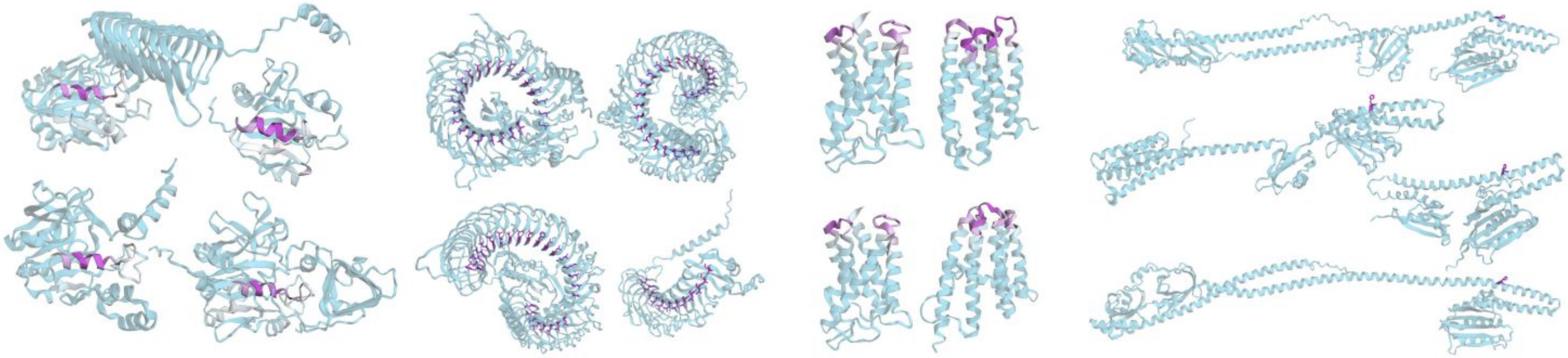
InterPLM: Discovering Interpretable Features in Protein Language Models via Sparse Autoencoders



Gökçe Uludoğan
PhD Candidate
LifeLU Reading Group | 28 November 2024

Elena Simon, James Zou

bioRxiv preprint
doi.org/10.1101/2024.11.14.623630

What exactly do protein language models (PLMs) ***learn***?

What exactly do *protein language models* (PLMs) learn?

> pLM ***interpretability***

pLM interpretability

- Analyzing these pLM internal representations
    - by probing the hidden states at different layers
    - by examining the patterns of attention between amino acids.

- Studies show that
    - Attention maps can reveal protein contacts and binding pockets
    - Hidden state representations from different layers can be probed to predict structural properties like secondary structure states

Many neurons are polysemantic.

Neurons don’t map cleanly to individual concepts, but instead exhibit superposition

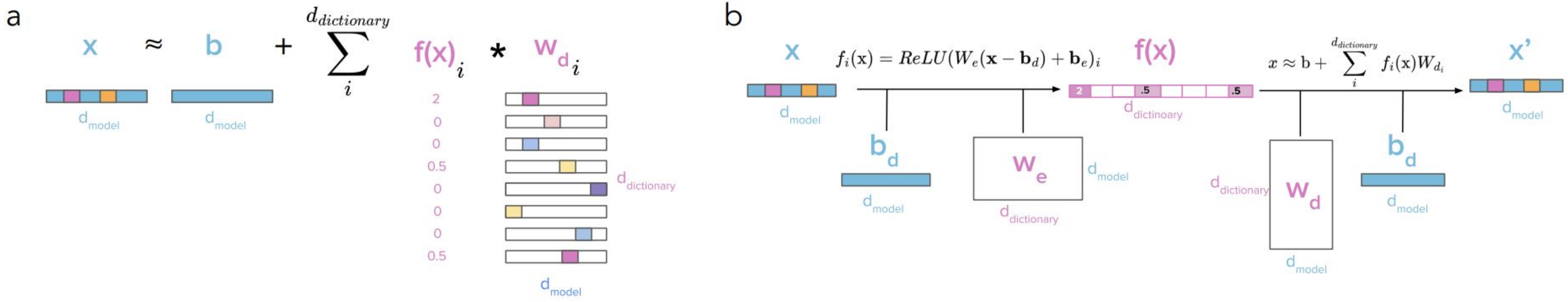
- multiple unrelated concepts are encoded by the same neurons

Can we ***decompose*** these neurons into ***basic interpretable features***?

Sparse Autoencoders

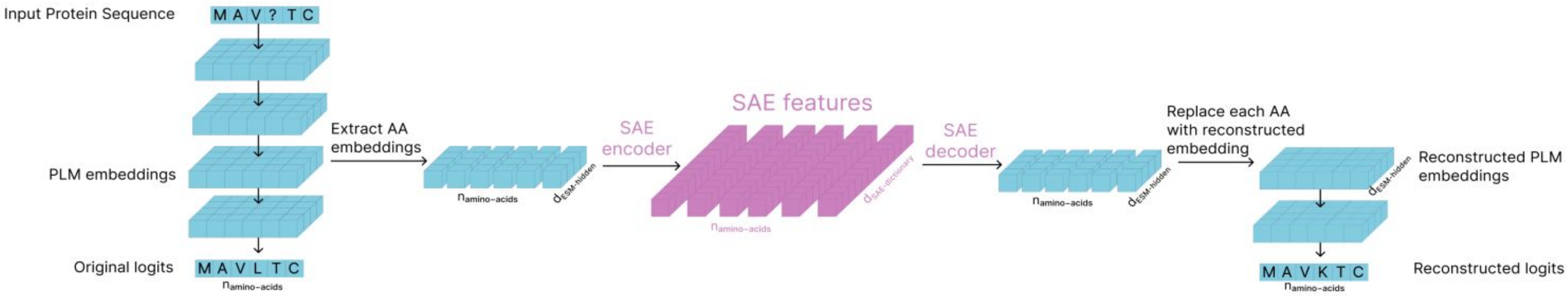
Sparse Autoencoders (SAEs) are a dictionary learning approach that transforms each neuron’s activation into a larger but sparse hidden layer.

