bish-bash-fold: what are protein structure prediction models learning?

Anonymous Authors[1]

Abstract

Protein structure prediction models rely on coevolutionary signals, intrinsically limiting their efficacy for de novo protein design. Here, we investigate this failure regime by training sparse autoencoders on the diffusion modules of AlphaFold3 and Boltz-2 to recover interpretable features along the denoising trajectory. By ablating MSA depth and sequence input, we categorise features into MSA-activated, MSA-silenced, sequence-driven, and model-prior classes. We find that AlphaFold3 is less sensitive to input degradation while Boltz-2 is more conditioning-dependent. Using the Garcia et al. (2025) dataset, we then investigated whether these features could provide mechanistic insight to designability failure modes. Linear probes trained on feature activations predict experimental outcome (measured by expression, solubility, monomericity, and CD) with higher accuracy and precision than pLDDT, and recover known design heuristics. As the first interpretability study of structure prediction denoising trajectories, this work establishes that internal representations capture critical designability signals beyond standard model outputs.

1. Introduction

Protein structure prediction models are integral to protein design. Optimisation-based methods such as Boltzdesign (Cho et al., 2025a) and Protein Hunter (Cho et al., 2025b) directly use the outputs of structure prediction models to guide design (Anishchenko et al., 2021; Wang et al., 2022; Goverde et al., 2023). Structure models also serve as core components of generative models, enabling tasks such as co-folding, motif scaffolding and binder design (Watson et al., 2023; Ingraham et al., 2023; Pacesa et al., 2025). Design models can themselves be built on structure predictors (Krishna et al., 2024), and structure-based sequence design models such as ProteinMPNN (Dauparas et al., 2022) and ESM-IF (Hsu et al., 2022) use Alphafold predicted structures for validation. Downstream, self-consistency and confidence metrics from structure prediction models are also used as in silico proxies for designability to reduce the number of sequences to validate experimentally (Wang et al., 2022; Watson et al., 2023; Bennett et al., 2023).

State-of-the-art models can predict 3D protein structures from amino acid sequences with atomic-level accuracy, but rely heavily on coevolutionary information. These models share a common two-stage design: the trunk learns coevolutionary signal from multiple sequence alignments (MSAs), and the resulting representations condition the diffusion module that generates 3D coordinates (Abramson et al., 2024; Chai Discovery team et al., 2024; Passaro et al., 2025). Under limited homology, models struggle to accurately predict targets (Bryant & Noe, 2024), evidenced by their poor ability to resolve the structural impacts of mutations (Buel & Walters, 2022), capture different conformation states (Saldaño et al., 2022; Chakravarty & Porter, 2022) or isoforms (Yang et al., 2024). These failure modes structurally manifest as incorrect chirality, steric clashes, and the misfolding of unstructured regions into rigid helices (Buttenschoen et al., 2024; Abramson et al., 2024). However, *de novo* proteins, by design, lack evolutionary history, placing them in the realm of where these models are known to fail (Anishchenko et al., 2021; Goverde et al., 2023). This discrepancy is often tolerable for small and well-constrained proteins (Frank et al., 2024), but is becoming increasingly acute as the field moves towards larger, more functionally complex designs.

Diffusion has been crucial in achieving current performance for the state-of-the-art models. However, under shallow homology, it inherits the trunk's weak coevolutionary signals and the resulting hallucinations are not reliably flagged by confidence metrics (Abramson et al., 2024). While the limitations of MSA are well-described, how that shapes the generative diffusion process remains underexplored. What does the diffusion module encode during denoising, and how does its representations respond to different conditioning?

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the 2026 Workshop on Generative and Agentic AI for Biology (ICML 2026). Do not distribute.

Probing these internals may be key to understanding how these models resolve structure, and what the implications are for *de novo* proteins.

Mechanistic interpretability offers a natural lens into these model internals. Here, we used sparse autoencoders to probe the diffusion modules of AlphaFold3 (Abramson et al., 2024) and Boltz-2 (Passaro et al., 2025), and decomposed the broader question of what the diffusion module encodes into three parts. First, we examined what the module encodes during denoising and how they evolve across the denoising trajectory. Second, we asked how the representations depend on the conditioning signal. By ablating MSA depth and sequence information, we isolated which features are driven by which inputs. Third, we asked what these models understand about designed proteins beyond the model output. To the best of our knowledge, this is the first mechanistic interpretability study of the denoising trajectory in state-of-the-art protein structure prediction models.

Our contributions are as follows:

1. **Interpretable features across denoising.** Sparse autoencoders recover interpretable features from the diffusion module throughout the denoising trajectory.

2. **Trajectory-dependent feature composition.** Feature composition evolves across denoising along model-specific axes, differing between AlphaFold3 and Boltz-2 despite their architectural similarity.

3. **Input-dependent feature expression.** Titrating MSA depth and ablating sequence information reveals which features are driven by which conditioning inputs across denoising.

4. **Features outperform pLDDT for designability.** Probes built on diffusion-module features predict experimental outcomes on designed *de novo* proteins more accurately than pLDDT, and expose interpretable success and failure modes.

2. Background

2.1. Diffusion in structure prediction models

In AlphaFold3 and Boltz-2, the Pairformer trunk processes sequence and MSA into single and pair representations, and a diffusion module generates coordinates conditioned on those representations. The diffusion modules are architecturally similar across these models with comparable noise schedules (Karras et al., 2022). Briefly, they differ in that Boltz-2 enabled additional conditioning and inference-time Boltz-steering whereas AlphaFold3 was trained with cross-distillation with AlphaFold-Multimer predictions (Evans et al., 2021).

2.2. MSA in structure prediction models

Multiple sequence alignments (MSAs) capture how amino acids co-evolve under 3D structural constraints, enabling structure prediction models to learn spatial proximity (Marks et al., 2011; Morcos et al., 2011). Although model performance is dependent on alignment depth, a body of work actively exploits this sensitivity by manipulating MSAs at inference (Roney & Ovchinnikov, 2022; Del Alamo et al., 2022; Wayment-Steele et al., 2024; Stein & Mchaourab, 2022; Kalakoti & Wallner, 2026; Suzuki & Amagasa, 2026). While these subsampling approaches have been applied to design and exploring alternative conformations, how MSA conditioning specifically propagates through downstream diffusion remains underexplored.

