Identifying function-relevant, sequence-agnostic features in a protein model using sparse autoencoders

Isha Harris[1,2], Gary Abel[1,3]
ishaharris99@gmail.com, gary@fourtheon.bio

[1] Fourth Eon Bio
[2] University of Cambridge
[3] Johns Hopkins Center for Health Security

Supervised Program for Alignment Research — Spring 2026

Abstract

Biological AI models are powerful tools for detecting biological threats, but it remains unclear whether their predictions generalize to highly non-natural proteins. This is increasingly important as AI-enabled biodesign produces novel sequences that may not be reliably captured by direct sequence-comparison approaches used in DNA synthesis screening. Here, we use sparse autoencoder (SAE) features from protein language model activations to test whether interpretable internal representations can identify structural similarity across generated proteins. We profile residue-level SAE activations across proteins with varying sequence identity and predicted structural similarity to a wild-type reference, and evaluate whether aggregated feature representations distinguish predicted structure-preserving from structure-altered designs. The results show that SAE features can identify structure-preserving proteins even at low sequence similarity, with wild-type-like activation patterns persisting despite substantial sequence divergence. This suggests that interpretability methods may help recover biologically meaningful signals from artificially designed proteins, supporting more robust, function-aware approaches to biological threat detection.

1 Introduction

DNA synthesis screening is an important safeguard against misuse of synthetic biology, but current screening methods rely heavily on direct sequence matching or simple similarity-based approaches [1]. These work best when a risky sequence is already known, or when a novel sequence remains in the sequence neighbourhood of a known example. However, newer generative models may be able to produce proteins that preserve important functional or biophysical properties while differing substantially at the sequence level [2], [3]. This means that AI-generated analogs of sequences of concern could potentially evade sequence-based tools [4]. It is therefore increasingly important to identify protein features that are more directly relevant to structure and function than to direct sequence similarity alone.

While one approach is to directly use the outputs from protein language models (PLMs), another promising avenue is to understand what PLMs have learnt internally. However, their learned representations are not easy to interpret directly, because biological concepts are often represented in superposition rather than cleanly localized to single neurons, making individual units polysemantic [5]. SAEs help address this by decomposing dense model activations into a larger set of sparse features, only a few of which are active for any given input. SAE features have been shown to recover interpretable structural and functional concepts in a range of settings [5], [6], [7]. However, these methods have so far been studied primarily on natural sequences, leaving it unclear whether they remain informative for highly non-natural proteins generated by biological language models.

This project asked whether SAE features capture patterns or signals that generalize across generated proteins with varying degrees of sequence identity and predicted structural similarity to a wild-type (WT) reference. We profiled SAE activations across generated sequence sets of structural analogs of a model protein, and evaluated whether aggregated feature representations could distinguish similarity classes and predict structure- or function-related properties.

2 Methods

We used datasets of foundation model-generated protein variant sequences derived from gpD, a head decoration protein from lambda bacteriophage, a benign virus.

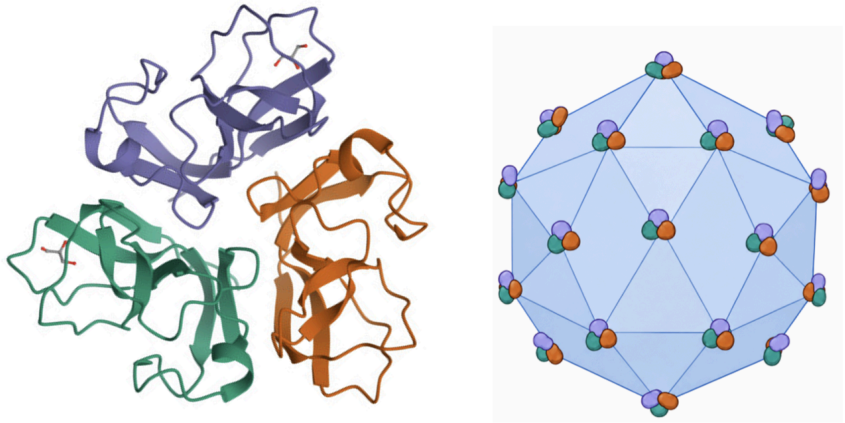

Figure 1. gpD as a model system for generated protein variants. *Wild-type gpD, a head decoration protein from lambda bacteriophage, was used as a benign reference protein. The left panel shows the gpD protein structure; the right panel illustrates its position at the vertices of the phage capsid, for context.*

Variants were generated with ESM3 [8], a masked generative protein language model that jointly models protein sequence, structure, and function, allowing new protein sequences to be sampled while conditioning on specified structural or functional constraints. This produced a set of engineered proteins spanning a range of sequence and structural distances from wild-type gpD.

For each engineered protein sequence, predicted structural similarity was quantified using backbone root mean square deviation (RMSD) in angstroms between the predicted variant structure and the wild-type gpD crystal structure [9]. Sequence identity was defined as the fraction of amino acid positions identical to the wild-type sequence.

