nature methods

Supplementary information

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

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

In the format provided by the authors and unedited

Contents

List of Figures

List of Tables

A Additional background on Protein Language Models

Protein language models (PLMs) adapt techniques from natural language processing to learn representations of protein sequences that capture both structural and functional properties. These models typically use transformer architectures, treating amino acids as discrete tokens. While natural language modeling typically uses an autoregressive training method, many modern PLMs employ masked language modeling objectives similar to BERT, as protein sequences fundamentally exhibit bidirectional dependencies - the folding and function of any given amino acid depends on both N-terminal and C-terminal context.

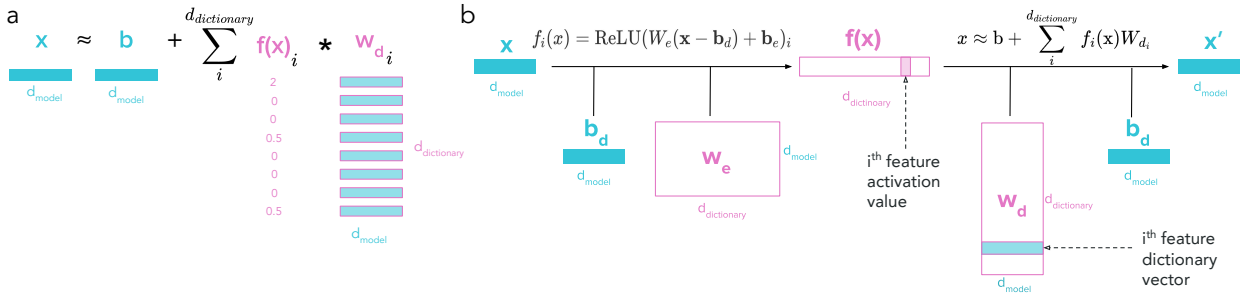
These models learn impressively rich protein representations through self-supervised training on large sequence databases without requiring structural annotations. The learned embeddings capture hierarchical information ranging from local physicochemical properties to global architectural features. Notably, these representations have proven crucial for protein structure prediction - they serve as the input embeddings for dedicated folding models like ESMFold [1], and even AlphaFold [2], while not explicitly a language model, dedicates most of the computational resources in a given prediction to learning protein representations from multiple sequence alignments in a conceptually similar manner.

Many efforts have demonstrated that the representations learned through masked language modeling alone contain remarkable structural and functional information, enabling state-of-the-art performance on tasks ranging from structure prediction to protein engineering [3]. This success appears to stem from the models' ability to capture the underlying patterns in evolutionary sequence data that reflect physical and biological constraints.

B Extended SAE Details and Analysis

B.1 Additional background

Sparse Autoencoders (SAEs) transform the latent vector for a single amino into a new vector with increased size and sparsity. When a specific position in this vector has a non-zero value, that corresponds to the presence of a specific pattern in the amino acid's neuron embedding. Ideally these model patterns correspond to human interpretable features that we can understand based on patterns by which the feature activates and the impact it has on the model when activated. Specifically, when we perform analysis on feature activation levels, these are using the $f_i(x)i$ values, while dictionary value analysis uses the learned weights in W_{d_i} as visualized in 1.



Supplementary Figure 1: **Overview of SAE decomposition and training.**
(a) Decomposition of embedding vector into weighted sum of dictionary elements. (b) Architecture for the SAE

While the SAEs trained in this paper use the architecture above, there are many other alternate SAE methods that vary weight initializations, nonlinearities, loss function, and other training details in aims to simultaneously increase sparsity and increase the reconstruction accuracy [4][5][6].

B.1.1 SAE metrics

While the goal of SAE training is to learn features that are maximally interpretable and accurate, these qualities are challenging to explicitly optimize for so during training we optimize for sparsity, hoping that more sparse features are more interpretable, and reconstruction quality.

Specifically, during training we calculate our loss as a weighted sum of an L1 norm calculating the absolute sum of all feature values, and mean squared error calculating how close the reconstructed x' is to the original x. Then, when

training SAEs or comparing different implementations, people evaluate L0, the average number of nonzero elements in the latent representation per token and Percent Loss Recovered, the percent of the original cross entropy of the model that is achieved when the model's embeddings are replaced by reconstructions. This last metric measures the amount 'recovered' by comparing the cross entropy of the model using reconstructed embeddings ($CE_{\text{Reconstruction}}$) to the original cross entropy (CE_{Original}) and the cross entropy when the specified embedding layer is instead replaced with all zeros(CE_{Zero}) but all later layers remain identical.

Metrics described are calculated per these equations:

$$L_1(f(x)) = \sum_{i=1}^{d_{\text{dictionary}}} |f_i(x)|$$

$$MSE(x, x') = \frac{1}{d_{\text{dictionary}}} \sum_{i=1}^{d_{\text{model}}} (x_i - x'_i)^2$$

$$L_0(f(x)) = \sum_{i=1}^{d_{\text{dictionary}}} 1(f_i(x) > 0)$$

$$\%\text{ Loss Recovered} = \left(1 - \frac{CE_{\text{Reconstruction}} - CE_{\text{Original}}}{CE_{\text{Zero}} - CE_{\text{Original}}}\right) * 100$$

While initial experiments used standard SAE metrics (L_0 and % Loss Recovered) to determine reasonable hyperparameter ranges, final model selection prioritized biological interpretability through Swiss-Prot concept associations rather than pure sparsity. Below are the hyperparameters and evaluation metrics for our selected models:

ESM-2 Model	Layer	Learning Rate	L1 penalty	L0	% Loss Recovered
8M	L1	1.0e-6	0.1	128	99.73
8M	L2	1.0e-6	0.08	163	99.72
8M	L3	1.0e-6	0.1	100	99.40
8M	L4	1.0e-6	0.1	106	98.94
8M	L5	1.0e-6	0.1	134	99.55
8M	L6	1.0e-6	0.09	178	100.00
650M	L1	1.0e-6	0.1	50	99.83
650M	L9	1.0e-5	0.06	211	99.49
650M	L18	1.0e-5	0.05	190	94.28
650M	L24	5.0e-5	0.05	121	92.42
650M	L30	1.0e-5	0.06	182	96.24
650M	L33	5.0e-5	0.06	148	100.00

