# From Mechanistic Interpretability to Mechanistic Biology: Training, Evaluating, and Interpreting Sparse Autoencoders on Protein Language Models

**Gökçe Uludoğan**
PhD Candidate

**LifeLU Reading Group | 27 February 2025**

**Etowah Adams, Liam Bai, Minji Lee,**
**Yiyang Yu, Mohammed AlQuraishi**

bioRxiv preprint
**doi.org/10.1101/2025.02.06.636901**

# Motivation & Background

**Proteins and Protein Language Models (pLMs)**

- pLMs learn from massive protein sequence databases
- They capture structure/function information in their internal representations

**Mechanistic interpretability**

- Understanding how models encode information
- Potential to reveal new biological insights

**Sparse Autoencoders (SAEs)**

- A technique previously used for large language models (natural language)
- Goal: Extract “human-interpretable” features from hidden activations

# Why Interpret pLMS?

**Widespread adoption of pLMs**

- ESM-2, others used for many downstream prediction

**Current black-box nature**

- We know pLMs can **succeed on** downstream tasks
- But which features or motifs drive predictions r**emain unclear**

**Potential for new discoveries**

- Unlike natural language, protein language is **cryptic**.
- Mechanistic interpretability in natural language has uncovered **concept neurons**
- Similar approaches in protein space might reveal previously **unknown biology**

# Overview of Approach

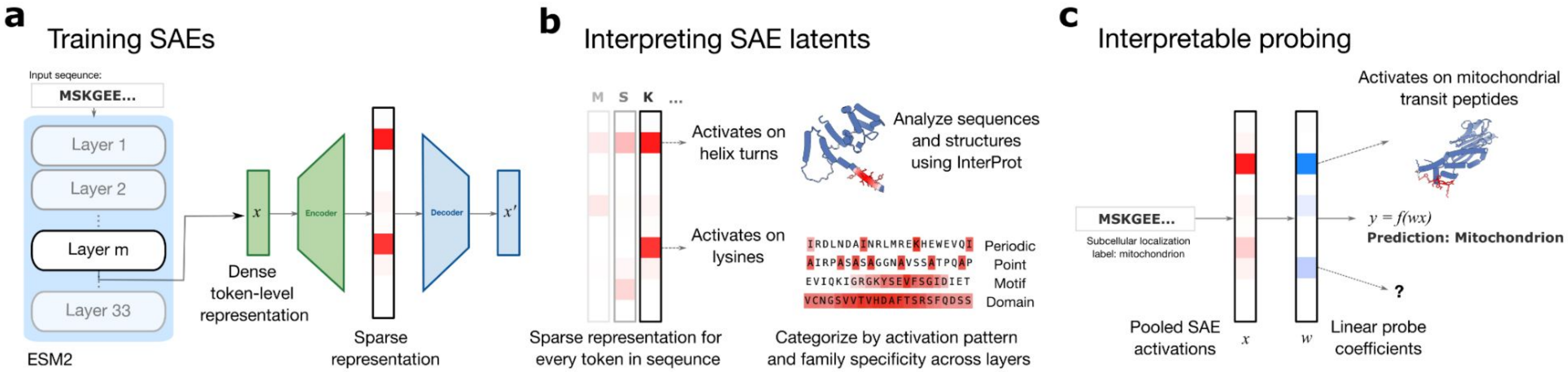


*Figure 1.* Overview of our paper. **a**. We train SAEs on the output of ESM intermediate layers. **b**. We interpret SAE latents using our latent visualizer InterProt and categorize features based on their activation pattern and family specificity. **c**. By interpreting the weights of linear models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

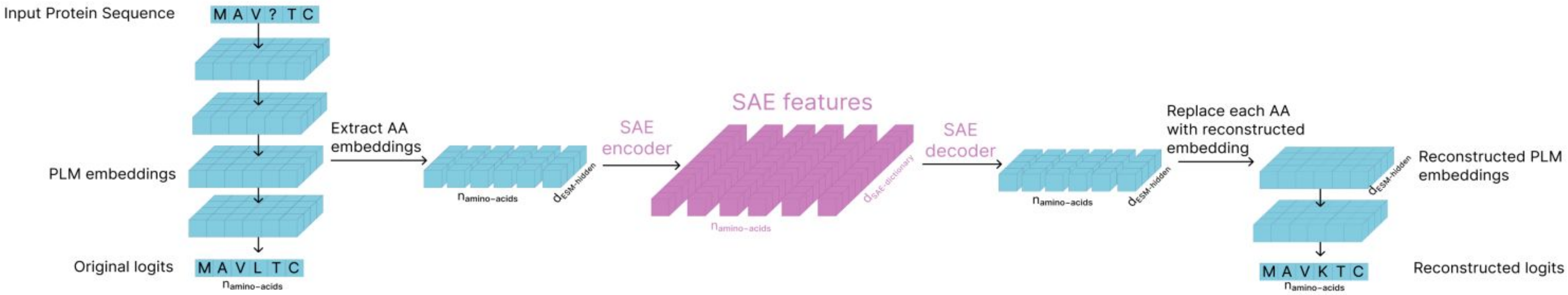
# Sparse Autoencoders

$$z = \mathrm{TopK}\left(W_{\mathrm{enc}}\left(x - b_{\mathrm{pre}}\right)\right)$$

$$\hat{x} = W_{\mathrm{dec}} z + b_{\mathrm{pre}}$$

## Architecture:

- Encoder: Linear projection of ESM's residual stream
- Activation: TopK (enforces sparsity by only keeping K largest latent activations)
- Decoder: Projects latent back to the original dimension



Simon, Elana, and James Zou. "InterPLM: Discovering Interpretable Features in Protein Language Models via Sparse Autoencoders." bioRxiv (2024): 2024-11.

