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

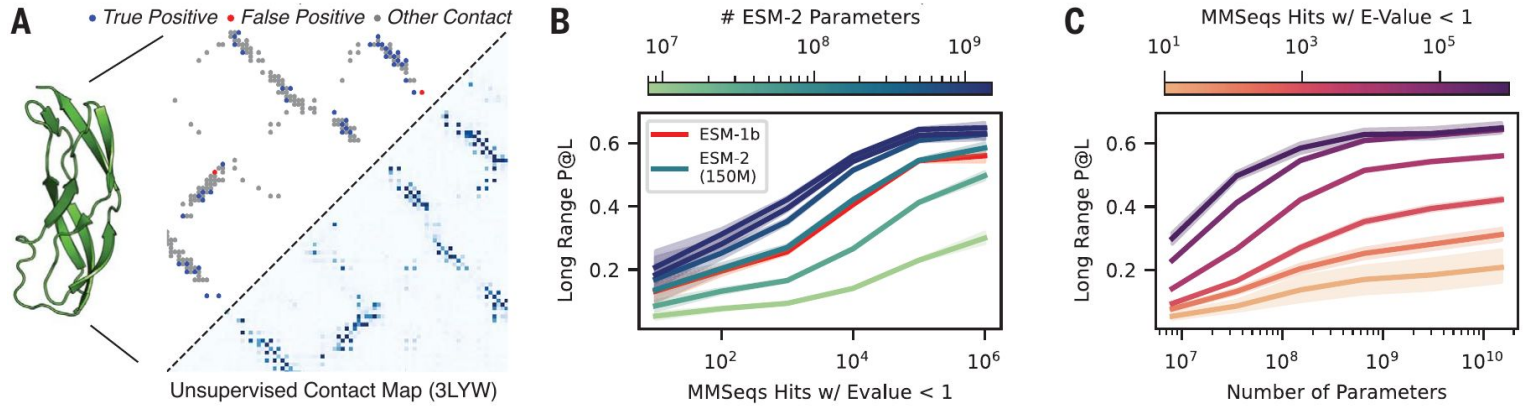
Elana Simon and James Zou (Stanford University)

Nature Methods 22, 2107–2117 (2025)

Jina Kim

Motivation

- **PLMs learn structure from sequence alone:** ESM-2 attention matches 3D contacts, improving with scale
- **Open question:** probing shows information is present, not *which concepts* the model represents
- **Physics or evolution?** Zhang et al. (PNAS 2024): PLMs mainly store coevolution statistics of motifs
- **Why it matters:** debugging models, biological discovery, controllable protein design



Lin et al., *Science* 2023, Fig. 1A–C: unsupervised contact map; contact precision vs. evolutionary depth and vs. parameters

Problem: Neurons Are Polysemantic

- **Polysemantic neurons:** in Inception v1, one neuron fires on cat faces and car fronts; in a small language model, one neuron fires on academic citations, English dialogue, HTTP requests and Korean text
- **Superposition hypothesis:** a network represents more features than it has neurons, each feature a direction (a combination of neurons); feature sparsity makes this possible
- **Sparse autoencoders (SAEs)** learn an overcomplete, sparse dictionary of these directions; in a one-layer transformer, SAE features were far more interpretable than neurons and could steer generation (Bricken et al. 2023)

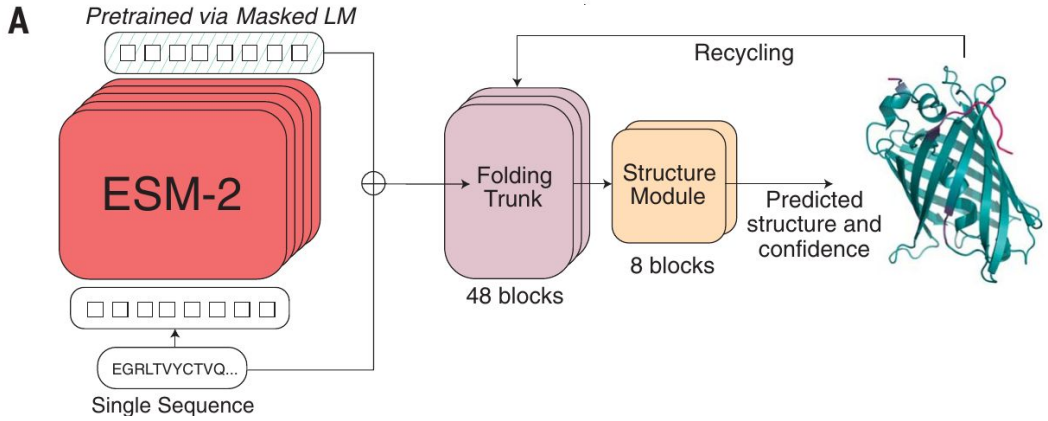
InterPLM's research questions

1. Do ESM-2 neurons show superposition, and can SAEs recover biologically meaningful features?
2. How can interpretability be measured against known biology?
3. Can features be described automatically, reveal missing annotations, and steer the model?

