nature methods

Article

https://doi.org/10.1038/s41592-025-02836-7

InterPLM: discovering interpretable features in protein language models via sparse autoencoders

Received: 26 November 2024

Accepted: 20 August 2025

Published online: 29 September 2025

Check for updates

Elana Simon & James Zou

Despite their success in protein modeling and design, the internal mechanisms of protein language models (PLMs) are poorly understood. Here we present a systematic framework to extract and analyze interpretable features from PLMs using sparse autoencoders. Training sparse autoencoders on ESM-2 embeddings, we identify thousands of interpretable features highlighting biological concepts including binding sites, structural motifs and functional domains. Individual neurons show considerably less conceptual alignment, suggesting PLMs store concepts in superposition. This superposition persists across model scales and larger PLMs capture more interpretable concepts. Beyond known annotations, ESM-2 learns coherent patterns across evolutionarily distinct protein families. To systematically analyze these numerous features, we developed an automated interpretation approach using large language models for feature description and validation. As practical applications, these features can accurately identify missing database annotations and enable targeted steering of sequence generation. Our results show PLM representations can be decomposed into interpretable components, demonstrating the feasibility and utility of mechanistically interpreting these models.

Language models of protein sequences have revolutionized our ability to predict protein structure and function[1,2]. These models achieve their impressive performance by learning rich representations from the vast diversity of naturally evolved proteins. Their success raises a fundamental question: what exactly do protein language models (PLMs) learn? Advances in interpretability research methods now enable us to investigate this question and understand at least some of the representations they have learned in service of this challenging task.

Gaining insights into the internal mechanisms of PLMs is crucial for model development and biological discovery[3]. Understanding how these models process information enables more informed decisions about model design, beyond merely optimizing performance metrics. By analyzing the features driving predictions, we can identify spurious correlations or biases, assess the generalizability of learned representations and potentially uncover novel biological insights. For instance, models might learn to predict certain protein properties based on subtle patterns or principles that have eluded human analysis, opening up new avenues for experimental investigation. Conversely, detecting biologically implausible features can guide improvements to the models' inductive biases and learning algorithms. This systematic analysis of prediction mechanisms offers opportunities to enhance model performance and reliability while extracting biological hypotheses from learned representations. Furthermore, PLM interpretation can illuminate how much these models learn genuine physical and chemical principles governing protein structure, as opposed to simply memorizing structural motifs. As PLMs continue to advance, systematic interpretation frameworks may uncover increasingly sophisticated biological insights, and it will be important to learn how to extract this knowledge.

In this work, we create a versatile framework using sparse autoencoders (SAEs) to interpret latent features learned by PLMs.

Stanford University, Stanford, CA, USA. e-mail: epsimon@stanford.edu; jamesz@stanford.edu

To demonstrate this framework, we analyze amino acid embeddings across all layers of ESM-2[1]. Using several scalable methods we developed for identifying and annotating protein SAE features, we discovered that the latent features of ESM-2 accurately capture many known biophysical properties—including catalytic active sites, zinc finger domains, mitochondrial targeting sequences, phosphorylated residues, disulfide bonds and disordered regions. By quantitatively comparing these identified features to known concept annotations, we establish new evaluation metrics for SAEs. We further use Claude-3.5 Sonnet (new)[4] to provide automated annotations of latent features and use PLM feature activation patterns to identify missing and potentially new protein annotations. We show how PLM features can be used to steer model outputs in interpretable ways. To facilitate exploration of these features, we provide an interactive visualization platform https://InterPLM.ai, along with code for training and analyzing features at https://github.com/ElanaPearl/interPLM.

Related works and background

PLM interpretability. PLMs are deep learning models (typically transformers) that perform self-supervised training on protein sequences, treating amino acids as tokens in a biological language to learn their underlying relationships and patterns[5]. These models consist of multiple transformer layers, each containing attention mechanisms and neural networks that progressively build up intermediate representations of the protein sequence. Prior work on interpreting PLMs has focused on analyzing these internal representations—both by probing the hidden states at different layers and by examining the patterns of attention between amino acids. Studies have demonstrated that hidden state representations can be probed to identify secondary structure states, attention maps can reveal protein contacts and binding pockets[6,7], and integrated gradients analysis of attention patterns can uncover active sites and transmembrane regions[8]. Recent evidence suggests that rather than learning fundamental protein physics, PLMs primarily learn and store coevolutionary patterns—coupled sequence patterns preserved through evolution[3]. This finding aligns with traditional approaches that explicitly modeled coevolutionary statistics[9] and with modern deep learning methods that leverage evolutionary relationships in training[10].

However, even if PLMs are memorizing evolutionarily conserved patterns in motifs, key questions remain about their internal mechanisms: How do they identify conserved motifs from individual sequences? What percentage of learned features actually focus on these conserved motif patterns? How do they leverage these memorized patterns for accurate sequence predictions? What additional computational strategies support these predictions? Answering these questions could both reveal valuable biological insights and guide future model development.

SAEs. In attempts to reverse-engineer neural networks, researchers often analyze individual neurons—the basic computational units that each output a single activation value in response to input. However, work in mechanistic interpretability has shown that these neurons do not map cleanly to individual concepts but instead exhibit superposition—where multiple unrelated concepts are encoded by the same neurons[11,12]. SAEs are a dictionary learning approach that addresses this by transforming each neuron's activation into a larger but sparse hidden layer[13–15].

At their core, SAEs learn a 'dictionary' of sparsely activated features that can reconstruct the original neuron activations. Each feature i is characterized by two components: its dictionary vector ($\mathbf{d}_i$) in the original embedding space (stored as rows in the decoder matrix) and its activation value (f_i) that determines its contribution. The reconstruction of an input activation vector $\mathbf{x}$ can be expressed as

$$\mathbf{x} \approx \mathbf{b} + \sum_{i=1}^{d_{\text{dict}}} f_i(\mathbf{x})\mathbf{d}_i,$$

where $\mathbf{b}$ is a bias term and d_{dict} is the size of the learned dictionary. This decomposition allows us to represent complex neuron activations as combinations of more interpretable features (see Supplementary Fig. 1 for more details).

SAE analysis has advanced our understanding of how language and vision models process information[14,16]. Neural network behavior can be understood through computational circuits—interconnected neurons that collectively perform specific functions. While traditional circuit analysis uncovers these functional components (such as edge detectors[12] or word-copying mechanisms[17]), using SAE features instead of raw neurons has improved the identification of circuits responsible for complex behaviors[18].

Researchers can characterize these features through multiple approaches: visual analysis[19], manual inspection[15] and large language model assistance[20]. Feature functionality can be verified through intervention studies, where adjusting feature activation values steers language model outputs toward specific behaviors, demonstrating their causal role[21].

One of the field's main challenges lies in evaluation. Although technical metrics such as how well autoencoders reconstruct their inputs are straightforward, assessing feature interpretability remains subjective. Recent work has made progress by using text-based games, where feature activations can be mapped to labeled game states in chess and Othello[22]. Protein studies offer similar potential through their structural and functional annotations, though interpreting biological features requires domain expertise that makes evaluation more challenging than in language or vision domains.

Our work with InterPLM demonstrates that while PLM neurons exhibit polysemantic behavior, SAE analysis reveals more interpretable features. We present a comprehensive framework of quantitative metrics and methods for visualizing, analyzing, describing, steering and learning from these PLM-derived SAE features.

Results

SAEs find interpretable concepts in PLMs

To understand the internal representations of PLMs, we trained SAEs on amino acid embeddings from ESM-2 as in Fig. 1a. We expanded each of the 6 layers of ESM-2-8M from 320 neurons to 10,420 latent features and 6 of ESM-2-650M's 33 layers from 1,280 neurons to 10,420 latent features. The quantity and biological domain-specificity of these features necessitated developing both qualitative visualization methods and quantitative evaluation approaches to enable interpretation. See Methods for training and analysis details.

By examining clusters of features based on their dictionary vectors, we can identify groups that detect related biological structures with distinct specializations, from subtle variations in kinase binding site recognition to hierarchical relationships between specific and general beta-barrel detectors. When evaluated with our quantitative measures of biological interpretability, we find substantially more known protein concepts in SAE features than ESM neurons. We also show that this feature interpretation can be automated with a large language model, and once a feature has an interpretation, this can be used to suggest missing and new protein annotations. We further show that targeted interventions on specific feature values enable controlled manipulation of the model's representations, producing interpretable changes that propagate to influence predictions even for nonintervened amino acids.