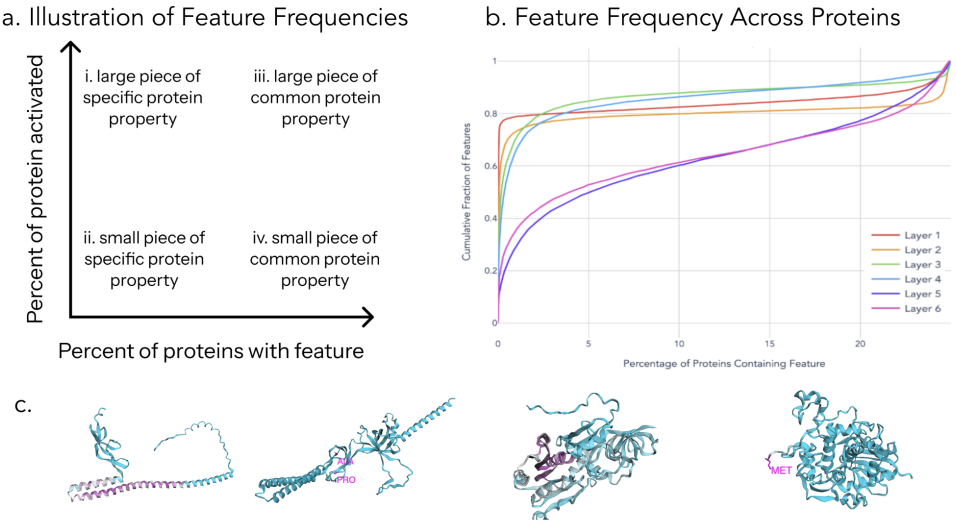
Supplementary Table 1: Layer-wise learning parameters and SAE performance metrics

C Extended SAE Feature Activation Analysis

C.1 Features with Different Activation Patterns Across and Within Proteins Reveal Distinct Roles

To systematically analyze protein language model features, we categorized them based on protein coverage and activation depth. Supplementary Figure 2 illustrates this classification framework across ESM-8M's Layer 4 features and highlights example features in each category.

Features in quadrant (i) activate strongly on specific protein families, such as feature L4/8921 which identifies alpha-helical regions in GrpE proteins. Quadrant (ii) features (40%) exhibit highly selective activation on precise motifs, like L4/9992 targeting specific 1-2 amino acid configurations in ¡5% of proteins. Quadrant (iii) features (12%) represent broader structural concepts present across diverse protein families, with feature L4/1547 consistently identifying Rossmann fold structures across multiple protein classes. Features in quadrant (iv) (33%) detect common but localized elements, such as specific secondary structures or sequence motifs like C-terminal alanine residues. 2c demonstrates examples of proteins in each quadrant, clarifying their differences. In (i) we have A and B. In (ii) we have A and B, in (iii) we have A and B, and in (iv) we have a and B. As seen 2b with the last two layers showing more features in quadrants (iii) and (i), suggesting increasing abstraction toward protein-wide concepts.