Contributions

1. **Method:** SAEs trained on every layer of ESM-2-8M and six layers of ESM-2-650M
2. **Evaluation:** domain-adjusted F1 against 433 Swiss-Prot concepts, with neuron and shuffled-weight baselines
3. **Main finding:** concepts are stored in superposition: up to 2,309 aligned features vs. 46 neurons per layer; 143 vs. 15 concepts; ESM-2-650M captures 427 concepts
4. **Interpretation at scale:** LLM-written feature descriptions, validated by predicting held-out activations (median r = 0.72)
5. **Applications:** identifying missing database annotations and steering masked-token predictions
6. **Resources:** InterPLM.ai dashboard, open-source code, and pretrained SAEs

Background: ESM-2



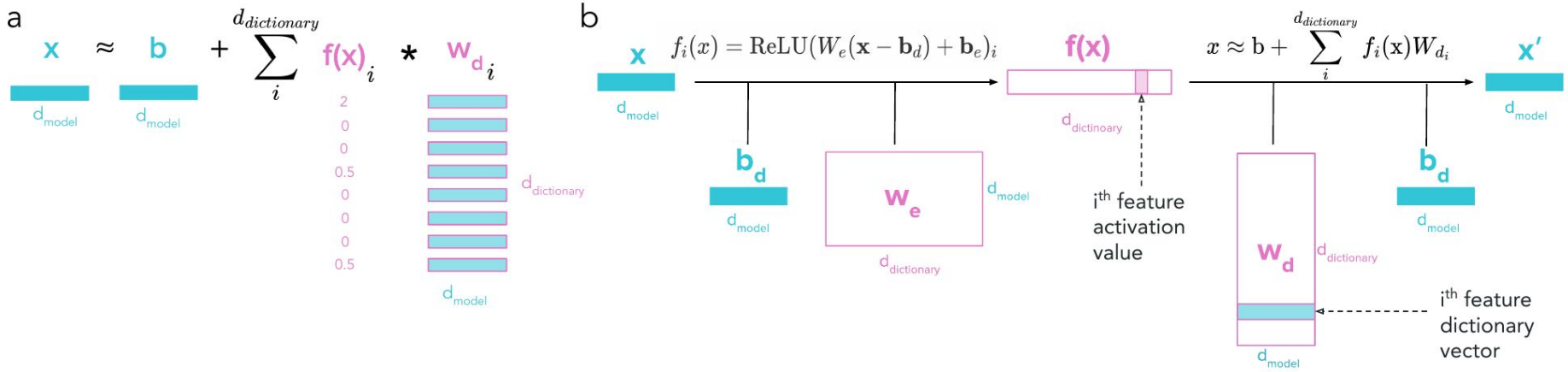
Lin et al., *Science* 2023, Fig. 2A: ESMFold is built on ESM-2

- Masked language model: predicts randomly masked residues (15%) from their context
- Trained on UniRef: ~43M UniRef50 clusters sampled from ~138M UniRef90 sequences
- Family spans 8M to 15B parameters

Model	Layers	Embedding width	Layers with SAEs in InterPLM
ESM-2-8M	6	320	1–6 (main analysis)
ESM-2-650M	33	1,280	1, 9, 18, 24, 30, 33

Feature IDs (f/#####) refer to ESM-2-8M layer 4 unless a layer is given (L#).

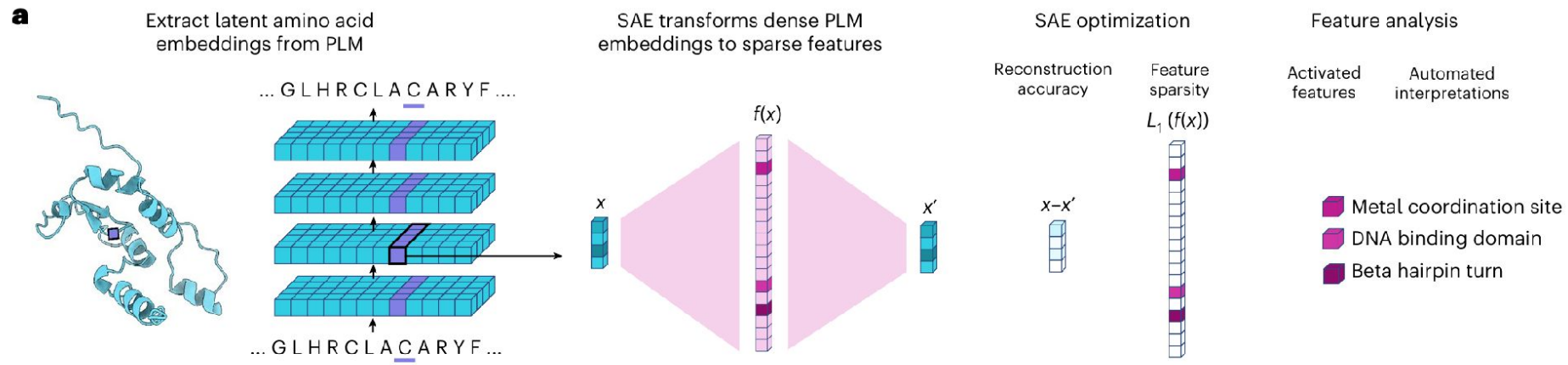
Background: Sparse Autoencoders



Simon & Zou, Supplementary Fig. 1: (a) decomposition into dictionary elements; (b) SAE architecture

- **Decomposition:** $x \approx b + \Sigma_i f_i(x) \cdot d_i$, where $f_i(x) \geq 0$ is feature i's sparse activation and d_i its dictionary vector
- **Loss:** reconstruction MSE + $\lambda \cdot L1(f(x))$; architecture from Bricken et al. 2023
- **Size:** ESM-2-8M 320 → 10,240 features (32×); ESM-2-650M 1,280 → 10,240 (8×)