2.3. Sparse Autoencoders and interpretability

Sparse Autoencoders (SAEs) aim to disentangle polysemantic model activations into sparse, linear combinations of learned feature vectors (Elhage et al., 2022; Bricken et al., 2023; Cunningham et al., 2023). An SAE encodes activations $\mathbf{x} \in \mathbb{R}^n$ into a high-dimensional latent space $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^m$, where $m \gg n$, and computes a reconstruction $\hat{\mathbf{x}}$ via a linear decoder. Top-K SAEs decomposes activations into interpretable features by enforcing sparsity in the latent space (Gao et al., 2024) where the encoder hard selects the k largest pre-activation values and zeros the remainder:

$$\begin{aligned} z &= \mathrm{TopK}(W_{\text{enc}}\mathbf{x} + \mathbf{b}_{\text{enc}}) \\ \hat{\mathbf{x}} &= W_{\text{dec}} z \end{aligned} \tag{1}$$

where $W_{\text{enc}} \in \mathbb{R}^{m \times n}$ is the encoder weight matrix that projects the input activation into the high-dimensional latent space, $\mathbf{b}_{\text{enc}} \in \mathbb{R}^m$ is the encoder bias, $z \in \mathbb{R}^m$ is the resulting sparse latent containing k non-zero entries, and $\hat{\mathbf{x}} \in \mathbb{R}^n$ is the reconstructed activation produced by the decoder weight matrix $W_{\text{dec}} \in \mathbb{R}^{n \times m}$.

The primary training objective minimizes the mean squared error (MSE) between the input and its reconstruction, $\mathcal{L}_{\text{MSE}} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2$. To prevent features from becoming permanently inactive, an auxiliary loss $\mathcal{L}_{\text{aux}} = \|\mathbf{e} - \hat{\mathbf{e}}\|_2^2$ can be added, where $\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}$ is the residual error and $\hat{\mathbf{e}}$ is its reconstruction using the most active dead latents (see Appendix A.2). This provides a gradient signal that pulls inactive features back The total loss is $\mathcal{L} = \mathcal{L}_{\text{MSE}} + \alpha\mathcal{L}_{\text{aux}}$, where auxiliary penalty α weights the auxiliary term.

2.4. Interpretability in protein models

Previous work in mechanistic interpretability of protein models has largely been confined to protein language models. Recent efforts have successfully trained SAEs on ESM-2 embeddings to identify features mapping to known properties such as functional domains, structural motifs, and

binding sites (Simon & Zou, 2024), and have demonstrated how these networks balance generic and family-specific representations (Adams et al., 2025). Extending these interpretability principles, FoldSAE (Zarzecki et al., 2025) applied SAEs to RFdiffusion (Watson et al., 2023).

2.5. Interpretability in diffusion

Previous work for interpretability in diffusion has predominantly been applied to vision models. Previous works demonstrated that SAEs can decompose denoising activations into interpretable features that highlight functional specialization across transformer blocks (Surkov et al., 2024; Cywiński & Deja, 2025). Timestep-resolved interpretability has also been achieved in diffusion language models (Wang et al., 2025). While FoldSAE (Zarzecki et al., 2025) has introduced SAEs to diffusion, it leaves the temporal dynamics of structural denoising trajectory uncharacterized.

3. Method

To probe the denoising trajectory, we train TopK SAEs on the final (layer 24) diffusion transformer activations of Boltz-2 and AlphaFold3. By extracting features immediately before the final LayerNorm leading to the atom decoders, we capture the richest token-level representations prior to coordinate generation. We train 10 independent SAEs total (one per model per timestep) at $t \in \{0, 50, 100, 146, 199\}$ to sample across high, mid, and low σ regimes (Karras et al., 2022) (see Appendix A.1 for training details). We evaluate each SAE on an unseen test set (see Appendix B.1 for details on data split and evaluation metrics) with the following metrics: 1) explained variance (EV) to measure reconstruction fidelity; 2) dead features to measure dictionary utilization; 3) probe accuracy to measure the linear separability of biologically meaningful properties from the learned features, using principal component analysis (PCA) of matched dimensionality as a baseline.

4. Experiments

4.1. The Diffusion Module Encodes Interpretable Structural Features

TopK SAEs achieve high explained variance (EV) with a near-zero percentage of dead features across all denoising timesteps (Table 1). High EV across k demonstrates a favourable sparsity-fidelity trade off, and low dead features indicate high dictionary utilisation. Automated interpretability reveals features corresponding to well-defined structural and biochemical properties, including residue charge, secondary structure, functional annotations and motifs (Figure 1). Although trained independently, AlphaFold3 and Boltz-2 converge on shared representations where identical features activate at homologous positions ($t = 199$) as well as protein-specific variations. Linear probes trained on the features consistently outperform the PCA baseline, indicating that the diffusion module's activations encode these structural concepts and are linearly separable. See model ablations detailed in the Appendix C.

Table 1. Summary of the best performing TopK sparse autoencoders derived from AlphaFold3 and Boltz-2 activations used for downstream analysis. Models from across extraction timesteps (t) using explained variance (EV), dead feature percentages, and probe accuracy (SAE vs. PCA). See Appendix sections B.1 and C for evaluation and ablations.

Model	t	k	Expansion Factor	EV ↑	Dead (%) ↓	Probe Acc (SAE) ↑	Probe Acc (PCA) ↑
AlphaFold3	0	64	64	0.9054	0.0000	0.8284	0.7015
	50	64	64	0.9049	0.0000	0.8251	0.6595
	100	64	64	0.8838	0.0000	0.8166	0.6417
	146	64	64	0.8852	0.0000	0.7981	0.6728
	199	64	64	0.9517	0.0061	0.8176	0.8176
Boltz-2	0	16	16	0.9842	0.0202	0.7749	0.6552
	50	32	32	0.9617	0.0014	0.7777	0.6549
	100	32	32	0.9617	0.0013	0.7858	0.6272
	146	32	32	0.9908	0.0160	0.8267	0.7229
	199	32	32	0.9946	0.1763	0.7437	0.6622