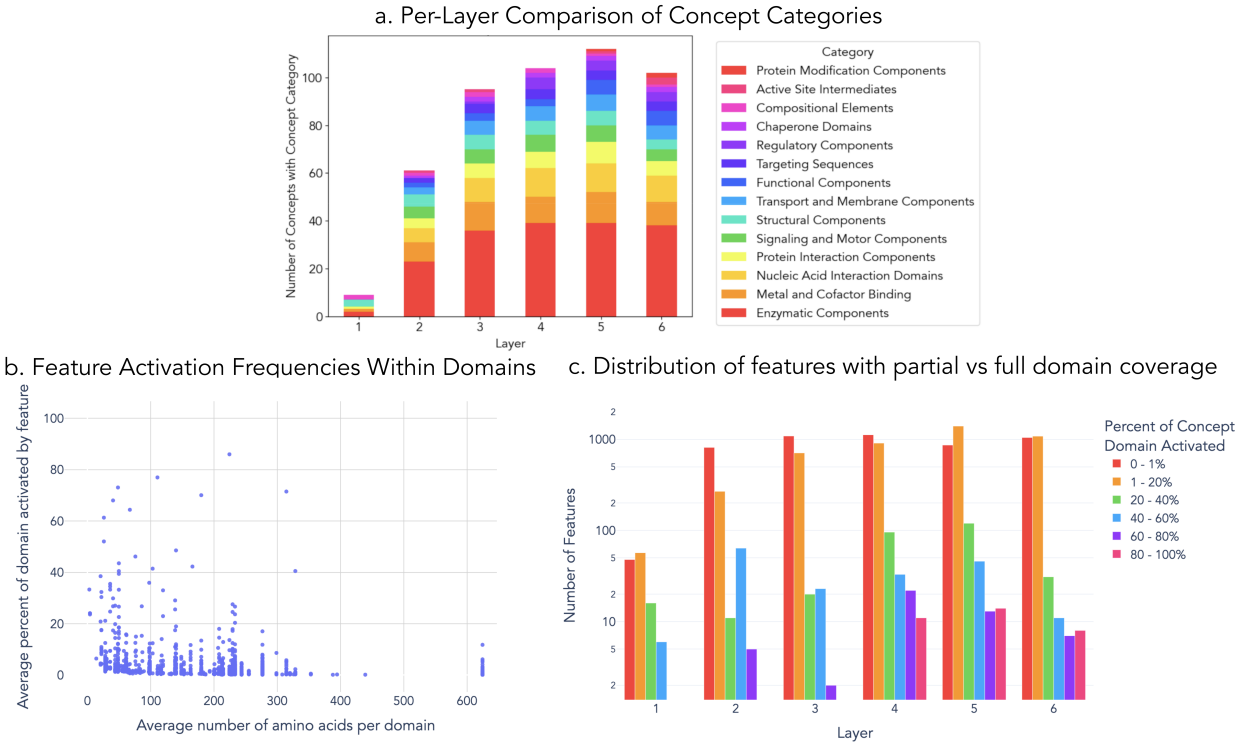


Supplementary Figure 2: **Categorization of protein features based on activation patterns.** (a) Classification framework for features based on two key dimensions: the percentage of proteins containing a feature (x-axis, specific vs. common) and the percentage of each protein activated by the feature (y-axis, granular vs. broad). This framework yields four distinct categories of features with different functional implications. (b) This graph displays cumulative distribution curves for six neural network layers, showing how features from layers 1-4 activate with relatively few proteins (as evidenced by their steep initial curves), while features in deeper layers 5-6 require a higher percentage of proteins to reach equivalent activation levels, indicating a shift from specific to more generalized feature detection throughout the network hierarchy. (c) Representative structures displaying examples from each quadrant (ESM-2-8M Layer 4). From top-left (clockwise): (i) Large-specific: Alpha-helical stretch in heat shock protein GrpE (f/8921, UniProtKB: B2V8C9), representing localized secondary structure elements; (ii) Small-specific: Two proximal beta-sheet residues in pHBA efflux pump (f/9992, UniProtKB: B1JKI2), capturing ubiquitous micro-structural motifs; (iii) Large-general: Alpha-beta-alpha sandwich within Rossmann-fold domains (f/1547, UniProtKB: Q28719); (iv) Small-general: N-terminal methionine (f/8395, UniProtKB: P07062), identifying complete structural elements characteristic of specific protein families.

Layer	# Structural-Only	Structural Examples	Sequential Examples	Both Structural & Sequential Examples
1	129	Pairs of cysteines (9041, 9861, 280, 9327, 7221)	Short beta strands, fragments with amino acid biases (2920, 226, 5598)	Consecutive stretches forming helices where pairs are structurally close (5334, 2456, 1379)
3	179	Leucine-rich repeats, cysteine pairs in zinc fingers (8888, 9347, 4870, 3761, 1497)	Adjacent amino acid pairs, conserved motifs, GG repeats (7289, 7461, 421, 3553, 5932, 1591, 4537)	Nearby helical bundles (5932, 1591, 4537)
6	359	Cysteine pairs, leucine-rich repeats, ANK repeats (22, 8259, 6023, 5502, 3371)	Adjacent amino acid pairs, conserved strands and helix fragments (61, 7114, 14509)	Helical bundles, beta strands in barrel plug, HTH fragments (722, 183, 204)

Supplementary Table 2: **Examples of Structural and Sequential Patterns Across ESM2-8M Layers**

C.2 Feature Activation Frequencies Across Proteins and Within Domains



Supplementary Figure 3: **Analysis of feature activation patterns across model layers.** (a) Per-layer comparison of concept categories showing increasing concept associations through deeper layers with peak associations at layers 4-5. (b) Feature activation frequencies within domains in ESM-2-8M Layer 6, plotting domain size (x-axis) against the average percentage of domain activated by features (y-axis). (c) Distribution of features with partial vs full domain coverage across layers, showing most features activate only small portions of domains ($< 1\%$ or $1 - 20\%$), with this pattern maintained across all layers and features with high domain coverage ($> 80\%$) only emerging at layer 4.

When looking at the concept category differences per layer (Figure 3a), we observe a general increase in the number of concept associations from layers 1-5, with a slight decrease at layer 6. Enzymatic Components (red) form the largest category across all layers, showing notable growth from layers 1-4. All functional categories follow similar growth trajectories through the network.

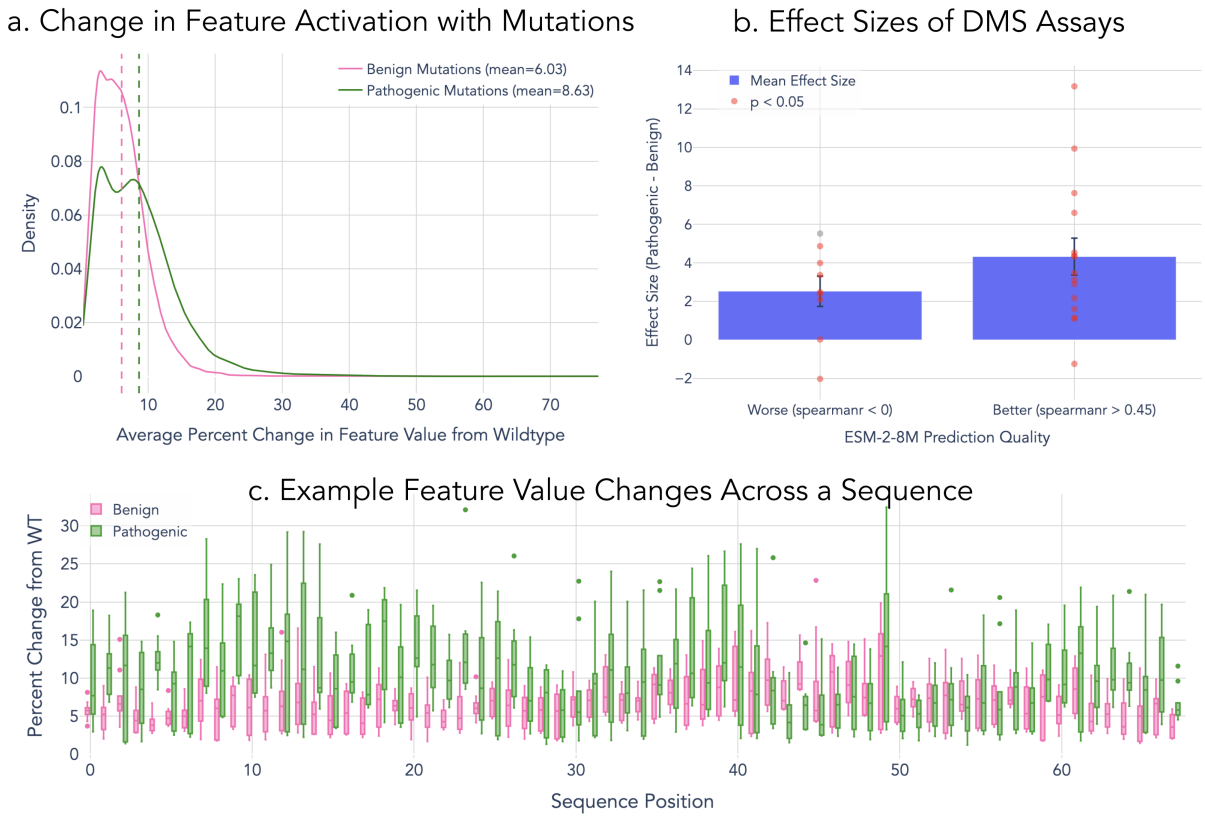
In Figure 3b, we observe that the percentage of a domain activated by features varies substantially based on domain size and type. Smaller domains (<100 amino acids) like Core-Binding domains show higher activation percentages (8-15%), while larger domains like TBDR beta-barrel exhibit lower activation percentages (0.5-2%). This inverse relationship between domain size and activation percentage is consistent across different concept categories.