We focus primarily on ESM-2-8M for computational efficiency and because we can comprehensively study every layer of the model. Our analyses show that the last three layers of ESM-2-8M contain the most biological concepts and structural features. For consistency and clarity, our visualizations primarily focus on features from these final three layers of ESM-2-8M, denoted f/ < sae-feature-number >, visualizing layer 4 unless otherwise specified (via L < layer-numer >) as we performed the most in-depth analysis of this layer.

https://doi.org/10.1038/s41592-025-02836-7

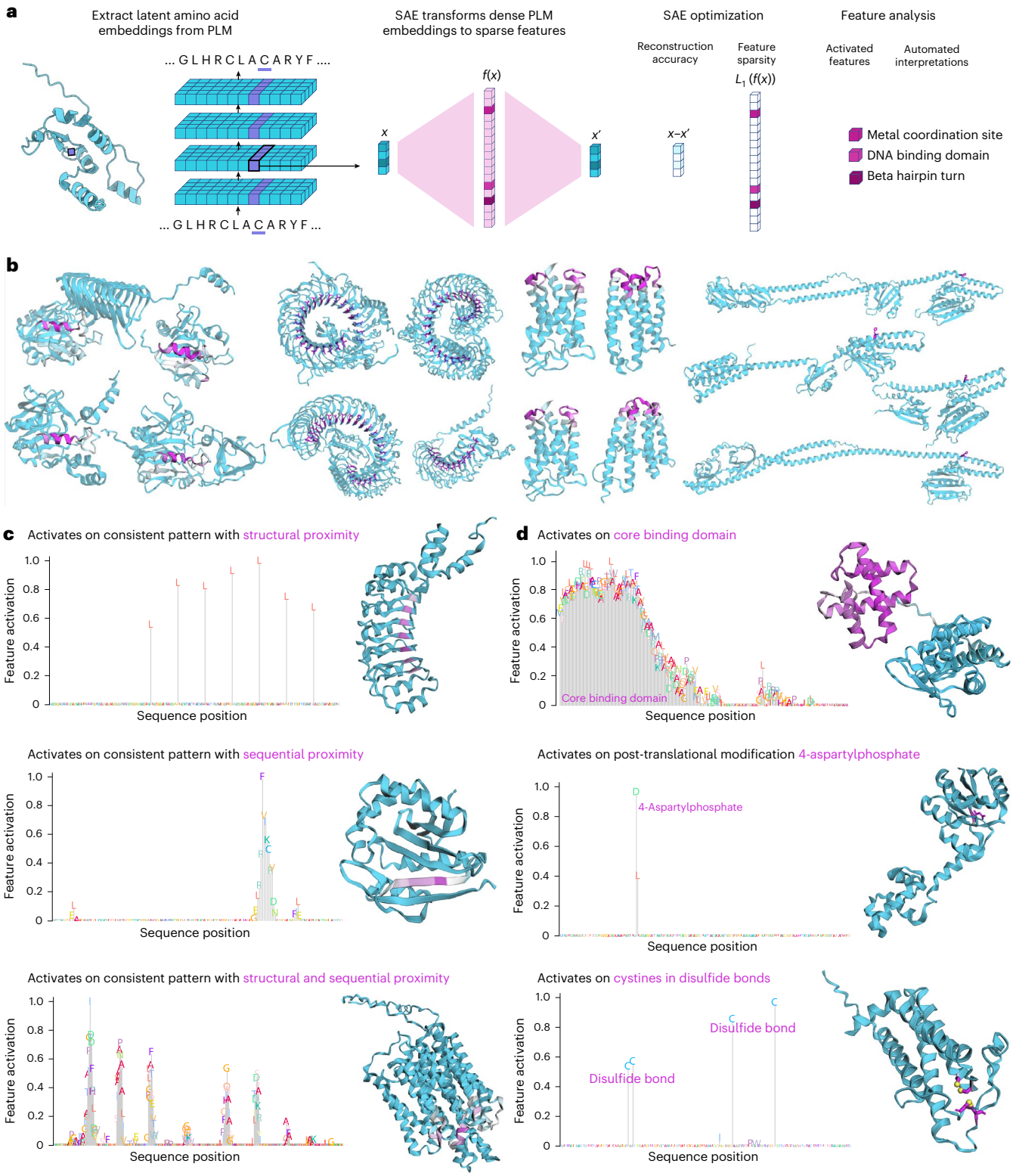


Fig. 1 | Overview of SAE methodology and representative SAE features revealed through automated activation pattern analysis. a, A pipeline for extracting interpretable features from a PLM with a SAE ($L_1 = L_1$ norm). **b**,**c**, Examples of features from ESM-2-8M exhibiting interpretable activation patterns, both structural and conceptual. Each feature is visualized using a protein where maximal activation occurs. Feature activation intensities are displayed both along the protein sequence (amino acid height indicates activation magnitude) and protein structure (darker pink indicates stronger activation). **b**, A structural visualization of four representative features containing multiple highly activating examples from ESM-2-8M layer 4 demonstrating the similarities in protein selection and positions of highly activating residues. Feature IDs and representative UniProt IDs from left to right: f/1854 (Q9NR50), f/10230 (F4HTV4), f/8144 (Q57N57), f/8128 (Q92H24) mapped onto example proteins. **c**, The features selected to demonstrate activation patterns in structurally proximate amino acids, sequentially proximate amino acids or both. **d**, The features selected based on strong associations between activation patterns and known Swiss-Prot biological concept annotations. Feature identifiers and UniProt IDs for visualized protieins (top to bottom) for **c**: L4-f/3147 (Q9QZH7), L4-f/10091 (P13857), L4-f/67 (E1ACQ6); for **d**: L4-f/8098 (Q1BAI5), L5-f/7125 (Q7A0I0), L5-f/746 (Q0I309).

https://doi.org/10.1038/s41592-025-02836-7

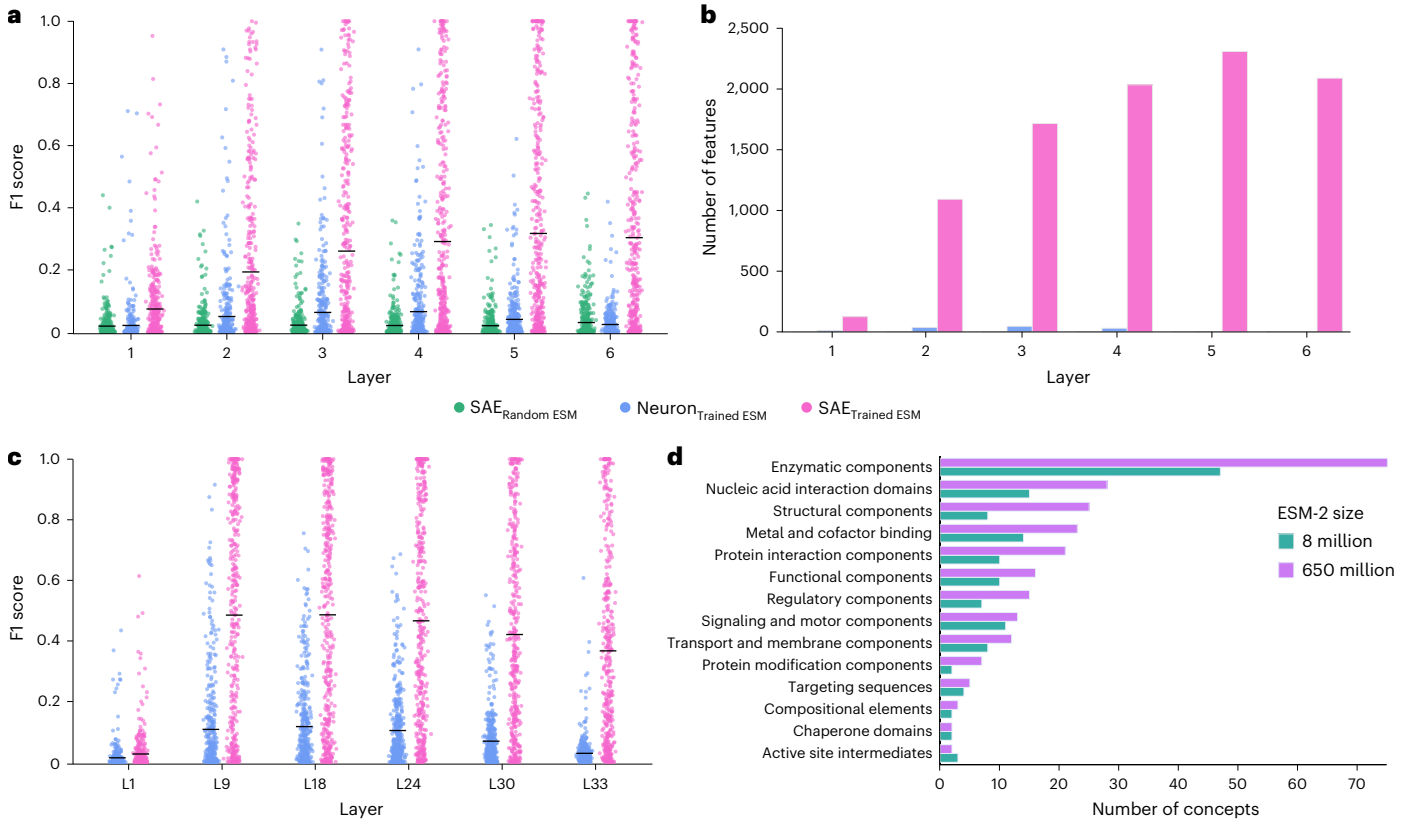


Fig. 2 | SAE features have stronger associations with Swiss-Prot concepts than ESM neurons across model scales. a, F1 score distributions for Swiss-Prot concept associations across ESM-2-8M layers. SAE features from trained models (pink) show higher F1 scores than original neurons (blue) or SAE features from shuffled weight controls (green) on heldout validation data. **b**, The number of features achieving F1 >0.5 with any Swiss-Prot concept across ESM-2-8M layers. The feature counts increase from layer 1 to layer 5, peaking at 2,309 features in layer 5. **c**, F1 score distributions comparing neurons (blue) and SAE features (pink) across six layers of ESM-2-650M. SAE features consistently achieve higher maximum F1 scores (0.95–1.0) compared with neurons (0.6–0.7) across all layers tested. **d**, The number of Swiss-Prot concept categories captured by ESM-2-8M (teal) versus ESM-2-650M (purple) models. ESM-2-650M captures more concepts across all 14 categories, with the largest differences in enzymatic components, nucleic acid interaction domains and structural components.