# Sparse Autoencoders

**Architecture:**

- Encoder: Linear projection of ESM’s residual stream
- Activation: TopK (enforces sparsity by only keeping K largest latent activations)
- Decoder: Projects latent back to the original dimension

**Loss:** Simple mean squared error reconstruction

**Sparsity:** Why it matters

- Encourages more **disentangled** / interpretable features
- Each latent **fires rarely but strongly**

# SAE Training

**Model:** ESM-2 (650M parameters)

**Data:**

- 1M sequences from UniRef50, each ≤1022 residues
- Minimizes redundancy but spans diverse protein families

**Hyperparameters:**

- k in TopK (controls how many latents can be active)
- Varying hidden dimension (expansion factor)
- Trained across different ESM layers (early/mid/late

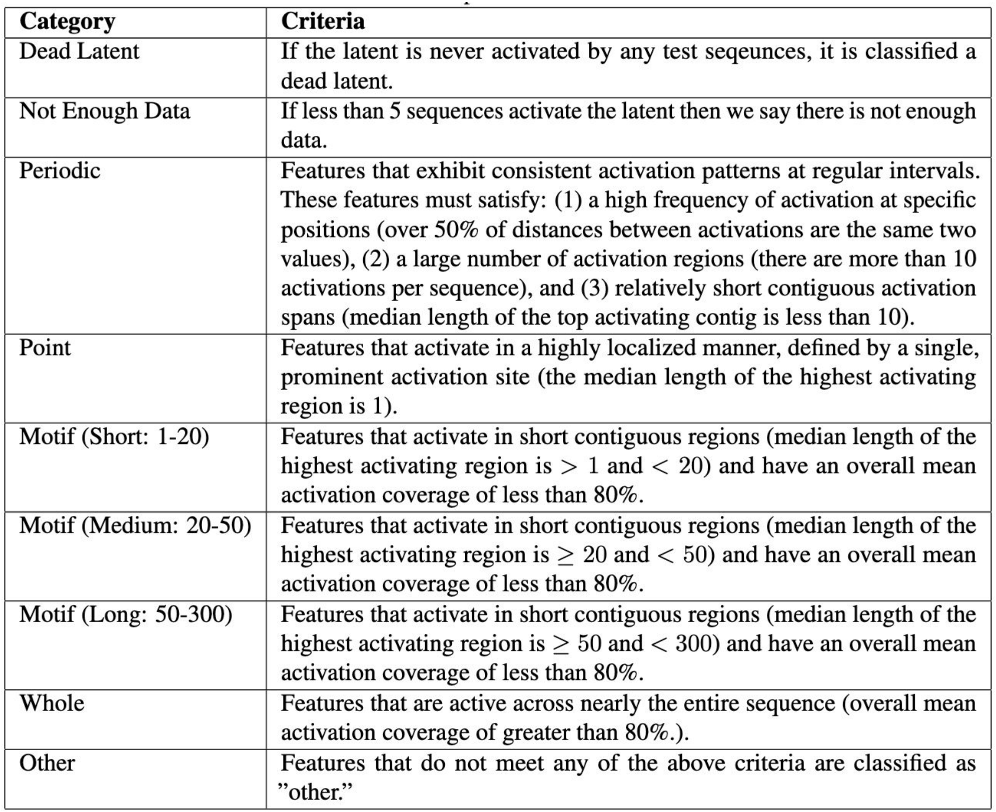
# SAE Latent Categorization

**Activation Pattern**

- Latents exhibit **distinct activation** patterns across their top sequences
- Categories
  - **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)
  - **domain** (activates over nearly the entire sequence).

**Family-specificity**

- **family-specific** if F1 score > 0.7 at a **certain activation** threshold.
- Swiss-Prot (clustered as 30% sequence identity) and categorized into **InterPro protein families**



| **Category** | **Criteria** |
|---|---|
| Dead Latent | If the latent is never activated by any test seqeunces, 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 $\geq 20$ 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 $\geq 50$ 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.” |

# Interpretable Probing on SAE latents

## Linear Probing

- Compare performance of linear probes on ESM embeddings vs. SAE embeddings

## Downstream tasks

- Secondary Structure (residue-level classification, alpha helix, beta strand, or other.)
- Subcellular Localization (protein-level classification)
- Thermostability (protein-level regression, melting temperature)
- Mammalian Cell Expression (protein-level  classification, expressed or not)

# SAEs uncover interpretable features

*Many discovered features correspond to recognizable biological and biochemical concepts*