They learn a "dictionary" of sparsely activated features that can reconstruct the original neuron activations.



Extracting Sparse Autoencoders Features from pLMs

$$\bar{\mathbf{x}} = \mathbf{x} - \mathbf{b}_d$$
$$\mathbf{f} = \mathrm{ReLU}(W_e\bar{\mathbf{x}} + \mathbf{b}_e)$$
$$\hat{\mathbf{x}} = W_d\mathbf{f} + \mathbf{b}_d$$
$$\mathcal{L} = \frac{1}{|X|}\sum_{\mathbf{x}\in X} ||\mathbf{x} - \hat{\mathbf{x}}||_2^2 + \lambda||\mathbf{f}||_1$$



Sparse Autoencoder Training

Data

- 5M random protein from UniRef50 (ESM2 training subset)

Representations

- ESM2-8M-UR50D Layers 1-6, embedding vectors of size 320

SAEs

- Feature dictionaries of size 10240 (x32)
- 20 SAEs per layer
- Normalize activation values using a scan across 50,000 proteins from SwissProt

Swiss-Prot Concept Evaluation Pipeline

Dataset

50,000 proteins (<1024 residues) from Swiss-Prot, split equally into validation and test

Annotations

Converted protein-level annotations into binary amino acid-level, preserving domain-level relationships.

Concept Filtering

Retained concepts with >10 unique domains or >1,500 amino acids in the validation set.

Swiss-Prot Concept Evaluation Pipeline

Binary Labels

Created feature-on/off labels using thresholds (0, 0.15, 0.5, 0.6, 0.8).

Evaluation

- Feature-concept associations were scored using modified precision and recall.
- Selected the threshold with the highest F1 for each feature-concept pair.
- Identified the top feature for each concept and averaged their F1 scores to select the best model per layer.

$$\text{precision} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalsePositives}}$$

$$\text{recall} = \frac{\text{DomainsWithTruePositive}}{\text{TotalDomains}}$$

$$\text{F1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Swiss-Prot Concept Evaluation Pipeline

Binary Labels

Created feature-on/off labels using thresholds (0, 0.15, 0.5, 0.6, 0.8).

Evaluation

- Feature-concept associations were scored using modified precision and recall.
- Selected the threshold with the highest F1 for each feature-concept pair.
- Identified the top feature for each concept and averaged their F1 scores to select the best model per layer.

$$\text{precision} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalsePositives}}$$

$$\text{recall} = \frac{\text{DomainsWithTruePositive}}{\text{TotalDomains}}$$

$$\text{F1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Evaluation

Test Metrics Calculation

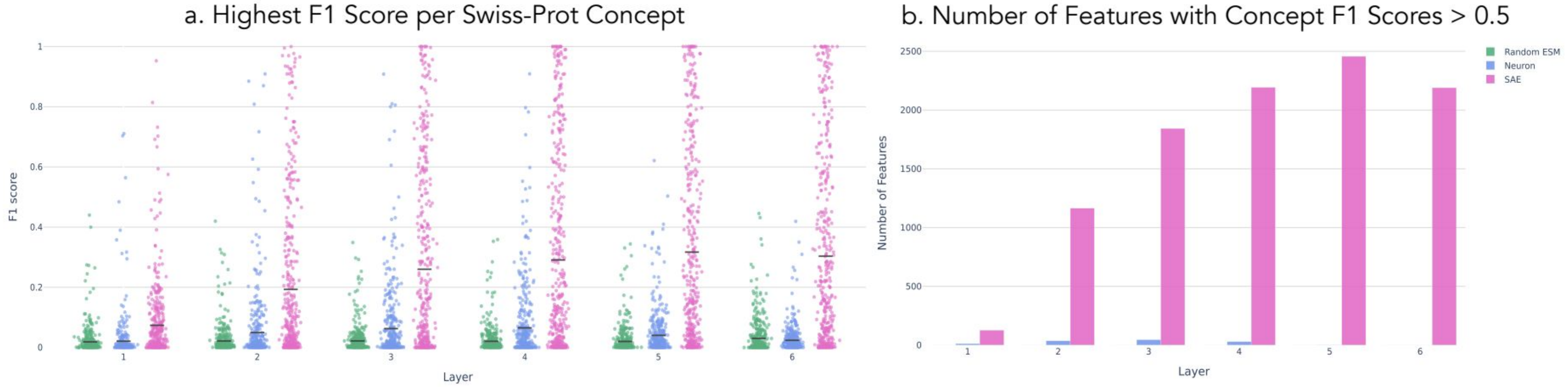
- Identify the top feature per concept (highest F1 on validation set), calculate its F1 on the test set, and report the scores.
- Select all feature-concept pairs with F1 > 0.5 in validation, calculate their F1 on the test set, and report how many retain F1 > 0.5.

Baseline

- Randomized baseline models were trained by shuffling ESM2-8M weights and biases.
- Followed identical training, model selection (6 hyperparameter choices per layer), and metric calculation pipelines.