Protein feature exploration through InterPLM reveals collectively diverse yet internally consistent feature activation patterns

Through our interactive dashboard https://InterPLM.ai you can both identify interesting features within any layer of a PLM based on quantitative patterns in feature activation and also visually explore each feature's activation across proteins, revealing that the features are all both very diverse from one another yet often internally consistent in what they are identifying.

For individual features, we can understand their function by examining the proteins that activate them to varying degrees (0–1 activation scale). In Fig. 1b, we visualize four different features from ESM-2-8M layer 4 and the proteins that most highly activate them, with amino acids colored by activation strength. Each feature demonstrates remarkable consistency in activation patterns both across proteins and within specific protein regions. While some features identify homogeneous collections of proteins, such as the bundle of transmembrane helices in f/8144, others activate on proteins that share specific conserved domains within otherwise structurally diverse proteins, such as the alpha helix in bacterial enzymes in f/1854.

As shown in Fig. 3a, our quantitative analysis across all features in a given layer reveals a diverse spectrum of feature behaviors across multiple dimensions. First we note that features vary widely in both proteome-wide prevalence and protein-specific activation strength, with some features activating across many proteins while others show high specificity to particular protein families. Across layers, we observe distinct patterns in how these activation properties vary, with deeper layers containing features with higher proteome-wide prevalence (Supplementary Fig. 2).

Further analysis of structural versus sequential activation patterns revealed three distinct modes of amino acid recognition: structural patterns (coordinated activation between spatially proximal residues), sequential patterns (activation across nearby residues in the primary sequence), and combinations of both mechanisms. Figure 3b demonstrates how we can quantitatively compare how strongly residues are clustered in sequential versus structural patterns, with the structural effect size plotted against sequential effect size. Features above the diagonal line exhibit stronger structural than sequential character, while those below show the opposite tendency.

Figure 1c shows typical examples of features selected from ESM-2-8M L5 based on high structural, sequential or combined proximity patterns. Features with predominantly structural but minimal sequential character typically identify repeat patterns across spatially proximate but sequentially distant residues. Conversely, features with strong sequential but minimal structural character often detect small motifs or beta strands. Features exhibiting both structural and sequential properties can identify more complex shapes, such as turns in separate alpha helices that are bundled together, leading to 3D proximity in the turns (Supplementary Table 2).

The sequential visualization of feature activations (Fig. 1c,d) reveals that some features exhibit bimodal activation patterns, being either fully on or off for specific amino acids (for example, the leucine-rich repeat in L4-f/3147 and disulfide bonds in L5-f/746). Others display broader coverage with smooth activation gradients around the

highest-activated regions (for example, the beta strand in L4-f/10091 and core binding domain in L4-f/8098), resulting in varied activation levels based on position even within activated regions.

As shown in Supplementary Fig. 4, these features demonstrate robustness to benign mutations, with highest-activated features per protein varying their activation level by an average of 6% across mutations that do not impact protein stability, compared with 8.6% variation in mutations that do impact stability. However, this differential sensitivity between pathogenic and benign mutations depends on how effectively the underlying PLM can predict mutation pathogenicity.

Features can be further characterized based on their associations with known biological annotations and clustering with other features, which we explore in greater depth in 'SAE features capture more biological annotations than neurons' and 'Features form clusters based on shared functional and structural roles' sections.

SAE features capture more biological annotations than neurons

When comparing binarized feature activation values to biological concept annotations extracted from Swiss-Prot (dataset and methodological details in Methods), we observed that our learned features are much more specific than many concept annotations. The 433 Swiss-Prot concepts we evaluated span structural patterns, biophysical properties, binding sites and sequential motifs. Whereas some concepts are specific to individual amino acids, others can cover broad domains that include many different physical regions, resulting in standard classification metrics overly penalizing features that activate on individual subsets of larger domains. For example, f/7653 activates on two conserved positions in tyrosine recombinase domains. While this feature has perfect precision on the 45 tyrosine recombinases in our test set, it has a recall of 0.011 because tyrosine recombinase domains are hundreds of amino acids long, so it 'misses' most examples. To address this, we calculate precision of feature prediction per amino acid but calculate recall per domain. In the case of the tyrosine recombinase feature, because it correctly identifies two amino acids in every domain tested, it now has a domain-adjusted recall of 1.0. When we use these adjustments, we find thousands of feature–concept matches per layer, with most features activating on <20% of a given concept domain, with full-domain features (>80%) only emerging at layer 4 (Supplementary Fig. 3c).

While individual neurons in ESM2-8M showed at most 46 clear concept associations per layer, SAE analysis uncovered far more conceptual structure, revealing up to 2,309 features with strong concept alignment per layer (Fig. 2). This difference is explained by both the SAE features' ability to capture more specific qualities within each concept category and their detection of a broader range of concepts overall, expanding from 15 distinct concepts identified by neurons to 143 identified by SAE features (identified concepts listed in Supplementary Table 4).

The dramatic difference in concept detection between SAEs and neurons persisted across model layers and disappeared in control experiments with randomized model weights. Notably, although the randomized models extract zero features corresponding to biological Swiss-Prot concepts (Fig. 2), they have hundreds of features strongly associated with individual amino acid types (Supplementary Fig. 5b).

We investigated how model scale influences concept representation by analyzing ESM-2-650M, a model with 650M parameters and a fourfold larger embedding dimension (1,280 versus 320) compared with ESM-2-8M. SAE features from ESM-2-650M consistently achieve higher concept detection precision compared with individual neurons (Fig. 2c), indicating that neurons maintain polysemantic representations despite the increased embedding capacity. When comparing six layers from ESM-2-650M against all layers of ESM-2-8M, the larger model identifies over 1.7× more concepts (427 versus 143) (Fig. 2d). This substantial increase in concept detection from just a subset of ESM-2-650M layers suggests that increased model capacity enables richer concept representations.

The enhanced concept capture spans all functional and structural categories, with notable expansions in enzymatic activities, nucleic acid interactions, structural elements and molecular binding features. Among the concepts uniquely captured by ESM-2-650M are enzyme classes such as protein palmitoyl transferase (L33, f/9015) and methionine sulfoxide reductase (L9, f/1312); DNA/RNA-binding elements such as the RNA recognition motif (L18, f/10233) and topoisomerase-primase domain (L24, f/8542); and structural features including immunoglobulin framework positions (L9, f/2303; L18, f/7511) and coiled-coil domains (L18, f/6676; L30, f/8603).

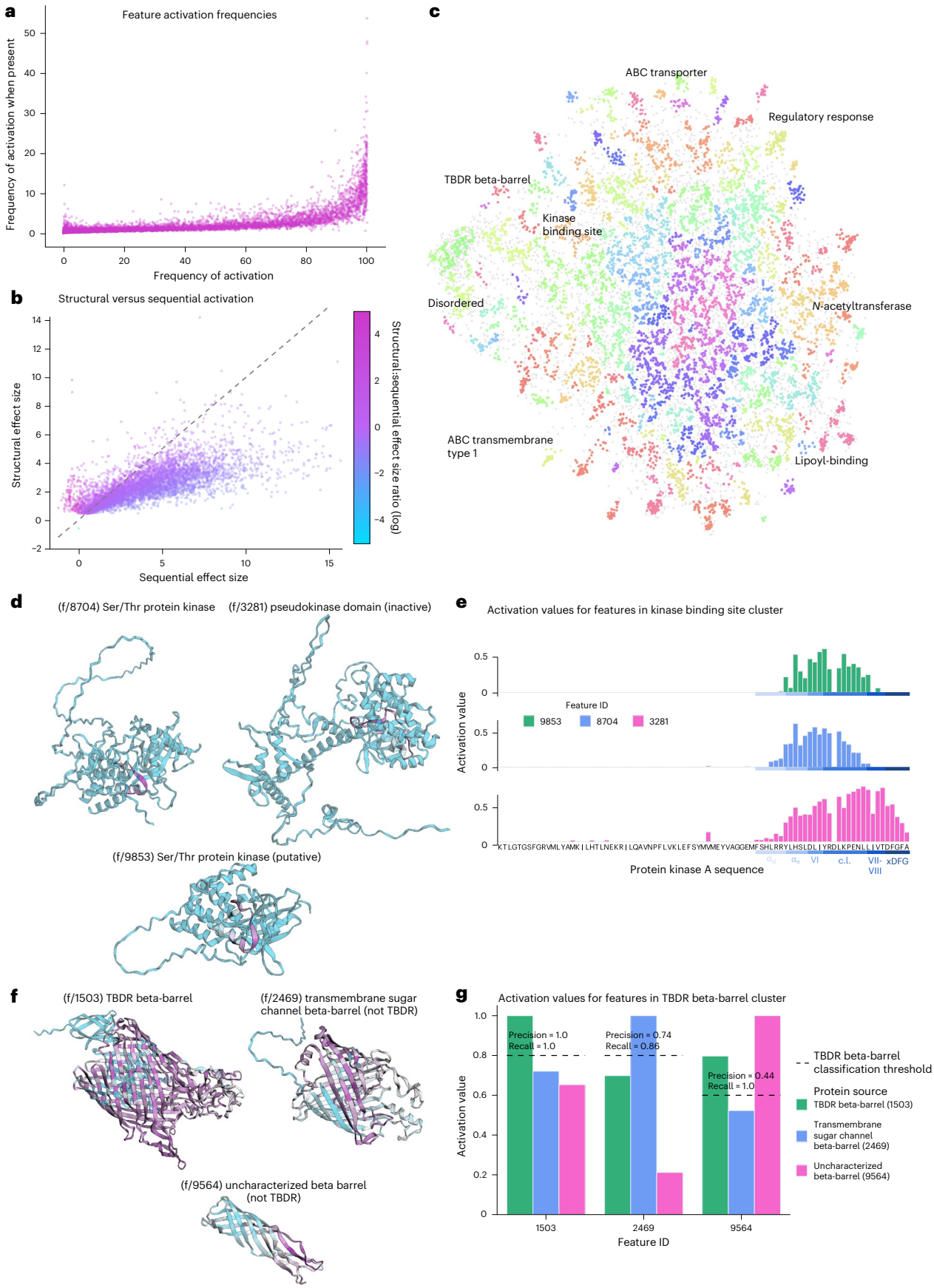
Features form clusters based on shared functional and structural roles