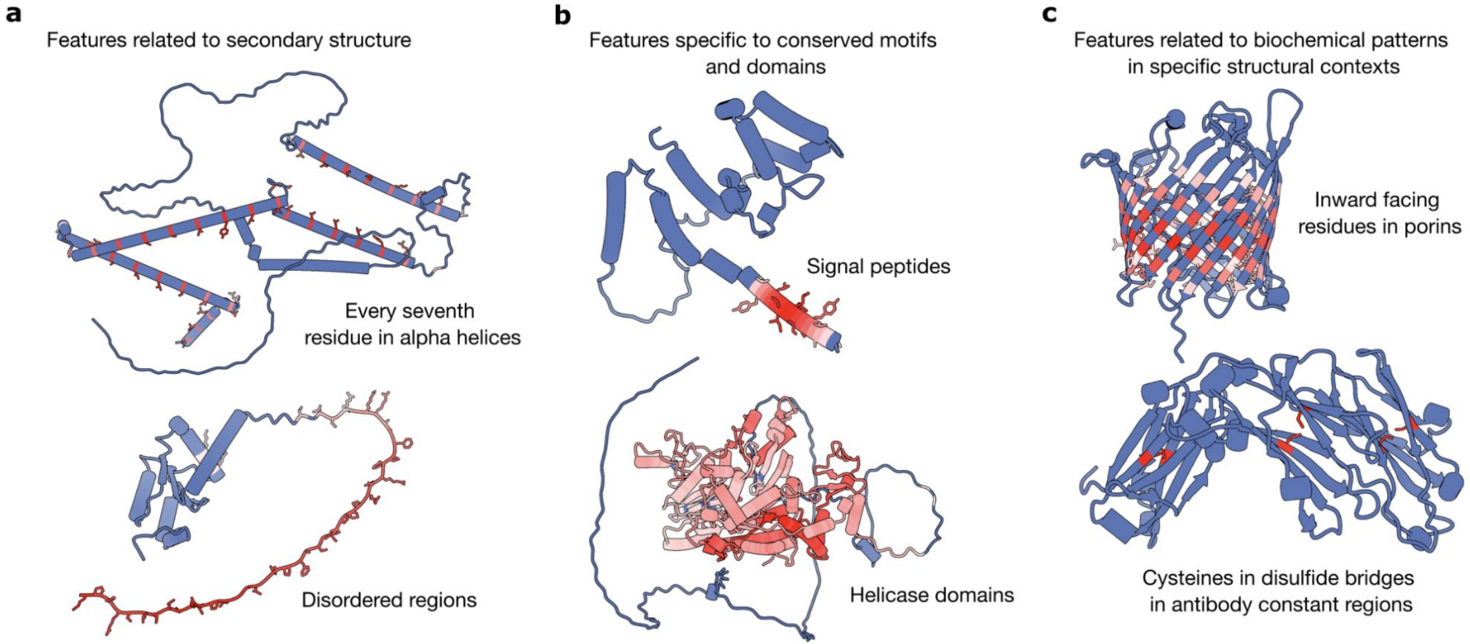
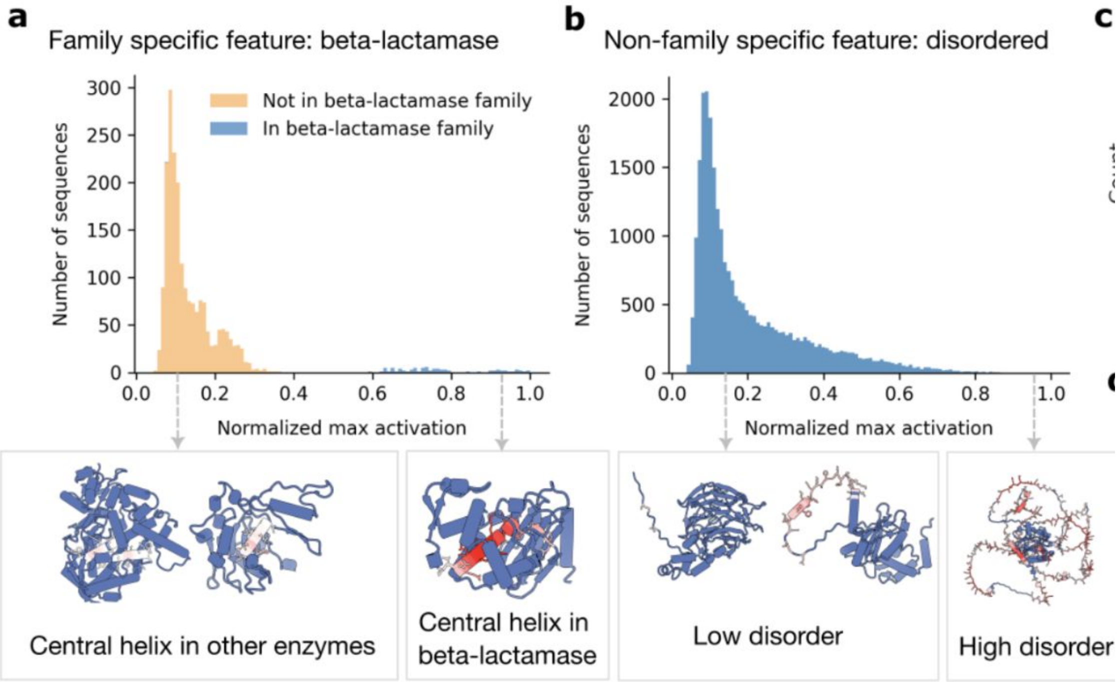

a
Features related to secondary structure
Every seventh residue in alpha helices
Disordered regions
b
Features specific to conserved motifs and domains
Signal peptides
Helicase domains
c
Features related to biochemical patterns in specific structural contexts
Inward facing residues in porins
Cysteines in disulfide bridges in antibody constant 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).

