A186 · Sparse Autoencoders Recover Reproducible Structural Signal in Protein Language Model Latent Space Representations

Presenter: Bridget Liu & Andrew Meng — Columbia University

Authors: Vignesh Karthik, Andrew Meng, Yuna Stechert, Davud Skenderi, Lyla Prasad, Osheen Abraham, Leela Iyer, Adit Anand, AJ Sillato

Session: Day 2 · Fri Oct 2, 2026 · 2:15–4:15 PM

Generative protein-design models produce candidate binders without revealing what internally distinguishes a strong binder from a weak one, limiting rational improvement and failure diagnosis. We apply sparse autoencoders (SAEs), a mechanistic-interpretability technique, to residue-level ESM-C activations of Boltz-designed binder candidates against Vilip-1, a blood biomarker of acute neurological injury, to identify human-interpretable features in these latent representations without modifying any binder sequences. Mixing real, evolutionarily distinct natural-protein sequences into SAE training was the largest lever for generalization beyond synthetic designs alone, raising held-out natural-protein reconstruction fidelity from 0.39 to 0.60 fraction of variance explained. To test whether learned features reflect genuine protein structure rather than training artifacts, we compared two independently trained SAE dictionaries: binder-alone encoding versus encoding with full binder-target cross-attention before target positions were discarded. Despite different encodings, each dictionary's most generic features converge on the same real-protein residues at a rate chance cannot explain (hypergeometric test against the ~7M-residue candidate space, $p \approx 7 \times 10^{-28}$), confirming that cross-attention preserves the learned signal. This convergence survives a shared-seed confound check (5 of 6 agreeing cases use different learned feature indices per dictionary) and is corroborated by InterPro domain annotations and calibrated LLM-assisted labeling that withholds a description when evidence is scattered. Together, these results show that SAE features recovered from protein language model activations carry real, reproducible structural signal, a necessary validation step before using such features to inform or steer binder design.

Sparse Autoencoders Recover Reproducible Structural Signal in Protein Language Model Latent Space Representations

Bridget Liu*†, Vignesh Karthik†, Andrew Meng†, Yuna Stechert, Davud Skenderi, Lyla Prasad, Osheen Abraham, Leela Iyer, Adit Anand, AJ Sillato

[1]Columbia University, New York, NY, USA

*Presenting author, †These authors contributed equally to this work

Motivation

Engineered protein biosensors offer real-time disease detection, but the current practice to develop these biosensors requires extensive human effort to design and validate them. Protein language models (PLMs) are an evolving methodology that alleviates the manual overhead involved in the design process, but generative binder-design models remain black boxes. In their current state, PLMs produce candidate binders without characterizing the internal reasoning that separates a high-affinity binder from a poor one. This opacity limits rational design improvement, failure-mode diagnosis, and cross-disease adaptation.

Approach

We train sparse autoencoders (SAEs) on ESM-C PLM's residue-level activations that we extracted from Boltz-designed binder candidates, to recover human-readable features present in the PLM's underlying latent representation. As proof-of-concept, we validated this pipeline in silico on Vilip-1, a blood-circulating biomarker of acute neurological injury.

Results

- **Best lever: natural-protein data mixing.** Folding real, evolutionarily distinct natural-protein sequences into SAE training alongside synthetic Boltz designs raised held-out natural-protein reconstruction fidelity from 0.39 to 0.60 fraction of variance explained (FVE) – the largest generalization gain of any lever tested, including added binder-design data from additional targets (UCH-L1, B-FABP). The selected checkpoint keeps high fidelity on synthetic designs (0.97 FVE) while nearly doubling generalization to real, unseen proteins.
- **Independent convergence on the same real residues.** Two independently trained SAE dictionaries – one encoding binders alone, one incorporating binder-target co-attention via ESM-C's native chain-break token – converge on the same real-protein residues at a rate statistically inconsistent with chance (hypergeometric test, $p \approx 7 \times 10^{-28}$). Both dictionaries stay healthy at scale: under 0.2% dead features out of 16,384.
- **Convergence maps onto known biology.** Agreeing hotspot residues fall in real InterPro domains (e.g., Collagen, Peptidase_A22A, and Myosin motor/cargo-binding domains), corroborated by InterPro annotations and calibrated LLM-assisted labeling. In one case (Myosin-Vb/MYO5B), both dictionaries independently flag two functionally distinct real sites within the same protein: its motor domain and its cargo-binding domain. Both dictionaries were trained from the same default random seed, so we checked for that confound directly: in 5 of 6 agreeing cases, they rely on *different* learned feature indices to detect the same real residues.

Ongoing & Future Work

This interpretability evaluation is a necessary precursor to our broader goal of steering generative design toward high-affinity binding features, analogous to concept-guided steering in text-to-image diffusion models, rather than blindly sampling the latent space. We plan to extend this pipeline to three clinical targets (UCH-L1, S100B, and B-FABP) and evaluate whether SAE features predict binding-quality metrics (binding confidence, ipTM, ipSAE) directly. We are also wet-lab validating our top 200 binders, selected via our own diversity-ranking method, helping establish the value of mechanistic interpretability in accelerating PLM-informed biosensor design.