Clustering features by their dictionary vectors exposes groups with coherent biological functions and distinct specializations. Within a kinase-associated feature cluster, three features exhibited activation on binding site regions with subtle variations in their spatial preferences (Fig. 3). One feature activated specifically on the conserved alpha helix and beta sheet (α_e, VI) preceding the catalytic loop, while others concentrated on distinct regions near the catalytic loop, as evidenced by their maximally activated examples in both structural and sequence representations (Fig. 3d,e). Though their peak activation positions varied, all features maintained high activation levels (>0.8) on their respective maximum examples across the cluster, suggesting they identify similar kinase subtypes.

Analysis of a beta-barrel-associated cluster revealed a distinct generalization pattern. While all three features in the cluster were labeled as TonB-dependent receptor (TBDR) beta-barrels, only one feature exhibited specificity for TBDR beta-barrels, whereas the other two identified beta-barrel structures more broadly, including TBDRs, as demonstrated by their maximally activating examples (Fig. 3f,g). All three features exhibit F1 associations with TBDR beta-barrels, yet they differ markedly in their specificity. f/1503 shows exceptional specificity (F1 = 0.998), functioning as a true TBDR detector. The other features, though capable of identifying TBDR structures (high recall),

Fig. 3 | SAE features reveal diverse activation patterns and functional clustering with varying specificity levels. a, Feature activation frequency distribution showing the relationship between proteome-wide prevalence and protein-specific activation strength as measured by the percentage of proteins that activate a feature (*x* axis) and the average percentage of residues that are activated in a protein when any are activated (*y* axis), revealing both ubiquitous and selective features. **b**, A comparison of structural versus sequential clustering for features with significant structural organization (Bonferroni-corrected structural $P < 0.05$). Each point represents one SAE feature, with axes showing Cohen's *d* effect sizes for sequential (±2 positions) and structural (<6 Å) clustering relative to shuffled controls. The point colors indicate the structural-to-sequential effect size ratio, with pinker shades representing features that cluster more strongly in 3D space than in sequence. **c**, A UMAP embedding of SAE features clustered based on their dictionary values, with several clusters of features associated with a shared Swiss-Prot concept annotated. The coloring is based on HDBSCAN cluster groups. **d**, The structures of maximally activating examples from three features in the kinase binding site cluster showing slightly varied preference in kinase types and positions near the catalytic loop. **e**, The activation value comparison of three kinase binding site cluster features (9853, 8704 and 3281) across protein kinase A sequence, demonstrating how different features within the same functional cluster activate at distinct sequence positions. **f**, Examples from the TBDR beta-barrel cluster showing three features with varying levels of specificity and their maximum activating protein structures. **g**, An activation value comparison showing how TBDR beta-barrel features (1503, 2469 and 9564) vary in their TBDR specificity, with feature classifications ranging from broad beta-barrel to specific transmembrane sugar channel subtypes.

https://doi.org/10.1038/s41592-025-02836-7



https://doi.org/10.1038/s41592-025-02836-7

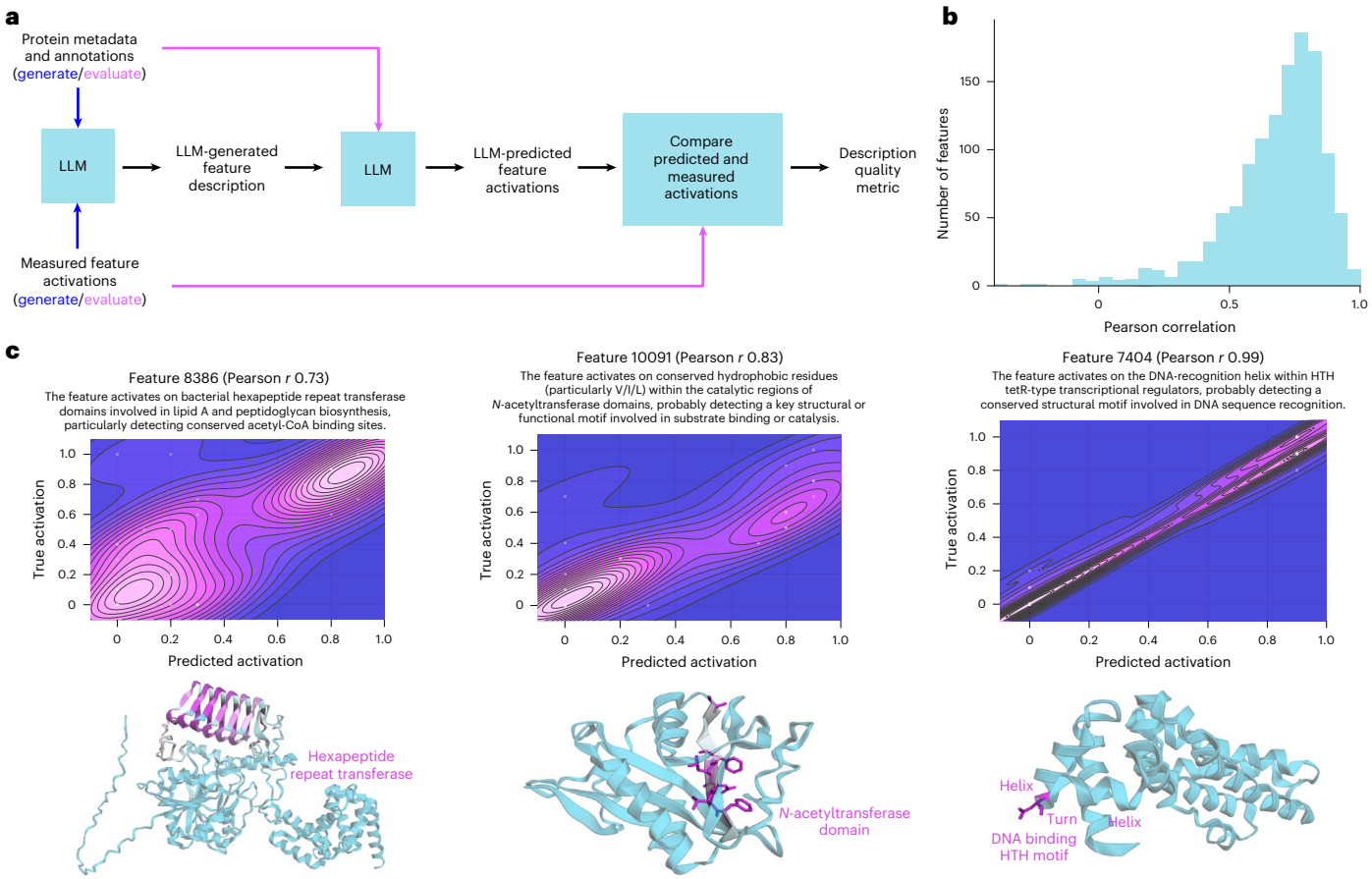


Fig. 4 | Language models can generate automatic feature descriptions for SAE features. a, A workflow for generating and validating descriptions with Claude-3.5 Sonnet (new). **b**, Comparing measured maximum activation values in proteins to predicted maximum activation values via Pearson *r* correlation across 1,240 features. **c**, Example features descriptions with a range of performance quality visualized via generated feature descriptions and maximally activated proteins of each feature. The predicted activation quality is visualized via kernel density estimation. The text is Claude's description summary of each feature. Elements of description are present in maximum examples annotated next to structures. Examples from ESM-2-8M layer 4, from left to right: the features and representative UniProt IDs are 8386 (P47823), 10091 (P13857) and 7404 (O34643). HTH, helix-turn-helix.

also recognize various other beta-barrel proteins (lower precision), resulting in lower but still meaningful F1 scores (0.793, 0.611). This clustering demonstrates how the model's embedding space captures the natural hierarchy of beta-barrel structures, where specialized TBDR detectors and general beta-barrel features maintain similar representations despite operating at different levels of structural specificity.