SAE features have stronger associations with Swiss-Prot concepts than ESM neurons

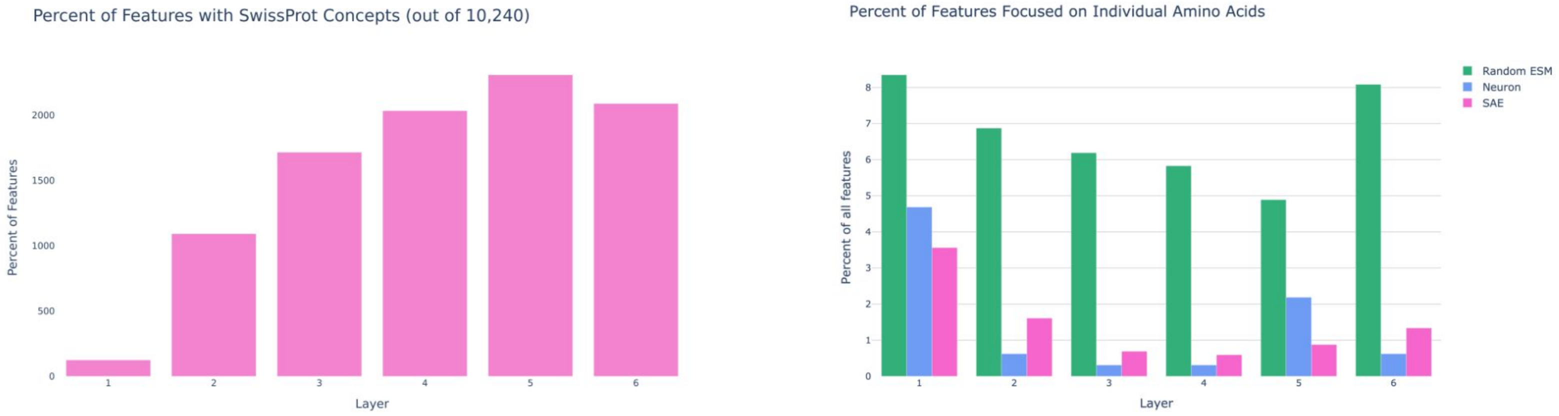
Using Swiss-Prot concept annotations to evaluate biological feature interpretability



a. Highest F1 Score per Swiss-Prot Concept

b. Number of Features with Concept F1 Scores > 0.5

SAE features have stronger associations with Swiss-Prot concepts than ESM neurons

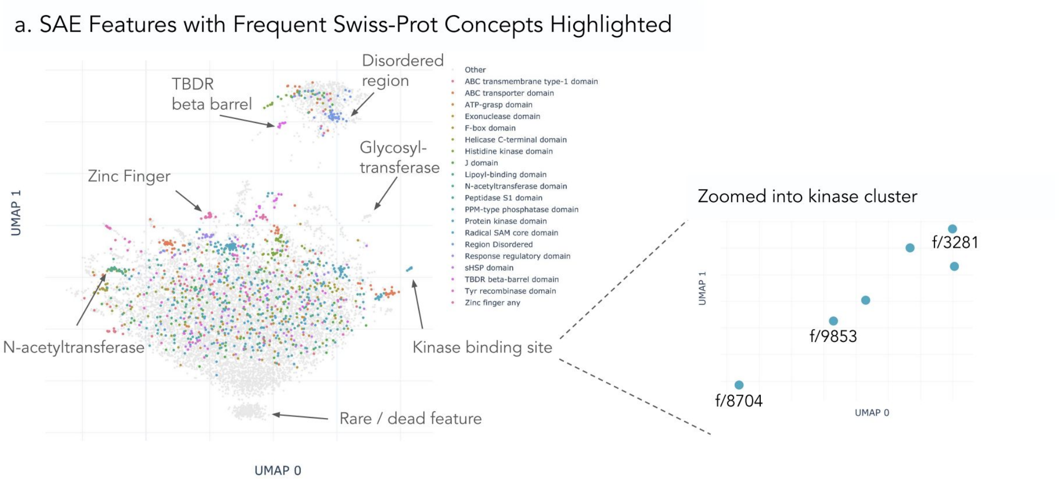


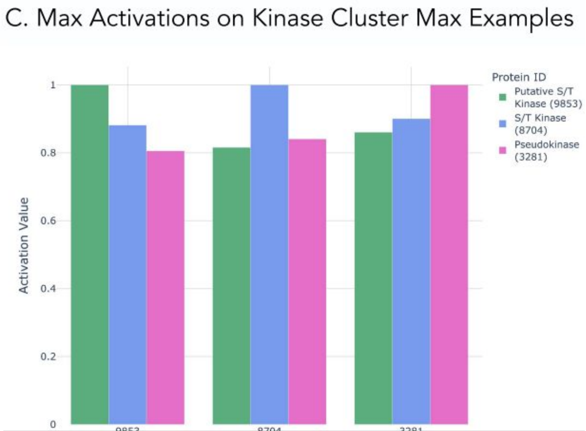
(a) Percent of SAE features per layer that are associated with any Swiss-Prot concept with F1 > 0.5

(b) Percent of features (or neurons) in each layer with F1 > 0.5 to an individual amino acid type.