Method: InterPLM Pipeline



Simon & Zou, *Nat. Methods* 2025, Fig. 1a

- Extract each residue's embedding after a transformer block (excluding <cls> and <eos>)
- Train one SAE per layer to reconstruct it through a wide, sparse feature layer
- Analyze features: activation patterns, Swiss-Prot concept associations, automated LLM descriptions

Method: SAE Training Details

- **Data:** 5M random UniRef50 sequences (part of ESM-2's training data); batch size 2,048; 20 SAEs per layer
- **Model selection:** best mean top-F1 on Swiss-Prot concepts (20% of validation, 135 concepts), not just sparsity
- **Normalization:** each feature scaled to [0, 1] by its max over 50,000 Swiss-Prot proteins

Setting	ESM-2-8M	ESM-2-650M
Expansion	32× (320 → 10,240)	8× (1,280 → 10,240)
Training steps	500,000	200,000
L1 penalty range	0.07–0.2	0.04–0.2
L0 of selected SAEs	100–178	50–211
% loss recovered	98.94–100	92.42–100

L0 and % loss recovered: Supplementary Table 1.

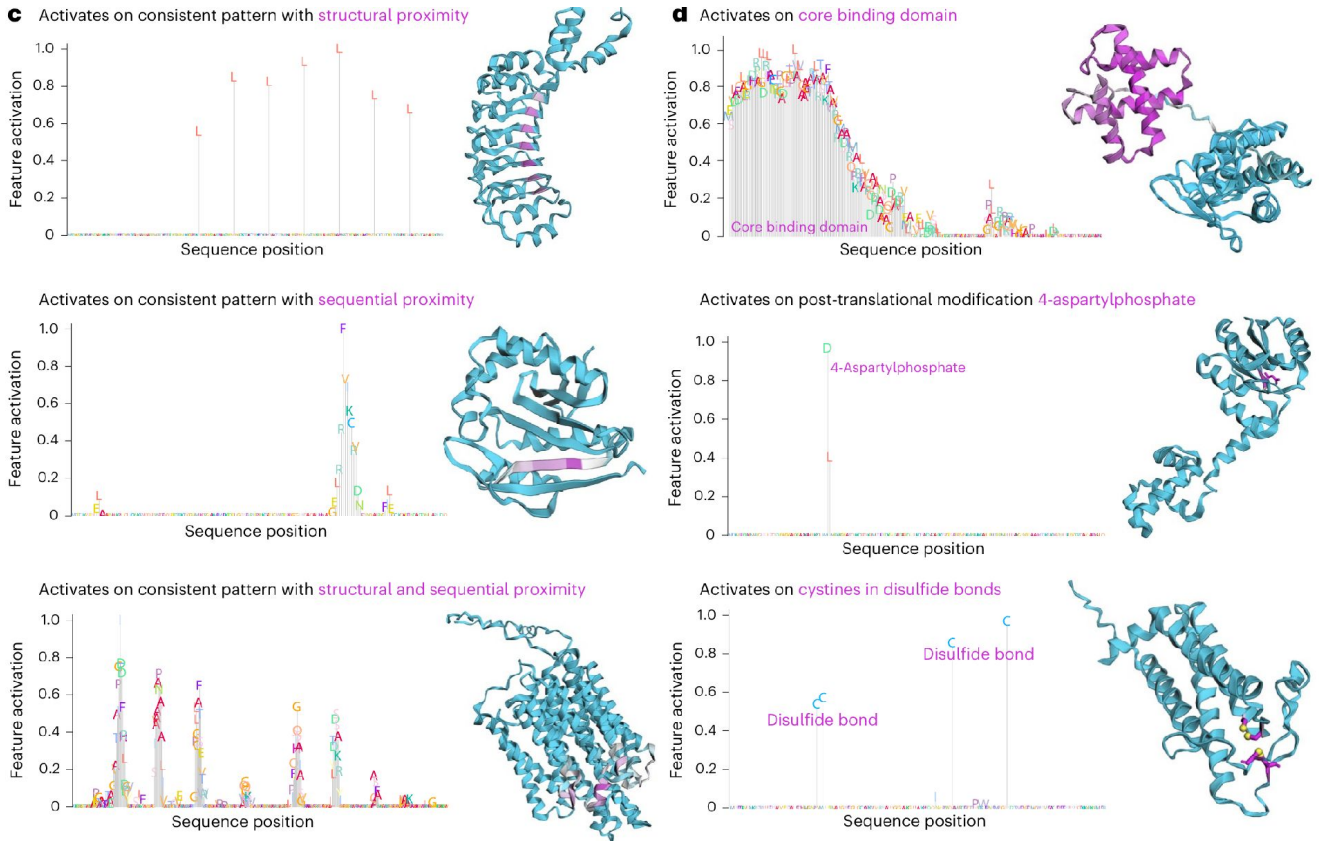
Method: Evaluating Features Against Swiss-Prot

- 50,000 Swiss-Prot proteins (< 1,024 aa), split 25k validation / 25k test
- Annotations converted to residue-level binary labels; 433 concepts retained
- Features binarized at thresholds {0, 0.15, 0.5, 0.6, 0.8}; best threshold per feature–concept pair
- A feature counts if F1 > 0.5 on both validation and test
- **Baselines:** raw neurons; SAE on ESM-2 with shuffled weights

Domain-adjusted F1

- Precision = TP / (TP + FP), per residue
- Recall = domains with ≥ 1 TP / all domains
- **Why:** features are often finer than annotations. f/7653 marks 2 conserved positions in tyrosine recombinase domains: per-residue recall 0.011, domain recall 1.0