As shown in Figure 3c, when a concept is associated with a domain, the amount of the domain activated varies by layer. The vast majority of features activate on less than 1% or between 1-20% of the domain, with this pattern consistent across all layers. Features activating across 20-40% or ¿40% of domains are relatively rare, with layer 2 showing the highest proportion of such broadly-activating features.

C.3 Robustness of concepts with model retrains

We evaluated the robustness of our SAE approach by re-training models on three different random subsets of UniRef50 (each containing 1 million proteins sampled from a pool of 4 million). The identified concepts proved largely consistent across these independently trained models. Each SAE discovered a similar number of interpretable features (98, 98, and 103 respectively), with approximately 91% of concepts showing comparable F1 scores (difference ¡0.25) across models. However, only 75% of features met our significance threshold (F1 ¿ 0.5) in all three models, indicating that while the overall feature landscape remains stable, the specific set of features that cross our interpretability threshold varies with training data.

C.4 Robustness of features to sequence mutations



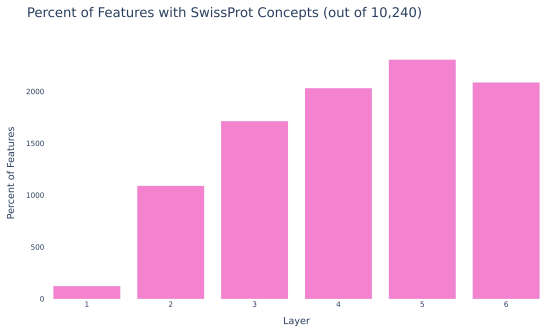
Supplementary Figure 4: **SAE features show simultaneous robustness and sensitivity to mutations.** All plots based on ESM-2-8M Layer 6. a) Distribution of feature activation changes in 14,940 amino acids across 24 protein stability DMS assays shows benign mutations (pink, mean=6.03%) causing smaller disruptions than pathogenic mutations (green, mean=8.63%). b) Comparing mutation effect sizes across all amino acids in an assay (mean(pathogenic) - mean(benign)) for assays grouped by ESM-2-8M prediction quality, with higher performance datasets exhibiting larger differences between mutation types (N=15 better quality, N=9 worse quality, Welch t-test p-value=0.08). Coloring of dots indicates p-value from paired wilcoxon signed rank test for each assay after multiple hypothesis adjustment. c) Position-by-position visualization of an example ProteinGym DMS Assay (PSAE-PICP2-Tsuboyama-2023-1PSE, $N_{benign} = N_{pathogenic} = 299$) demonstrates higher disruption from pathogenic mutations while maintaining overall robustness.

To test the robustness of our SAE features, we leveraged deep mutational scanning (DMS) data, which contains experimental measurements of how mutations affect protein stability and function. We analyzed 24 different protein assays with $\Delta\Delta G$ measurements from [7], specifically selecting 15 where ESM-2-8M performs well (Spearmanr $>$ 0.45) and 10 where it performs poorly (Spearmanr $<$ 0) in predicting mutational effects as identified in [8]. Our analysis reveals dual properties in our SAE features: benign mutations show small changes in feature activation (mean=6.03%); pathogenic mutations still small, but show significantly larger changes (mean=8.63%), as shown in Figure 4a. Across all 24 assays, when we use a wilcoxon signed rank test to compare the benign to pathogenic mutations (and multiply this by the number of assays to adjust for multiple testing), all assays but one show larger distribution in pathogenic vefsus benign mutations. This makes sense because pathogenic mutations, by definition, disrupt protein function, which our features appear to capture through larger perturbations in the learned representation space.

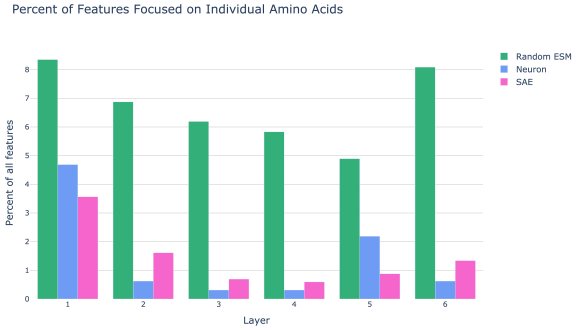
Assays where ESM-2-8M performs well as a predictor (Spearman $>$ 0.45) show larger effect sizes differentiating pathogenic from benign mutations compared to poorly predicted assays (Figure 4b). A Welch's t-test comparing effect sizes between high-quality and low-quality prediction groups yields a p-value of 0.081, suggesting a trend toward larger feature changes in well-predicted assays. This suggests our interpretable features may provide insight into when and why protein language models succeed or fail at variant effect prediction tasks. Position-specific analysis

demonstrates that mutation impacts vary across protein sequences, potentially highlighting functionally important regions where our features show heightened sensitivity while maintaining overall robustness to sequence variations.