Large language models can generate meaningful feature descriptions

To extend beyond Swiss-Prot concepts, which label less than 20% of features across layers (Supplementary Fig. 5a), we developed an automated pipeline using Claude to generate feature descriptions. By providing Claude-3.5 Sonnet (new) with the Swiss-Prot concept information including text information not applicable for classification, along with examples of 40 proteins with varying levels of maximum feature activation, we generate descriptions of what protein and amino acid characteristics activate the feature at different levels. As validation, the model-generated descriptions and Swiss-Prot metadata are used to predict feature activation levels on separate proteins and showed high correlation with actual feature activation (median Pearson *r* correlation = 0.72) across diverse proteins. As shown in Fig. 4, the descriptions accurately match specific highlighted protein features, with density plots revealing distinct clusters of correctly predicted high and low feature activations. The two examples with higher correlations identify specific conserved motifs directly tied to consistent protein structure and function, while the third example (f/8386) detects a structural motif (hexapeptide beta helix) that does not have Swiss-Prot annotations and appears across multiple functionally diverse proteins, potentially explaining the increased difficultly and lower performance at protein-level activation prediction. However, we note that across the distribution of tested features, there is only a weak dependence between a feature's Swiss-Prot concept prediction performance and its LLM description accuracy (Pearson $r = 0.11$), suggesting that LLM-generated descriptions can effectively characterize many protein features regardless of whether they are easily categorized by annotation associations (Supplementary Fig. 6).

Feature activations identify missing and novel protein annotations

Analysis of features with strong concept associations revealed that apparent 'false positive' activations often indicate missing database annotations rather than model errors. This insight suggests our features can detect functional elements overlooked by conventional annotation pipelines. For example, f/939 highlights a single conserved amino acid in a Nudix box motif. Among proteins with high activation of this feature, we identified one example (B2GFH1) lacking a Swiss-Prot label, while all other highly activated proteins carry this annotation. As Fig. 5 demonstrates, the region surrounding this activated position closely resembles labeled Nudix motifs. The presence of a Nudix motif in this protein is independently confirmed by hidden Markov model-based

https://doi.org/10.1038/s41592-025-02836-7

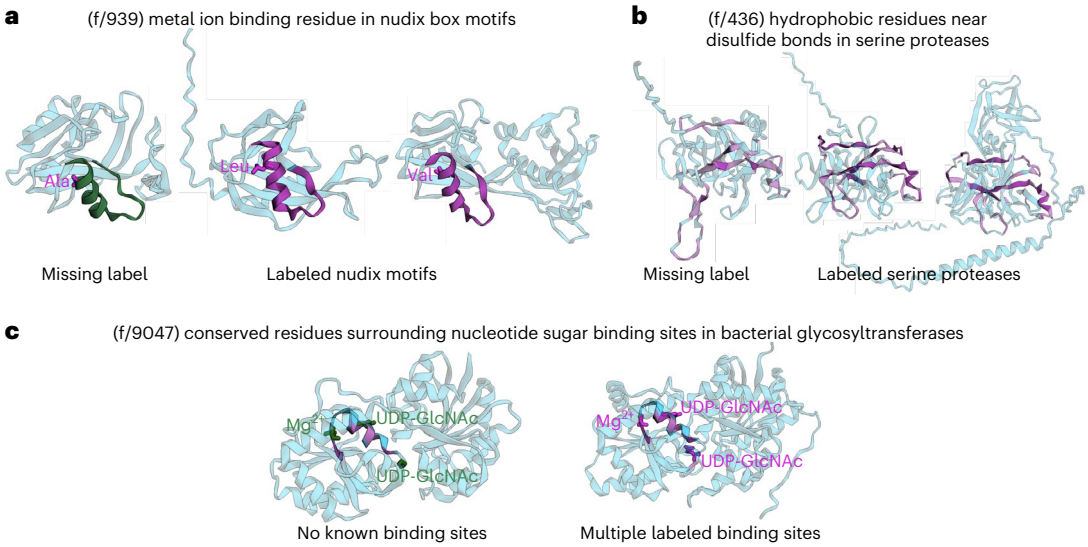


Fig. 5 | Feature activation patterns can be used to identify missing and new protein annotations. a, f/939 identifies missing motif annotation for Nudix box. It activates on a single amino acid in conserved position which is labeled in each structure. The right two (Q8ZHU8, A6VQT1) proteins are examples with high activations and pink coloring on the region with Nudix label. The left protein (B2GFH1), which does not have a Nudix motif annotation in Swiss-Prot, has implied feature annotation highlighted in green. The presence of a Nudix motif somewhere in this protein is confirmed by InterPro. **b**, f/436 identifies missing domain annotation for peptidase S1. It activates on a span of 80 amino acids with shared structural pattern. Right two proteins (Q99895, Q9H3S3) have high activation on confirmed peptidase S1 domains, with higher activation highlighted in pink. Left protein (Q7JIG6), which does not have peptidase S1 domain annotation in Swiss-Prot, is probably a missing annotation. The presence of an S1 domain somewhere in this protein is confirmed by InterPro. **c**, f/9047 suggests missing binding site annotations for sugar nucleotide binding sites within bacterial glycosyltransferases. In both structures, higher activation indicated with darker pink. The right protein (Q6NJL3) has Swiss-Prot annotations for binding sites labeled in pink. The left protein (O05083), which has no known binding site annotations but does have glycosyltransferase activity, has implied binding site annotations labeled in green.

annotations in InterPro[23]. Similar patterns were observed with feature f/436, which identifies proteins that Swiss-Prot correctly classifies as serine proteases but are missing the more specific peptidase S1 domain label that is present in InterPro. These examples serve as validation that our approach can identify missing annotations, reducing concerns about false positives and suggesting potential for discovering more notable functional features.

Our features detect functional similarities missed by both Swiss-Prot annotations and current automated annotation pipelines. Feature 9047 exemplifies this by identifying conserved nucleotide sugar and Mg^{2+} binding sites across diverse glycosyltransferase families. While the LLM-generated description correctly identifies UDP-dependent glycosyltransferases, most high-activating proteins lack formal binding site annotations in UniProtKB. The top ten activating proteins span multiple families with low sequence similarity to each-other (11–22% identity) despite high structural conservation (TM-scores of 0.74–0.78). Structural alignment reveals conserved binding site architecture with sequence variations accommodating different nucleotide sugar substrates. This feature-based approach complements traditional sequence methods by revealing functional relationships across evolutionary distances, potentially enhancing protein annotation systems with mechanistic insights from PLMs, though these more novel annotations will require additional experimental validation.

Protein sequence generation can be steered by activating interpretable features

To validate that features capture causally meaningful patterns, we performed targeted interventions in the model's predictions. Unlike in language models where features can be meaningfully steered by clamping their values across entire sequences, protein features rarely maintain consistent activation across domains. This makes it crucial to test whether localized feature manipulation can still influence model behavior at a distance. While many features capture complex biological concepts, quantitatively demonstrating specific interventions' effects is challenging for patterns lacking clear sequence-based validation. We therefore focused on a simple, measurable example: showing how activating specific features can steer glycine predictions in periodic patterns.

We tested three features that activate on periodic glycine repeats (GXXGXX) within collagen-like domains. Given a sequence with a glycine followed by a mask three positions later, increasing these features' activation values on the first glycine position increased the probability of glycine at both that position and the masked position—consistent with the expected periodic pattern (Fig. 6 and Supplementary Fig. 7). Importantly, these features were originally identified by their characteristic activation pattern occurring every third amino acid in natural sequences. The fact that activating them on just a single glycine position—rather than their standard periodic distribution—still produced interpretable effects demonstrates the robustness of their learned patterns. As expected, highly glycine-specific features (F1 scores: 0.995, 0.990, 0.86) only influenced the directly modified position. By contrast, the periodic glycine features demonstrated a more sophisticated capability: they successfully steered predictions for unmodified positions, even propagating this effect for multiple subsequent periodic repeats with diminishing intensity (Supplementary Fig. 8), revealing their capture of higher-order sequence patterns.

These results demonstrate that features activating on interpretable biological patterns can be capable of causally influence model behavior in predictable ways, even at positions beyond direct manipulation. However, further research is needed to understand the scope and limitations of feature-based steering across different sequence contexts and more complex biological patterns.