# SAE latents reveal a learned notion of protein families

*A large subset of latents activates weakly on most sequences but strongly on proteins belonging to a particular family.*



# Steering family-specific latents

*Steered sequence retains awareness of family-level conservation, as it changes most minimally at evolutionarily conserved positions*

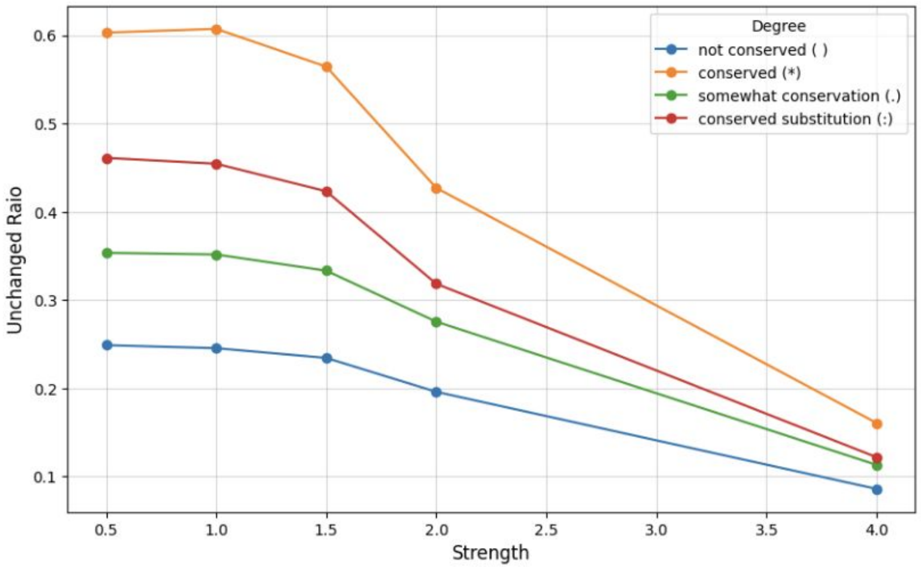


*Figure 6.* The fraction of residues that remains unchanged after steering, averaged for 168 family specific features that fraction of unchanged exceeds 20% for strength=0.5.

# Influence of SAE training choices on latent classification

***Effect of k***

*Lower k (higher sparsity) increases family-specific features, suggesting they are key for reconstruction.*

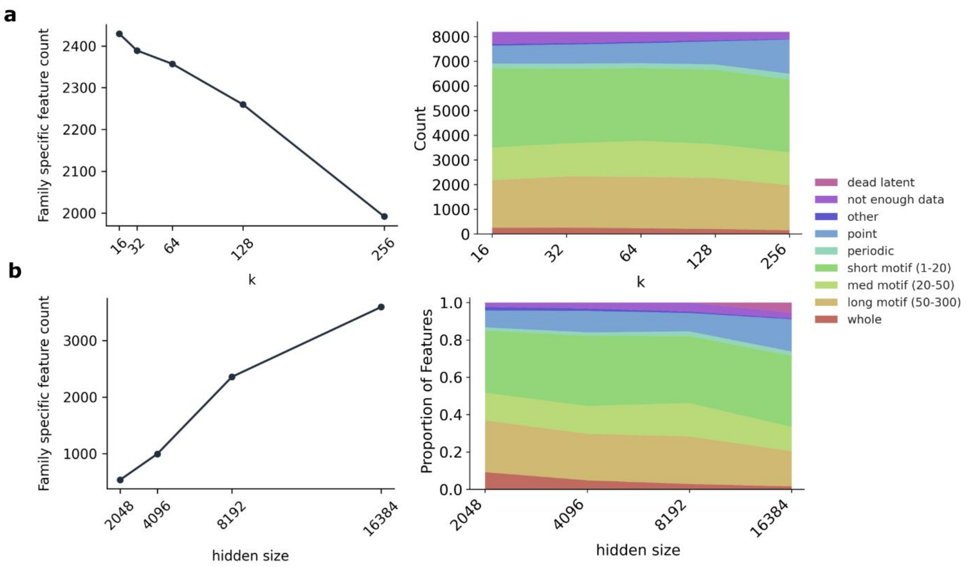


*Figure 7.* Testing SAE hyperparameters. **a**. *k* parameter sweep (other hyperparameters held constant, layer=24, hidden size=8096) showing the number of family specific features (left) and the features categorized by activation pattern (right). **b**. SAE hidden size hyperparameter sweep (other hyperparameters held constant, layer=24, k=64) showing the number of family specific features (left) and the features categorized by activation pattern (right).