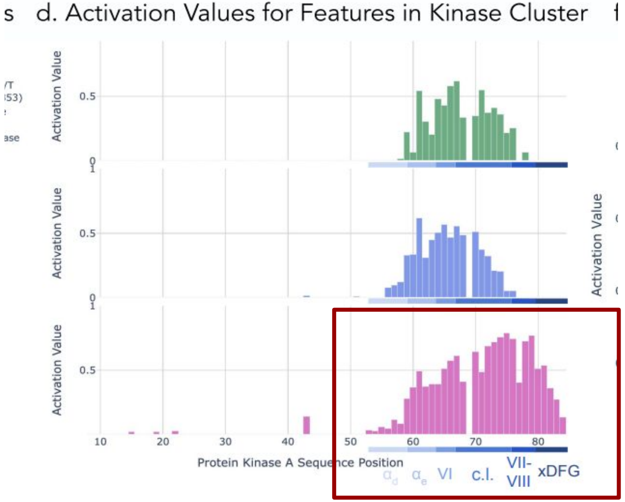
Figure 9: Additional feature-concept analysis across layers

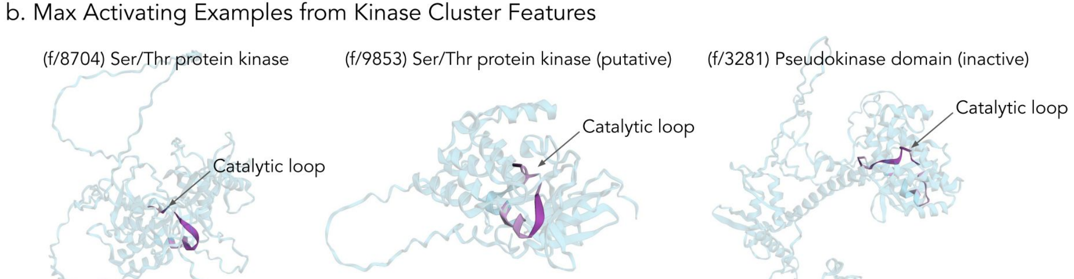
Clustering reveals groups of features with similar functional and structural roles but subtle differences in activation patterns.





variance in peak activation positions



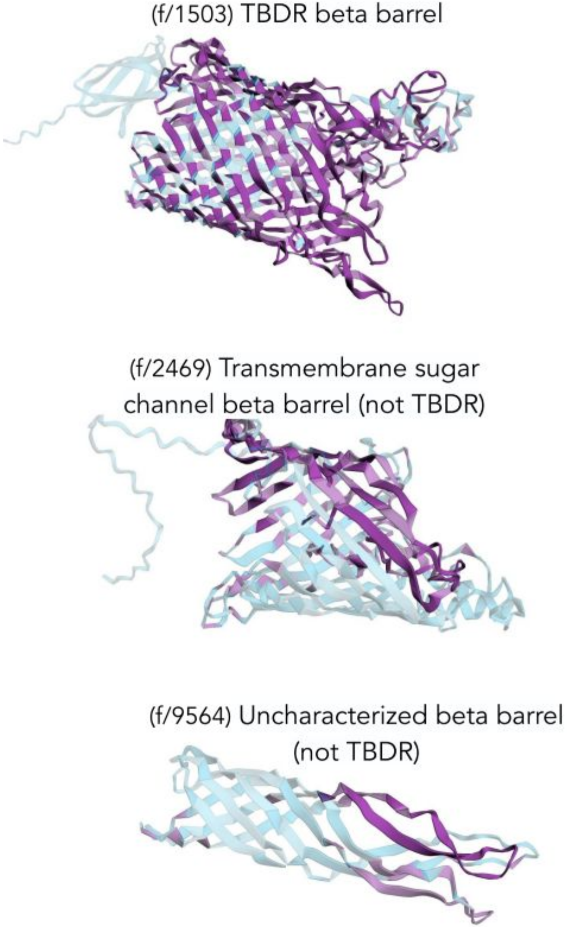

b. Max Activating Examples from Kinase Cluster Features
(f/8704) Ser/Thr protein kinase
(f/9853) Ser/Thr protein kinase (putative)
(f/3281) Pseudokinase domain (inactive)
Catalytic loop
Catalytic loop
Catalytic loop

subtle variations in their spatial preferences

Clustering reveals groups of features with similar functional and structural roles but subtle differences in activation patterns.

TBDR beta barrels features with varying specificity

e. Max Activating Examples from TBDR Beta Barrel Cluster Features

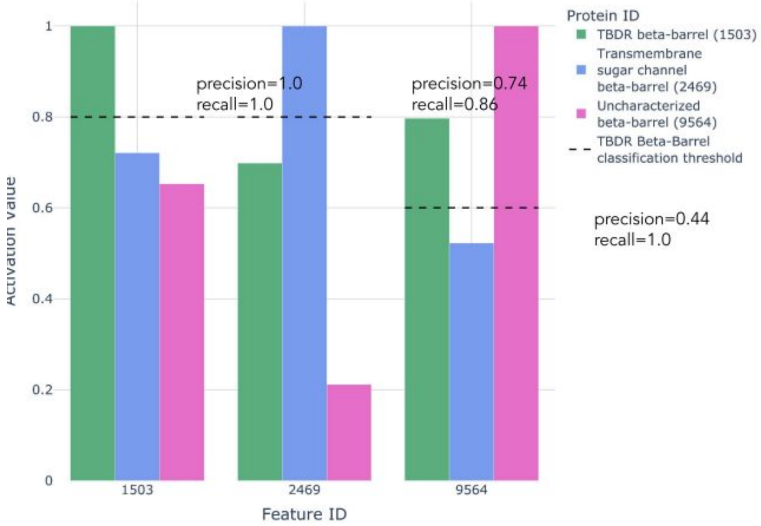


(f/1503) TBDR beta barrel

(f/2469) Transmembrane sugar channel beta barrel (not TBDR)