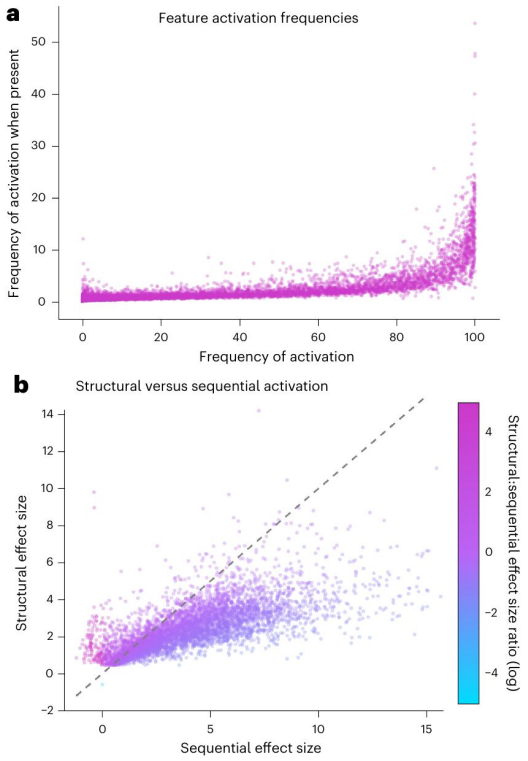
Results: What Features Look Like



Simon & Zou, *Nat. Methods* 2025, Fig. 1c,d

- **c** Activation patterns: structural proximity (L4-f/3147, leucine-rich repeat), sequential proximity (L4-f/10091, β strand) or both (L4-f/67)
- **d** Swiss-Prot concepts: core-binding domain (L4-f/8098), 4-aspartylphosphate (L5-f/7125), cystines in disulfide bonds (L5-f/746)
- Bars: activation per residue; structures: darker pink = stronger

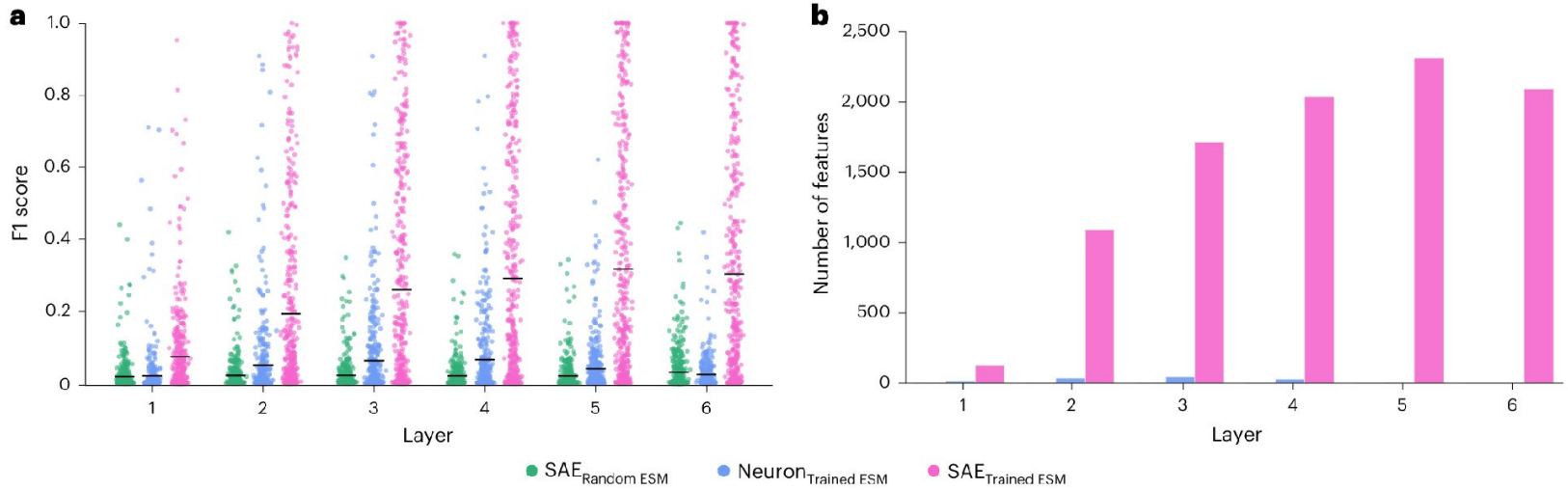
Results: Quantifying Activation Patterns



Simon & Zou, *Nat. Methods* 2025, Fig. 3a,b

- **a** Features range from rare, family-specific to proteome-wide; deeper layers hold more prevalent features
- **b** For each feature's peak residue: mean activation within ±2 positions (sequential) vs. within 6 Å (structural), as Cohen's d against shuffled activations, Bonferroni-corrected
- Points above the diagonal cluster more in 3D than in sequence
- Structural-only features in ESM-2-8M: 129 (layer 1), 179 (layer 3), 359 (layer 6)

Results: SAE Features vs. Neurons (ESM-2-8M)