# Influence of SAE training choices on latent classification

***Effect of expansion factor***

*A larger latent dimension does not change activation patterns (e.g. point, motif) but increases family-specific features.*

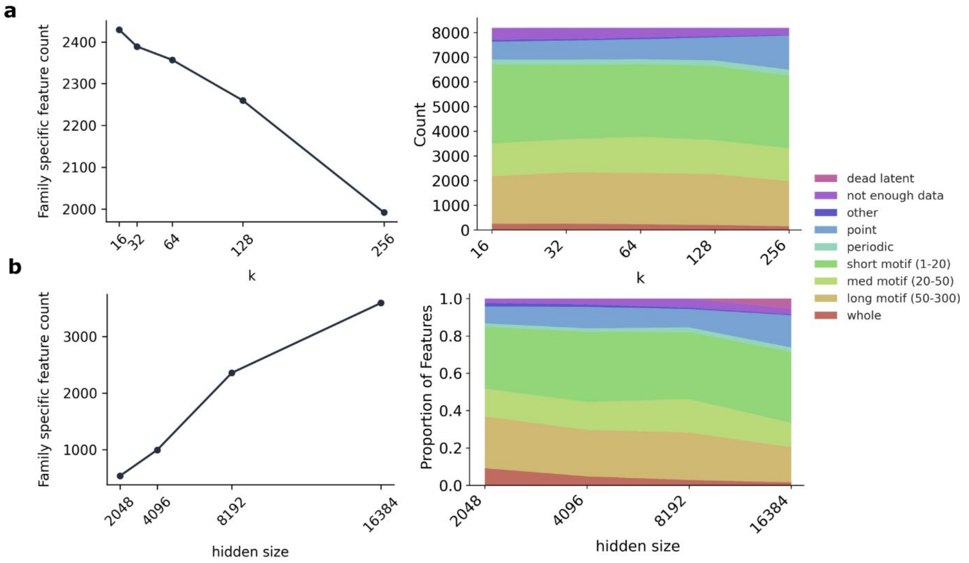


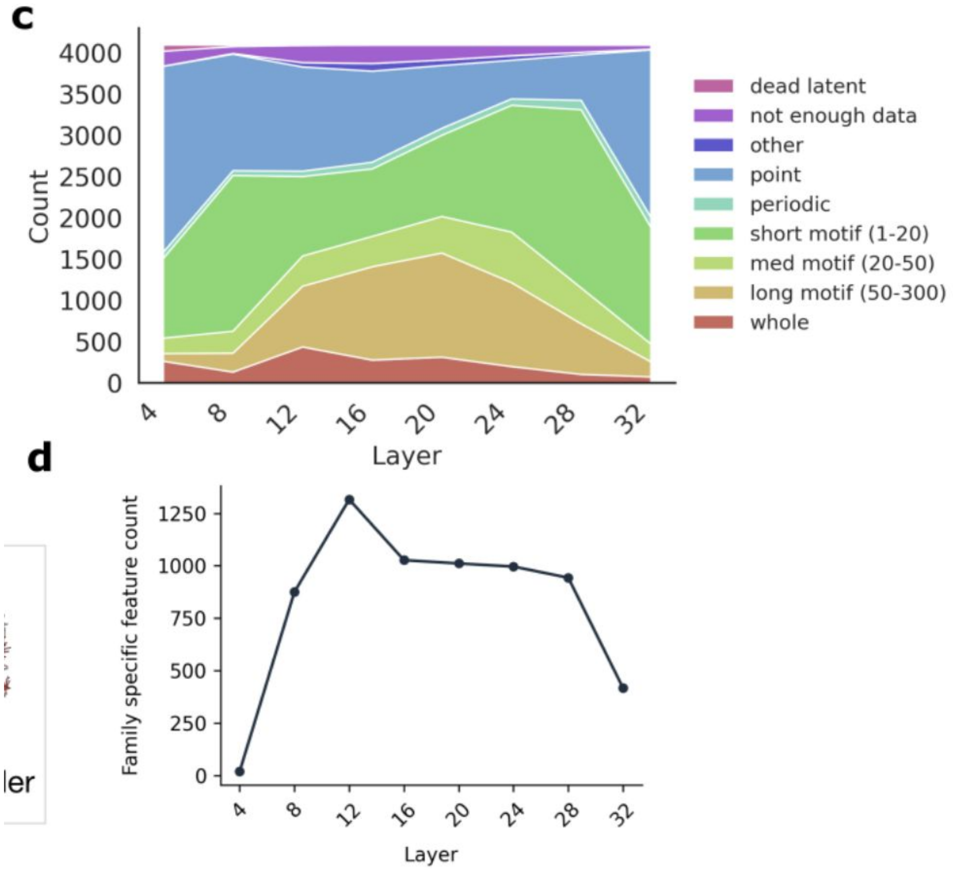
*Figure 7.* Testing SAE hyperparameters. **a**. $k$ parameter sweep (other hyperparameters held constant, layer=24, hidden size=8096) showing the number of family specific features (left) and the features categorized by activation pattern (right). **b**. SAE hidden size hyperparameter sweep (other hyperparameters held constant, layer=24, k=64) showing the number of family specific features (left) and the features categorized by activation pattern (right).