(f/9564) Uncharacterized beta barrel (not TBDR)

f. Max Activations on TBDR Beta Barrel Max Examples



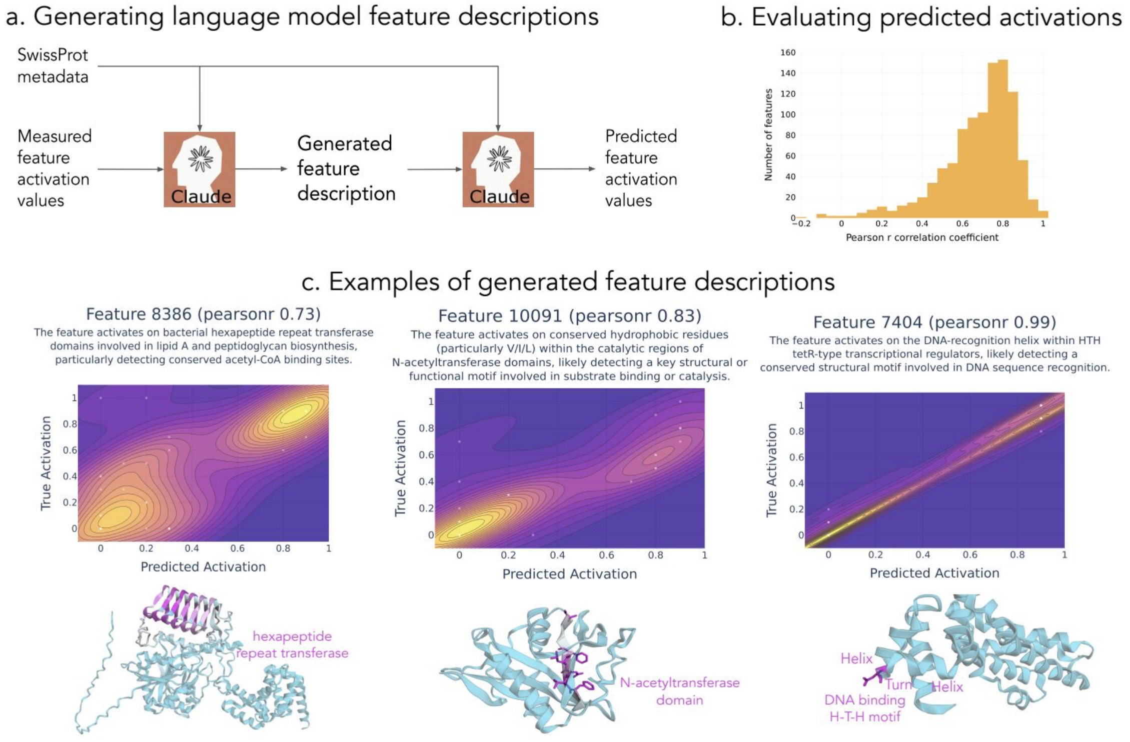
Language models can generate automatic feature descriptions for SAE features.

Generating Descriptions

Used Claude-3.5 Sonnet (new) with Swiss-Prot concept data and protein examples with varying feature activation levels to generate descriptions of protein and amino acid traits driving feature activation.

Validation

Predicted feature activation levels on separate proteins using model-generated descriptions and Swiss-Prot metadata, achieving a high correlation with actual activation (median Pearson r = 0.72).



a. Generating language model feature descriptions

b. Evaluating predicted activations

c. Examples of generated feature descriptions

Feature 8386 (pearsonr 0.73)

The feature activates on bacterial hexapeptide repeat transferase domains involved in lipid A and peptidoglycan biosynthesis, particularly detecting conserved acetyl-CoA binding sites.

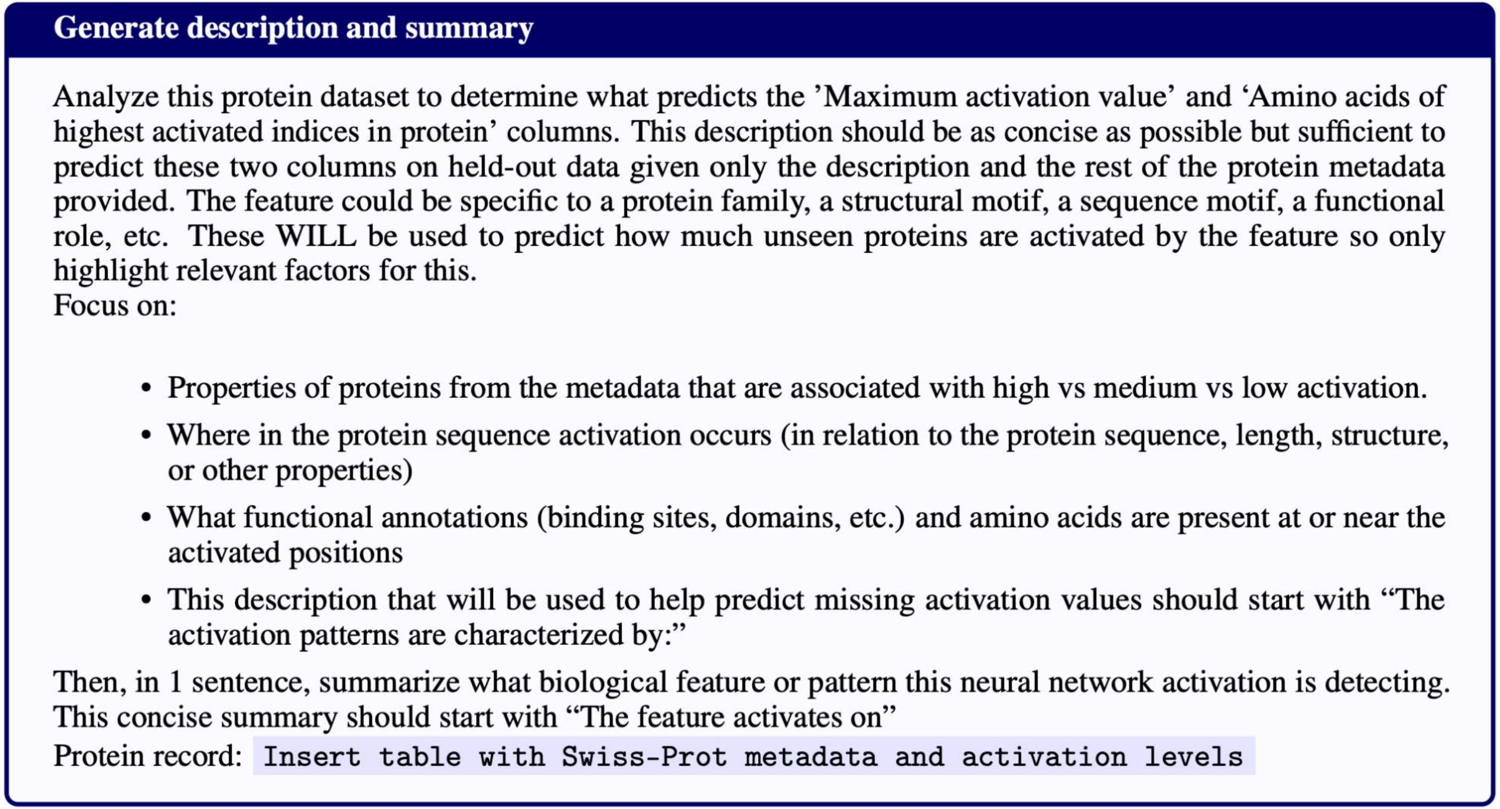
Feature 10091 (pearsonr 0.83)