Discussion

Our work demonstrates that SAEs can extract thousands of interpretable features from PLM representations. Through quantitative evaluation against Swiss-Prot annotations, we showed that these features

https://doi.org/10.1038/s41592-025-02836-7

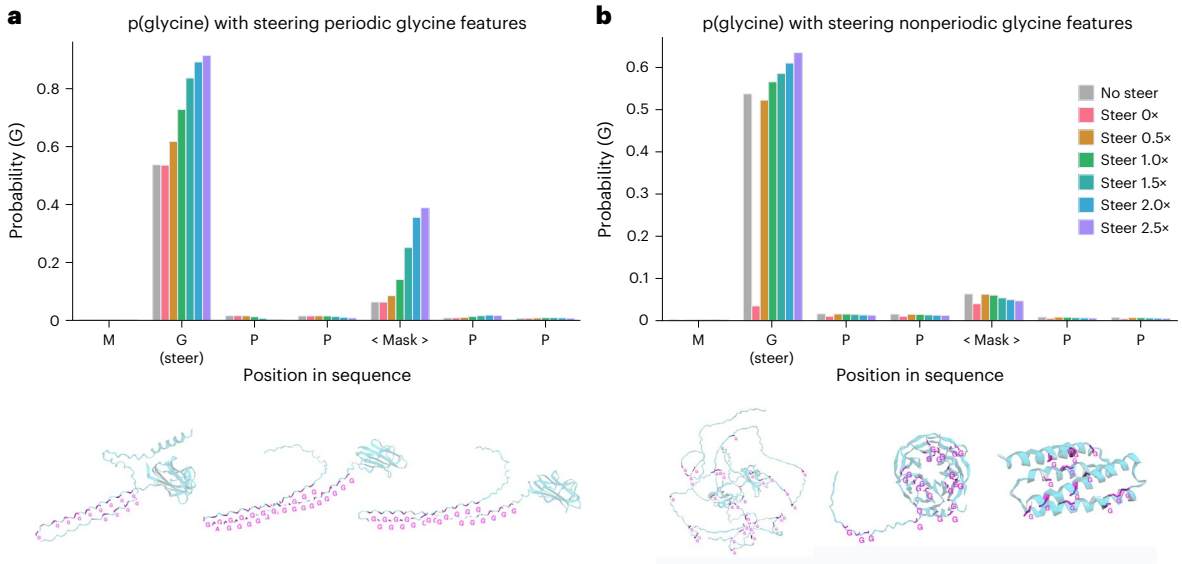


Fig. 6 | Steering feature activation on single amino acid additionally influences protein generation for nearby amino acids in interpretable way for group of periodic glycine features. a, Steering periodic glycine features. Steering activation of features that activate on glycines in periodic repeats of GXX in collagen-like regions by only steering one G and measuring impact on nearby positions. The result of steering all three periodic features on the predicted probability values for glycine (G) at each position in the sequence 'MGPP < mask > PP'. The feature steering was only applied to the unmasked G, not to the masked token or any other positions. The steer amount corresponds to the value the feature was clamped to, where 1× is the maximum observed activation value of this feature observed. Maximally activating examples for the three periodic features (f/4616, f/4970, f/10003) are shown below. **b**, Steering nonperiodic glycine features. The steering effects and maximally activated examples as in **a** but features selected to have the highest F1 score for glycine (f/6581, f/781, f/5381).

capture meaningful biological patterns that exist in superposition within the model's neurons. Our analysis framework combines automated interpretation through large language models with biological database annotations, enabling scalable feature characterization. Beyond identifying features that activate on annotated patterns, we showed that feature activation patterns can identify missing database annotations and enable targeted control of model predictions through interpretable steering.

Implications for PLM interpretability

The dramatic difference between neuron and SAE feature interpretability (SAEs extract 3× the concepts found in the 8-million-parameter ESM-2 neurons and 7× the concepts found in the 650-million-parameter ESM-2 neurons) provides strong evidence for information storage in superposition within PLMs. The persistence of superposition in larger models, despite their increased representational capacity, reveals an important aspect of how PLMs scale: they maintain distributed representations while dramatically increasing the breadth of biological concepts they capture. This expanded concept coverage aligns with broader evidence that larger PLMs learn richer protein representations, as shown by their enhanced performance on both masked token prediction and downstream tasks[1]. Interestingly, we found that SAEs trained on randomized PLMs still extract features specific to individual amino acids but fail to capture any complex biological concepts we tested for, suggesting that meaningful feature extraction requires learned model weights. This phenomena of identifying amino acids from randomized embeddings aligns with recent observations on randomized language models[21] showing that SAEs capture both properties of the model and the underlying data distribution.

By identifying feature–concept associations through domain-level F1 scores, our biological concept evaluation framework provides a new quantitative approach to assessing interpretability methods, which handles concepts that are annotated more coarsely than the learned features. Furthermore, our LLM-generated feature descriptions achieve strong predictive power at identifying highly activated proteins, even when there are no Swiss-Prot annotation associations, offering a complementary approach to Swiss-Prot annotations for understanding protein features. While these metrics are only approximations of the true biological interpretability of a model, these metrics clearly distinguish between different approaches and could enable systematic comparison of interpretability techniques. The discovery that 'false positive' predictions often indicate missing annotations demonstrates how interpretability tools can provide immediate practical value.

Analysis of feature activation patterns and dictionary vectors reveals both capabilities and open questions about model representations. While we identify many known conserved motifs and 3D patterns, the proportion of features dedicated to motif memorization versus other computational strategies remains unclear. Our framework provides a potential method for quantifying this balance, extending recent work showing PLMs primarily learn coevolutionary statistics[3]. In addition, our steering experiments demonstrate that features can influence both local and neighboring amino acid predictions, though the precise mechanisms by which features contribute to masked token prediction require further investigation.

Applications of interpretable PLM features

Our SAE framework offers valuable applications across model development, biological discovery and protein engineering. For model development, these features enable systematic comparison of learning patterns across different PLMs, revealing how architectural choices influence biological concept capture. Feature tracking during training and fine-tuning could provide insights into knowledge acquisition, while analysis of feature activation during failure modes could highlight systematic biases and guide improvements. Beyond confirming known patterns, these interpretable features may reveal novel biological insights about unannotated proteins and features that don't align with current annotations but show consistent activation patterns across protein families could point to unrecognized biological motifs or relationships that have escaped traditional analysis methods. Notably, our functional annotation approach based on SAE features

offers constant-time inference as a rapid complementary source to traditional methods, eliminating the need to scan each new protein against thousands of hidden Markov models or perform computationally expensive sequential or structural database comparisons. Since all protein clustering and feature description can be precomputed, this enables considerably faster initial annotation of novel sequences. Moreover, we have demonstrated that feature-based steering can be used to influence sequence generation in targeted ways. While periodic glycine steering may not revolutionize protein engineering itself, it demonstrates a promising new direction for controlling sequence generation during protein design tasks.

Limitations

Our feature interpretation methodology relies primarily on correlating feature activations with known properties, which provides valuable insights but has inherent constraints. While these correlations help identify what features might represent, they do not definitively establish causal relationships between features and model behavior and rely on the presence of existing localized annotations. Our LLM-based annotation approach expands beyond traditional database annotations, offering more flexible descriptions, yet the interpretability remains constrained by existing biological knowledge frameworks. This presents an opportunity to develop methods that might identify completely novel patterns that transcend current understanding, which the features can identify due to their unsupervised nature, but our current methods likely struggle to describe.

A key challenge is differentiating between features we cannot currently understand based on existing knowledge, features that appear uninterpretable due to limitations in our correlation-based approach and features that may be fundamentally uninterpretable. While LLM-based annotations help mitigate this by providing more nuanced descriptions, they still rely on text-based associations, which are often sparse and cannot fully capture the richness of protein biology let alone identify truly novel patterns outside current biological understanding. Further, this correlation-based approach raises the question of whether we are truly understanding what features represent within the model or merely finding post hoc explanations that align with our existing knowledge.

Our steering experiments successfully demonstrate causal relationships for specific features with periodic patterns, providing proof-of-concept validation that features can influence model's outputs. However, extending this approach to more complex biological concepts presents challenges, particularly with ESM-2, which as a masked language model was not designed for comprehensive sequence generation. Evaluating whether steering achieves intended biological changes without unintended consequences requires more sophisticated validation methods. Finally, our current focus on unmasked token representations leaves unexplored the information encoded in masked positions and special tokens at the start and end of sequences; special tokens may capture important protein-level information complementary to our residue-level analysis and analysis of features in masked positions will be important for steering masked amino acids.

Future directions

Improving interpretability methods. Future work should focus on enhancing feature extraction from increasingly complex and capable biological models, which potentially contain richer biological knowledge with reduced evolutionary constraints but present additional challenges through their complicated architectures. This advancement will necessitate developing quantitative metrics that extend beyond standard dictionary learning reconstruction and sparsity measures to evaluate which dictionaries will prove valuable for downstream applications such as knowledge discovery and steering. Moreover, we need to move beyond correlation-based descriptions of features toward mechanistic understanding of their biological significance. In addition, we must refine our methods for interpreting features and identifying those that could be interpreted but have not yet been. A crucial next step will also involve integrating dictionary learning approaches with previous attention-based interpretation analyses and tracking how these features combine into interpretable circuits for specific capabilities like contact prediction or binding site identification, thereby better illuminating how models extract structural and functional information from sequences.

Using model interpretations for model development. Future work should leverage interpretability to improve PLMs in three key areas. First, we need to analyze how different representation choices impact biological understanding at specific architectural stages, comparing embeddings across model variants and tracing biological concept circuits to inform design decisions. Second, developing methods to systematically compare different models throughout training would reveal which architectural approaches best capture specific biological concepts, helping resolve tradeoffs in model design. Third, for protein design applications, research should identify features that control biologically important properties that are difficult to manipulate through traditional methods-such as binding specificity or solubility-enabling precise steering of generation toward functionally desirable outcomes. This will be more important to prioritize when working with PLMs that can perform more complicated sequence design such as ProGen2[2] or ESM-3[24].