# Influence of SAE training choices on latent classification

***Comparison across layers***

*Motifs/domains dominate early layers and shorter activations appear later.*

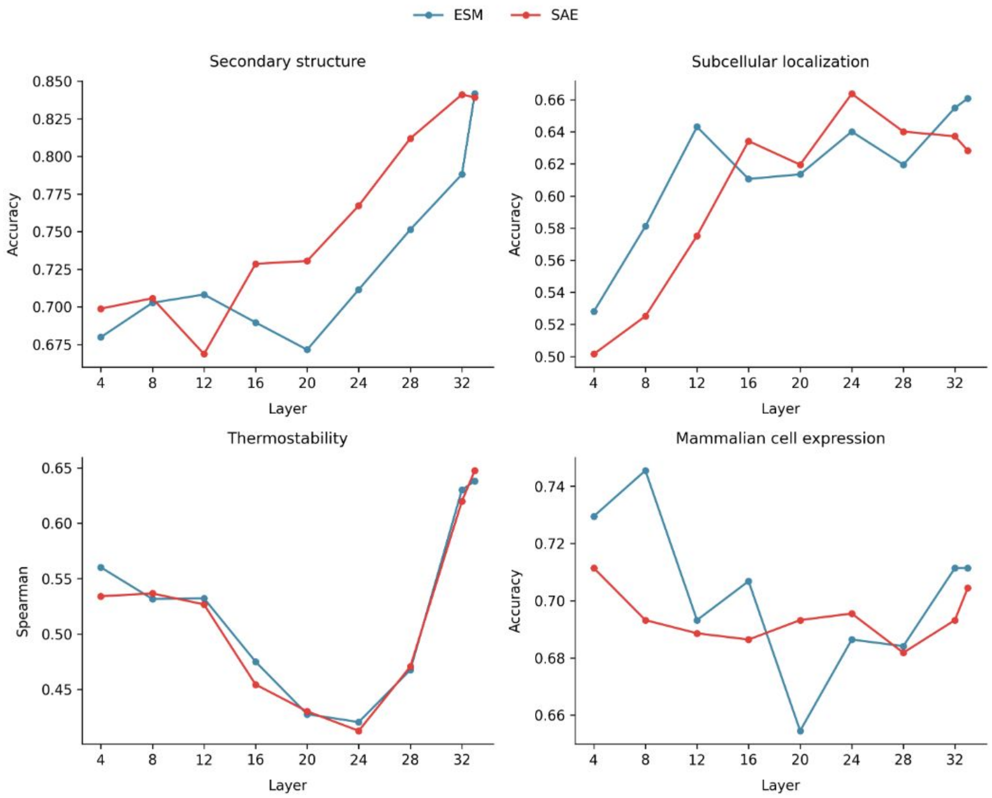
*Family-specific features peak in early-to-mid ESM layers.*



# SAE probes perform competitively with ESM probes

*Linear probes on SAEs achieve performance similar to their ESM baselines across all layers.*

*For secondary structure prediction, the SAE probe consistently outperforms the ESM probe.*

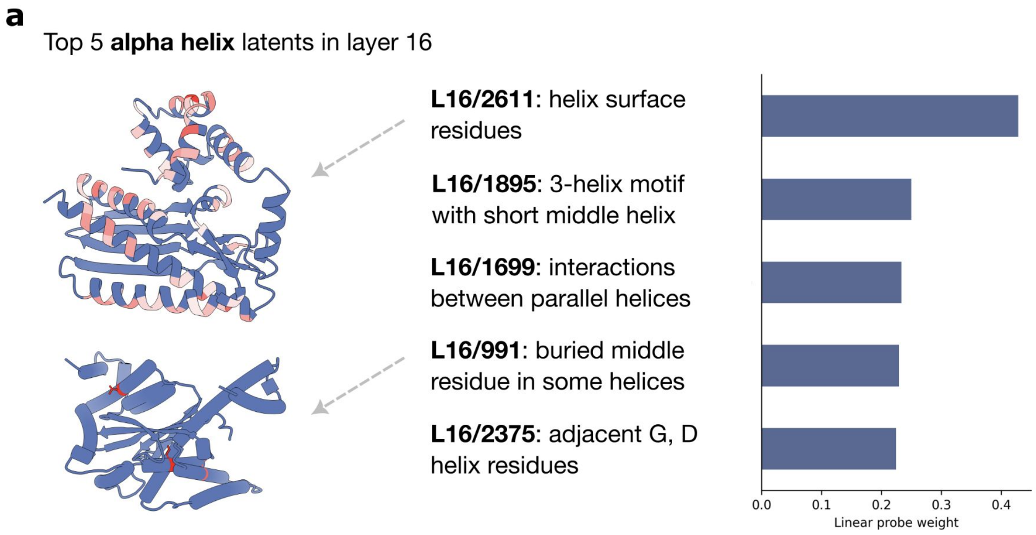


# SAE probes uncover interpretable latents corresponding to known mechanisms

## *Secondary Structure*

*The top 5 alpha helix latents from the secondary structure prediction task. These latents* ***activate mostly exclusively*** *on alpha helices.*

*They range from specific helix residues, to entire helices, to helix-helix interactions.*


Top 5 **alpha helix** latents in layer 16

*Similar trends for beta strands and other, where other tends to correlate with features that activate on disordered regions.*

