and set activation level of all sequences to 0.5×, 1×, 1.5×, 2×, and 4× the maximum value. As shown in Figure 5, steering tends to keep the residues with high consensus as determined by Clustal Omega, effectively pushing the different sequences of a family to the consensus sequence.

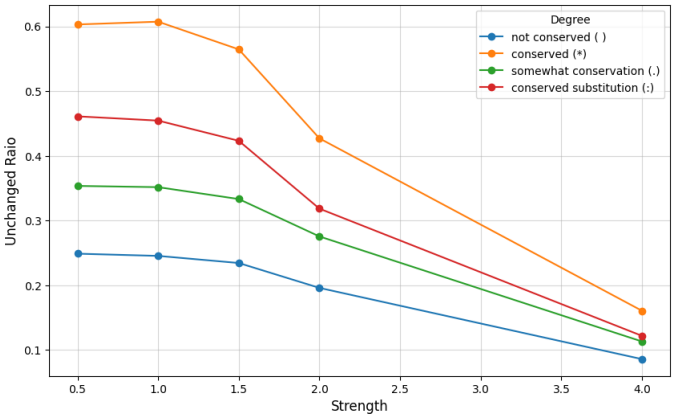

Figure 5: The fraction of residues that remains unchanged after steering, averaged for 168 family specific features that fraction of unchanged exceeds 20% for strength=0.5. Clustal Omega determines the degree of conservation by comparing residues at each position across sequences using a substitution scoring matrix.

A.4 Classifying latents by family specificity

We evaluate the ability of a single SAE latent to classify a protein as belonging to a specific InterPro-defined protein family (Paysan-Lafosse et al., 2023). To do this, we determine the optimal normalized activation threshold that maximizes the F1 score for binary classification of protein families. Specifically, we sweep threshold values from 0.1 to 0.9 in 0.1 increments and select the threshold that results in the highest F1 score. Our final evaluation is conducted on a held-out test set.

For our evaluation, we use protein sequences from Swiss-Prot (Boeckmann et al., 2003). To reduce sequence redundancy, we cluster all Swiss-Prot sequences with a length of 1022 residues or fewer at 30% sequence identity using MMseqs2 (`mmseqs easy-cluster swissprot.fasta session tmp --min-seq-id 0.3 -c 0.9 -s 8 --max-seqs 1000000 --cluster-mode 1`) (Steinegger & Söding, 2017). Given the low sequence identity between clustered sequences, we perform a random train-test split for classification evaluation.

To determine which protein family to assess for each latent, we identify the InterPro family of the sequence that exhibits the highest activation for that latent. The F1 score is then computed for this family using the thresholding approach described above.

A.5 Classifying latents by activation pattern

Table 2: Activation pattern classification criteria

Category	**Criteria**
Dead Latent	If the latent is never activated by any test sequence, it is classified a dead latent.
Not Enough Data	If less than 5 sequences activate the latent then we say there is not enough data.
Periodic	Features that exhibit consistent activation patterns at regular intervals. These features must satisfy: (1) a high frequency of activation at specific positions (over 50% of distances between activations are the same two values), (2) a large number of activation regions (there are more than 10 activations per sequence), and (3) relatively short contiguous activation spans (median length of the top activating contig is less than 10).
Point	Features that activate in a highly localized manner, defined by a single, prominent activation site (the median length of the highest activating region is 1).
Motif (Short: 1-20)	Features that activate in short contiguous regions (median length of the highest activating region is > 1 and < 20) and have an overall mean activation coverage of less than 80%.
Motif (Medium: 20-50)	Features that activate in short contiguous regions (median length of the highest activating region is ≥ 20 and < 50) and have an overall mean activation coverage of less than 80%.
Motif (Long: 50-300)	Features that activate in short contiguous regions (median length of the highest activating region is ≥ 50 and < 300) and have an overall mean activation coverage of less than 80%.
Whole	Features that are active across nearly the entire sequence (overall mean activation coverage of greater than 80%.).
Other	Features that do not meet any of the above criteria are classified as "other."