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.

Feature 7404 (pearsonr 0.99)

The feature activates on the DNA-recognition helix within HTH tetR-type transcriptional regulators, likely detecting a conserved structural motif involved in DNA sequence recognition.

Language models can generate automatic feature descriptions for SAE features.



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.

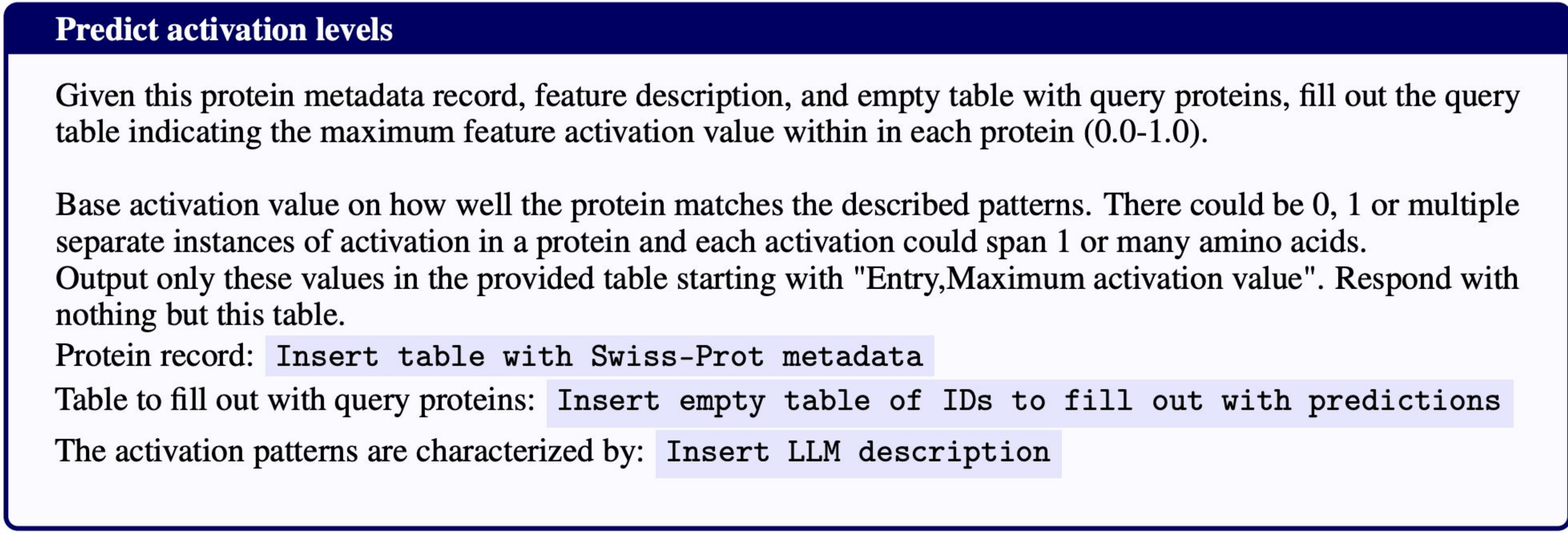
Focus on:

- 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`

Language models can generate automatic feature descriptions for SAE features.



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`

C.2.2 Structural Features

Field Name	Full Name
ft_act_site	Active Sites
ft_binding	Binding Sites
ft_disulfid	Disulfide Bonds
ft_helix	Helical Regions
ft_turn	Turns
ft_strand	Beta Strands
ft_coiled	Coiled Coil Regions
ft_non_std	Non-standard Residues
ft_transmem	Transmembrane Regions
ft_intramem	Intramembrane Regions