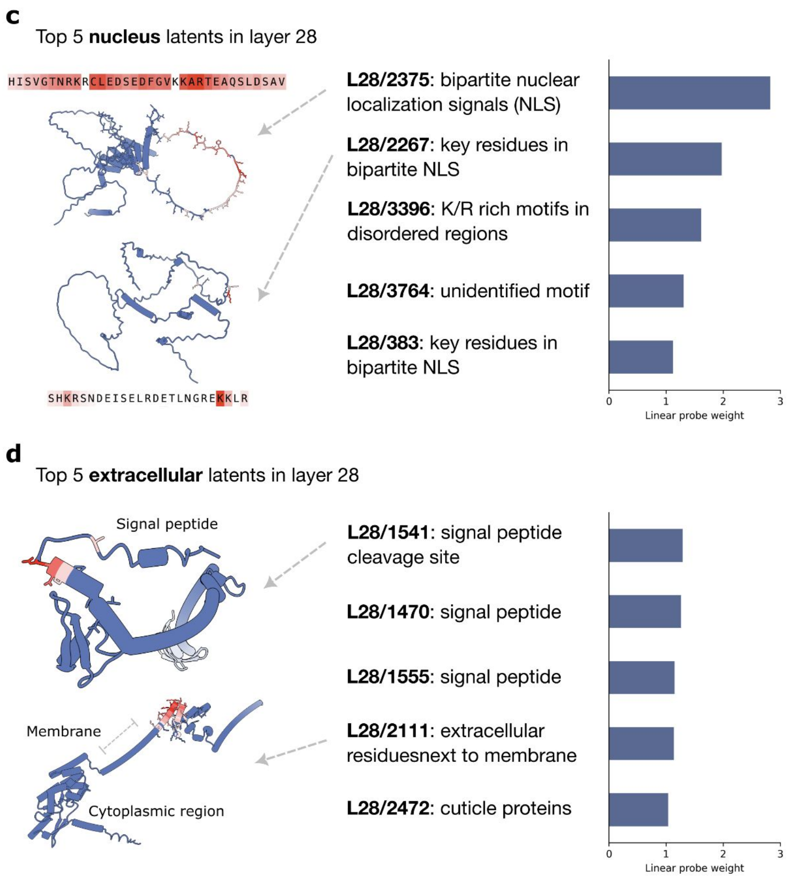
# SAE probes uncover interpretable latents corresponding to known mechanisms

***Subcellular Localization***

*Transport mechanisms rely on sequence motifs.*

***Nuclear Localization.*** *Highly weighted latents align with nuclear localization signals (NLS).* ***Example:*** *Layer 28 (L28/2375) activates on K/R-rich motifs and bipartite NLS (R/K(X)$_{10\text{–}12}$KRXK), recognizing both ends and the variable linker.*

***Extracellular Localization.*** *Top latents detect signal peptides (short N-terminal sequences triggering secretion).* ***Example:*** *L28/1541 activates on cleavage sites, while L28/1470 & L28/1555 activate across peptides.*



c Top 5 **nucleus** latents in layer 28

d Top 5 **extracellular** latents in layer 28

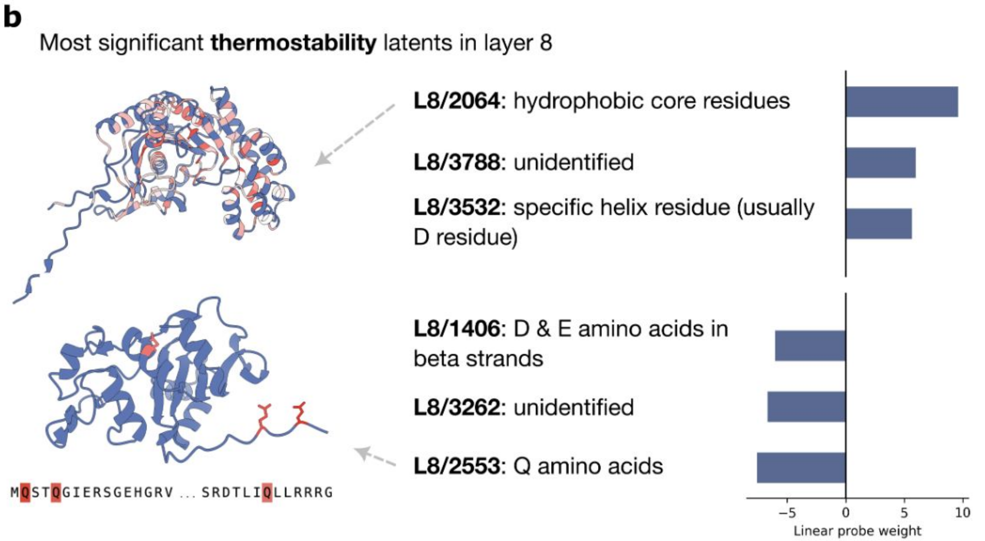
# SAE probes uncover interpretable latents corresponding to known mechanisms

***Thermostability***

***Does not scale** with larger models.*

*SAE probes reveal **reliance on amino acid composition** rather than complex features.*

*Early layers contain more relevant features: L8/2064 activates on **hydrophobic residues**, key for **stability.***