A.6 TESTING SAE HYPERPARAMTERS

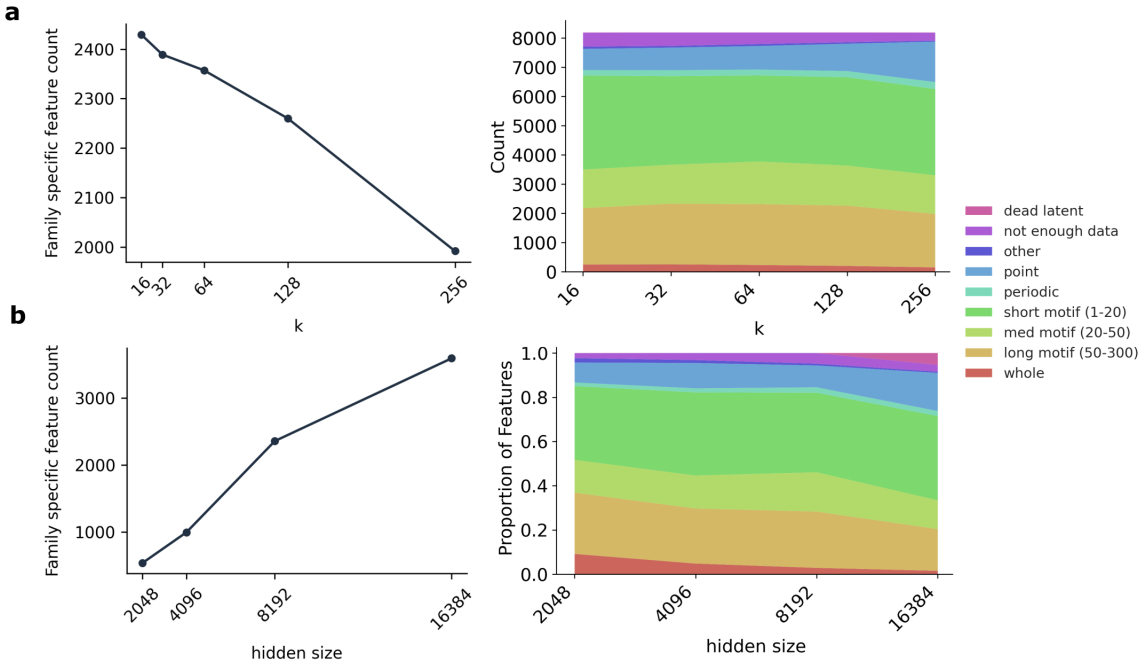

Figure 6: Testing SAE hyperparameters. **a**. k parameter sweep (other hyperparameters held constant, layer=24, hidden size=8096) showing the number of family specific features (left) and the features categorized by activation pattern (right). **b**. SAE hidden size hyperparameter sweep (other hyperparameters held constant, layer=24, k=64) showing the number of family specific features (left) and the features categorized by activation pattern (right).

A.7 ESM vs SAE probes

We find that linear probes on SAEs achieve performance similar to their ESM baselines across all layers (Figure 7). For secondary structure prediction, the SAE probe consistently outperforms the ESM probe, a result we discuss in the next section and attribute to the abundance of SAE latents that correspond to secondary structure.

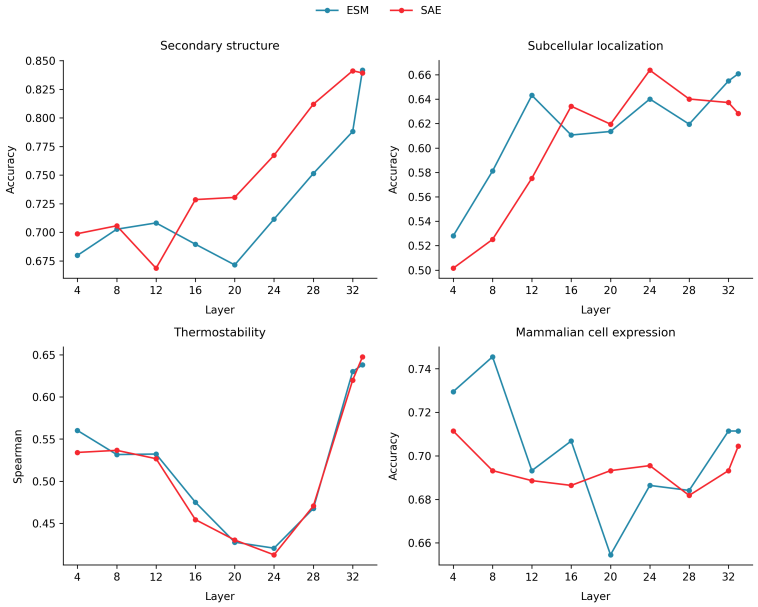

Figure 7: Linear probe performance on ESM vs. SAE embeddings across ESM layers. In all 4 downstream tasks, SAE performs competitively with ESM. For secondary structure prediction, SAE consistently outperforms ESM in most layers. All classifications tasks are measured using accuracy; thermostability is measured using Spearman's rank correlation.

A.8 THE CHO EXPRESSION TASK & ADDITIONAL VISUALIZATIONS

We were interested to see if our SAE probing method could be used to help interpret a novel task with relevance to drug development: binary human protein expression in CHO cells. The two latents most predictive of expression activate on terminus-specific long motifs (Figure 8b). The latent most predictive of failed expression recognizes an ATP binding site, suggesting that its interactions with ATP may disrupt essential metabolic processes in the host cell.

While we found many latents hard to interpret, this task demonstrates the potential of our method in a practical setting. If SAEs can help discover biological mechanisms that affect expression in CHO cells, then can guide us in designing protein therapeutics with better developability profiles.