C.5 Additional Concept Coverage Statistics



(a) Percent of SAE features per ESM-2-8M layer that are associated with any Swiss-Prot concept with $F1 > 0.5$



(b) Percent of features (or neurons) in each ESM-2-8M layer with $F1 > 0.5$ to an individual amino acid type.

Supplementary Figure 5

D Swiss-Prot Concepts Information

D.1 Swiss-Prot Metadata Categories

These tables contain all of the metadata used for quantitative feature-concept assoications and LLM feature descriptions and validation, organized into groups with similar themes. The last two columns specify whether each field was used in quantitative concept analysis, LLM descriptions, or both.

D.2 Swiss-Prot Metadata Categories

These tables contain all of the metadata used for quantitative feature-concept associations and LLM feature descriptions and validation, organized into groups with similar themes. The last two columns specify whether each field was used in quantitative concept analysis, LLM descriptions, or both.

D.2.1 Basic Identification Fields

Supplementary Table 3: Swiss-Prot Metadata Categories Used for Feature Descriptions

Field Name	Full Name	Description	Quant.	LLM
accession	Accession Number	Unique identifier for the protein entry in UniProt	N	Y
id	UniProt ID	Short mnemonic name for the protein	N	Y
protein_name	Protein Name	Full recommended name of the protein	N	Y
gene_names	Gene Names	Names of the genes encoding the protein	N	Y
sequence	Protein Sequence	Complete amino acid sequence of the protein	N	Y
organism_name	Organism	Scientific name of the organism the protein is from	N	Y
length	Sequence Length	Total number of amino acids in the protein	N	Y

D.2.2 Structural Features

Field Name	Full Name	Description	Quant.	LLM
ft_act_site	Active Sites	Specific amino acids directly involved in the protein's chemical reaction	Y	Y

Field Name	Full Name	Description	Quant.	LLM
ft_binding	Binding Sites	Regions where the protein interacts with other molecules	Y	Y
ft_disulfid	Disulfide Bonds	Covalent bonds between sulfur atoms that stabilize protein structure	Y	Y
ft_helix	Helical Regions	Areas where protein forms alpha-helical structures	Y	Y
ft_turn	Turns	Regions where protein chain changes direction	Y	Y
ft_strand	Beta Strands	Regions forming sheet-like structural elements	Y	Y
ft_coiled	Coiled Coil Regions	Areas where multiple helices intertwine	Y	Y
ft_non_std	Non-standard Residues	Non-standard amino acids in the protein	N	Y
ft_transmem	Transmembrane Regions	Regions that span cellular membranes	N	Y
ft_intramem	Intramembrane Regions	Regions located within membranes	N	Y

D.2.3 Modifications and Chemical Features

Field Name	Full Name	Description	Quant.	LLM
ft_carbohyd	Carbohydrate Modifications	Locations where sugar groups are attached to the protein	Y	Y
ft_lipid	Lipid Modifications	Sites where lipid molecules are attached to the protein	Y	Y
ft_mod_res	Modified Residues	Amino acids that undergo post-translational modifications	Y	Y
cc_cofactor	Cofactor Information	Non-protein molecules required for protein function	N	Y

D.2.4 Targeting and Localization

Field Name	Full Name	Description	Quant.	LLM
ft_signal	Signal Peptide	Sequence that directs protein trafficking in the cell	Y	Y
ft_transit	Transit Peptide	Sequence guiding proteins to specific cellular compartments	Y	Y

D.2.5 Functional Domains and Regions

Field Name	Full Name	Description	Quant.	LLM
ft_compbias	Compositionally Biased Regions	Sequences with unusual amino acid distributions	Y	Y
ft_domain	Protein Domains	Distinct functional or structural protein units	Y	Y
ft_motif	Short Motifs	Small functionally important amino acid patterns	Y	Y
ft_region	Regions of Interest	Areas with specific biological significance	Y	Y
ft_zn_fing	Zinc Finger Regions	DNA-binding structural motifs containing zinc	Y	Y
ft_dna_bind	DNA Binding Regions	Regions that interact with DNA	N	Y
ft_repeat	Repeated Regions	Repeated sequence motifs within the protein	N	Y
cc_domain	Domain Commentary	General information about functional protein units	N	Y

D.2.6 Functional Annotations

Field Name	Full Name	Description	Quant.	LLM
cc_catalytic_activity	Catalytic Activity	Description of the chemical reaction(s) performed by the protein	N	Y
ec	Enzyme Commission Number	Enzyme Commission number for categorizing enzyme-catalyzed reactions	N	Y
cc_activity_regulation	Activity Regulation	Information about how the protein's activity is controlled	N	Y
cc_function	Function	General description of the protein's biological role	N	Y
protein_families	Protein Families	Classification of the protein into functional/evolutionary groups	N	Y
go_f	Gene Ontology Function	Gene Ontology terms describing molecular functions	N	Y

E LLM-based Autointerpretability

E.1 Prompts

Generate description and summary

Analyze this protein dataset to determine what predicts the 'Maximum activation value' and 'Amino acids of highest activated indices in protein' columns. This description should be as concise as possible but sufficient to predict these two columns on held-out data given only the description and the rest of the protein metadata provided. The feature could be specific to a protein family, a structural motif, a sequence motif, a functional role, etc. These WILL be used to predict how much unseen proteins are activated by the feature so only highlight relevant factors for this.
Focus on:

- Properties of proteins from the metadata that are associated with high vs medium vs low activation.
- Where in the protein sequence activation occurs (in relation to the protein sequence, length, structure, or other properties)
- What functional annotations (binding sites, domains, etc.) and amino acids are present at or near the activated positions
- This description that will be used to help predict missing activation values should start with "The activation patterns are characterized by:"

Then, in 1 sentence, summarize what biological feature or pattern this neural network activation is detecting. This concise summary should start with "The feature activates on"
Protein record: `Insert table with Swiss-Prot metadata and activation levels`

Predict activation levels

Given this protein metadata record, feature description, and empty table with query proteins, fill out the query table indicating the maximum feature activation value within in each protein (0.0-1.0).