Translating to biological insights. Looking forward, a key opportunity lies in converting interpretability insights into biological discoveries. By identifying shared feature activation patterns, we could transfer functional annotations to unannotated proteins, though this requires developing scalable methods to identify truly informative features and robust evaluation frameworks. More ambitiously, by analyzing features that don't match existing annotations yet remain interpretable in their function, we can potentially discover novel biological motifs and develop validation approaches for these computational hypotheses. These computational methods, learning directly from raw data without the constraints of existing human knowledge frameworks, could even identify patterns in understudied or unannotated proteins. As we explore increasingly sophisticated models, the potential for discovering novel biological insights from interpretable features will continue to expand, offering new perspectives on protein structure and function that complement traditional approaches.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-025-02836-7.

References

1. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
2. Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N. & Madani, A. ProGen2: exploring the boundaries of protein language models. *Cell Syst.* **14**, 968–978.e3 (2023).
3. Zhang, Z. et al. Protein language models learn evolutionary statistics of interacting sequence motifs. *Proc. National Acad. Sci. USA* **121**, e2406285121 (2024).
4. *The Claude 3 Model Family: Opus, Sonnet, Haiku* (Anthropic, 2024); https://www.anthropic.com/news/claude-3-family
5. Simon, E., Swanson, K. & Zou, J. Language models for biological research: a primer. *Nat. Methods* **21**, 1422–1429 (2024).
6. Vig, J. et al. BERTology meets biology: interpreting attention in protein language models. In *Proc. International Conference on Learning Representations* (ICLR, 2021).

7. Rao, R., Meier, J., Sercu, T., Ovchinnikov, S. & Rives, A. Transformer protein language models are unsupervised structure learners. In *Proc. International Conference on Learning Representations* (ICLR, 2021).
8. Wenzel, M., Grüner, E. & Strodthoff, N. Insights into the inner workings of transformer models for protein function prediction. *Bioinformatics* **40**, btae031 (2024).
9. Marks, D. S. et al. Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6**, e28766 (2011).
10. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
11. Arora, S. et al. Linear algebraic structure of word senses, with applications to polysemy. *Trans. Assoc. Comput. Linguist.* **6**, 483–495 (2018).
12. Olah, C. et al. Zoom in: an introduction to circuits. *Distill* **5**, e00024.001 (2020).
13. Yun, Z., Chen, Y., Olshausen, B. & LeCun, Y. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. In *Proc. Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures* (eds Agirre, E. et al.) 1–10 (Association for Computational Linguistics, 2021).
14. Cunningham, H., Ewart, A., Riggs, L., Huben, R. & Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *Proc. International Conference on Learning Representations* (ICLR, 2024).
15. Bricken, T. et al. Towards monosemanticity: decomposing language models with dictionary learning. *Transformer Circuits Thread* https://transformer-circuits.pub/2023/monosemantic-features/index.html (2023).
16. Gorton, L. The missing curve detectors of InceptionV1: applying sparse autoencoders to InceptionV1 early vision. In *Proc. ICML 2024 Workshop on Mechanistic Interpretability* (ICML, 2024).
17. Olsson, C. et al. In-context learning and induction heads. *Transformer Circuits Thread* https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html (2022).
18. Marks, S. et al. Sparse feature circuits: discovering and editing interpretable causal graphs in language models. In *Proc. International Conference on Learning Representations* (ICLR, 2025).
19. McDougall, C. SAE visualizer. *GitHub* https://github.com/callummcdougall/sae_visualizer (2024).
20. Bills, S. et al. Language models can explain neurons in language models. *OpenAI* https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html (2023).
21. Templeton, A. et al. Scaling monosemanticity: extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread* https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html (2024).
22. Karvonen, A. et al. Measuring progress in dictionary learning for language model interpretability with board game models. In *Proc. 38th Annual Conference on Neural Information Processing Systems* (NeurIPS, 2024).
23. Paysan-Lafosse, T. et al. InterPro in 2022. *Nucleic Acids Res.* **51**, D418–D427 (2023).
24. Hayes, T. et al. Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).

Methods

SAEs training

Dataset preparation. We selected 5 million random protein sequences from UniRef50, part of the training dataset for ESM-2. For each protein, we extracted hidden representations after transformer block layers 1–6 from ESM-2-8M-UR50D and layers 1, 9, 28, 24, 30 and 33 from ESM-2-650M-UR50D, excluding < cls > and < eos > tokens. The datasets were each sharded into groups of 1,000 proteins each, with tokens shuffled within these groups to ensure random sampling during training. We downloaded the ESM model and weights from HuggingFace[25] (v0.24) and trained with Pytorch[26] (v2.1).

Architecture and training parameters. We trained SAEs following[15], using a 32× expansion factor for ESM-2-8M (320–10,240 features) and 8× for ESM-2-650M (1,280–10,240 features). For each layer, we trained 20 SAEs using a batch size of 2,048. ESM-2-8M models trained for 500,000 steps with learning rates from 1×10^{-4} to 1×10^{-8} in 10× increments, while ESM-2-650M trained for 200,000 steps with learning rates from 5×10^{-4} to 5×10^{-6} in 5× increments, due to computational constraints. L1 penalties ranged from 0.07 to 0.2 for ESM-2-8M and 0.04 to 0.2 for ESM-2-650M. Both parameters increased linearly from 0 during training, with learning rate reaching maximum within the first 5% of steps. The parameters used for SAEs in study listed in Supplementary Table 1.

Feature normalization. To standardize feature comparisons, we normalized activation values using a scan across 50,000 proteins from Swiss-Prot. For each feature, we identified the maximum activation value across this dataset and used it to scale the encoder weights and decoder weights reciprocally, ensuring all features were scaled between 0 and 1 while preserving the final reconstruction values.

Swiss-prot concept evaluation pipeline

Dataset construction. From the reviewed subset of UniprotKB (Swiss-Prot), we randomly sampled 50,000 proteins with lengths under 1,024 amino acids. We converted all binary and categorical protein-level annotations into binary amino acid-level annotations, maintaining domain-level relationships for multi-amino-acid annotations. The dataset was split equally into validation and test sets of 25,000 proteins each. We retained only concepts present in either more than ten unique domains or more than 1,500 amino acids within the validation set. Concepts were extracted from the list of annotations listed in Supplementary Table 3.

Feature–concept association analysis. For each normalized feature, we created binary feature-on/feature-off labels using activation thresholds of 0, 0.15, 0.5, 0.6 and 0.8. We evaluated feature–concept associations using modified precision and recall metrics

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \quad (1)$$

$$\text{Recall} = \frac{\text{domains with true positive}}{\text{total domains}} \quad (2)$$

$$\text{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \quad (3)$$

For each feature–concept pair, we selected the threshold yielding the highest F1 score for final evaluation.

Model selection and evaluation. We conducted initial evaluations on 20% of the validation set to compare hyperparameter configurations (only including the 135 concepts that had more than ten domains or 1,500 amino acids in this subset alone). For each concept, we identified the feature with the highest F1 score and used the average of these top scores to select the best model per layer. These six models (one per ESM2-8M layer) were used for subsequent analyses and the InterPLM dashboard.

To calculate the test metrics we (1) identify the feature with highest F1 score per-concept on the full validation set, then for each concept, calculate the F1 score of the selected feature on the test set and report these values then (2) identify all feature–concept pairs with F1 >0.5 in the validation set, calculate their F1 scores on the test set and report the number of these that have F1 >0.5 in the test set.

Baselines. To train the randomized baseline models, we shuffled the values within each weight and bias of ESM2-8M, calculated embeddings on the same datasets and repeated the same training (using six hyperparameter choices per layer), concept-based model selection and metric calculation processes. To compare with neurons, we scaled all neuron values between 0 and 1 (based on the minimum and maximum values found in our Swiss-Prot subset), then input these into an SAE with expansion factor of 1× that has an identity matrix for the encoder and decoder, such that all other analysis can be performed identically.

LLM feature annotation pipeline

Example selection. An analysis was performed on a random selection of 1,240 (10%) features. For each feature, representative proteins were selected by scanning 50,000 Swiss-Prot proteins to find those with maximum activation levels in distinct ranges. Activation levels were quantized into bins of 0.1 (0–0.1, 0.1–0.2, …, 0.9–1.0), with two proteins selected per bin that achieved their peak activation in that range, except for the highest bin (0.9–1.0), which received ten proteins. Moreover, ten random proteins with zero activation were included to provide negative examples. For features where fewer than 20 proteins reached peak activation in the highest range (0.9–1.0), additional examples were sampled from the second-highest range (0.8–0.9) to achieve a total of 24 proteins between these two bins, split evenly between training and evaluation sets. Features were excluded if fewer than 20 proteins could be found reaching peak activation across the top three activation ranges combined.

Description generation and validation. For each feature, we compiled a table containing protein metadata, quantized maximum activation values, indices of activated amino acids and amino acid identities at these positions. Using this data, we prompted Claude-3.5 Sonnet (new) to generate both a detailed description of the feature and a one-sentence summary that could guide activation level prediction for new proteins.