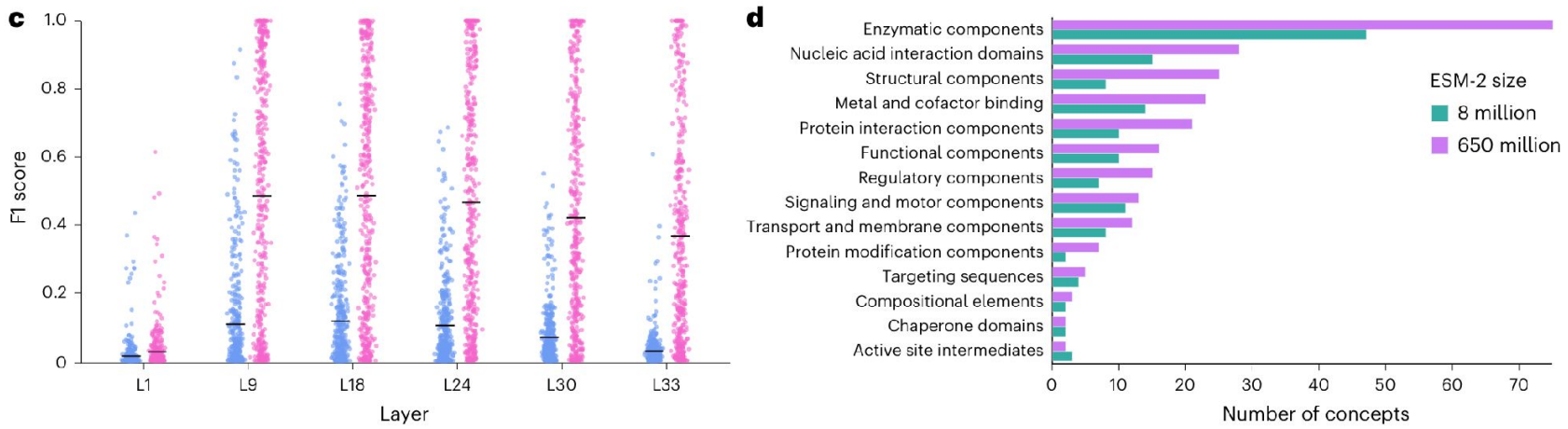


Simon & Zou, *Nat. Methods* 2025, Fig. 2a,b: pink = SAE features, blue = neurons, green = SAE on shuffled-weight ESM-2

- Up to 2,309 features per layer (layer 5) vs. at most 46 neurons reach F1 > 0.5; 143 vs. 15 distinct concepts
- SAE on shuffled-weight ESM-2: zero concept features, though hundreds track amino-acid identity

arXiv v1 (2024) reported up to 2,548 features; numbers here are from the published version.

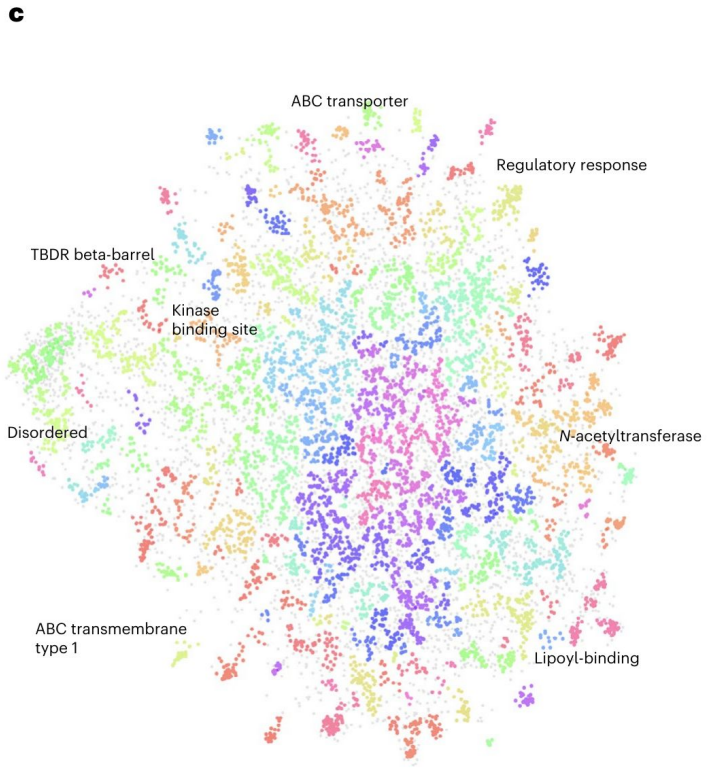
Results: Scaling to ESM-2-650M



Simon & Zou, *Nat. Methods* 2025, Fig. 2c,d

- **c** In every 650M layer, SAE features reach max F1 0.95–1.0 vs. 0.6–0.7 for neurons: superposition persists despite 4× wider embeddings
- **d** Six 650M layers capture 427 concepts vs. 143 in all 8M layers (over 1.7×), with gains in all 14 categories
- Only in 650M: RNA recognition motif (L18), topoisomerase-primase domain (L24), coiled coils (L18, L30)