We used thresholds on these two metrics to assign variants to cohorts differing in sequence similarity and structural preservation. This design allowed controlled comparisons in which one axis was held approximately constant while the other varied;

for example, comparing structurally preserved and structurally altered variants at similar sequence identity. We refer to these cohorts throughout using compound labels: **highseq/highstruc**, **highseq/lowstruc**, **lowseq/highstruc**, and **lowseq/lowstruc.**

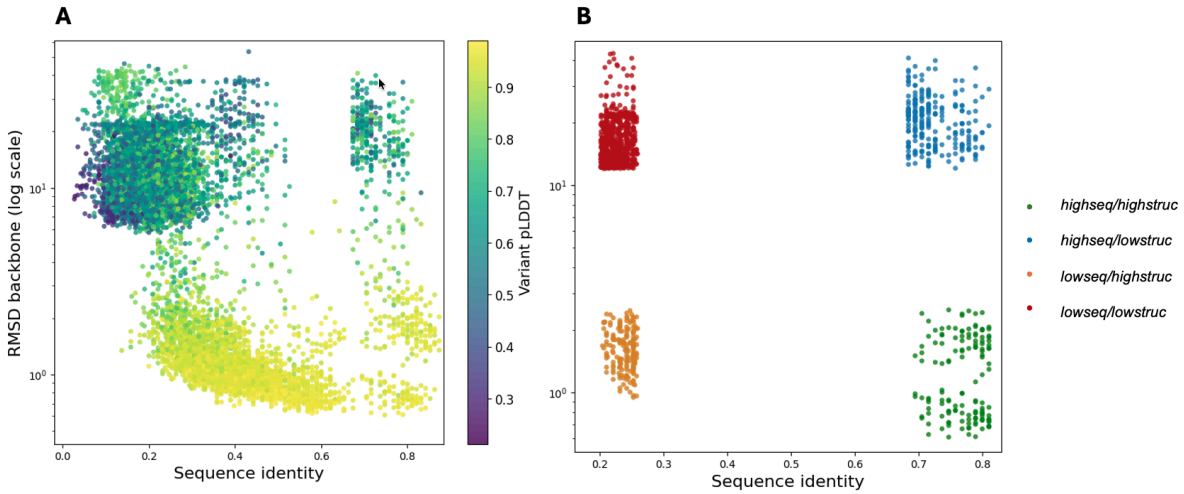

Figure 2. Generated gpD variants decouple structural and sequence similarity. *Each point represents a generated gpD variant, plotted by sequence identity to wild-type gpD and backbone RMSD from the predicted wild-type structure (Å).* ***A:*** *Variants coloured by Predicted Local Distance Difference Test (pLDDT), a per-residue confidence score assigned by the structure prediction model, averaged across the predicted structure.* ***B:*** *Variants assigned to four analysis cohorts using thresholds on sequence identity and backbone RMSD. High-sequence cohorts were defined as 0.68 < sequence identity < 0.82, and low-sequence cohorts as 0.20 < sequence identity < 0.26. High-structure cohorts had backbone RMSD < 2.5 Å, whereas low-structure cohorts had backbone RMSD > 12 Å. These thresholds define the* ***highseq/highstruc****,* ***highseq/lowstruc****,* ***lowseq/highstruc****, and* ***lowseq/lowstruc*** *cohorts, enabling comparisons that vary sequence similarity and structural preservation independently.*

Each sequence was first embedded using ESM-2 [10]. From a specified ESM-2 layer, we extracted the residue-level hidden states after removing special tokens, giving an $L \times d$ embedding matrix for each protein, where L is sequence length and d is the embedding dimension.

To convert these dense residue embeddings into sparse, interpretable representations, we passed them through the pretrained InterPLM [5] sparse autoencoder. InterPLM

provides SAEs matched to specific ESM-2 models and layers: for ESM-2 8M, SAEs are available for all six transformer layers, mapping 320-dimensional residue embeddings to 10,420 sparse features; for ESM-2 650M, SAEs are available for selected layers — layers 1, 9, 18, 24, 30, and 33 — mapping 1,280-dimensional residue embeddings to 10,420 sparse features.

For each residue, the SAE encodes the dense ESM-2 embedding as a high-dimensional sparse activation vector, then reconstructs the original embedding as a linear combination of learned dictionary elements. The SAE is trained to minimize reconstruction error with an L1 penalty on the latent activations, encouraging each residue representation to be captured by a small number of active features.

For each protein variant, this produced a feature-by-position activation matrix, with each entry representing the activation strength of a given SAE feature at a given amino acid position. For the ESM-2 8M InterPLM model used here, this yielded 10,420 SAE features, each evaluated across all 95 residue positions in gpD.

To obtain sequence-level representations suitable for downstream analysis, each SAE feature was summarized across the residue axis. We first computed position-agnostic summaries that captured overall activation magnitude, including maximum activation, mean activation, top-k means, and the fraction of positions active above a range of thresholds.

To retain information about residue-wise activation patterns, we also computed reference-aware summaries by comparing each variant's per-feature activation trace with the corresponding wild-type trace. For each SAE feature, we calculated the Pearson correlation between the variant and wild-type traces across aligned residue positions. This captured the extent to which a feature showed the same positional pattern as in wild-type gpD, rather than simply measuring its overall activation level.