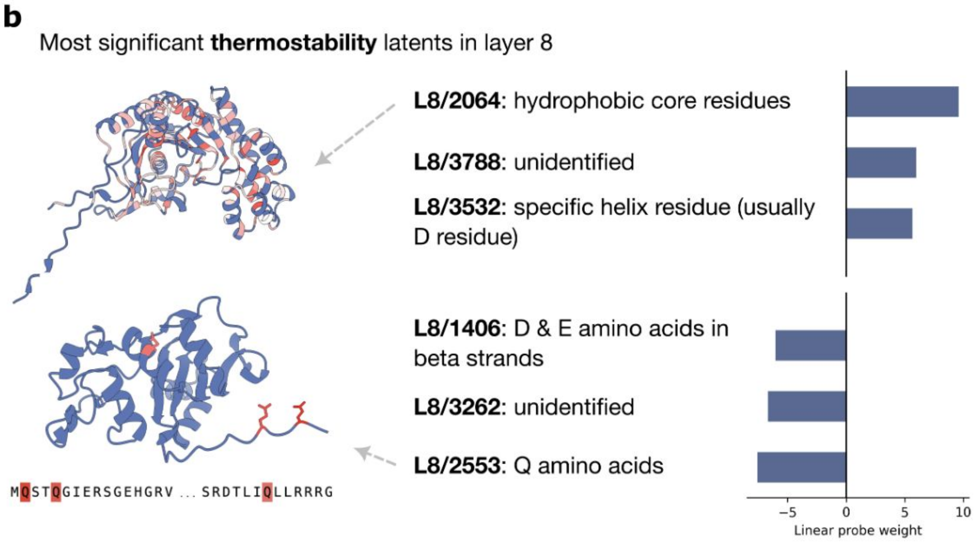


# SAE probes uncover interpretable latents corresponding to known mechanisms

## *Thermostability*

*SAE probes reveal **reliance on amino acid composition** rather than complex features.*

*Unlike the other tasks, **the middle layers perform worse** than early or late layers, detecting specific amino acids:*
***most positive latents:**Arginine (R), Tyrosine (Y), and Leucine (L);*
***most negative latents:** Glutamine (Q), Aspartic acid (D), and Threonine (T)*
***Absence of Glutamine** correlates with **increased thermostability.***



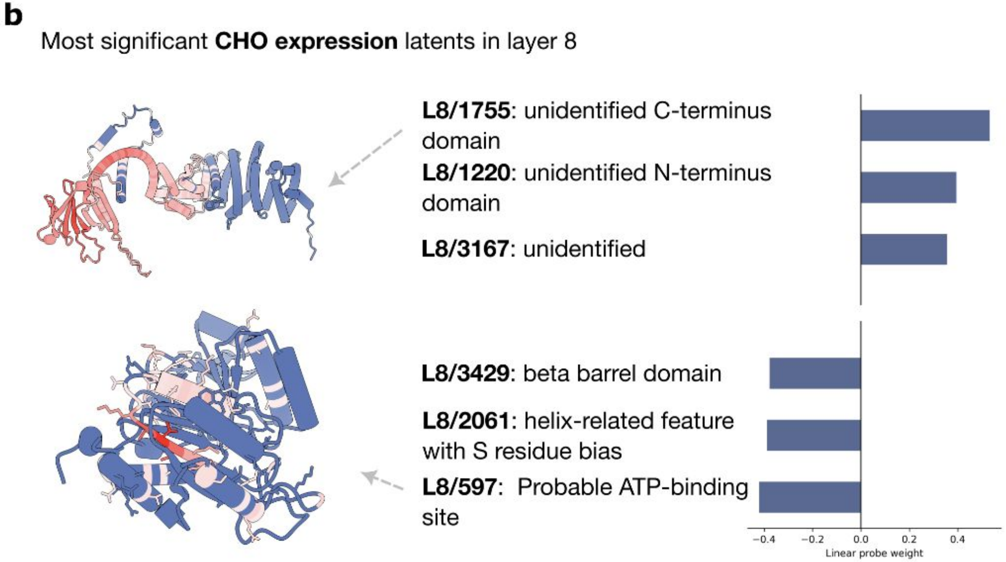
# SAE probes uncover interpretable latents corresponding to known mechanisms

***Mammalian cell expression***

***drug development:*** *binary human protein expression in CHO cells*

*The top expression latents activate on* ***terminus-specific motifs.***

*The latent for* ***failed expression*** *detects an* ***ATP binding site****, possibly* ***disrupting*** *host metabolism.*



**SAEs help us see inside pLMs' black box:**

- Expose interpretable features like motifs, structural elements, family signatures
- Probing shows which latents matter for tasks like secondary structure or localization

**Biological significance**

- Family-specific latents confirm pLMs store homology-like patterns
- SAEs can recover known motifs (e.g., NLS, signal peptides) and possibly undiscovered ones

**Limitations**

- Only tested ESM-2 (650M); what about bigger or smaller pLMs?
- Sparse autoencoder features can vary if data or hyperparameters change
- Not all latents are easily interpretable or might reflect unknown biology

# Future Directions

**Scaling up:**

- Larger pLMs, different architectures

**Task-specific training data:**

- Fine-tuning or re-training SAEs on certain families for deeper mechanistic insights

**New biological discoveries:**

- Hard-to-interpret but predictive latents might point to undiscovered motifs or structure–function relationships

**Steering for protein engineering:**

- Adjusting activation of certain latents to "edit" a protein while preserving other characteristics