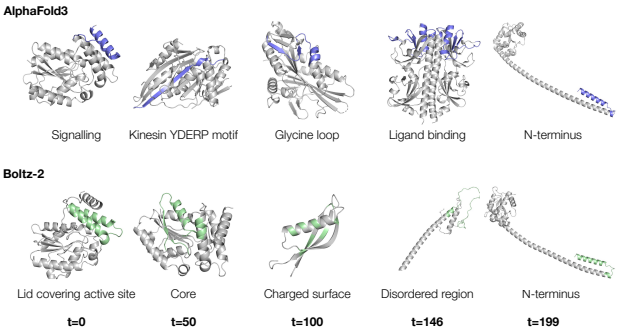
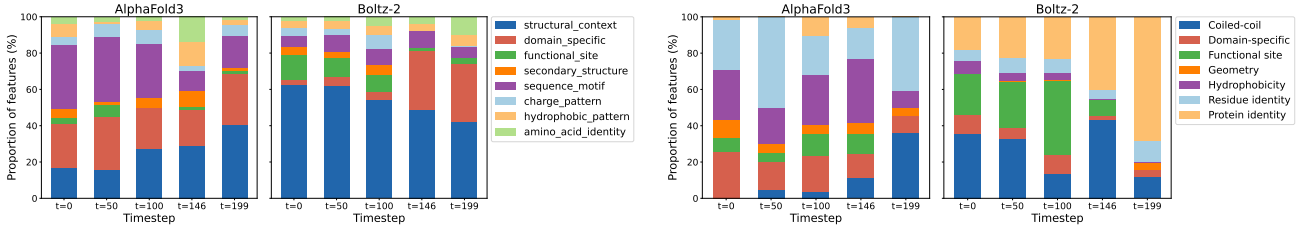


Figure 1. Feature activations for AlphaFold3 (top) and Boltz-2 (bottom) across denoising timesteps. Text annotations are from UniProt (The UniProt Consortium, 2023) and these results are from automated interpretability with Claude Opus 4.6 accessed through their API (Anthropic, 2026). (UniProt IDs: top row, left to right: A0A024SC78, P48467, A8BKD1, G3XD24, A0A009IHW8; bottom row, left to right: P00590, P22894, B0R5M0, O62479, A0A009IHW8).

4.2. Diffusion Module Encodes Interpretable Features Across Denoising Timesteps in sequence

To understand how structural information evolves during denoising, we categorised the active features at each timestep into eight feature categories derived from UniProtKB/SwissProt (The UniProt Consortium, 2023). AlphaFold3 maintains a relatively balanced category composition throughout denoising, with structural-context and domain-specific features each contributing meaningfully at every timestep (Figure 2a). Structural-context features rise progressively across denoising, while sequence-motif features are front-loaded until t=100. Boltz-2 instead



(a) UniProt category trace

(b) Alpha helix context trace

Figure 2. Evolution of feature activation during denoising. (a) Comparative category composition of active features for AlphaFold3 (Abramson et al., 2024) and Boltz-2 (Passaro et al., 2025) across the denoising trajectory. (b) Fine-grained trace of alpha helix motifs, illustrating the transition from local physicochemical features to global structural context. While AlphaFold3 follows a bottom-up resolution path, Boltz-2 demonstrates an early-stage commitment to protein-identity and domain-specific features, highlighting divergent computational strategies for structural synthesis despite similar architectures.

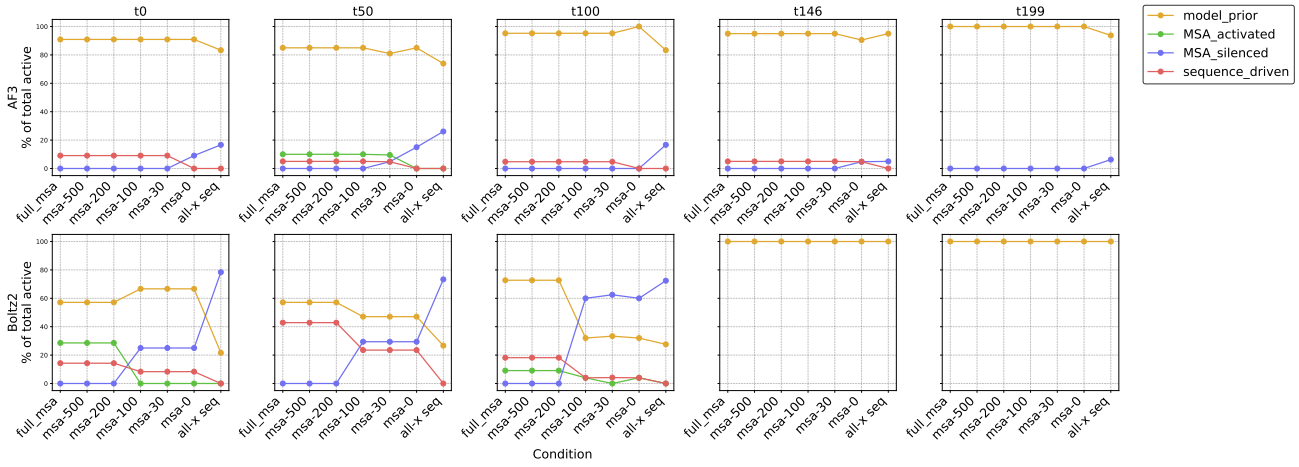


Figure 3. Feature category composition across denoising timesteps for activations from AlphaFold3 (Abramson et al., 2024) (top) and Boltz-2 (Passaro et al., 2025) (bottom) under varying MSA depth and sequence ablations. Each active feature was categorised as model-prior, MSA-activated, MSA-silenced or sequence-driven at each timestep, see Appendix D. AlphaFold3 maintains a stable composition across conditions, whereas Boltz-2 exhibits three distinct regimes of MSA dependence with a critical transition between MSA depths of 100-200.

shows structural-context features contributing progressively less, with late timesteps dominated by domain-specific and structural-context features. Boltz-2 also draws less on sequence motifs throughout denoising than AlphaFold3.

The category-level view established which feature types dominate at each stage; tracing a single structural motif, such as the alpha helix, revealed how feature usage specialises within a category as denoising proceeds (Figure 2). In AlphaFold3, residue-identity and hydrophobicity features together account for over half of the active representation at every timestep. Functional-site features such as disulfide bonds and binding-site residues nearly double at t=100. At t=199, coiled-coil features rise sharply, which may indicate late-stage placement of the helix into a tertiary context. In Boltz-2, by contrast, residue-identity and hydrophobicity features remain marginal throughout, and the model draws instead on coiled-coil and functional-site features from early timesteps. Here, protein-identity features progressively rise. Taken together with the increase of domain specific feature expression, this may suggest that the model reocognises and commits to protein identity instead of resolving the structure. Despite architectural similarities, AlphaFold3 and Boltz-2 traverse denoising differently.