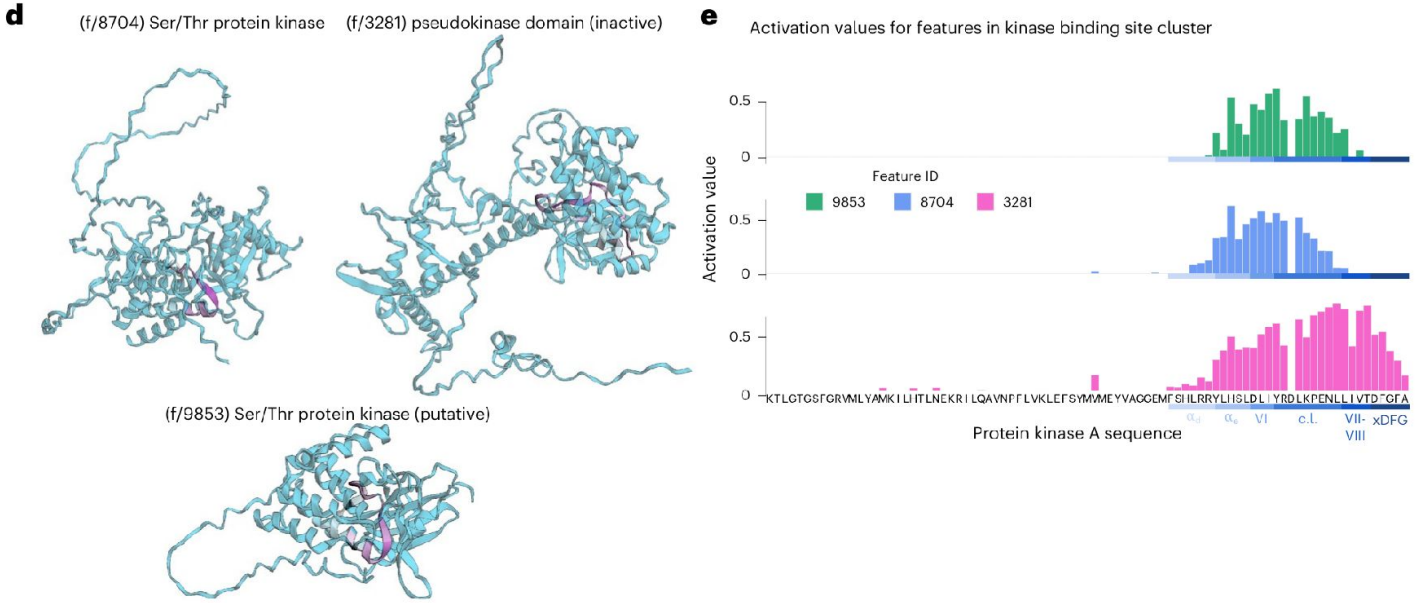
Results: Feature Clusters



Simon & Zou, *Nat. Methods* 2025, Fig. 3c

- UMAP of normalized decoder (dictionary) vectors, cosine metric; HDBSCAN clusters
- Clusters gather features that share a Swiss-Prot concept: kinase binding site, TBDR β-barrel, ABC transporter, disordered regions, N-acetyltransferase
- Within a cluster, features specialize

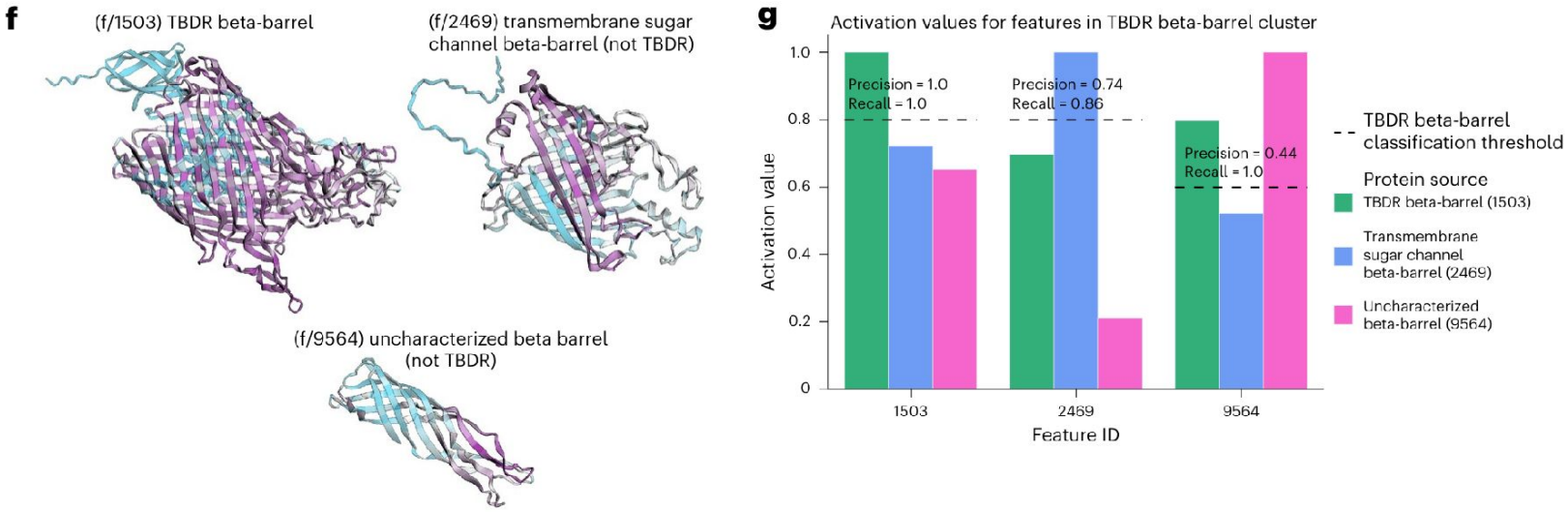
Results: Kinase Binding-Site Cluster



Simon & Zou, *Nat. Methods* 2025, Fig. 3d,e

- Features 9853, 8704 and 3281 fire at different positions around the catalytic loop of protein kinase A
- Each keeps activation above 0.8 on the cluster's top examples: same kinase subtypes, different sub-sites