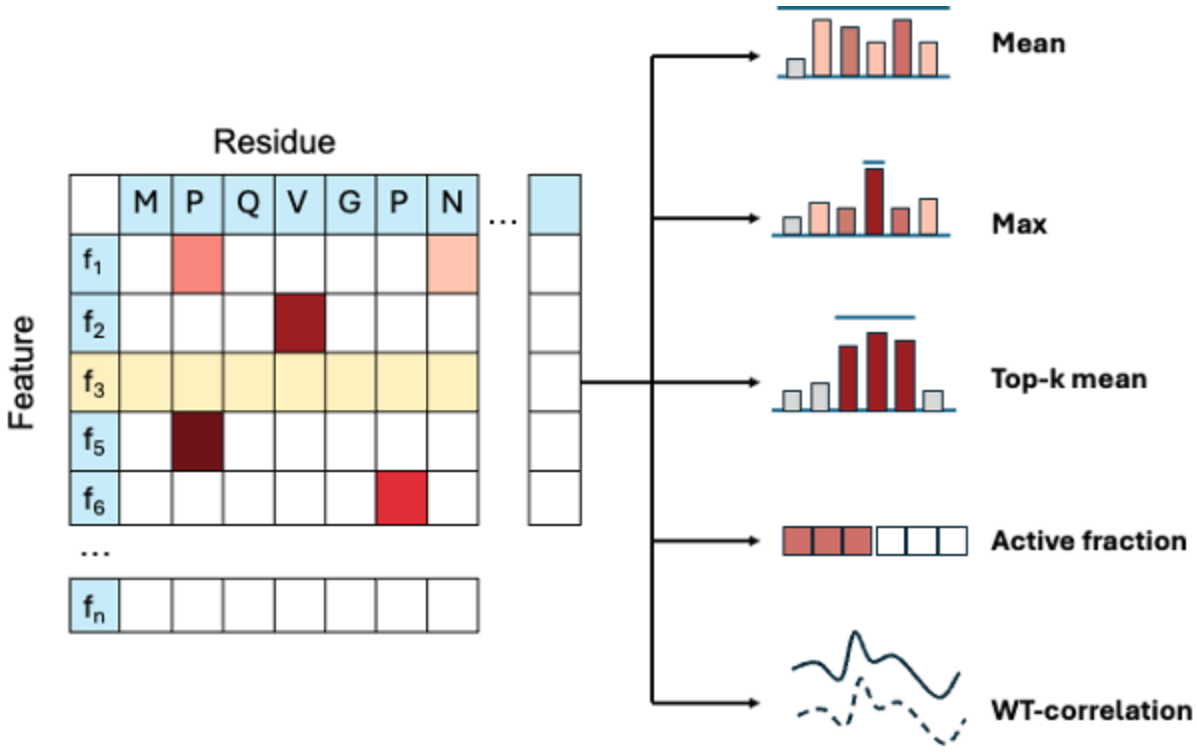

Figure 3. Aggregating residue-level SAE activations into protein-level features. *InterPLM produces a feature-by-residue activation matrix for each protein, where each row corresponds to an SAE feature and each column to an aligned amino acid position. Each feature trace was summarised using activation magnitude metrics, including mean, maximum, top-k mean, active fraction and correlation to wild-type.*

An L1-regularized logistic regression model was trained to distinguish two predefined cohorts, with the structure-preserving cohort treated as the positive class and the structure-disrupted cohort treated as the negative class. In this study, the contrast was restricted to **lowseq/highstruc** versus **lowseq/lowstruc**, allowing comparison of variants with similarly low sequence identity to wild type but contrasting structural outcomes. This tested whether InterPLM activation features contained information associated with preservation of gpD-like structure beyond sequence identity alone.

We evaluated these feature summaries across a range of analysis choices, including the ESM-2 model size, ESM-2 layer, and feature aggregation strategy. The input features were the InterPLM feature summary vectors described above, with separate model-fitting runs performed for each configuration. Features were standardized within

each training fold. The logistic regression model learned a linear combination of SAE-derived features whose weighted sum predicted whether a variant belonged to the positive class. The L1 penalty encouraged sparsity in the fitted coefficients, so that only a subset of informative InterPLM features contributed to the classifier. Model performance was evaluated using repeated stratified cross-validation with 5 folds and 2 repeats. Within each training split, the strength of the L1 penalty was tuned by inner cross-validation, selecting the value that gave the best validation performance.

After model fitting, each variant was assigned a signature score corresponding to the logistic regression predicted probability for the positive class. Scores were computed from the fitted logistic regression model, and averaged across cross-validation folds. Scores ranged from 0 to 1, with higher values indicating greater predicted alignment with the learned InterPLM feature signature of the structure-preserving reference class.

The trained signature was then applied, without retraining, to variants from additional cohorts to test whether the SAE-derived signal generalized beyond the original training contrast. To provide an external background comparison, we also scored a random set of length-matched proteins retrieved from GenBank, allowing signature scores from engineered gpD variants to be compared against unrelated proteins of the same length.

3 Results

WT-correlation captures preservation of residue-level activation patterns

To check whether WT-correlation was a meaningful summary of residue-level feature preservation, we first visually inspected representative activation traces spanning different correlation values. High-correlation variants retained the main WT-like peaks and troughs across the sequence, whereas lower-correlation variants showed visibly disrupted or absent activation structure. This supports using WT-correlation as a compact summary of how closely a feature's positional activation pattern is preserved relative to wild-type gpD.

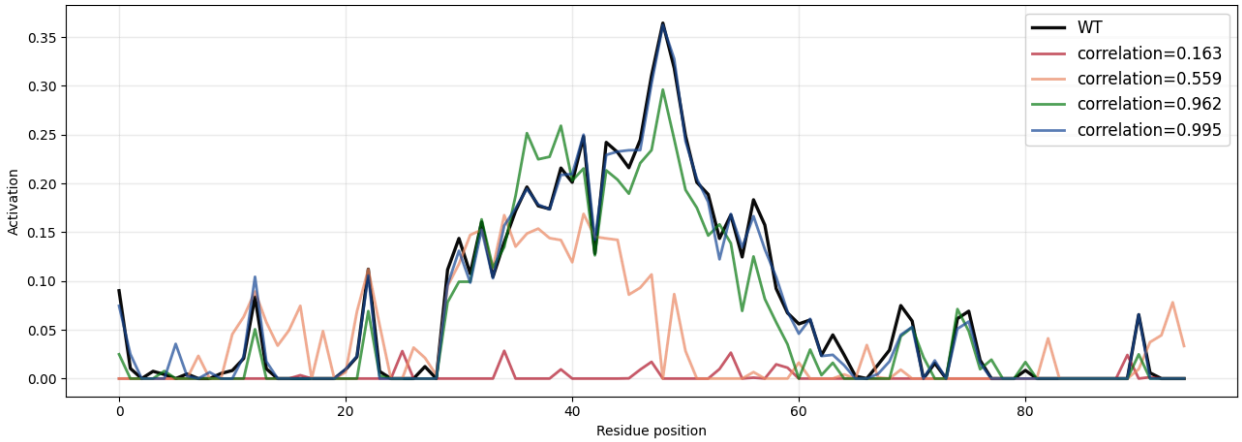

Figure 4. WT-correlation captures preservation of residue-level activation patterns. *Representative activation traces are shown for one SAE feature across generated variants with different correlations to the wild-type trace. Variants with high WT-correlation preserve the main residue-wise peaks and troughs of the wild-type activation pattern, whereas low-correlation variants show visibly disrupted or absent activation structure.*