4.3. Features Decompose into Four Categories by Input Dependence

To investigate how features depend on input evidence at inference, we titrated from a deep multiple sequence alignment (MSA raw depth 1000) down to a completely ablated MSA, alongside an all-X (unknown) sequence condition (Figure 3; dataset details in Appendix D). For each residue, we recorded feature activations across six ablation levels and classified features into four categories (see Appendix E.1 for details on classification):

- **MSA-activated**: active with MSA, dead without.
- **MSA-silenced**: dead with MSA, active when both

MSA and sequence are absent.

- **Sequence-driven**: unaffected by MSA, inactive when the sequence is ablated.
- **Model-prior**: active under all conditions, including all-X.

Notably, AlphaFold3 retains a high proportion of model-prior features across all conditions, demonstrating robust insensitivity to MSA depth. Boltz-2 is comparatively conditioning-sensitive, exhibiting three distinct regimes of MSA dependence: deep (MSA $\geq$ 200), shallow ($0 <$ MSA < 200), and information-free (all-X), with MSA 200 to 100 range acting as a critical transition zone. In both models, the early denoising timesteps ($t = 146$ and $t = 199$) remain largely MSA-independent, varying little across input conditions. Conversely, MSA information is primarily integrated during early timesteps ($t = 0$–100).The high proportion of sequence-motif features observed in Figure 2a, juxtaposed with the low proportion of sequence-driven features identified here, raises an apparent tension: motif features dominate the representation, yet most are not recruited directly by sequence input. This suggests that feature expression is largely decoupled from input degradation at this stage. This likely arises either from a strong learned diffusion prior or from rigid pre-conditioning by the pairformer upstream. This is consistent with hallucination behaviour where the denoiser proceeds as if the representation reflects valid input regardless of conditioning quality. Crucially, we emphasize that this decoupling concerns feature *expression*, not correct *placement*.

4.4. Internal Representations Predict Experimental Success Beyond Model Confidence

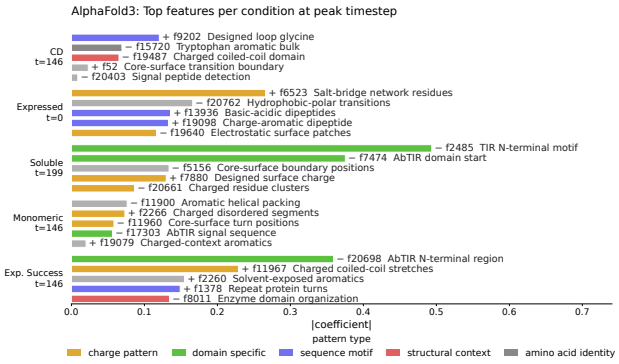


Figure 4. Labelled L1 coefficients identifying the AlphaFold3 SAE features most predictive of each designability criterion. The best performing SAE was used to illustrate utility of SAEs.

Beyond extracting mechanistic insights into model internals, we investigated whether these features hold downstream utility for *de novo* protein design. As an exploratory case study, we evaluated their ability to predict experimental designability. Using a curated dataset from (Garcia et al., 2025),

Table 2. Best-performing model per property, comparing pLDDT-only baselines against the best SAE-based model for each backbone (AF3 and Boltz). Dataset from (Garcia et al., 2025), n=614. Bold values indicate the highest AUROC/AUPRC within each property.

Property	Model	AUROC	AUPRC
Expressed	AF3 pLDDT Only	0.6799	0.9470
	AF3 t=0 L1-LogReg	0.7636 ± 0.07	0.9680
	Boltz pLDDT Only	0.7595	0.9652
	Boltz t=199 L1-LogReg	**0.8069**	**0.9738**
Solubility	AF3 pLDDT Only	0.5876	0.8747
	AF3 t=199 L1-LogReg	**0.7669**	**0.9381**
	Boltz pLDDT Only	0.7030	0.9230
	Boltz t=199 L1-LogReg	0.7423	0.9369
Monomeric	AF3 pLDDT Only	0.5539	0.6926
	AF3 t=146 L1-LogReg	**0.7150**	**0.8080**
	Boltz pLDDT Only	0.6395	0.7562
	Boltz t=146 L1-LogReg	0.6658	0.7793
CD	AF3 pLDDT Only	0.5908	0.8910
	AF3 t=146 L1-LogReg	**0.78279**	**0.9505**
	Boltz pLDDT Only	0.5936	0.9002
	Boltz t=199 L1-LogReg	0.7585	0.9423
Experimental Success	AF3 pLDDT Only	0.6222	0.5271
	AF3 t=146 L1-LogReg	**0.7648**	**0.6864**
	Boltz pLDDT Only	0.7129	0.6417
	Boltz t=146 L1-LogReg	0.7354	0.6596

we trained L1-regularized logistic regression probes on feature activations across denoising timesteps to predict binary outcomes across four key experimental stages: expression, solubility, monomericity, and circular dichroism (CD). Evaluated via AUROC and AUPRC against a pLDDT baseline (see Appendix F.2 for training details), probes trained on SAE features surprisingly outperform pLDDT across all four stages for both AlphaFold3 and Boltz-2 (Table 2). Notably, AlphaFold3 SAE features exhibit a particularly large margin of improvement. This aligns with our earlier finding that AlphaFold3 relies on stronger learned priors under shallow conditioning, whereas Boltz-2 encodes more protein-identity and domain-specific features. This robust prior provides AlphaFold3 with a critical advantage when evaluating *de novo* sequences, which inherently lack coevolutionary signals and known domain identities.

Crucially, examining the regression coefficients allows us to identify the specific structural features driving these predictions (Figure 4). Several recovered features mirror established design heuristics. For example, well-folded CD is positively predicted by core-surface boundaries and negatively by bulky aromatics. Expression is largely driven by internal packing, with salt bridges scoring positively and localized electrostatic patches scoring negatively. Solubility is dominated by surface-charge and N-terminal motif features, and monomericity is associated with charged disordered segments. Overall experimental success strongly correlates with charged coiled-coils, solvent-exposed aromatics, and repeat-protein turns, establishing that diffusion module representations encode designability signals beyond standard

confidence metrics.