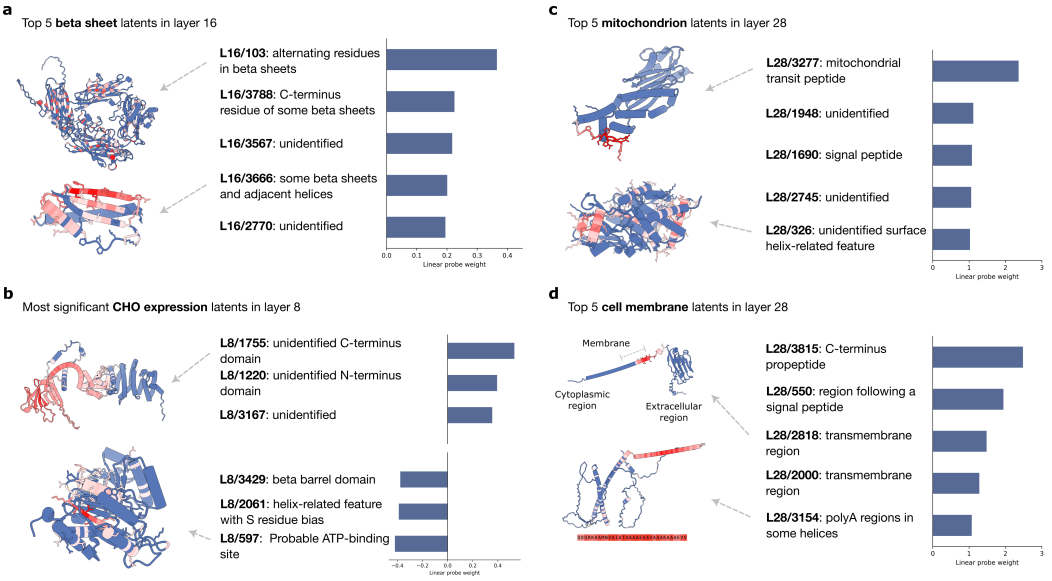

Figure 8: Linear probes identify predictive SAE latents for downstream tasks. **a**. The top 5 beta sheet latents from the secondary structure prediction task (SAE trained on ESM layer 16). **b**. The latents corresponding to the 3 most positive and 3 most negative coefficients from the linear probe on CHO cell line expression. The most positive latent activates on an unidentified motif; the most negative latent activates on an ATP binding site (SAE trained on ESM layer 8). **c-d**. Top 5 mitochondrion and cell membrane latents from the subcellular localization prediction task. The top nucleus latents identify transit peptides. The top cell membrane latents correlate with signal peptides and transmembrane regions (SAE trained on ESM layer 28).

A.9 DISTRIBUTION OF LINEAR PROBE COEFFICIENTS

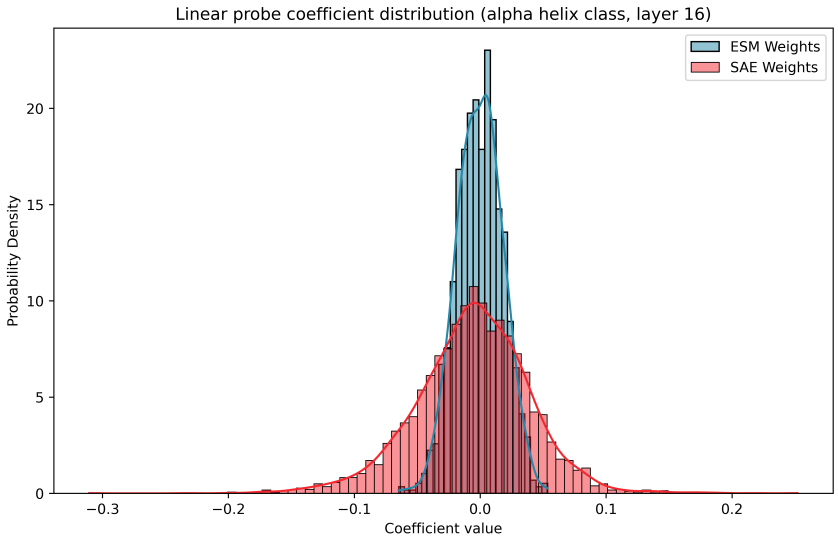

Figure 9: Example distribution of ESM vs. SAE linear probe coefficients. Shown here are coefficients for the alpha helix class from the secondary structure prediction task, using layer 16. The SAE coefficients have higher variance ($2.03\text{e}{-3}$) than the ESM coefficients ($3.24\text{e}{-4}$) because more latents are individually predictive of the label. The SAE coefficients also have a larger negative skew ($-2.05\text{e}{-1}$) compared to ESM ($-6.51\text{e}{-2}$). This can be explained by the SAE latents being more monosemantic: they sort more distinctly into helix vs. non-helix classes, and there are more non-helix examples than helix ones.