WT-relative summaries produced coherent classifier signatures

Across the tested ESM-2 configurations, discrimination was consistently high. Mean AUROC values were concentrated near ceiling performance, spanning 0.917-1.00 across the run summaries, with most regimes above 0.98. This first-pass screen therefore showed that many combinations of model layer, aggregation method and cohort contrast could separate the cohorts robustly.

We next asked whether the learned signatures were coherent, that is, whether selected features formed a consistent cohort-level pattern rather than a set of individually predictive but noisy features. In ESM-2 650M layer 30, aggregation by WT-correlation gave similar cohort discrimination to simple activation summaries (mean AUROC 0.969, compared with 0.964 for max and 0.972 for mean), but produced a clearer cohort-level pattern, with stronger separation between groups and less variation within each group. This suggests that preserving the residue-wise organisation of SAE feature activations is more informative than activation strength alone.

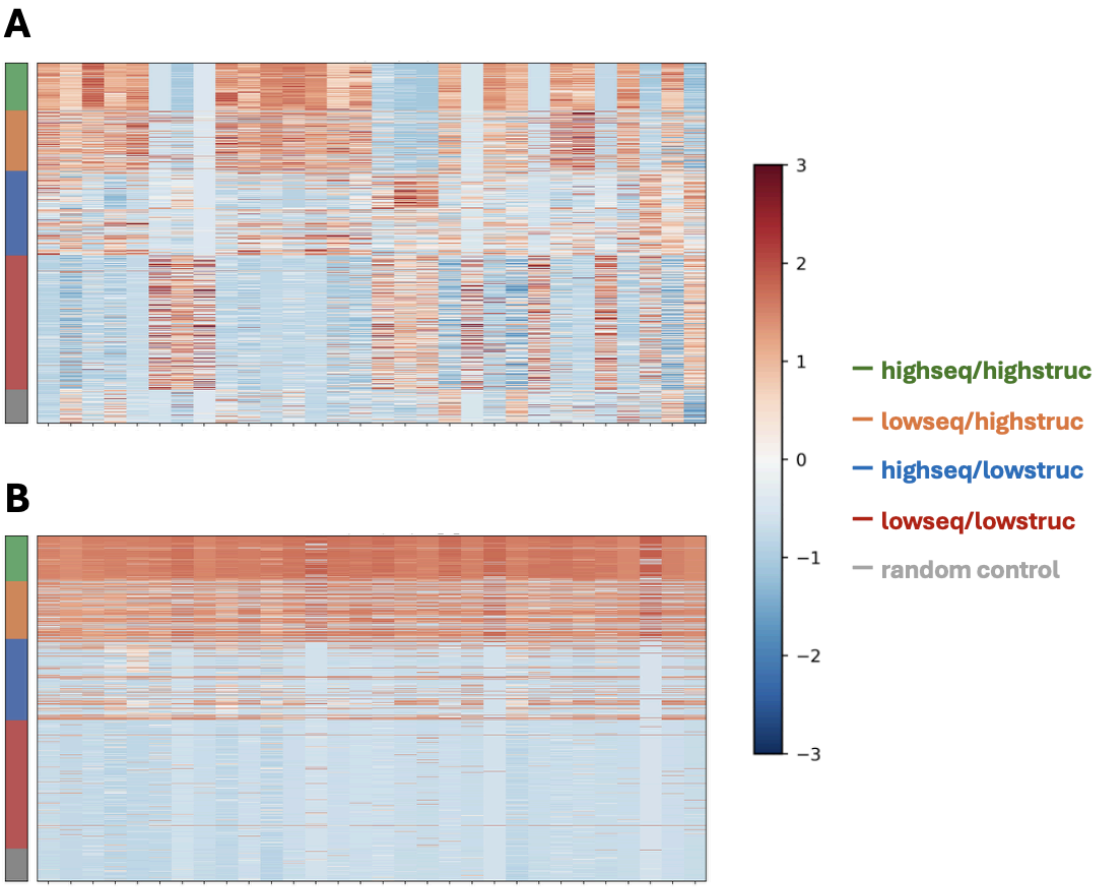

Figure 5. WT-correlation produces a more coherent cohort-level signature than activation magnitude alone. *Heatmaps show the 30 most predictive InterPLM features, selected by the logistic regression classifier across gpD variant cohorts and random length-matched controls. Rows represent variants grouped by cohort, with cohort membership indicated by the coloured side bars; columns represent selected SAE features.* ***A:*** *Max-pooled feature activations across residue positions.* ***B:*** *Correlation between each variant's residue-level feature activation pattern and the corresponding wild-type gpD pattern. Values are z-scored across variants and clipped at ±3 for visualisation. Both summaries separate the training cohorts, but max activation produces a noisier and less structured pattern across variants. In contrast, WT-correlation yields a clearer cohort-level signature, with structurally similar variants showing more consistently WT-like feature patterns and structurally divergent or random-control proteins showing lower alignment.*

On this basis, subsequent mechanistic analyses focused on WT-correlation features from layer 30 of ESM-2 650M.

A WT-relative signature trained on low-sequence variants transfers to high-sequence structure-preserving variants

We applied the learned signature, defined as the logistic regression predicted probability for the positive class, to the cohorts not seen during training. The aim was to test whether a score learned from low-sequence variants could recover the same structural signal in variants with substantially higher sequence similarity to wild-type. The signature transferred strongly to **highseq/highstruc** proteins, which received near maximal scores, in fact even higher than the positive training cohort (median 0.998 for **highseq/highstruc**; 0.996 for **lowseq/highstruc**). In contrast, **lowseq/lowstruc** variants and control length-matched GenBank proteins scored low (medians 0.0845 and 0.0762, respectively), while **highseq/lowstruc** proteins showed an intermediate distribution (median 0.295). This suggests that features learned from low-sequence proteins capture a structural-conservation signal that transfers to unseen sequences, including proteins that occupy distinct regions of sequence space.

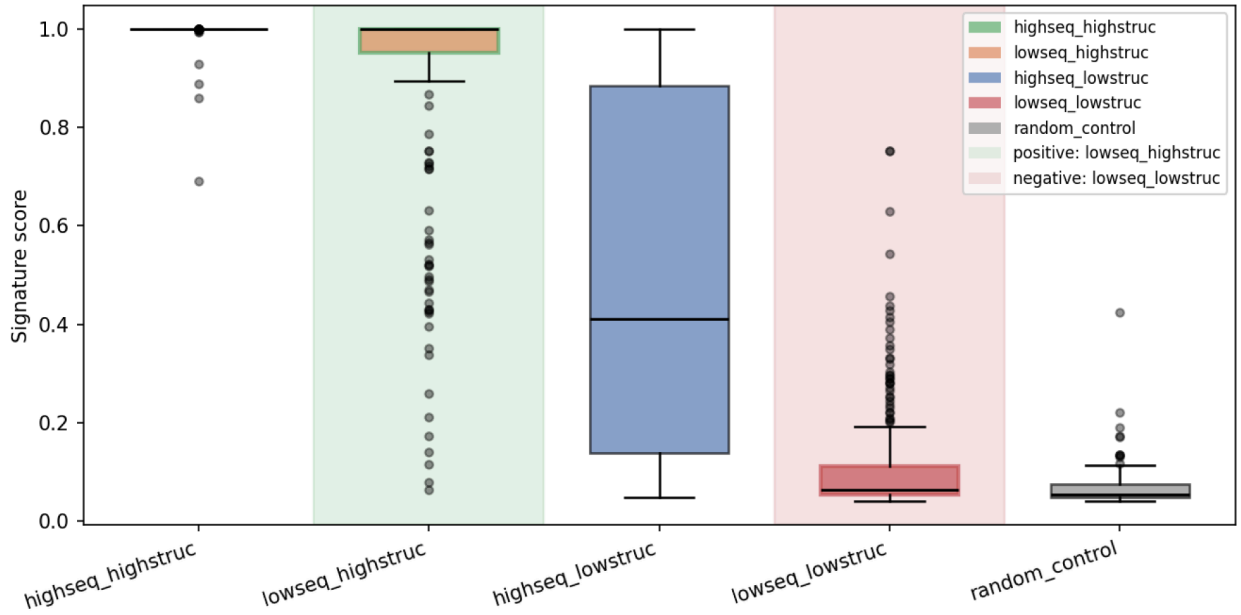

Figure 6. Distribution of the learned structural signature score across sequence/structure cohorts. *The model was trained to distinguish* ***lowseq/highstruc*** *from* ***lowseq/lowstruc*** *proteins using WT-correlated SAE activation features. Scores are fold-averaged logistic regression outputs on a 0–1 scale: higher scores indicate a more* ***lowseq/highstruc****-like activation signature, while lower scores indicate a more* ***lowseq/lowstruc****-like signature. Shaded regions mark the positive and negative training cohorts.*

Feature-level inspection suggests the signal is not driven by conserved amino acids

To examine what the classifier-associated features were detecting, we inspected residue-level activation traces for representative SAE features. A selected feature showed a conserved activation peak across wild-type gpD and structure-preserving variants, even though the local amino acid sequence underlying the peak differed between variants. This suggests that the feature is not simply detecting a conserved linear sequence motif.

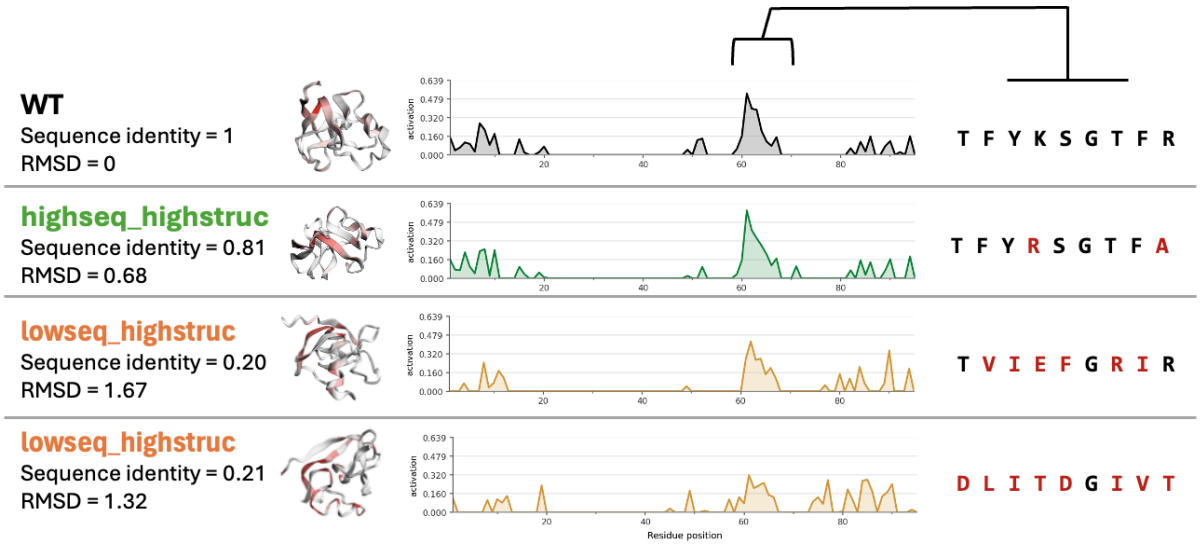

Figure 7. SAE feature 5940 shows a conserved activation peak despite local sequence divergence. *Activation traces for feature 5940 are shown across wild-type gpD and structurally similar variants. The main activation peak occurs in a similar positional region across variants, despite substantial differences in the local amino acid sequence.*

By contrast, structurally disrupted variants lost the characteristic activation peak, including cases where sequence identity remained relatively high. This supports the interpretation that the feature is sensitive to local structural or biophysical context rather than sequence identity alone.

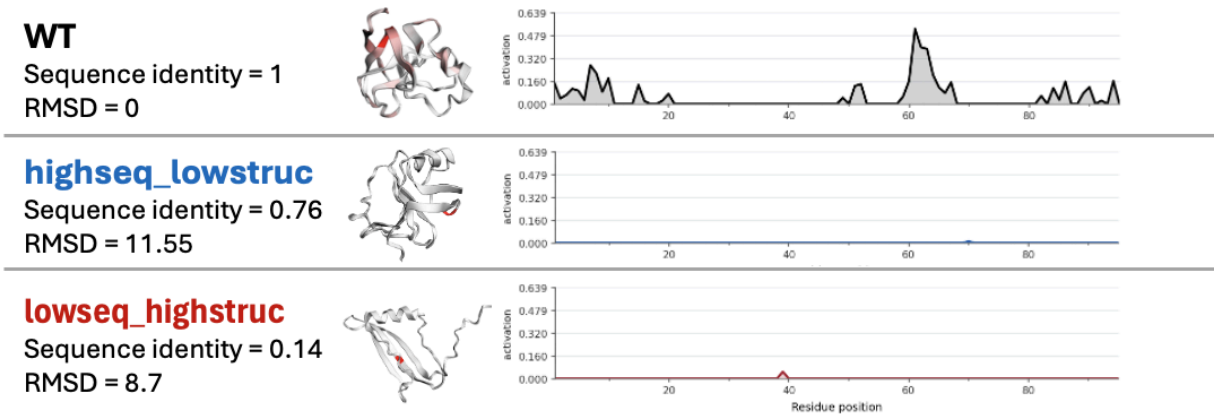

Figure 8. SAE feature 5940 is lost in structurally altered variants. *Feature 5940 shows a clear activation peak in wild-type gpD but little or no activation in variants with high backbone RMSD from wild type.*

Targeted perturbations support a structure-context interpretation

We next tested whether the activation pattern depended on absolute sequence position or on local structural context. Adding a 20-amino-acid unstructured segment to the N-terminus shifted the activation trace along the sequence but did not abolish the main peak. In contrast, inserting an unstructured coil within the peak region disrupted the activation itself. This suggests that the feature is not tied to an absolute residue index. Instead, it appears to depend on the local context that gives rise to the activation peak.

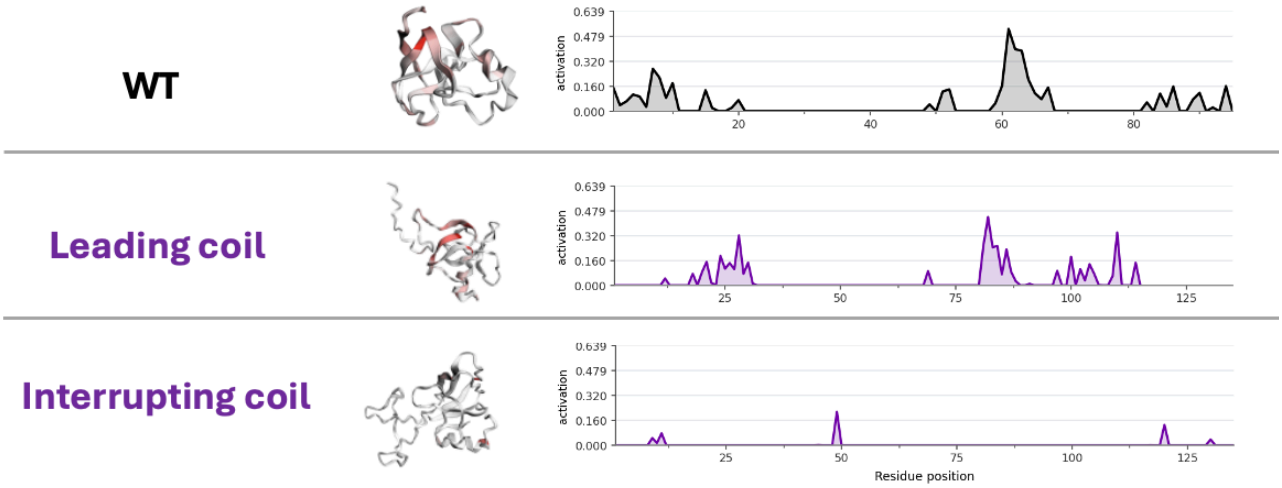

Figure 9. Feature activation is lost in structurally disrupted variants and altered by targeted perturbations. *Adding an unstructured N-terminal segment shifts the activation pattern along the sequence without abolishing the main peak, whereas inserting a coil within the peak region disrupts the activation. This suggests that the feature depends on local structural context rather than absolute sequence position alone.*

Signatures from deeper layers are more structure-selective than from early layers

Finally, we compared WT-relative signatures across model layers. At layer 1, the signature separated the training contrast well, but appeared less structurally specific: **highseq/lowstruc** proteins received near-maximal scores, despite being structurally distinct from the WT, suggesting that the early-layer signature was driven largely by sequence-level similarity rather than predicted structural preservation.

At layer 30, the same contrast was learned with similar accuracy, but the resulting score appeared to track structure more closely. **Highseq/highstruc** proteins remained high-scoring, whereas **highseq/lowstruc** proteins scored lower and more variably. This suggests that deeper-layer SAE features encode a signal more tied to preserved fold than sequence similarity.

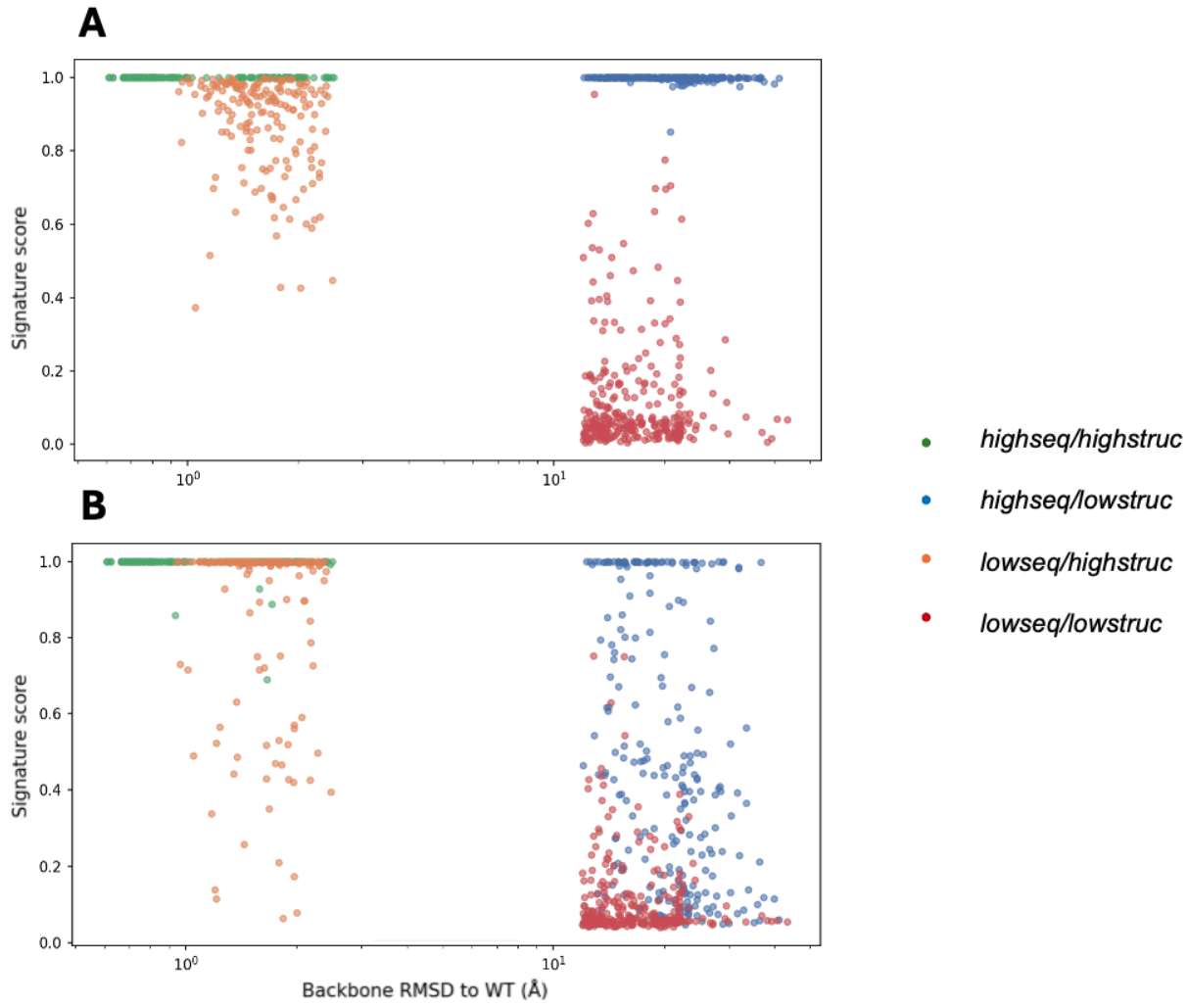

Figure 10. Signatures become more structure-selective in deeper model layers. *Classifiers were trained on the* ***lowseq/highstruc*** *versus* ***lowseq/lowstruc*** *contrast using WT-correlation aggregated SAE features, and the resulting signature scores were applied to all cohorts. Each point is a protein, plotted by backbone RMSD and signature score. Points are colored by cohort.* ***A:*** *At layer 1, the signature separates the training contrast but also remains high for highseq/lowstruc, indicating a sensitivity to sequence proximity.* ***B:*** *Layer 30 scores are more selectively elevated for structurally preserved proteins, and lower and more variable for structurally altered proteins.*

Importantly, the residual high-scoring **highseq/lowstruc** proteins at layer 30 may not simply represent model errors. These proteins are labelled as structurally distant based on RMSD from ESM3-predicted structures, but manual follow-up cases showed that some have much lower RMSDs when structures are predicted with AlphaFold. Some apparent false positives may therefore reflect uncertainty in the structural labels rather than poor signature performance.

Taken together, these results suggest the interpretation that early ESM-2 SAE layers encode local or sequence-proximal similarity to WT, while deeper layers capture more abstract, structure-relevant properties.

4 Discussion

This project tested whether SAE features from protein language models can identify predicted structural preservation in generated proteins beyond sequence similarity. Using gpD as a proof-of-principle system, we found that SAE-derived signatures separated variants by structural similarity to wild-type, and transferred to unseen proteins from distinct regions of sequence space. This suggests that InterPLM features can capture structure-relevant signals beyond simply reflecting sequence identity or memorized motifs.

Manual inspection supported this interpretation. Representative features showed similar activation peaks across structurally similar variants despite substantial local sequence differences. Perturbation experiments further suggested that these activations depend on local structural context: an N-terminal unstructured segment shifted the trace without abolishing the peak, whereas inserting a coil within the peak region abolished the activation.

Layer comparisons suggested that this signal becomes more structure-selective in deeper model layers. Early-layer signatures were more dominated by sequence proximity, with high-sequence-similarity but structurally divergent variants still scoring highly. By contrast, deeper-layer signatures better separated structurally preserved

from structurally altered variants, consistent with deeper layers encoding more abstract structure-relevant properties.

Future work should test other protein families and variants produced by varying design methods, such as structure-conditioned sequence design or backbone diffusion. A limitation of the present study is that the generation and interpretation steps both used ESM-family masked protein language models, with the variants generated by ESM3, and the features used for analysis derived from ESM-2 activations through InterPLM. While ESM3 and ESM-2 are distinct models with different architectures and objectives, both belong to the broader ESM family of protein language models. More work is needed to determine the extent to which these results generalize to sequences generated using other design methods, or to other protein types.

Testing variants of a more diverse set of proteins, and generated using a broader range of design methods, would address the limitations above. If signatures learned from one set of variants transfer to protein variants designed by other methods, this would provide stronger evidence that they reflect genuine structural conservation rather than ESM-family biases or generator-specific artefacts.

Another limitation is that sequence and structural similarity were discretized into threshold-defined cohorts. This simplifies the analysis and allows controlled comparisons between groups, but discards information about variation within each cohort and makes the results sensitive to the chosen cut-offs. Future work should therefore model structural and sequence similarity as continuous variables when measuring SAE feature responses and signature scores.

Relatedly, the structural labels themselves depend on predicted structures and RMSD thresholds, which introduces potential label noise, especially for generated proteins whose predicted folds may be uncertain or model-dependent. In manual follow-up cases, proteins labelled as structurally distant from wild-type gpD using ESM3-predicted structures received much lower RMSD values when re-predicted with AlphaFold. This suggests that some apparent false positives may reflect uncertainty in the structural labels rather than failure of the learned signature. Future work should assess label robustness by comparing RMSD values across multiple structure-prediction

methods, incorporating confidence metrics such as pLDDT, and flagging cases where predictors disagree.

Overall, these results suggest that SAE features can expose structure-preservation signals inside protein language models that may be biologically meaningful. In this proof-of-principle system, WT-relative SAE signatures identified plausible gpD-like protein variants even at low sequence identity, and feature-level inspection suggested that the signal was not reducible to conserved amino acid sequence motifs. With broader validation, this approach could contribute to more interpretable, structure- and function-aware screening methods for synthetic biology.

References

[1] G. R. Abel, Jr. et al., "Beyond sequence similarity: Toward function-based screening of nucleic acid synthesis," SSRN Electron. J., Mar. 25, 2026, doi: 10.2139/ssrn.6444478.

[2] S. H. King, C. L. Driscoll, D. B. Li, D. Guo, A. T. Merchant, G. Brixi, M. E. Wilkinson, and B. L. Hie, "Generative design of novel bacteriophages with genome language models," *bioRxiv*, Sep. 12, 2025, doi: 10.1101/2025.09.12.675911.

[3] S. P. Ikonomova, B. J. Wittmann, F. Piorino, D. J. Ross, S. W. Schaffter, O. Vasilyeva, E. Horvitz, J. Diggans, E. A. Strychalski, S. Lin-Gibson, and G. J. Taghon, "Experimental evaluation of AI-driven protein design risks using safe biological proxies," *bioRxiv*, May 15, 2025, doi: 10.1101/2025.05.15.654077.

[4] B. J. Wittmann *et al.*, "Strengthening nucleic acid biosecurity screening against generative protein design tools," *Science*, vol. 390, no. 6768, pp. 82–87, Oct. 2025, doi: 10.1126/science.adu8578.

[5] E. Simon and J. Zou, "InterPLM: Discovering interpretable features in protein language models via sparse autoencoders," *Nature Methods*, vol. 22, pp. 2107–2117, 2025, doi: 10.1038/s41592-025-02836-7.

[6] E. Adams, L. Bai, M. Lee, Y. Yu, and M. AlQuraishi, "From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models," *bioRxiv*, 2025, doi: 10.1101/2025.02.06.636901.

[7] A. Maiwald, P. Jedryszek, F. Draye, G. M. Morris, and O. M. Crook, "Decode-gLM: Tools to interpret, audit, and steer genomic language models," *bioRxiv*, 2025, doi: 10.1101/2025.10.31.685860.

[8] T. Hayes *et al*., "Simulating 500 million years of evolution with a language model," *Science*, vol. 387, no. 6736, pp. 850–858, 2025, doi: 10.1126/science.ads0018.

[9] C. Chang, A. Plückthun, and A. Wlodawer, "Crystal structure of a truncated version of the phage λ protein gpD," *Proteins*, vol. 57, no. 4, pp. 866–868, 2004, doi: 10.1002/prot.20254.

[10] Z. Lin *et al*., "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023, doi: 10.1126/science.ade2574.