C.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

C.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

C.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

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

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

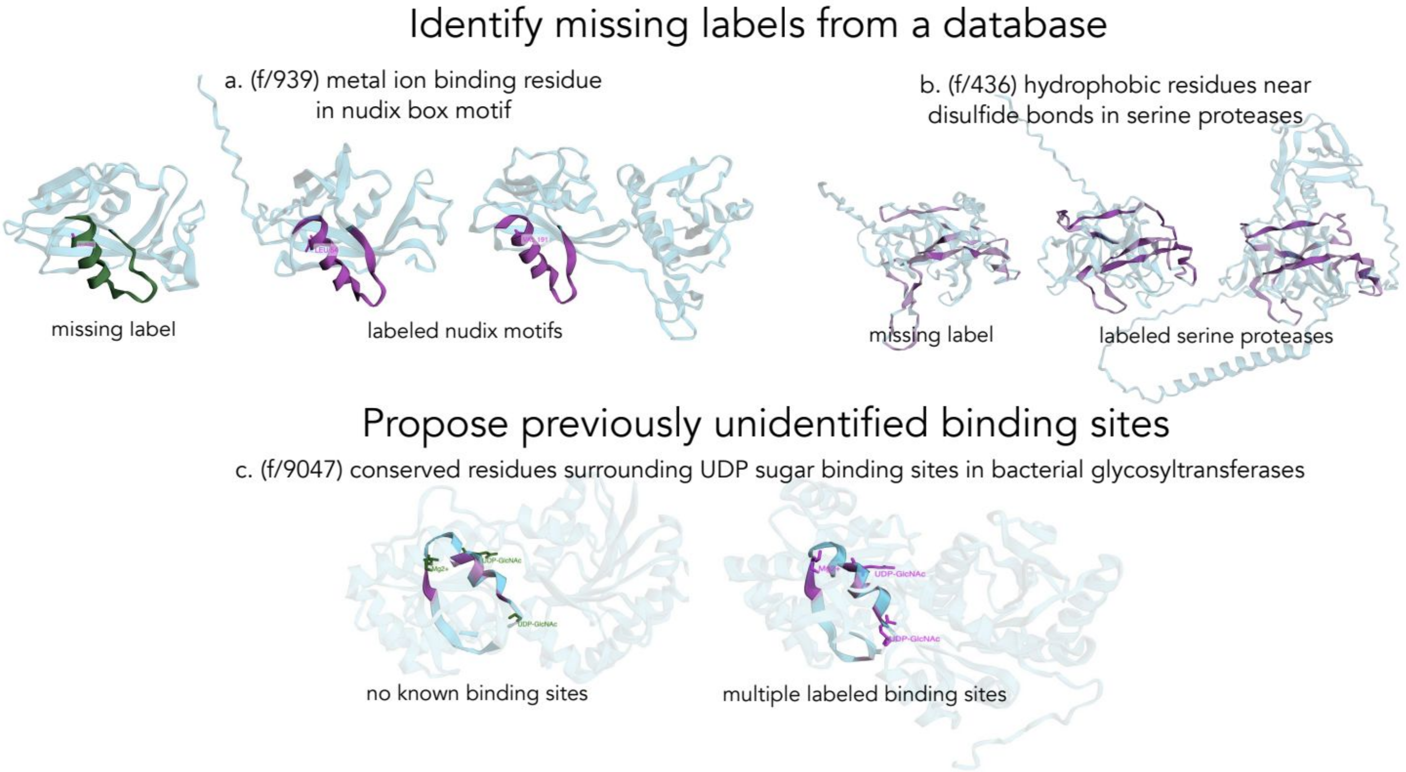
Table 9: Swiss-Prot Concepts associated with SAE features in any layer (Part B). * Indicates concept that is also associated with a neuron in any layer.

Domain [FT]

- 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

Table 8: Swiss-Prot Concepts associated with SAE features in any layer (Part A). * Indicates concept that is also associated with a neuron in any layer.

Feature activation patterns can be used to identify missing and new protein annotations.



Identify missing labels from a database

a. (f/939) metal ion binding residue in nudix box motif

missing label

labeled nudix motifs

b. (f/436) hydrophobic residues near disulfide bonds in serine proteases

missing label

labeled serine proteases

Propose previously unidentified binding sites

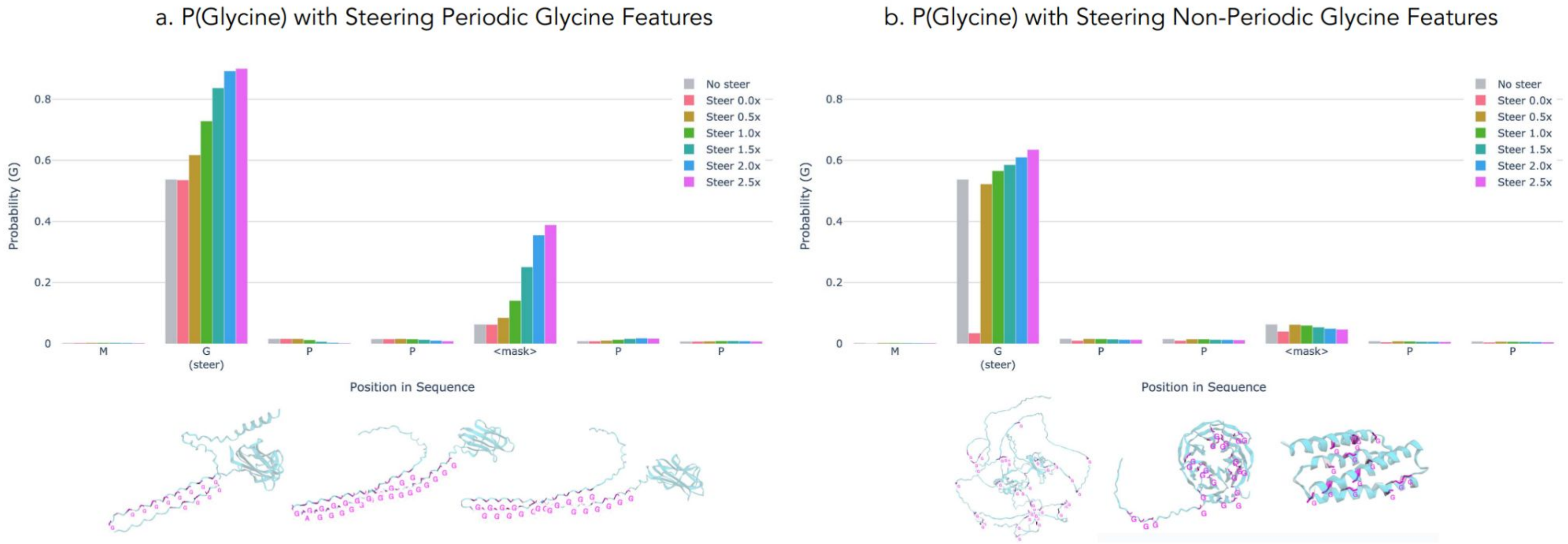
c. (f/9047) conserved residues surrounding UDP sugar binding sites in bacterial glycosyltransferases