5. Conclusion

We showed that sparse autoencoders recover interpretable features from the diffusion modules of AlphaFold3 and Boltz-2, and that these features partition by conditioning dependence into MSA-activated, MSA-silenced, sequence-driven, and model-prior classes - with AlphaFold3 expressing far more input-insensitive priors than Boltz-2. Probes on these features predict experimental outcomes for designed proteins more accurately than pLDDT, suggesting internal representations carry designability signal absent from output confidence.

As an exploratory study, this work has several limitations. Our analysis was restricted to a subset of diffusion timesteps and focused exclusively on single-domain proteins; the extent to which these features generalize to large, multi-chain complexes remains to be characterized. Additionally, our MSA titration analysis suggests a promising future direction: coupling SAE feature activation analysis upstream with Pairformer outputs and downstream with ground-truth accuracy metrics like RMSD. Finally, we aim to leverage designability features as targets in protein design, potentially steering generative models toward more well-folded sequence spaces than is possible using current output-based scoring functions.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630 (8016):493–500, 2024.

Adams, E., Bai, L., Lee, M., Yu, Y., and AlQuraishi, M. From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models. *bioRxiv*, pp. 2025–02, 2025. doi: 10.1101/2025.02.06.636901.

Ahdritz, G., Bouatta, N., Kadyan, S., Jarosch, L., Berenberg, D., Fisk, I., Watkins, A., Ra, S., Bonneau, R., and AlQuraishi, M. Openproteinset: Training data for structural biology at scale. *Advances in Neural Information Processing Systems*, 36:4597–4609, 2023.

Anishchenko, I., Pellock, S. J., Chidyausiku, T. M., Ramelot, T. A., Ovchinnikov, S., Hao, J., Bafna, K., Norn, C., Kang, A., Bera, A. K., et al. De novo protein design by deep network hallucination. *Nature*, 600(7889):547–552, 2021.

Anthropic. Claude Opus 4.6 system card. Technical report, Anthropic, February 2026. URL https://www.anthropic.com/claude-opus-4-6-system-card.

Bennett, N. R., Coventry, B., Goreshnik, I., Huang, B., Allen, A., Vafeados, D., Peng, Y. P., Dauparas, J., Baek, M., Stewart, L., et al. Improving de novo protein binder design with deep learning. *Nature Communications*, 14 (1):2625, 2023.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bryant, P. and Noe, F. Dissecting AlphaFold's capabilities with limited sequence information. *bioRxiv*, pp. 2024–03, 2024.

Buel, G. R. and Walters, K. J. Can alphafold2 predict the impact of missense mutations on structure? *Nature structural & molecular biology*, 29(1):1–2, 2022.

Buttenschoen, M., Morris, G. M., and Deane, C. M. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024.

Chai Discovery team, Boitreaud, J., Dent, J., McPartlon, M., Meier, J., Reis, V., Rogozhonikov, A., and Wu, K. Chai-1: Decoding the molecular interactions of life. *BioRxiv*, pp. 2024–10, 2024.

Chakravarty, D. and Porter, L. L. AlphaFold2 fails to predict protein fold switching. *Protein Science*, 31(6):e4353, 2022.

Cho, Y., Pacesa, M., Zhang, Z., Correia, B. E., and Ovchinnikov, S. Boltzdesign1: Inverting all-atom structure prediction model for generalized biomolecular binder design. *BioRxiv*, pp. 2025–04, 2025a.

Cho, Y., Rangel, G., Bhardwaj, G., and Ovchinnikov, S. Protein hunter: exploiting structure hallucination within diffusion for protein design. *bioRxiv*, pp. 2025–10, 2025b.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Cywiński, B. and Deja, K. SAeUron: Interpretable concept unlearning in diffusion models with sparse autoencoders. *arXiv preprint arXiv:2501.18052*, 2025.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I., Courbet, A., de Haas, R. J., Bethel, N., et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378 (6615):49–56, 2022.

Del Alamo, D., Sala, D., Mchaourab, H. S., and Meiler, J. Sampling alternative conformational states of transporters and receptors with alphafold2. *elife*, 11:e75751, 2022.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., and Olah, C. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Evans, R., O'neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., et al. Protein complex prediction with alphafold-multimer. *biorxiv*, pp. 2021–10, 2021.

Frank, C., Schiwietz, D., Fuß, L., Ovchinnikov, S., and Dietz, H. Alphafold2 refinement improves designability of large de novo proteins. *bioRxiv*, pp. 2024–11, 2024.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Garcia, M., Dixit, S. M., and Rocklin, G. J. Evaluating zero-shot prediction of protein design success by alphafold, esmfold, and proteinmpnn. *bioRxiv*, pp. 2025–07, 2025.

Goverde, C. A., Wolf, B., Khakzad, H., Rosset, S., and Correia, B. E. De novo protein design by inversion of the AlphaFold structure prediction network. *Protein Science*, 32(6):e4653, 2023.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, pp. 8946–8970. PMLR, 2022.

Ingraham, J. B., Baranov, M., Costello, Z., Barber, K. W., Wang, W., Ismail, A., Frappier, V., Lord, D. M., Ng-Thow-Hing, C., Van Vlack, E. R., et al. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023.

Kalakoti, Y. and Wallner, B. Afsample3: Generating and selecting multiple conformational states with alphafold3. *bioRxiv*, pp. 2026–01, 2026.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35: 26565–26577, 2022.

Krishna, R., Wang, J., Ahern, W., Sturmfels, P., Venkatesh, P., Kalvet, I., Lee, G. R., Morey-Burrows, F. S., Anishchenko, I., Humphreys, I. R., et al. Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, 384(6693):eadl2528, 2024.

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., and Sander, C. Protein 3d structure computed from evolutionary sequence variation. *PloS one*, 6(12):e28766, 2011.