Base activation value on how well the protein matches the described patterns. There could be 0, 1 or multiple separate instances of activation in a protein and each activation could span 1 or many amino acids.
Output only these values in the provided table starting with "Entry,Maximum activation value". Respond with nothing but this table.
Protein record: `Insert table with Swiss-Prot metadata`
Table to fill out with query proteins: `Insert empty table of IDs to fill out with predictions`
The activation patterns are characterized by: `Insert LLM description`

Supplementary Table 4: Swiss-Prot Concepts associated with SAE features in any layer of ESM-2 8M. * Indicates concept that is also associated with a neuron in any layer.

Active Site

- Acyl-ester intermediate
- O-(3'-phospho-DNA)-tyrosine intermediate
- Tele-phosphohistidine intermediate

Coiled Coil

Compositional Bias

- Acidic residues
- Pro residues

Disulfide Bond

Modified Residue

- 4-aspartylphosphate
- O-(pantetheine 4'-phosphoryl)serine

Signal Peptide

- Tat-type signal
- any

Transit Peptide

- Mitochondrion
- any

Zinc Finger

- CR-type
- PHD-type
- RING-type
- any

Domain

- 2Fe-2S ferredoxin-type
- 4Fe-4S ferredoxin-type 1
- 4Fe-4S ferredoxin-type 2
- AB hydrolase-1
- ABC transmembrane type-1*
- ABC transporter*
- ATP-grasp
- C-type lectin
- C2
- CBS 1
- CBS 2
- CN hydrolase
- CP-type G
- Carrier
- CheB-type methylesterase
- Core-binding (CB)
- DPCK
- DRBM
- Disintegrin
- EngA-type G 2
- EngB-type G
- Era-type G
- Exonuclease
- F-box*
- FAD-binding FR-type
- FAD-binding PCMH-type
- Fe2OG dioxygenase
- Fibronectin type-III 1
- Fibronectin type-III 2
- G-alpha
- GH16
- GH18
- GMPS ATP-PPase
- GST C-terminal
- GST N-terminal
- Glutamine amidotransferase type-1*
- HD
- HTH araC/xylS-type*
- HTH cro/C1-type
- HTH luxR-type
- HTH lysR-type
- HTH marR-type
- HTH tetR-type
- Helicase ATP-binding
- Helicase C-terminal
- Histidine kinase
- Ig-like
- J*
- KH
- KH type-2
- Kinesin motor
- LIM zinc-binding 1
- LIM zinc-binding 2
- Lipoyl-binding
- MPN
- MTTase N-terminal
- N-acetyltransferase*
- NodB homology
- Nudix hydrolase
- Obg
- PDZ
- PH
- PPIase FKBP-type
- PPM-type phosphatase
- Peptidase A1
- Peptidase M12B
- Peptidase M14
- Peptidase S1*
- Peptidase S8*
- Protein kinase*
- RNase H type-1
- Radical SAM core*
- Response regulatory*
- Rhodanese
- Rieske
- S1 motif
- S1-like
- SH3
- SIS
- Sigma-54 factor interaction
- SpoVT-AbrB 1
- SpoVT-AbrB 2
- TBDR beta-barrel
- TGS
- TIR
- Thioredoxin
- TrmE-type G
- Tyr recombinase*
- Tyrosine-protein phosphatase
- Urease
- VWFA
- YjeF N-terminal
- YrdC-like
- bHLH
- bZIP
- sHSP
- tr-type G

Motif

- Beta-hairpin
- DEAD box
- Effector region
- Histidine box-2
- JAMM motif
- NPA 1
- Nudix box
- PP-loop motif
- Q motif
- Selectivity filter

Region

- 3-hydroxyacyl-CoA dehydrogenase
- A
- Adenylyl removase
- Adenylyl transferase
- Basic motif
- Disordered*
- Domain II
- FAD-dependent cmnm(5)s(2)U34 oxidoreductase
- Framework-3
- Interaction with substrate tRNA
- Large ATPase domain (RuvB-L)
- N-acetyltransferase
- NBD2
- NMP
- Pyrophosphorylase
- Ribokinase
- Small ATPAse domain (RuvB-S)
- Uridylyl-removing
- Uridylyltransferase

E.2 Example Descriptions and Summaries

Full Description Layer 4 Feature 9047

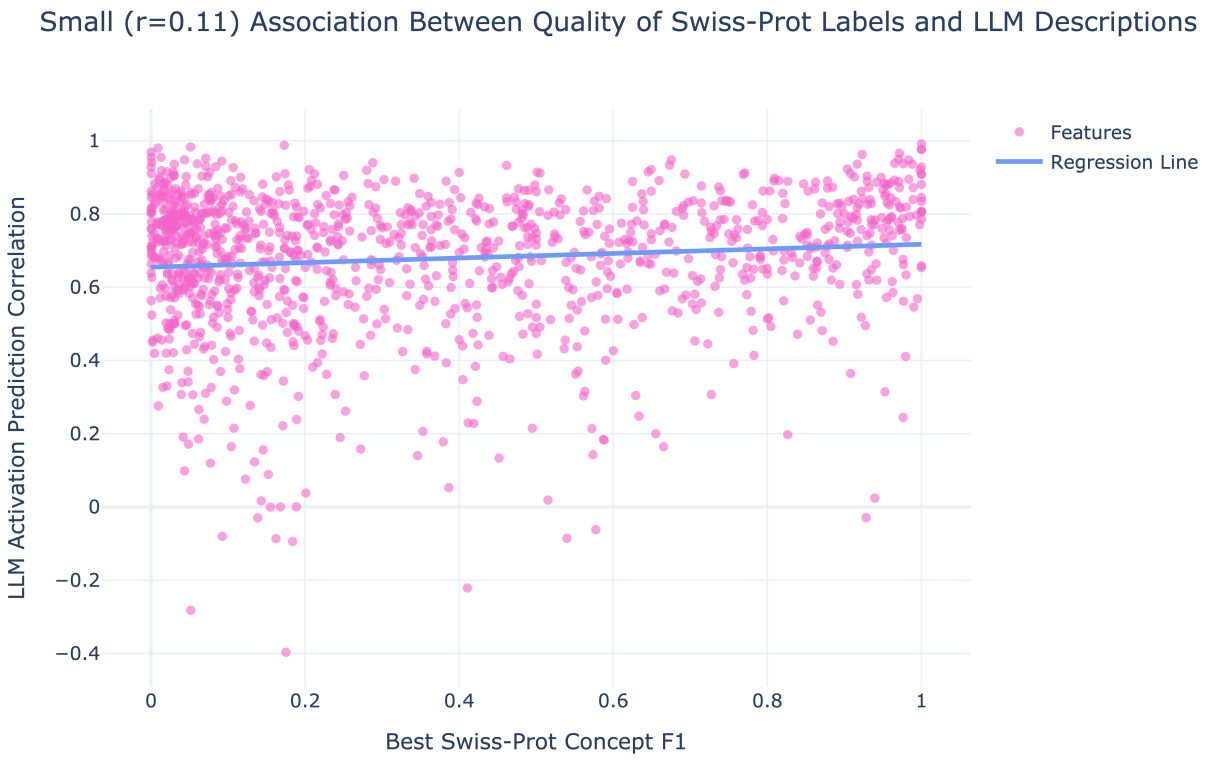
The activation patterns are characterized by:

- Highest activations (0.9-1.0) occur in glycosyltransferase proteins, particularly glycogen synthases (GlgA) and similar enzymes that transfer sugar molecules
- Activated positions consistently occur around amino acid positions 280-450 in these proteins, specifically involving glycine (G), alanine (A), or proline (P) residues
- The activated sites frequently overlap with substrate binding regions, particularly nucleotide-sugar binding sites (e.g., ADP-glucose, UDP-glucose, GDP-mannose)
- Medium activations (0.3-0.8) are seen in other transferases and synthetases with similar substrate binding patterns
- Proteins without sugar/nucleotide binding functions show no activation (0.0)

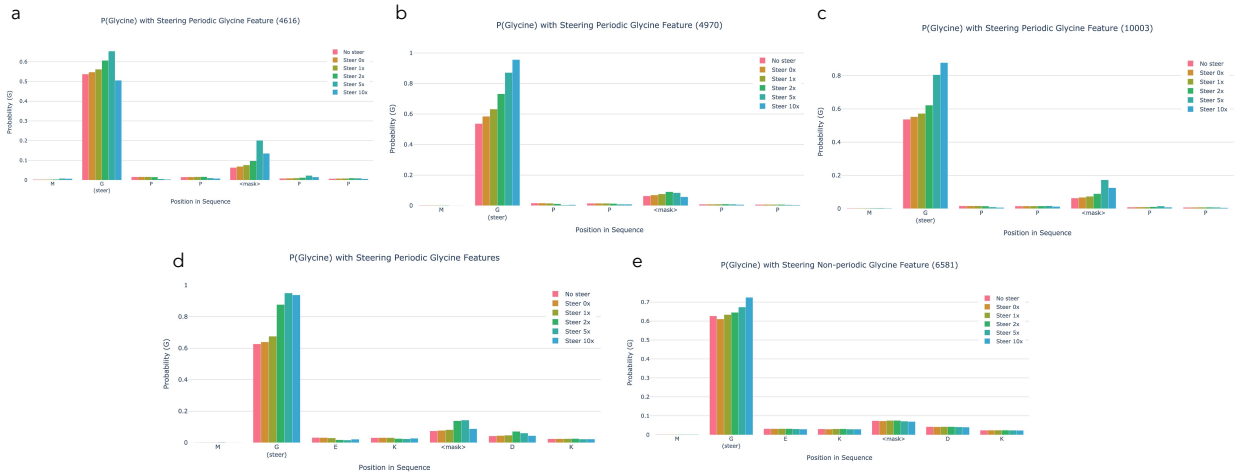
F Additional steering experiments

In Figure 7, we see that each periodic glycine feature can be used to steer the model to predict a periodic glycine pattern. We also evaluate a variant of the original sequence on the combination of periodic Glycine features and the non-periodic Glycine feature, observing that both increase the probability of Glycine on the steered position. As with the original sequence, we again see that steering has no positive impact on the effect of the Non-periodic Glycine (rather it has a small, negative impact).

Again, in Figure 8, we now test the same 3 periodic Glycine features (4616,4970, 10003), along with the 3 features with highest F1 to Glycine (6581, 781, 5489), and 3 randomly selected features (0, 1,000, 10,000). Here we calculate the slope of p(Glycine) across steering increasing [0, 0.5,0.75, 1, 1.5, 1.75, 2]. Alternate variants of the originally steered sequence maintain similar results and again, we see that even when applying 3 features at a time, the periodic Glycine features have a stronger effect on the masked position's Glycine probability than random or Glycine-specific features. This is also maintained in the longer sequence with multiple the masked periodic patterns being steered until this effect diminishes around the 5th masked position.



Supplementary Figure 6: Comparing quality of Swiss-Prot concept labels and accuracy of LLM predicted feature activation patterns reveals low correlation.