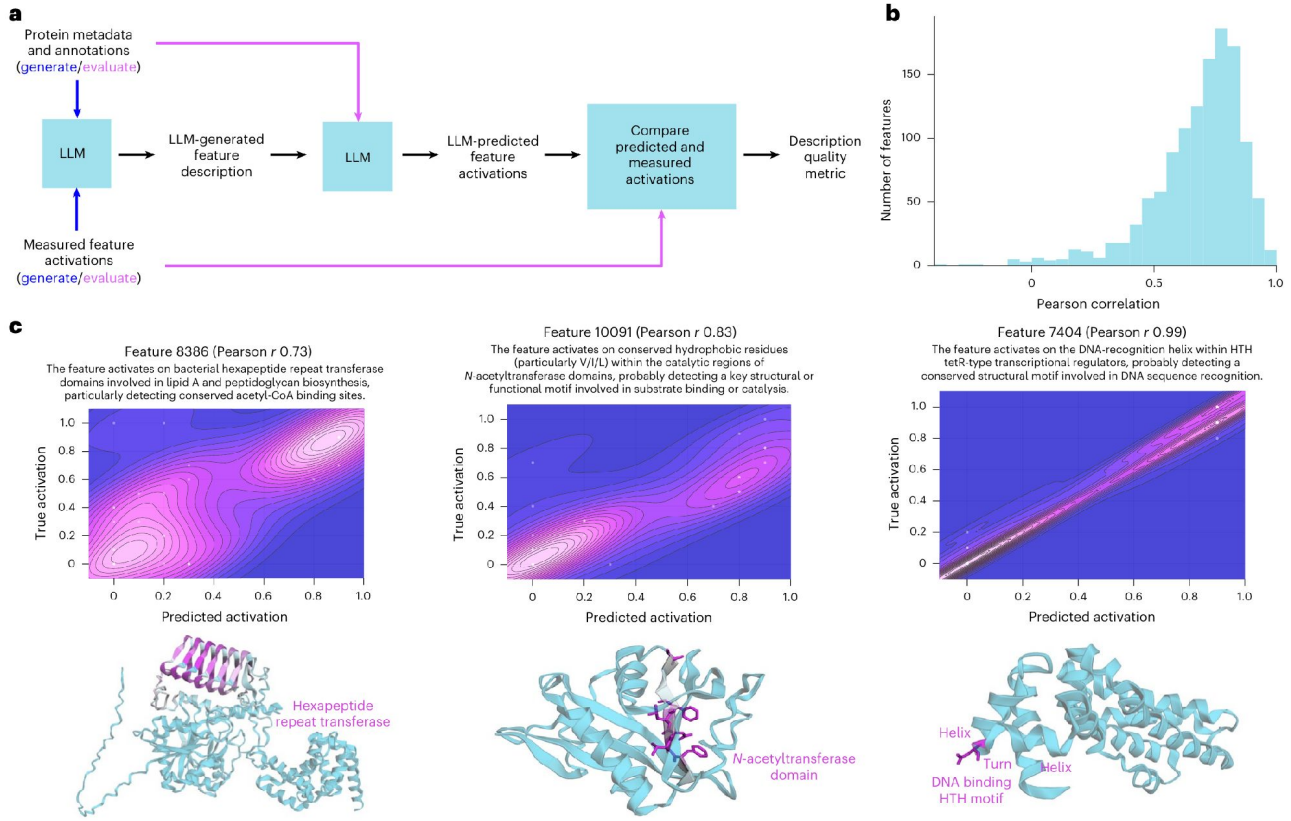
Results: β-Barrel Cluster Hierarchy



Simon & Zou, *Nat. Methods* 2025, Fig. 3f,g

- f/1503 is a specific TonB-dependent receptor (TBDR) barrel detector (F1 = 0.998)
- f/2469 and f/9564 also catch TBDRs (high recall) but fire on other β-barrels (F1 = 0.793, 0.611)
- Specific and general detectors sit close in dictionary space: a hierarchy of β-barrels

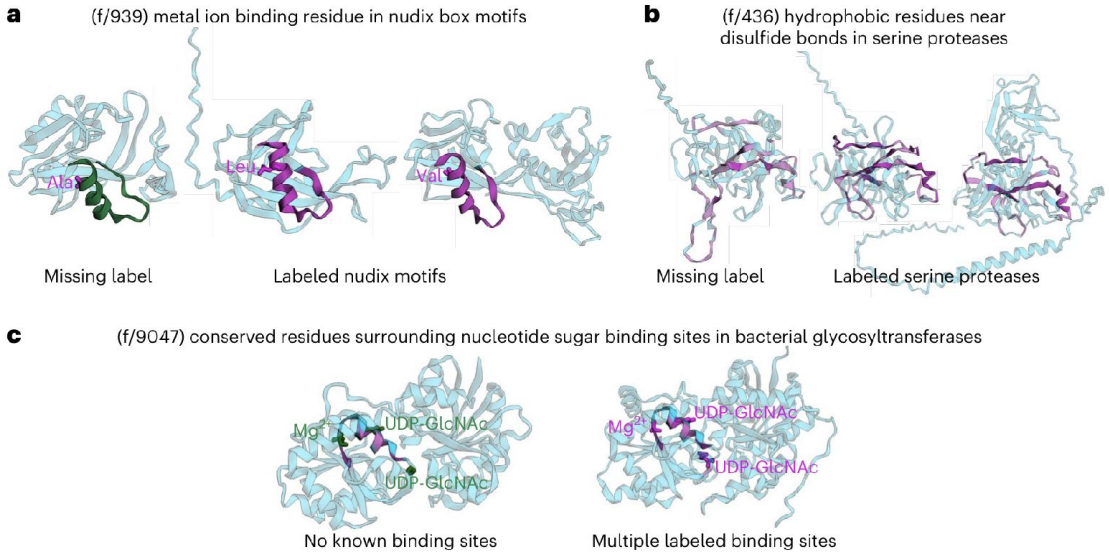
Results: Automated Feature Descriptions



Simon & Zou, *Nat. Methods* 2025, Fig. 4

- Swiss-Prot labels fewer than 20% of features
- 1,240 random features (10%): Claude 3.5 Sonnet (new) sees metadata and activations for 40 proteins, writes a description
- It then predicts max activation on held-out proteins from description + metadata
- **Median Pearson r = 0.72** (b)
- Weakly related to Swiss-Prot F1 (r = 0.11)