Mistry, J., Bateman, A., Bork, P., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Bordin, N., Andreeva, A., Necci, M., Khanna, T., Sangrador-Vegas, A., Lazaridis, T., Llinares-López, F., Salvatore, M., Piñeiro, Á., Bridge, A., Britto, R., Punta, M., and Finn, R. D. The Pfam protein families database: embracing AI/ML. *Nucleic Acids Research*, 53(D1):D523–D534, 2025. doi: 10.1093/nar/gkae997.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

Pacesa, M., Nickel, L., Schellhaas, C., Schmidt, J., Pyatova, E., Kissling, L., Barendse, P., Choudhury, J., Kapoor, S., Alcaraz-Serna, A., et al. One-shot design of functional protein binders with bindcraft. *Nature*, 646(8084):483–492, 2025.

Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Somnath, V. R., Getz, N., Portnoi, T., Roy, J., Stark, H., et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025.

Querino Lima Afonso, M., Pidruchna, I., Nair, S., Midlik, A., Lawal, D., Vollmar, M., Appasamy, S. D., Choudhary, P., Kunnakkattu, I. R., Bertoni, D., Escobar, C. A., Balasubramaniyan, B., Gaborova, R., Díaz Leines, G., Harrus, D., Gupta, D., Evans, G. L., Paramval, U., Magana, P., Tanweer, A., Todor, M., Thorpe, C. J., Tsenkov, M., Ganguly, S., Ellaway, J., Bueno, W. M., Bellaiche,

A., Sehnal, D., Svobodová, R., Fleming, J. R., and Velankar, S. PDBe: enhanced structural data exploration to facilitate discovery. *Nucleic Acids Research*, 54(D1): D440–D444, 2025. doi: 10.1093/nar/gkaf1120.

Roney, J. P. and Ovchinnikov, S. State-of-the-art estimation of protein model accuracy using alphafold. *Physical review letters*, 129(23):238101, 2022.

Saldaño, T., Escobedo, N., Marchetti, J., Zea, D. J., Mac Donagh, J., Velez Rueda, A. J., Gonik, E., García Melani, A., Novomisky Nechcoff, J., Salas, M. N., et al. Impact of protein conformational diversity on AlphaFold predictions. *Bioinformatics*, 38(10):2742–2748, 2022.

Simon, E. and Zou, J. InterPLM: Discovering interpretable features in protein language models via sparse autoencoders. *bioRxiv*, pp. 2024–11, 2024. doi: 10.1101/2024.11.14.623630.

Stein, R. A. and Mchaourab, H. S. Speach_af: Sampling protein ensembles and conformational heterogeneity with alphafold2. *PLoS computational biology*, 18(8):e1010483, 2022.

Steinegger, M. and Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017. doi: 10.1038/nbt.3988.

Surkov, V., Wendler, C., Mari, A., Terekhov, M., Deschenaux, J., West, R., Gulcehre, C., and Bau, D. One-step is enough: Sparse autoencoders for text-to-image diffusion models. *arXiv preprint arXiv:2410.22366*, 2024.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and The UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015. doi: 10.1093/bioinformatics/btu739.

Suzuki, S. and Amagasa, T. Steering conformational sampling in boltz-2 via pair representation scaling. *bioRxiv*, pp. 2026–01, 2026.

The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1): D523–D531, 2023. doi: 10.1093/nar/gkac1052.

Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J. L., Castro, K. M., Ragotte, R., Saragovi, A., Milles, L. F., Baek, M., et al. Scaffolding protein functional sites using deep learning. *Science*, 377(6604):387–394, 2022.

Wang, X., Jiang, B., Wan, Y., Yang, B., Kong, L., and Zou, D. DLM-Scope: Mechanistic interpretability of diffusion language models via sparse autoencoders. *arXiv preprint arXiv:2602.05859*, 2025.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976): 1089–1100, 2023.

Wayment-Steele, H. K., Ojoawo, A., Otten, R., Apitz, J. M., Pitsawong, W., Hömberger, M., Ovchinnikov, S., Colwell, L., and Kern, D. Predicting multiple conformations via sequence clustering and alphafold2. *Nature*, 625(7996): 832–839, 2024.

Yang, Y., Xie, Y., Li, Z., Diala, C. S., Ali, M. A., Li, R., Xu, Y., Wu, A., Kim, P., Hosseini, S.-R., et al. Systematic characterization of protein structural features of alternative splicing isoforms using alphafold 2. *bioRxiv*, pp. 2024–01, 2024.

Zarzecki, W. et al. FoldSAE: Learning to steer protein folding through sparse representations. *arXiv preprint arXiv:2511.22519*, 2025.

A. Training

A.1. Data

SAEs were trained on model activations generated during the forward pass of curated UniProt sequences (The UniProt Consortium, 2023). To simplify sequence composition and aid downstream feature interpretation, the dataset was restricted to single-domain proteins containing exactly one Pfam annotation (Mistry et al., 2025). The final training corpus consisted of 10 million tokens across 55,000 sequences.

A.2. Training details

We trained TopK Sparse Autoencoders (SAEs) on dense activations extracted at timesteps $t \in \{0, 50, 100, 146, 199\}$. Prior to input, the activations were layer-normalized. The models were trained to reconstruct the original dense activations over 40 epochs using a batch size of 1024. We used the Constrained Adam optimizer with a learning rate of 1e-4. The training objective included an auxiliary loss penalty term inverse of k. The decoder weights were unit-normalized after each optimization step.

During training, a feature was defined as dead if it failed to activate for 2000 consecutive steps. To prevent the collapse of the latent space and mitigate the emergence of these dead features, we adopted a periodic neuron resampling strategy. We applied a resample factor of 0.5 at specific training steps $t \in \{3000, 6000, 10000, 20000, 40000\}$.

B. Evaluation

B.1. Data

We construct two complementary evaluation sets, each suited to a different class of metric.

Test set We curated a time-based holdout set of 7,473 proteins from PDB entries released after July 1, 2024 (Querino Lima Afonso et al., 2025). We filtered for single-chain structures (50–300 residues) solved via X-ray crystallography or cryo-EM at ≤ 4.0 Å resolution. To prevent data leakage, we removed sequences sharing $> 30\%$ identity (≥ 0.8 coverage) with the training set using MMseqs2 (Steinegger & Söding, 2017). The final set was deduplicated by clustering at 30% sequence identity to remove internal redundancy.

Probing evaluation set We compiled 8,902 human-reviewed, single-domain UniProt proteins (2,632,217 tokens) (The UniProt Consortium, 2023; Mistry et al., 2025) that possess high-quality X-ray structures in the PDB (≤ 4.0 Å resolution, $R_{\text{free}} < 0.4$) (Querino Lima Afonso et al., 2025). As before, sequences sharing $> 30\%$ identity with the training set were excluded (Steinegger & Söding, 2017). Using UniProt per-residue annotations, we evaluated four features spanning local secondary structure to rare functional motifs: `HELIX`, `STRAND`, `DISULFID`, and `ZN_FING`. To prevent leakage between correlated residues, the data was partitioned into a strict 70/10/20 train/validation/test split at the protein level, which was kept consistent across all probe and SAE evaluations.

B.2. Metrics

Explained variance We measured activation-level reconstruction fidelity using explained variance,

$$\text{EV}(x, \hat{x}) = 1 - \frac{\text{Var}(x - \hat{x})}{\text{Var}(x)}$$

where x was the dense activation extracted from a target model timestep and $\hat{x}$ was its SAE reconstruction. Variance was computed per coordinate and averaged across the activation dimension, with expectations taken over tokens in the evaluation set. $\text{EV} = 1$ corresponded to exact reconstruction, while $\text{EV} = 0$ matched the trivial baseline of predicting the activation mean. We reported EV alongside the sparsity level k to characterise the sparsity–fidelity trade-off.

Linear probing To assess whether SAE features encoded meaningful concepts in a linearly accessible form, we trained independent logistic regression classifiers on four UniProt annotations (The UniProt Consortium, 2023) from our probing dataset. Sequences were passed through the structure prediction model with the target SAE hooked in, yielding per-residue feature tensors. Classifiers were trained on these features against per-residue binary labels using balanced class weights,

and evaluated on held-out test sets via AUPRC and AUROC. We reported scores alongside positive-class prevalence as a random-guessing baseline. To contextualise absolute values, we compared each SAE probe against a PCA baseline at matched dimensionality and reported 95% confidence intervals from 1,000 protein-level bootstrap resamples.

B.3. Automated Interpretability

Automated interpretability aims to generate natural-language descriptions of what each SAE feature represents by examining the contexts in which it activates, using a language model as an annotator in place of manual inspection. We capture activations by running forward passes through the structure prediction models on the corpus of evaluation proteins, with SAE checkpoints hooked in at their trained timesteps in the diffusion module. For each residue, we record the top-k with the matched k feature activations along with the corresponding UniProt residue-level annotations (The UniProt Consortium, 2023). These top-activating features and annotations are then passed to Claude Opus 4.6 (Anthropic, 2026) via the Anthropic API, which is prompted to synthesise a natural-language description of each feature from the patterns present in its activation profile.

C. Ablation

Table 3. Reconstruction quality and dead features across the K × expansion factor sweep at five diffusion timesteps, for both Boltz and AF3 backbones. Explained variance (↑) computed on held-out activations; dead features reported as percentage of total dictionary size. Aux penalty values shown are for Boltz; AF3 used a constant aux penalty of 0.25 across all runs.

				Boltz		AlphaFold3	
Timestep	K	Expansion	Aux Penalty	EV ↑	Dead % ↓	EV ↑	Dead % ↓
$t = 199$	16	16×	0.0625	0.9946	17.6	0.8843	0.0
	16	32×	0.0313	0.9947	52.8	0.8930	0.0
	16	64×	0.0156	0.9946	74.9	0.8978	0.0
	32	16×	0.0625	1.0000	96.7	0.9223	0.0
	32	32×	0.0313	1.0000	98.4	0.9250	0.0
	32	64×	0.0156	0.9984	82.4	0.9304	0.0
	64	16×	0.0625	1.0000	96.2	0.9486	0.0
	64	32×	0.0313	1.0000	98.0	0.9517	0.0
	64	64×	0.0156	1.0000	99.0	0.9540	0.0
$t = 146$	16	16×	0.0625	0.9769	0.0	0.7675	0.0
	16	32×	0.0313	0.9766	0.0	0.7765	0.0
	16	64×	0.0156	0.9764	49.3	0.7850	0.0
	32	16×	0.0625	0.9909	0.1	0.8283	0.0
	32	32×	0.0313	0.9908	1.6	0.8347	0.0
	32	64×	0.0156	0.9909	59.0	0.8409	0.0
	64	16×	0.0625	0.9973	53.4	0.8808	0.0
	64	32×	0.0313	0.9973	79.5	0.8852	0.0
	64	64×	0.0156	0.9973	88.9	0.8891	0.0
$t = 100$	16	16×	0.0625	0.9170	0.0	0.7633	0.0
	16	32×	0.0313	0.9111	0.0	0.7736	0.0
	16	64×	0.0156	0.9108	0.0	0.7836	0.0
	32	16×	0.0625	0.9588	0.0	0.8247	0.0
	32	32×	0.0313	0.9617	0.1	0.8313	0.0
	32	64×	0.0156	0.9587	2.5	0.8383	0.0
	64	16×	0.0625	0.9849	4.6	0.8790	0.0
	64	32×	0.0313	0.9849	30.4	0.8838	0.0
	64	64×	0.0156	0.9848	61.1	0.8881	0.0
$t = 50$	16	16×	0.0625	0.9170	0.0	0.8049	0.0
	16	32×	0.0313	0.9169	0.0	0.8152	0.0
	16	64×	0.0156	0.9166	0.1	0.8255	0.0
	32	16×	0.0625	0.9617	0.0	0.8526	0.0
	32	32×	0.0313	0.9617	0.1	0.8597	0.0
	32	64×	0.0156	0.9616	15.5	0.8684	0.0
	64	16×	0.0625	0.9852	21.7	0.8597	0.0
	64	32×	0.0313	0.9852	55.6	0.9049	0.0
	64	64×	0.0156	0.9852	76.0	0.9092	0.0
$t = 0$	16	16×	0.0625	0.9301	0.0	0.8078	0.0
	16	32×	0.0313	0.9300	17.6	0.8182	0.0
	16	64×	0.0156	0.9298	0.1	0.8285	0.0
	32	16×	0.0625	0.9686	2.0	0.8550	0.0
	32	32×	0.0313	0.9687	39.2	0.8623	0.0
	32	64×	0.0156	0.9687	24.6	0.8689	0.0
	64	16×	0.0625	0.9878	24.6	0.9008	0.0
	64	32×	0.0313	0.9879	55.3	0.9054	0.0
	64	64×	0.0156	0.9878	79.4	0.9100	0.0

D. Dataset construction input signal ablation

To investigate model reliance on distinct input modalities, we curated a dataset featuring controlled variations in both Multiple Sequence Alignment (MSA) depth and sequence identity. We selected representative sequences from UniRef50 clusters containing at least 1,000 members (Suzek et al., 2015) and retrieved their corresponding MSAs from OpenProteinSet (Ahdritz et al., 2023). The resulting set was filtered to approximately 2,000 proteins meeting a minimum raw MSA depth of 500 and a $N_{\text{eff}} \geq 200$.