no known binding sites

multiple labeled binding sites

Protein Sequence Generation Can be Steered by Activating Interpretable Features

Investigated the impact of steering features activating on periodic glycine (G) repeats (e.g., GXX in collagen-like regions).



a. P(Glycine) with Steering Periodic Glycine Features

b. P(Glycine) with Steering Non-Periodic Glycine Features

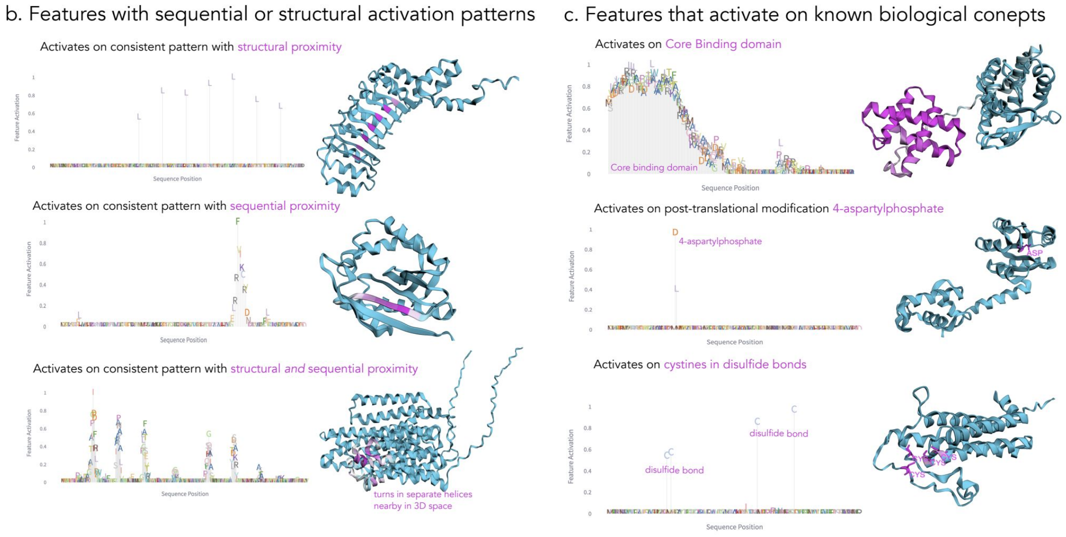
Exploring Features with InterPLM.ai

Sequential vs Structural Activation Patterns

Uncover how features capture local sequence motifs and long-range structural relationships.

Protein Coverage Analysis:

Differentiate between features highlighting specific local properties and those representing broader domain-level patterns.



b. Features with sequential or structural activation patterns

c. Features that activate on known biological conepts

Activates on consistent pattern with structural proximity

Activates on consistent pattern with sequential proximity

Activates on consistent pattern with structural *and* sequential proximity

Activates on Core Binding domain

Activates on post-translational modification 4-aspartylphosphate

Activates on cystines in disulfide bonds

Feature Similarity

Visualized through UMAP, enabling clustering and comparison of feature behavior.

Alignment with Swiss-Prot Concepts

Explore how features correspond to known Swiss-Prot annotations.

Characterizing features:
Sequential and Structural Properties

High-Activation Region Identification

Focused on regions with activation >0.6 in proteins with AlphaFold structures.

Sequential Clustering: Calculated mean activation within ±2 sequence positions of the highest activation residue.

Structural Clustering: Calculated mean activation of residues within 6Å in 3D space.

Null Distributions:

Generated by averaging 5 random permutations per protein.

Significance Testing:

Assessed clustering significance using paired t-tests and Cohen's *d* effect sizes, sampling 100 proteins per feature.

Features with fewer than 25 valid examples were excluded.

Conclusion

- Training SAEs on ESM-2 embeddings revealed up to **2,548 human-interpretable features per layer**, strongly ***correlating with 143 biological concepts*** (e.g., binding sites, structural motifs, functional domains).

- The disparity in interpretability (46 neurons vs. 2,548 SAE features per layer aligned with Swiss-Prot concepts) highlights evidence of information storage in ***superposition within pLMs.***

- ESM-2 captures coherent concepts ***beyond existing annotations.***
- Proposed pipeline uses language models to interpret novel latent features learned by SAEs.

Conclusion

- SAEs trained on **randomized pLMs extract amino acid-specific features** but *fail to identify complex biological concepts*, emphasizing the importance of learned weights.

- This aligns with findings that SAEs capture both the **underlying data distribution** and properties of the model, as shown in randomized language models.

- While SAEs reveal learned patterns, further work is needed to map **how features combine into interpretable circuits** for tasks like 3D contact prediction, binding site detection, and allosteric site identification.