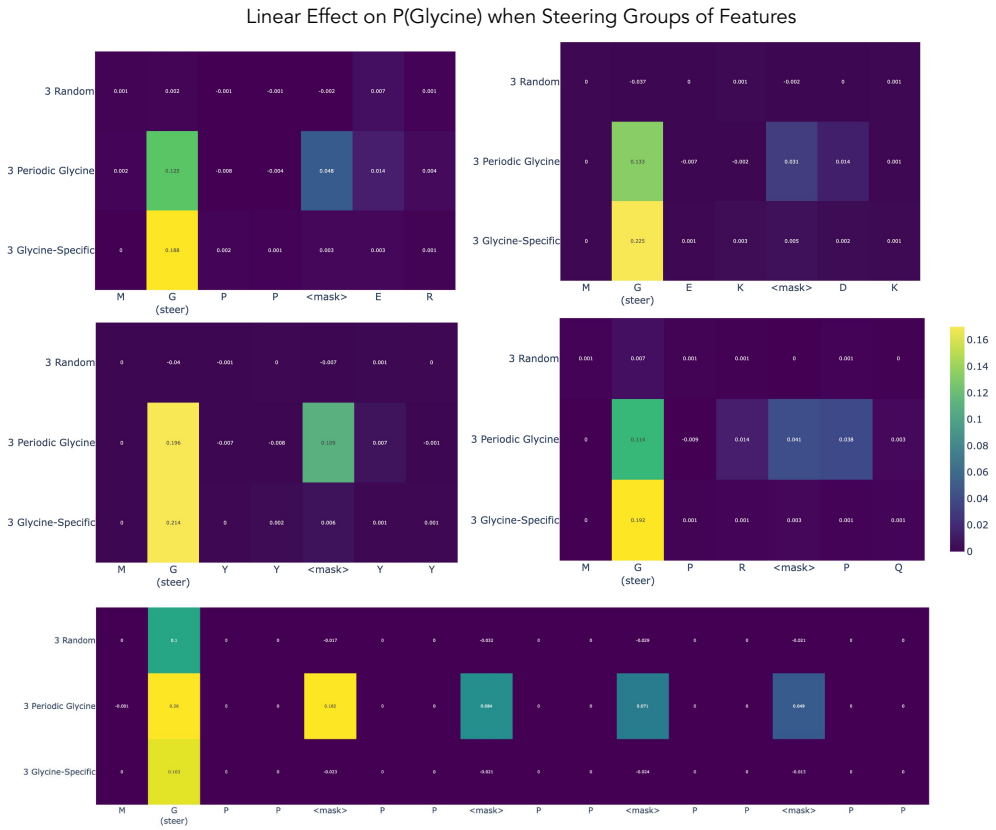


Supplementary Figure 7: Top row (a-c): Steering MGPP¡mask¿PP on each of the individual Periodic Glycine Features. Bottom row: Steering an alternate sequence, MGEK¡mask¿DK. (d) Steering all periodic glycine features (e) Steering non-periodic Glycine feature

Feature	Pearson r	Feature Summary
4360	0.75	The feature activates on interchain disulfide bonds and surrounding hydrophobic residues in serine proteases, particularly those involved in venom and blood coagulation pathways.
9390	0.98	The feature activates on the conserved Nudix box motif of Nudix hydrolase enzymes, particularly detecting the metal ion binding residues that are essential for their nucleotide pyrophosphatase activity.
3147	0.70	The feature activates on conserved leucine and cysteine residues that occur in leucine-rich repeat domains and metal-binding structural motifs, particularly those involved in protein-protein interactions and signaling.
4616	0.76	The feature activates on conserved glycine residues in structured regions, with highest sensitivity to the characteristic glycine-containing repeats of collagens and GTP-binding motifs.
8704	0.75	The feature activates on conserved catalytic motifs in protein kinase active sites, particularly detecting the proton acceptor residues and surrounding amino acids involved in phosphotransfer reactions.
9047	0.80	The feature activates on conserved glycine/alanine/proline residues within the nucleotide-sugar binding domains of glycosyltransferases, particularly at positions known to interact with the sugar-nucleotide donor substrate.
10091	0.83	The feature activates on conserved hydrophobic residues (particularly V/I/L) within the catalytic regions of N-acetyltransferase domains, likely detecting a key structural or functional motif involved in substrate binding or catalysis.
1503	0.73	The feature activates on extracellular substrate binding loops of TonB-dependent outer membrane transporters, particularly those involved in nutrient uptake.
2469	0.85	The feature activates on conserved structural and sequence elements in bacterial outer membrane beta-barrel proteins, particularly around substrate binding and ion coordination sites in porins and TonB-dependent receptors.

Supplementary Table 5: Example feature description summaries and their corresponding Pearson correlation coefficients when used along with more verbose descriptions to predict maximum activation levels.

It should be noted with all of these experiments that we have only tested a few, somewhat constrained sequences so far, and more work needs to be done to evaluate the contexts in which steering this feature can or cannot work.



Supplementary Figure 8: Measuring the linear effect on the predicted probability of Glycine. Evaluated p(Glycine) at each position in a sequence as groups of features were steered on first Glycine position at varying levels (0, 0.5,0.75, 1, 1.5, 1.75, 2). Calculated slope of p(Glycine) with respect to steering amount and visualize this for four sequences across each group of features. 3 Random features (f/0, f/1000, f/10000) were selected based on no activation-based information, 3 Periodic Glycine features (f/4616, f/4970, f/10003) selected based on maximum activation on periodic glycine patterns in collagen, and 3 Glycine-Specific features (f/6581, f/781, f/5389) selected to have the highest F1 scores to Glycine.

References

[1] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. Publisher: American Association for the Advancement of Science.

[2] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, June 2024. Publisher: Nature Publishing Group.

[3] Francesca-Zhoufan Li, Ava P. Amini, Yisong Yue, Kevin K. Yang, and Alex Xijie Lu. Feature Reuse and Scaling: Understanding Transfer Learning with Protein Language Models. June 2024.

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

[5] Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping Ahead: Improving Reconstruction Fidelity with JumpReLU Sparse Autoencoders, July 2024. arXiv:2407.14435 version: 1.

[6] Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders Find Interpretable LLM Feature Circuits, November 2024. arXiv:2406.11944 version: 2.

[7] Kotaro Tsuboyama, Justas Dauparas, Jonathan Chen, Elodie Laine, Yasser Mohseni Behbahani, Jonathan J. Weinstein, Niall M. Mangan, Sergey Ovchinnikov, and Gabriel J. Rocklin. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620(7973):434–444, August 2023. Publisher: Nature Publishing Group.

[8] Pascal Notin, Aaron W Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Hansen Spinner, Nathan Rollins, Ada Shaw, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Rose Orenbuch, Yarin Gal, and Debora S Marks. ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design.