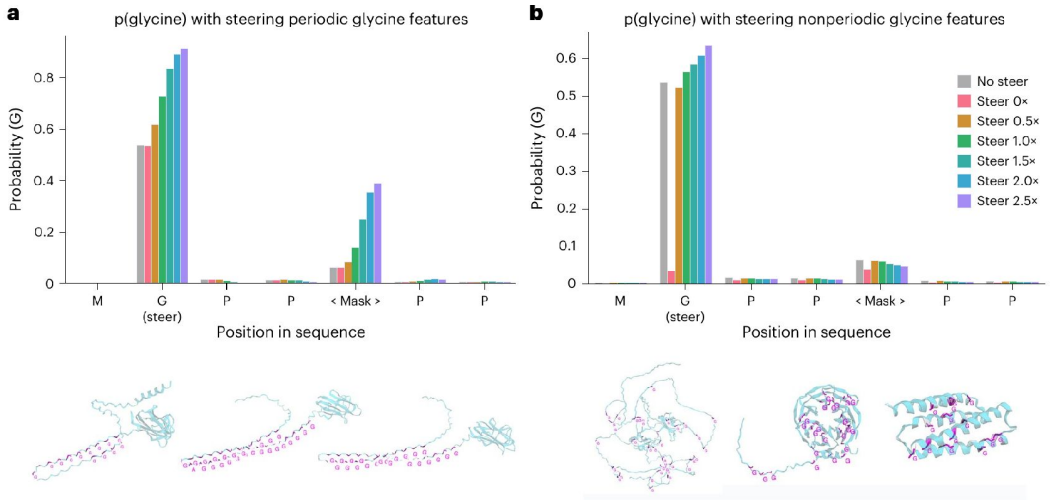
Results: Finding Missing Annotations



Simon & Zou, *Nat. Methods* 2025, Fig. 5: green = implied by the feature; pink = annotated

- **a, b** f/939 (Nudix box) and f/436 (serine proteases) flag proteins missing Swiss-Prot labels; InterPro confirms both
- **c** f/9047 proposes nucleotide-sugar and Mg^{2+} binding sites across distant glycosyltransferases; not yet validated

Results: Steering Model Predictions



Simon & Zou, *Nat. Methods* 2025, Fig. 6

- Clamp features only on the G of MGPP<mask>PP (0–2.5× max activation), keep the SAE error term, read P(G)
- **a** Periodic-glycine features (f/4616, f/4970, f/10003) raise P(G) at the steered G *and* the masked position; **b** glycine-specific features (F1 0.995, 0.990, 0.86) change only the steered position

Strengths

- **Biology-grounded metric:** domain-adjusted F1 against 433 Swiss-Prot concepts, with held-out test sets and neuron and shuffled-weight controls
- **Biological model selection:** SAE hyperparameters chosen by concept alignment, not only sparsity and reconstruction
- **Multiple lines of evidence:** annotation matching, validated LLM descriptions, database curation cases and a causal steering test
- **Robustness checks:** benign mutations change top features by 6.0% on average vs. 8.6% for destabilizing ones (Supp. Fig. 4)
- **Open resources:** InterPLM.ai dashboard, code, and pretrained SAEs for ESM-2-8M and 650M

Limitations and Open Questions

- **Correlation, not mechanism:** concept matches do not establish causal roles; no circuits yet
- **Small models, partial coverage:** mostly the 8M model; 6 of 33 layers for 650M; no structure models
- **Annotation-bound evaluation:** binary residue labels miss quantitative and protein-level properties
- **Stability across retrains:** only 75% of F1 > 0.5 features recur in all three retrained SAEs (Supp. C.3)
- **Unexplored inputs:** masked-token and special-token embeddings
- **Toy steering:** one simple periodic pattern on a few sequences

Related Work

- **SAEs for language models:** Towards Monosemanticity (Bricken et al. 2023); Scaling Monosemanticity (Templeton et al. 2024) introduced steering by clamping features
- **Other protein SAEs:** InterProt: TopK SAEs on ESM-2-650M (Adams et al. 2025); SAEs on ESM2-3B inside ESMFold (arXiv 2503.08764)
- **Other modalities:** SAEs on the DNA language model Evo 2
- **Downstream use:** InterPLM features reused as interpretable inputs, e.g., for neuropeptide precursor prediction (bioRxiv 2026)

Discussion Questions

1. Is “matches a Swiss-Prot label” the right definition of interpretable and what about the >80% of features with no label?
2. Do features reflect protein biology, or memorized family and coevolution statistics? What experiment would tell them apart?
3. What would convincing steering look like for protein design, and how would we validate it?
4. Should PLMs be built to be interpretable, or is post-hoc decomposition enough?

Takeaway: ESM-2 stores many biological concepts in superposition; SAEs plus a biology-grounded metric make them measurable, describable and, in simple cases, steerable.