For each protein, we performed an input signal titration by randomly subsampling the raw MSA to discrete sequence counts of $N \in \{500, 200, 100, 30, 15, 0\}$. The $N = 0$ condition served as the single-sequence limit, effectively depriving the model of co-evolutionary information. As a null baseline, we generated a fully masked ("all-X") analog for each sequence, where all canonical amino acids were replaced by the generic `X` token to ablate both sequence identity and evolutionary context while maintaining the original sequence length.

We executed forward passes across the complete titration ladder and the all-X baselines. During these passes, five SAEs were hooked into the model—one for each diffusion timestep $t \in \{0, 50, 100, 146, 199\}$. This setup enabled a granular analysis of how varying levels of co-evolutionary and sequence-specific information shaped the features decoded by the SAEs throughout the generation process.

E. Feature classification

E.1. Effective Rate Metric

For each feature f, protein p, and condition c, we compute an *effective rate*:

$$r_{f,p,c} = \frac{1}{L_p} \sum_{i=1}^{L_p} a_{f,p,c,i}$$

where $a_{f,p,c,i}$ is the SAE activation of feature f at residue i for protein p under condition c, and L_p is the sequence length of the protein. This captures both how frequently a feature fires (rate) and its activation magnitude when active.

We pool all nonzero effective rates across the feature, protein, and condition dimensions for the three core classification conditions: `all-X`, `titrate_0`, and `full_MSA`. We define the ON threshold (τ_{on}) and the OFF threshold (τ_{off}) as the 75th and 25th percentiles of the pooled distribution, respectively. Features exhibiting zero firings across all four classification conditions are excluded from this pooling step. This ensures that the large density of inactive features at higher timesteps does not artificially deflate the threshold percentiles.

For each feature, we calculate its mean effective rate across all proteins within each of the four classification conditions. This profile is encoded into a four-character signature over the condition set {`all-X`, `titrate_0` and `full_MSA`} as follows.

E.2. Category Assignment

For each feature, we calculated its mean effective rate across all proteins within each of the four classification conditions. This activation profile was encoded into a four-part signature corresponding to the condition set {`all-X`, `titrate_0` and `full_MSA`}. Specifically, a condition was assigned `ON` if the mean effective rate was $\geq \tau_{\text{on}}$, `OFF` if it was $\leq \tau_{\text{off}}$, and `_` otherwise to indicate intermediate activity.

These signatures were then mapped to discrete behavioral categories based on a defined priority order. A feature was classified as **model_prior** if all four positions were `ON`, indicating it remains active even without valid sequence input. It was labeled **low_activity** if all informative positions were `OFF`, meaning it exhibited some firings but never crossed τ_{on} in any classification condition. The **sequence_driven** category was assigned when the `all-X` condition was `OFF`, but the query sequence (`titrate_0`) and MSA-bearing conditions were `ON`, demonstrating a response to sequence identity but invariance to MSA depth. Features were designated as **MSA_activated** if they had at least one `ON` state in `full_MSA`) and no `ON` states in the low-information positions (`all-X`, `titrate_0`), meaning they are specifically recruited by co-evolutionary signal. Conversely, **MSA_silenced** features had at least one `ON` state in the low-information positions `titrate_0`) and`all-X`, but no `ON` states in the high-information positions, indicating active suppression by deep MSAs.

F. Linear probes for designability

F.1. Data

We used the dataset published by Garcia et al. (2025), in which each designed protein sequence is annotated with binary labels corresponding to four experimental outcomes - expression, solubility, monomericity, and circular dichroism (CD), as well as an aggregate label denoting overall experimental success. To prevent data leakage, the sequences were partitioned using a homology-based split at a 30% sequence identity threshold with MMseqs2 (Steinegger & Söding, 2017).

F.2. Probe training

To assess whether SAE latent features encode information predictive of experimental protein design outcomes, we trained ℓ_1-regularised logistic regression probes on mean-pooled SAE activations. For each protein, per-residue SAE activation vectors $\mathbf{a}_i \in \mathbb{R}^d$ were averaged across residue positions to obtain a single protein-level representation $\bar{\mathbf{a}} = \frac{1}{L}\sum_{i=1}^{L} \mathbf{a}_i$, where L is the sequence length and d is the SAE dictionary size.

F.3. Evaluation

Features were standardised to zero mean and unit variance before fitting a logistic regression classifier with ℓ_1 penalty (regularisation strength $C = 0.1$, coordinate-descent solver, maximum 2,000 iterations). The ℓ_1 penalty encourages sparse coefficient vectors, selecting a compact subset of SAE features most informative of the target label. We evaluated predictive performance via 10-fold stratified cross-validation, collecting out-of-fold predicted probabilities $\hat{p}_i$ for every protein i, from which we computed AUROC and AUPRC on the pooled held-out predictions.

We compared the SAE probe to the pLDDT baseline using a paired bootstrap test on the pooled out-of-fold predictions ($B = 10,000$ resamples), reporting the mean AUROC difference, its 95% percentile confidence interval, and a one-sided p-value. Resamples containing only one class were discarded.

F.4. Feature analysis

To identify which SAE latents drive predictive performance, we refit the L1-logistic regression within each of the 10 CV folds and ranked features by $|\beta_j|$ per fold. We report the top-k selection frequency across folds (with $k = 50$), the mean absolute coefficient, and the mean signed coefficient with its across-fold standard deviation; features selected in $\geq 7/10$ folds are considered stable. Retained features are cross-referenced against precomputed auto-interpretation summaries to obtain human-readable descriptions. This cross-fold criterion guards against features that are artefacts of a single partition.