To validate these descriptions, we provided Claude with an independent set of proteins (matched for size and activation distribution) along with their metadata but without activation information. Claude's predicted activation levels were compared with measured values using Pearson correlation.

Feature analysis and visualization

UMAP embedding and clustering. We performed dimensionality reduction on the normalized SAE decoder weights using uniform manifold approximation and projection (UMAP) (parameters: metric='cosine', neighbors=15, min dist=0.1). Clustering was performed using HDBSCAN[27] (min cluster size=5, min samples=3) for visualization in the InterPLM interface.

Sequential and structural feature analysis. We assessed whether SAE features exhibit meaningful spatial organization in protein sequences and structures using the following procedure:

- **Data preparation**: for each SAE feature and each data shard:

  Use the SAE to get feature activation values a_i for each amino acid position i
  Identify all positions where activations exceed threshold $a_i > 0.6$
  Retain only proteins with complete AlphaFold 3D coordinate data and at least 25 proteins where activations exceed threshold
  Randomly sample up to 100 proteins per feature for analysis

- **Clustering metric calculation**: for each selected protein:

  Find the position with maximum activation $\text{pos}_{\max} = \arg\max_i a_i$
  Sequential clustering: calculate mean activation of sequence neighbors

$$\text{seq\_score} = \frac{1}{|\text{seq\_neighbors}|} \sum_{j \in \text{seq\_neighbors}} a_j,$$

  where $\text{seq\_neighbors} = \{j : |j - \text{pos}_{\max}| \le 2, j \ne \text{pos}_{\max}\}$
  Structural clustering: calculate mean activation of 3D spatial neighbors

$$\text{struct\_score} = \frac{1}{|\text{struct\_neighbors}|} \sum_{j \in \text{struct\_neighbors}} a_j,$$

  where $\text{struct\_neighbors} = \{j : \text{distance}(j, \text{pos}_{\max}) \le 6\ \text{Å}, j \ne \text{pos}_{\max}\}$
  Distance calculated using C_α coordinates: $\text{distance}(i,j) = ||C_\alpha(i) - C_\alpha(j)||$
- **Null distribution generation**: for each protein:

  Create 5 random permutations of the activation values $(a_1, a_2, \ldots, a_n)$
  For each permutation, recalculate $\text{seq\_score}_{\text{null}}$ and $\text{struct\_score}_{\text{null}}$ using the same neighbor positions but shuffled activation values
  Average $\{\text{seq\_score}_{\text{null}}\}$ across the five permutations to get final null values for each protein
- **Statistical testing**: for each feature:

  Collect observed scores $\{\text{seq\_score}_{\text{observed}}\}$ and null scores $\{\text{seq\_score}_{\text{null}}\}$ across all proteins
  Perform paired *t*-test comparing observed versus null distributions
  Calculate Cohen's *d* effect size: $d = \frac{\text{mean(observed} - \text{null)}}{\text{s.d.(observed} - \text{null)}}$

  Repeat independently for structural clustering scores
- **Multiple testing correction**: apply Bonferroni correction by multiplying all *P* values by the total number of features tested per layer

This analysis was performed separately for each SAE layer, yielding statistical significance and effect size measures for both sequential and structural clustering tendencies of learned features.

For structural feature identification in InterPLM, we considered only proteins with Bonferroni-corrected structural *P* values< 0.05, with features colored based on the ratio of structural to sequential effect sizes. Structural-only features are defined as having structural *P* value <0.05 but sequential *P* value >0.05.

Steering experiments

Following the approach described in ref. 21, we decomposed ESM embeddings into SAE reconstruction predictions and error terms. For sequence steering, we (1) extracted embeddings at the specified layer, (2) calculated SAE reconstructions and error terms, (3) modified the reconstruction by clamping specified features to desired values, (4) combined modified reconstructions with error terms, (5) allowed normal model processing to continue and (6) extracted logits and calculated softmax probabilities for comparison across steering conditions. Steering experiments all conducted using NNsight[28].

Categorizing Swiss-Prot concepts

To analyze concept distributions across models, we developed a hierarchical categorization system for the 258 Swiss-Prot concepts identified. Each concept already had a broad type (for example, Domain) and specific subtype (for example, DRBM). We used Claude-3.5 Sonnet to generate detailed functional descriptions for each concept (for example, 'Domain DRBM' → 'Double-stranded RNA binding motif'). Using these descriptions, we prompted Claude to create biologically meaningful categories and assign each concept to its most appropriate category. This resulted in 14 high-level functional categories (for example, 'Nucleic Acid Interaction Domains') that we used to systematically compare concept distributions between different-sized PLMs.

Sequential and structural alignment

Sequence alignments and percent identity calculations were performed with Clustal Omega[29] via UniProt and structural alignments were computed with TM-Align[30] via RCSB[31].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data analyzed in this study were obtained from publicly available databases. The protein sequences were sourced from UniRef50 and Swiss-Prot databases (accessible via https://www.uniprot.org/help/downloads), while protein structures were retrieved from the AlphaFold Database using version 4 and isoform 1 of each protein (AFDB-F1-v4). Per-layer analysis results, along with an interactive visualization platform, are available via InterPLM.ai at https://InterPLM.ai.

Code availability

Core computational methods, including implementations of SAEs, key feature analysis pipelines and visualization tools, are publicly available via GitHub at https://github.com/ElanaPearl/InterPLM. This repository includes documentation and a guide for running these analyses and generating custom dashboards. Model weights for SAEs analyzed in the study are available at https://huggingface.co/Elana/InterPLM.

References

25. Wolf, T. et al. Transformers: state-of-the-art natural language processing. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (eds Liu, Q. & Schlangen, D.) 38–45 (Association for Computational Linguistics, 2020).
26. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Proc. 33rd International Conference on Neural Information Processing Systems* 8026–8037 (NeurIPS, 2019).
27. Campello, R. J. G. B., Moulavi, D. & Sander, J. in *Advances in Knowledge Discovery and Data Mining* (eds. Pei, J. et al.) Vol. 7819, 160–172 (Springer, 2013).
28. Fiotto-Kaufman, J. F. et al. NNsight and NDIF: democratizing access to open-weight foundation model internals. In *Proc. International Conference on Learning Representations* (ICLR, 2025).
29. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
30. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
31. Bittrich, S., Segura, J., Duarte, J. M., Burley, S. K. & Rose, Y. RCSB Protein Data Bank: exploring protein 3D similarities via comprehensive structural alignments. *Bioinformatics* **40**, btae370 (2024).

Acknowledgements

We thank T. McGrath, A. Tamkin, N. Joseph and the Zou lab members for helpful discussions and feedback. E.S. is supported by NSF GRFP (grant no. DGE-2146755), and J.Z is supported by funding from the CZ Biohub. We received no specific funding for this work.

Author contributions

E.S. designed and conducted the study, analyzed the data, and wrote the paper. J.Z. provided supervision throughout the project and feedback on the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-025-02836-7.

Correspondence and requests for materials should be addressed to Elana Simon or James Zou.

Peer review information *Nature Methods* thanks James Fraser, Jeffrey Ruffolo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Zou

Last updated by author(s): Jul 2, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Pipelines for downloading and processing data are available at github.com/ElanaPearl/interPLM.
Data analysis	Core computational methods, including implementations of sparse autoencoders, key feature analysis pipelines, and visualization tools, are publicly available at github.com/ElanaPearl/interPLM. This repository includes documentation and a guide for running these analyses and generating custom dashboards. Trained models are available at https://huggingface.co/collections/Elana/interplm-678bae7a162e0dc0a860c44d.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data analyzed in this study were obtained from publicly available databases. Protein sequences were sourced from UniRef50 and Swiss-Prot databases

(accessible via https://www.uniprot.org/help/downloads), while protein structures were retrieved from the AlphaFold Database and protein stability measurements were downloaded from ProteinGym. Our GitHub repository provides comprehensive links and code for downloading both protein annotations and AlphaFold-predicted structures. The complete set of per-layer analysis results, along with an interactive visualization platform, is available at https://InterPLM.ai

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data where this information has been collected, and consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	TWe trained our ESM2-8M SAEs on 5 million random protein sequences (>1 billion amino acids) based on dataset sizes used in prior literature in language models (e.g. Bricken et al (2023) used 8B tokens). Due to computational constraints we trained the ESM2-650M SAEs on .5 billion tokens.50,000 proteins randomly sampled from Swiss-Prot for feature evaluation. These sample sizes were chosen to balance comprehensive coverage of protein diversity with computational constraints and were sufficient to distinguish feature associations from a trained PLM with the null distribution of associations from an untrained PLM.
Data exclusions	From the Swiss-Prot dataset, we excluded proteins longer than 1,022 amino acids to accommodate model constraints. For feature analysis, we excluded concept annotations with fewer than 25 examples meeting activation criteria.
Replication	All computational analyses can be reproduced using the code provided in our GitHub repository. Key findings about feature activations and patterns can also be independently verified through interactive visualizations available at InterPLM.ai.
Randomization	Not applicable - this is a computational study analyzing existing data.
Blinding	Not applicable - this is a computational study analyzing existing data.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging