Biological Features Reveal Early Functional Commitment in Diffusion Protein Language Models

Chuyang Zhou[1], **Bingxin Zhou**[2], **Andi Han**[1], **Chang Xu**[1]

[1]The University of Sydney, [2]Shanghai Jiao Tong University

{chuyang.zhou,andi.han,c.xu}@sydney.edu.au

bingxin.zhou@sjtu.edu.cn

Abstract

Generative protein models are widely used for functional sequence design, yet the process by which biological information is organized during generation remains poorly understood. Diffusion protein language models provide a tractable setting for this question, as sequence generation proceeds along a denoising trajectory from highly corrupted states to complete proteins. This trajectory enables a temporal view of interpretability, revealing when biological signals emerge and when specific residues become committed. We analyze DPLM representations using sparse autoencoders trained across layers and noise levels. The resulting sparse features capture biologically meaningful information across residue and protein scales while preserving downstream predictive performance. Following these features along the denoising trajectory reveals a consistent functional ordering: catalytic-site-enriched features pre-activate at still-masked catalytic positions before residue identity is resolved, and catalytic residues are subsequently recovered earlier by the iterative denoiser. This functional prioritization is robust across multiple confound controls and depends on native protein sequence context. Across ProteinGym deep-mutational-scanning assays, residues recovered earlier during denoising are also more mutation-sensitive. Together, these findings reveal a biologically organized sequence of functional commitments during protein generation, highlighting denoising trajectories as a valuable source of biological information for understanding generative protein models and guiding future development of function-aware decoding strategies and constrained protein sequence design.

1 Introduction

Mechanistic analyses of protein language models have mostly studied representations of *fully observed protein sequences* — the model is run once on an unmasked native protein, and interpretability methods ask what biological structure is encoded in the resulting residue embeddings. A typical instance is ESM-2 [Lin et al., 2023] run on an observed sequence, with subsequent analyses operating on those embeddings [Simon and Zou, 2025, Adams et al., 2025, Gujral et al., 2025, Liu et al., 2026b]. This has been productive, but it leaves

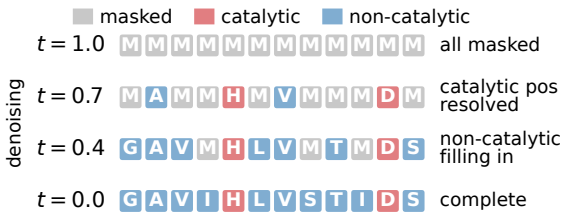

Figure 1: **Biologically ordered denoising.** DPLM recovers catalytic positions earlier on average than surrounding non-catalytic positions during denoising.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

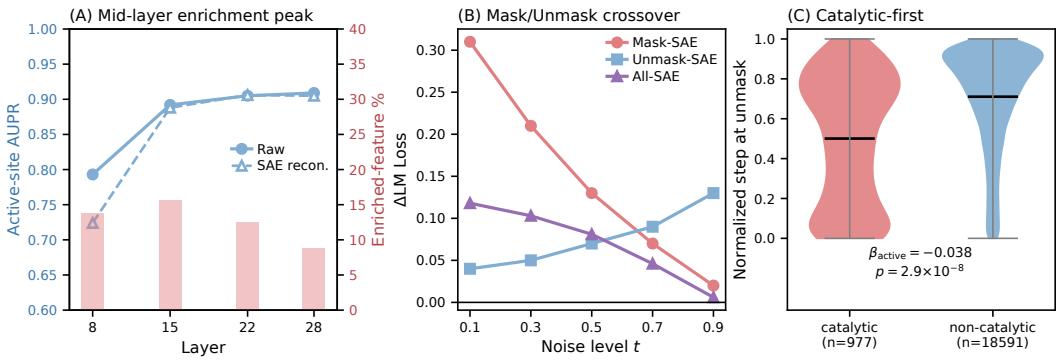

Figure 2: **Overview of main results.** (A) Active-site linear-probe AUPR is preserved under SAE reconstruction, while the fraction of active-site-enriched features peaks in earlier layers. (B) Mask-SAE and Unmask-SAE exhibit opposite noise-dependent reconstruction-loss trends at L28, while All-SAE remains nearly flat. (C) Catalytic positions are recovered earlier than non-catalytic positions during iterative denoising on the held-out active-sites test split.

out a temporal question: *when does biological information enter a generative computation before the sequence is fully known*?

Diffusion protein language models (DPLMs) [Wang et al., 2024, 2025] make this question explicit. They generate sequences through iterative masked denoising over a full noise schedule $t \in [0, 1]$, so each forward pass is one stage of a longer recovery process. Biological information may therefore be organized not only in the representations available at a given layer, but also in the order in which residues and constraints become resolved. This makes residue recovery time a mechanistic variable.

We analyze DPLM using sparse autoencoder features trained across layers and noise levels. The sparse features first provide a static readout: they preserve downstream biological signal under reconstruction and align with residue- and protein-level functional annotations (Figure 2A). The features exhibit biological class specificity, with class-enriched feature sets activating most strongly on their corresponding biological annotations.

The same features also provide a temporal readout of denoising. Feature activity changes systematically across the noise schedule, and biology-enriched features strengthen as sequence context becomes available (Figure 2B). Our central finding is *catalytic-first recovery*: catalytic-site-enriched features pre-activate at still-masked catalytic positions before their residue identities are resolved, and these positions are subsequently recovered earlier during DPLM's iterative denoising (Figure 2C). This temporal pattern persists under confound controls and is absent under a random-feature control group. The matched ESM-2 iterative-fill baseline shows a weaker version of the catalytic-first bias, suggesting that diffusion training amplifies a temporal tendency already latent in masked protein prediction.

Finally, we connect the denoising-time signal to experimentally measured functional constraints. Across ProteinGym deep-mutational-scanning assays, residues recovered earlier by DPLM exhibit greater mutation sensitivity, demonstrating that recovery order captures biologically meaningful information beyond catalytic-site annotations. These findings can further motivate trajectory-informed approaches to function-aware decoding and constrained protein sequence design.

2 Related Work

Interpreting protein language models. Protein language models encode information about structure, function, and evolutionary constraint in their internal representations [Lin et al., 2023, Zhou et al., 2024, Tan et al., 2025a,b]. Interpretability work has studied these representations through attention patterns, probing, and task-specific analyses, showing that biological variables can be localized in model internals rather than only in downstream predictions [Vig et al., 2021, Wenzel et al., 2024, Li et al., 2025, Liu et al., 2025, 2026a, Wang et al., 2026b]. This literature establishes the central premise that PLM computation can be compared against external biological annotations. Most analyses,

however, evaluate the model once on an observed (unmasked) sequence, so the biological question is framed as what information is present in the representation of a fully observed protein.

Sparse features in protein representations. Sparse autoencoders provide a feature-level decomposition of dense activations [Bricken et al., 2023, Cunningham et al., 2023, Gao et al., 2024, Templeton et al., 2024]. In protein models, InterPLM introduced residue-level feature analyses on ESM-2, with subsequent work extending SAE interpretability studies to larger protein models and identifying features aligned with sequence motifs, binding sites, structural domains, Gene Ontology terms, and downstream biological properties [Simon and Zou, 2025, Adams et al., 2025, Gujral et al., 2025, Liu et al., 2026b]. These studies provide practical validation approaches for feature interpretability, including feature localization, annotation enrichment, and reconstruction-based evaluation of biological signal. We adopt this feature-level view in a setting where the same residue can be represented while still masked and subsequently recovered during generation.

Diffusion language models and trajectory interpretation. Diffusion language models generate by iteratively denoising masked tokens rather than by a single left-to-right pass. DPLM trains this objective for proteins across all masking rates on UniRef50 [Wang et al., 2024], and DPLM-2 extends it to joint sequence-structure generation [Wang et al., 2025]. Text diffusion models such as LLaDA and Dream show that masked diffusion can scale as a general language-modeling paradigm [Nie et al., 2025, Ye et al., 2025]. DLM-Scope [Wang et al., 2026a] demonstrates that sparse features in text diffusion models vary over the noise schedule and can expose computation that single-pass analyses on fully observed inputs miss. Protein diffusion adds residue-resolved external structure to this trajectory: each denoising step can be aligned with catalysis, binding, conservation, structure, and mutation sensitivity. This makes it possible to connect feature-level biological representations with the order in which a generator commits to functionally constrained residues.

3 Background and Analysis Setup

Here we define the analysis objects used in the paper. Our analysis uses one representation-level object, sparse features, to evaluate what biological structure DPLM encodes, and two denoising-time objects, t_{feat} and t_{unmask}, to evaluate when that structure appears relative to discrete residue recovery.

3.1 DPLM as an iterative denoiser

We study DPLM-650M [Wang et al., 2024], a 33-layer transformer (d_{model}=1280) trained on UniRef50 with discrete absorbing-state masked diffusion. The forward (noising) process corrupts the fully observed sequence $\mathbf{x}_0 \in \mathcal{V}^L$ token-wise into a partially masked sequence $\mathbf{x}_t$,

$$q(x_t^i \mid x_0^i) = \mathrm{Cat}\big(x_t^i,\ (1-t)\,\mathbf{e}_{x_0^i} + t\,\mathbf{e}_{[\mathrm{M}]}\big), \qquad q(\mathbf{x}_t \mid \mathbf{x}_0) = \prod_{i=1}^{L} q(x_t^i \mid x_0^i), \tag{1}$$

where $t \in [0,1]$ is the masking rate, [M] is the absorbing mask token, and $\mathbf{e}_v$ is the one-hot vector at vocabulary entry v. The reverse model $p_\theta(x_0^i \mid \mathbf{x}_t)$ is trained with a weighted token-wise cross-entropy reconstruction loss across all noise levels, so unlike a fixed-rate masked language model DPLM is not restricted to a single masking fraction.

At inference, the denoiser is run iteratively: at each step the masked positions are scored by the current p_θ, then a subset is unmasked, and the mask rate is lowered. Each residue thus has a well-defined discrete recovery time $t_{\text{unmask}}(i) \in [0,1]$, the normalized step at which residue i is first fixed. This discrete schedule is the main observable for the catalytic-first analyses below.

3.2 Sparse features from residual-stream activations

For a residual-stream activation $\mathbf{h} \in \mathbb{R}^d$ we train Top-K sparse autoencoders [Gao et al., 2024]:

$$\mathbf{z} = \mathrm{TopK}(W_{\text{enc}}(\mathbf{h} - \mathbf{b}_{\text{dec}}) + \mathbf{b}_{\text{enc}}), \qquad \hat{\mathbf{h}} = W_{\text{dec}}\,\mathbf{z} + \mathbf{b}_{\text{dec}}, \tag{2}$$

with D=16,384 dictionary entries and K=64 active features per residue. The reconstruction loss is $\|\mathbf{h} - \hat{\mathbf{h}}\|_2^2$ and decoder columns are unit-normalized after each step. Activations are collected at layers $\ell \in \{3, 8, 15, 22, 28, 32\}$ from 50,000 UniRef50 sequences corrupted at $t \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

Unless stated otherwise, all SAEs are trained on the union of masked and unmasked positions across all noise levels ("All-SAE"). Full training hyperparameters are in Appendix A.

Sparse features are evaluated at two scales. At the residue level, VenusX [Tan et al., 2026] provides active-site, binding-site, and conserved-region annotations. Reconstruction fidelity is assessed by comparing linear-probe performance on raw DPLM activations and SAE-reconstructed activations across the classification tasks. Feature enrichment tests each SAE feature's top-activating residues for annotation overlap using Fisher's exact test with Benjamini and Hochberg FDR correction (α=0.05).

At the protein level an amino-acid (AA) identity-matched null distribution rules out trivial single-AA-detector readings (Appendix B.3). These readouts determine whether feature activations can be treated as biological variables in the denoising-time analyses below.

3.3 Denoising-time commitment measurements

We next define the temporal variables used to test biological commitment. From the VenusX active-sites *train* split we identify a catalytic-enriched feature set F_{cat} at each of L15/L22/L28 (top-20 active-site-enriched features per layer at BH-corrected $p < 0.05$). A layer-matched non-enriched set F_{rand} at fixed seed serves as a null. Held-out evaluation uses the disjoint 148-protein VenusX active-sites *test* split. During iterative denoising of a held-out protein we track, for each residue i, the discrete recovery time $t_{\text{unmask}}(i) \in [0, 1]$ (defined above) and the *latent feature time* $t_{\text{feat}}(i) \in [0, 1]$, the normalized step at which any feature in F_{cat} first crosses its train-pool q_{95} activation threshold at i. The latent preactivation indicator

$$P_i = \mathbf{1}[\, t_{\text{feat}}(i) < t_{\text{unmask}}(i) \,] \tag{3}$$

records whether a catalytic feature becomes active at residue i *before* the discrete denoiser commits to unmasking it. Discrete and latent schedules can then be tested with the same regression specification on different left-hand-side variables.

3.4 Statistical specification and controls

The catalytic-first analyses of Section 5 fit a cluster-robust OLS family:

$$y_i = \alpha + \sum_a \gamma_a \mathbf{1}[\text{AA}_i{=}a] + \beta_{\text{ent}}\, \text{entropy}_z(i) + \beta_{\text{active}} \mathbf{1}_{\text{cat}}(i) + \varepsilon_i$$

with standard errors clustered by protein (CR1, n_p=148). The response y_i is either t_{unmask} in discrete schedule or P_i in latent schedule, and the headline coefficient is β_{active}.

Three sequence-perturbation conditions hold the 50% random-mask pattern fixed across conditions: (i) shuffle only *non*-catalytic residues, keeping every catalytic position and its identity, (ii) full-sequence shuffle, and (iii) force-mask the ± 3 neighbours of every catalytic position. Additional controls include an AA-identity-matched null, 8 Å/12 Åstructural-centrality covariates, and DPLM-clean-input prediction entropy as a conservation proxy.

3.5 Noise-conditioned position roles

To test whether masked and unmasked positions are read out under a shared sparse dictionary, we additionally train two position-specialized variants alongside the All-SAE: a *Mask-SAE* on activations from masked positions only and an *Unmask-SAE* on activations from unmasked positions only, at each layer. They are compared via ΔLM Loss under SAE substitution and decoder-direction overlap with full details in Appendices C.2.

4 Sparse Features Recover Biological Structure in DPLM

We first establish that DPLM sparse features preserve biologically meaningful information and exhibit functional specificity, then examine how these features are organized across model layers and evolve during denoising.

4.1 Biological Signal Preservation and Feature Specificity

The sparse decomposition preserves biological signal under reconstruction. Inserting the SAE reconstruction $\hat{\mathbf{h}}$ in place of DPLM's residual stream incurs only a small cross-entropy cost, with the largest cost concentrated in the middle layers. More importantly, the same projection preserves task-level biological information. Figure 2A shows that active-site prediction remains nearly unchanged under SAE reconstruction at the middle and late layers. At L28, AUPR reaches 0.905 after reconstruction, closely matching the 0.909 obtained from raw activations. This preservation extends to binding-site and conserved-site prediction, as well as secondary structure, thermostability, and subcellular localization. The complete evaluations in Tables 3 and 4 show similarly small reconstruction gaps across these tasks. At the protein level, the sparse code alone also retains substantial predictive signal for Meltome thermostability and DeepLoc localization.

Sparse features are class-selective. We rank the top-20 active-, binding-, and conservation-enriched feature sets at L15/L22/L28 from disjoint VenusX-train pools, and measure the mean per-position max-over-set activation on positions annotated as each class on the held-out test split. Figure 3 shows that each feature set activates most strongly on its corresponding biological class. Own-class activation exceeds the strongest off-class entry by $2.1\times$, $4.2\times$, and $4.3\times$ for active-, binding-, and conservation-enriched features, respectively. Individual enriched features also localise near annotated catalytic positions in held-out enzymes, as illustrated by the residue-level activation profiles in Figure 8.

Several enriched features carry strong AA-identity signatures, so raw enrichment alone could overstate biological specificity. We therefore recompute enrichment against an AA-matched null that accounts for each feature's top-activating amino-acid composition. The active-site features remain strongly enriched under this matched background, and most binding-site examples do as well. Across 38 feature–annotation pairs with raw enrichment of at least $3\times$, the median AA-matched enrichment reaches $27.7\times$, with 76% significant under matched permutations. The full analysis in Appendix B.3 supports biological specificity beyond amino-acid composition for a substantial subset of enriched features.

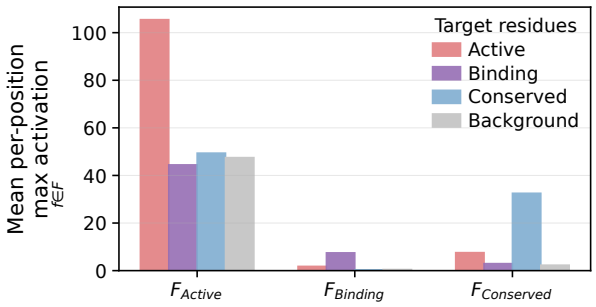

Figure 3: **Sparse features are class-selective.** Top active-, binding-, and conserved-enriched feature sets activate most strongly on held-out residues from the matching class. Values are averaged over L15/L22/L28.

4.2 Biological Features Across Layers and Denoising Time

Sparse biological enrichment peaks in the middle layers. Figure 2A reveals distinct layer-wise patterns for predictive accessibility and sparse biological factorization. The fraction of active-site-enriched SAE features peaks at L15 (15.6%), compared with 8.8% at L28, while raw active-site AUPR increases from 0.892 to 0.909 across the same layers. Reconstruction cost is also highest in the middle layers. These observations suggest that biologically enriched sparse features are more concentrated in the middle layers, even as later representations support stronger linear predictions.

Biology-enriched features change activity as sequence context becomes available. Across a controlled mask-fraction sweep, feature-set Jaccard similarity increases, activation entropy decreases, and feature turnover slows as masking is reduced, revealing progressively more stable sparse representations, as shown in Figure 9. This global pattern is also reflected in biologically enriched features. Table 1 shows that the mean activations of an active-site feature and a catalytic-motif feature at L15 increase by $2.35\times$ and $2.33\times$, respectively, between fully masked and clean inputs.

The denoising axis also changes how sparse representations are used at masked and unmasked positions. Figure 4 shows that Mask-SAE reconstruction loss decreases as masking increases, whereas Unmask-SAE loss rises in the middle and deep layers. All-SAE maintains relatively low loss

Table 1: **Biological feature activation across noise levels.** Mean SAE activation of two annotated L15 features under a controlled mask-fraction sweep. Both features strengthen as sequence context becomes available ($t = 1$ fully masked; $t = 0$ fully observed).

Feature	$t = 1.0$	0.8	0.6	0.4	0.2	0.0	Ratio
Active site	86.5	156.6	171.4	183.2	191.9	203.5	2.35×
Catalytic motif	66.7	103.4	121.3	138.9	149.4	155.7	2.33×

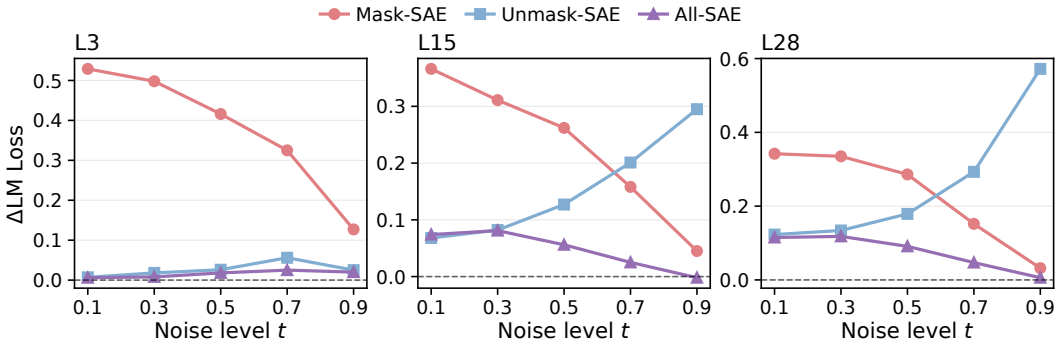

Figure 4: **Masked and unmasked positions use sparse features differently over time.** Mask-SAE loss falls as masking increases, whereas Unmask-SAE loss rises in middle and deep layers. All-SAE stays low across the schedule, indicating that a shared dictionary can cover both regimes while activation statistics differ.

across the noise schedule. Decoder-direction matching further reveals substantial overlap between the specialized dictionaries, with median best-match cosine similarities of 0.76 at L15 and 0.72 at L28, as detailed in Appendix C.2. These findings indicate that masked and unmasked positions access related sparse feature directions under different activation statistics.

These noise-dependent representations provide a basis for tracking catalytic-enriched features at still-masked positions during iterative denoising, where their activation timing can be compared with discrete residue recovery.

5 DPLM Commits Early to Catalytic Residues

During iterative denoising, residue identities are resolved at different steps. Catalytic residues provide a stringent test of biological ordering: they are sparse, functionally central, and embedded in diverse sequence contexts. We examine when these residues are recovered and whether their functional identity emerges earlier in the sparse representations.

5.1 Catalytic-First Recovery Is Biologically Specific and Context-Dependent

Figure 5A shows that catalytic residues are recovered substantially earlier than non-catalytic residues under native sequence context. The unadjusted difference corresponds to 21% of the normalized recovery trajectory. Prediction entropy accounts for much of this separation, yet catalytic identity retains an additional early-recovery association after entropy adjustment. The coefficient decreases from -0.210 to -0.038, remaining significant with a p-value of 2.9×10^{-8}. Thus, catalytic positions are recovered approximately 4% of the trajectory earlier than non-catalytic positions with comparable prediction entropy. The complete regression specifications are reported in Table 12.

Sequence perturbations separate global context from local completion. The perturbation experiments in Figure 5B reveal a strong dependence on native sequence context. Shuffling non-catalytic residues reduces the entropy-adjusted catalytic coefficient from -0.038 to -0.007, while full-sequence shuffling produces a similarly negligible effect of -0.009. In contrast, force-masking the ± 3 neighbours of catalytic positions preserves the early-recovery effect at -0.041. The persistence under local masking, together with its disappearance under sequence shuffling, suggests that

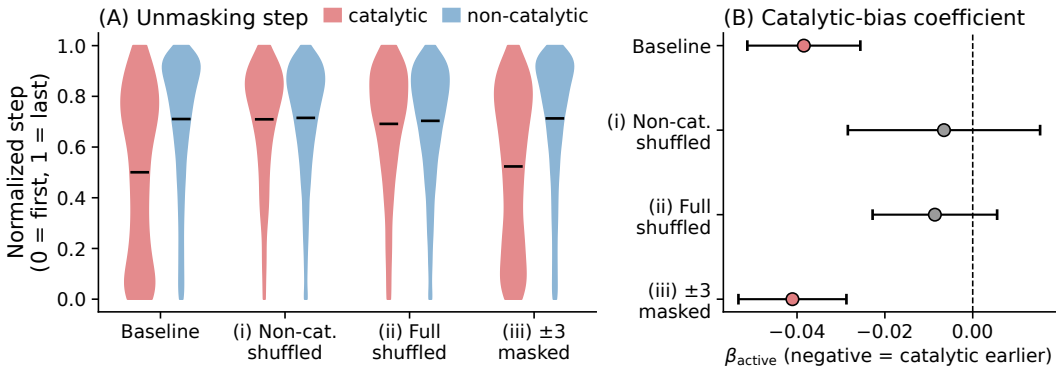

Figure 5: **Catalytic residues are recovered earlier.** Catalytic positions are resolved earlier than non-catalytic positions under native context. The effect disappears under sequence shuffling and persists when local neighbours are masked.

catalytic-first recovery draws on broader protein context. This pattern remains consistent across initial mask fractions of 30–70% and decoding schedules of 100–500 steps, as further documented in Appendix D.

Catalytic identity carries a distinct early-recovery signal. The association remains stable under additional controls for amino-acid identity, structural centrality, and conservation-related sequence information. Adding amino-acid fixed effects leaves the catalytic coefficient nearly unchanged at -0.038. Structural covariates derived from ESMFold and clean-input prediction entropy likewise preserve the early-recovery association, with the fully adjusted structural/conservation model yielding a coefficient of -0.042. The complete control analyses in Appendix D further examine whether this preference extends to other biological annotations. The cross-task evaluation on 120 proteins further examines whether early recovery extends to other biological annotations. Although catalytic, binding, and conserved residues all exhibit earlier unadjusted recovery, the association for catalytic residues remains five to seven-fold stronger after controlling for prediction entropy, while those for binding and conserved residues become negligible. This pattern provides further evidence that catalytic-first recovery reflects a functionally selective timing bias.

5.2 Catalytic Features Pre-Activate Before Residue Recovery

The catalytic-first recovery pattern is also reflected in sparse feature activations before residues are unmasked. We track the catalytic-enriched feature set F_{cat} during iterative denoising and record the preactivation indicator $P_i = \mathbf{1}[t_{\text{feat}}(i) < t_{\text{unmask}}(i)]$ defined in Section 3.3. Using the top-20 features per layer at the q_{95} activation threshold, preactivation occurs at 16.6% of catalytic positions, compared with only 5.8% of non-catalytic positions. Figure 6A shows that this association persists after controlling for amino-acid identity and within-protein prediction entropy, with an adjusted effect of 9.1 percentage points and a cluster-robust p-value of 2.5×10^{-4}. In contrast, a random non-enriched feature set F_{rand} exhibits the opposite pattern, activating less frequently before recovery at catalytic positions. This contrast indicates that the preactivation bias is associated with catalytic-enriched feature directions rather than a general tendency for sparse features to activate before unmasking.

The latent preactivation effect is robust across feature-set size, threshold, and layer. Expanding the catalytic feature set from top-10 to top-50 strengthens the preactivation bias, as does raising the activation threshold from q_{90} to q_{99}. Each of L15, L22, and L28 independently exhibits a positive association, with stronger effects when their features are combined. The complete sensitivity analyses are reported in Appendix D.3. The context perturbations reveal a similar pattern to discrete recovery: full-sequence shuffling eliminates the preactivation bias, whereas force-masking the ± 3 neighbours of catalytic positions preserves it, as shown in Figure 6. These findings connect the emergence of catalytic-associated information in sparse representations with the subsequent recovery of catalytic residues.

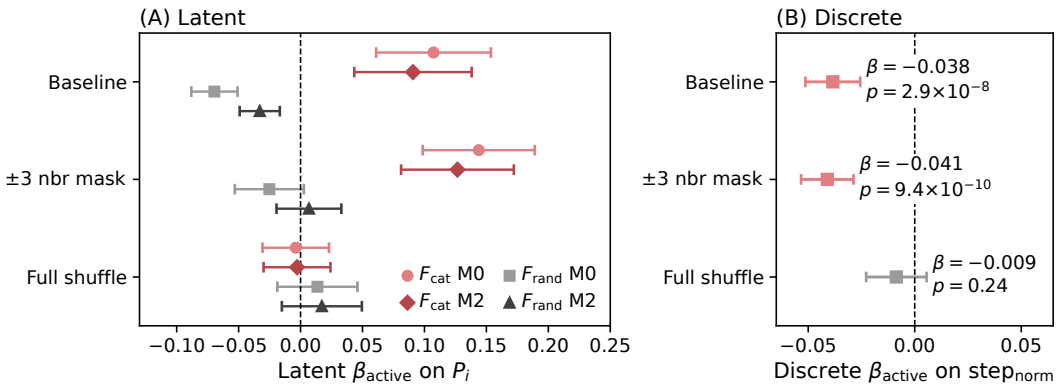

Figure 6: **Catalytic features pre-activate before recovery.** Active-site-enriched features fire at still-masked catalytic residues before unmasking, while random non-enriched features show the opposite pattern. (A) Latent preactivation coefficient β_{active} for F_{cat} vs. random F_{rand} on the indicator $P_i{=}\mathbf{1}[t_{\text{feat}}{<}t_{\text{unmask}}]$. Positive coefficients indicate greater preactivation at catalytic positions. (B) Discrete-schedule coefficient β_{active} on the per-residue normalized denoising step in $[0, 1]$. Negative coefficients indicate earlier catalytic recovery. Both panels report cluster-robust 95% CIs.

5.3 Diffusion Amplifies a Masked-Prediction Timing Bias

We further evaluate matched ESM-2 iterative-fill baseline, using SAEs trained on ESM-2 650M with the same dictionary size and sparsity and then running ESM-2 iteratively at the same initial mask rate. It produces the same direction at reduced magnitude (Appendix Table 15). Static per-feature discriminability is essentially the same between the two objectives at matched scale: the best individual SAE feature for VenusX active-site prediction reaches near-identical max-AUPR for DPLM and ESM-2 at L15 and L22, and decoder-direction overlap with matched-size ESM-2 SAEs is substantial but not saturated. Diffusion training amplifies a timing tendency that fixed-rate masked prediction already partially encodes.

6 Recovery Order Aligns with Mutation Sensitivity

The catalytic-first recovery pattern in Section 5 is established using active-site annotations, which also guide the selection of F_{cat}. To examine whether this temporal ordering reflects broader functional constraints, we turn to ProteinGym substitution deep-mutational-scanning (DMS) assays, which provide independent experimental measurements of mutation sensitivity.

6.1 Recovery Order Tracks Experimental Mutation Sensitivity

Within each assay, we quantify residue-level mutation sensitivity by averaging the rank-normalised deleteriousness of single substitutions at each wild-type position. We compare this measure with $-t_{\text{unmask}}$, where larger values indicate earlier recovery during DPLM denoising.

Figure 7A reveals a consistent relationship between recovery order and experimental mutation sensitivity. Residues resolved early in the denoising trajectory tend to be more sensitive to mutations, with sensitivity generally decreasing across later recovery-time deciles. Across 61 ProteinGym assays, 82% show positive correlations between earlier recovery and greater mutation sensitivity, reaching a median Spearman ρ of 0.287 with a one-sided Wilcoxon p-value of 7×10^{-7}. The relationship is particularly pronounced in the 11 enzyme-activity assays, where the median correlation reaches 0.46. These findings extend the catalytic-first observation beyond VenusX annotations, showing that DPLM recovery order carries information about experimentally measured functional constraints.

The SAE-derived preactivation signal also shows an association with experimental mutation sensitivity. We evaluate the preactivation indicator $\mathbf{1}[t_{\text{feat}}^{q95}{<}t_{\text{unmask}}]$ separately for feature sets enriched for different VenusX annotations. After accounting for recovery time and sequence-prior information, catalytic-enriched features retain a positive association with mutation sensitivity. Binding-enriched features show no detectable association, whereas conservation-enriched features exhibit the opposite direction.

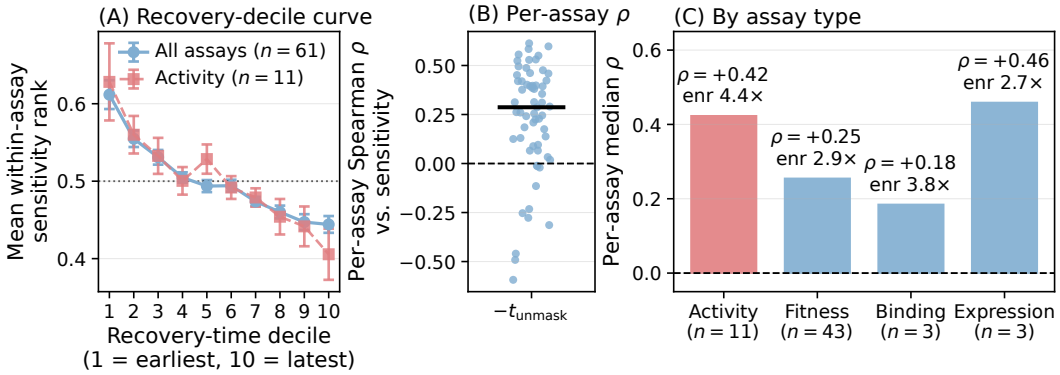

Figure 7: **Recovery order aligns with mutation sensitivity.** Earlier-recovered residues are more mutation-sensitive across ProteinGym assays, especially enzyme-activity assays. (A) Sensitivity by recovery-time decile. (B) Per-assay Spearman correlations. (C) Median correlation by assay type.

This distinction suggests that the functional information captured by latent preactivation depends on the biological specificity of the underlying features.

The pooled SAE preactivation association persists across partial-mask regimes from 30% to 70% and strengthens with stricter activation thresholds from q_{90} to q_{99}. Under full 100% masking, the association disappears, consistent with a dependence on protein-specific sequence context. A sequence-overlap audit identifies only one ProteinGym assay with substantial similarity to the VenusX feature-selection pool; excluding this assay leaves the SAE-derived association essentially unchanged.

6.2 Recovery Order Complements Sequence-Prior Signals

ESM-2 masked-marginal log probabilities provide stronger standalone predictions of mutation sensitivity than DPLM recovery time. Across the 61 ProteinGym assays, ESM-2 achieves a Spearman correlation of 0.471, compared with 0.287 for $-t_{\text{unmask}}$, while DPLM clean-input prediction entropy falls between the two. The masked marginal measures the compatibility of a wild-type residue with its sequence context, whereas recovery time records when the iterative generator resolves its identity.

Despite its lower standalone correlation, recovery order provides complementary information for identifying mutation-sensitive residues. Adding $-t_{\text{unmask}}$ to the sequence-prior baseline improves leave-one-assay-out top-10% sensitivity retrieval AUROC from 0.654 to 0.687 across all assays, with a larger improvement from 0.673 to 0.713 on enzyme-activity assays. Appendix Table 17 provides the full comparison, where SAE-derived timing variables contribute little additional retrieval performance. These results show that the denoising trajectory captures functional information that complements conventional sequence-based predictions.

7 Conclusion and Discussion

This paper presents a residue-level interpretability study of DPLM, revealing how biological information is progressively organized during denoising. Using SAEs across layers and noise levels, we show that DPLM representations contain biologically meaningful features that evolve during generation. Catalytic-enriched features pre-activate at still-masked catalytic positions, which are subsequently recovered earlier by the denoiser. The matched ESM-2 comparison suggests that diffusion amplifies this temporal bias. Recovery order also aligns with mutation sensitivity in ProteinGym assays, linking generative dynamics to experimentally measured functional constraints.

These findings highlight denoising trajectories as a source of biological information for understanding generative protein models. Recovery order could help identify functionally constrained residues during sequence optimization, while catalytic-feature preactivation may inform adaptive masking and function-aware decoding strategies. Together, these directions motivate trajectory-guided, constrained protein sequence design.

The main limitation is that these applications remain to be evaluated. Future work could investigate causal interventions on catalytic-enriched features and assess whether trajectory-informed generation improves functional preservation, sequence diversity, and generation quality.

Acknowledgments and Disclosure of Funding

This work was supported by the National Key Research and Development Program of China (2024YFA0917603); the National Natural Science Foundation of China (62302291); the Computational Biology Key Program of Shanghai Science and Technology Commission (23JS1400600); the AI for Science Program, Shanghai Municipal Commission of Economy and Informatization (2025-GZL-RGZN-BTBX-02009).

References

Etowah Adams, Liam Bai, Minji Lee, Yiyang Yu, and Mohammed AlQuraishi. From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *PMLR*, 2025.

José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. DeepLoc: Prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, 2017. doi: 10.1093/bioinformatics/btx431.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Onkar Gujral, Mihir Bafna, Eric Alm, and Bonnie Berger. Sparse autoencoders uncover biologically interpretable features in protein language model representations. *Proceedings of the National Academy of Sciences*, 122(34):e2506316122, 2025. doi: 10.1073/pnas.2506316122.

Anna Jarzab, Nils Kurzawa, Thomas Hopf, Matthias Moerch, Jana Zecha, Niels Leijten, Yangyang Bian, Eva Musiol, Melanie Maschberger, Gabriele Stoehr, Isabelle Becher, Charlotte Daly, Patroklos Samaras, Julia Mergner, Britta Spanier, Angel Angelov, Thilo Werner, Marcus Bantscheff, Mathias Wilhelm, Martin Klingenspor, Simone Lemeer, Wolfgang Liebl, Hannes Hahne, Mikhail M. Savitski, and Bernhard Kuster. Meltome atlas—thermal proteome stability across the tree of life. *Nature Methods*, 17(5):495–503, 2020. doi: 10.1038/s41592-020-0801-4.

Song Li, Yang Tan, Song Ke, Bingxin Zhou, et al. Immunogenicity prediction with dual attention enables vaccine target selection. In *International Conference on Learning Representations*, volume 2025, pages 67913–67925, 2025.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Chen Liu, Mingchen Li, Yang Tan, Wenrui Gou, Guisheng Fan, and Bingxin Zhou. Sequence-only prediction of binding affinity changes: a robust and interpretable model for antibody engineering. *Bioinformatics*, 41(8):btaf446, 2025.

Chen Liu, Bingxin Zhou, Xinyuan Wang, Ming Li, Guisheng Fan, and Liang Hong. Retrieval-augmented multimodal learning for enzyme-substrate interaction prediction under low-homology shift. *arXiv preprint arXiv:2606.22823*, 2026a.

Xiangyu Liu, Haodi Lei, Yi Liu, Yang Liu, and Wei Hu. ProtSAE: Disentangling and interpreting protein language models via semantically-guided sparse autoencoders. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI)*, 2026b.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

Pascal Notin, Aaron W. Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Hansen Spinner, Nathan Rollins, Ada Shaw, Rose Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora S. Marks. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems (NeurIPS Datasets and Benchmarks Track)*, volume 36, 2023.

Elana Simon and James Zou. InterPLM: Discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, 22:2107–2117, 2025. doi: 10.1038/s41592-025-02836-7.

Yang Tan, Ruilin Wang, Banghao Wu, Liang Hong, and Bingxin Zhou. From high-throughput evaluation to wet-lab studies: advancing mutation effect prediction with a retrieval-enhanced model. *Bioinformatics*, 41(Supplement_1):i401–i409, 2025a.

Yang Tan, Bingxin Zhou, Lirong Zheng, Guisheng Fan, and Liang Hong. Semantical and geometrical protein encoding toward enhanced bioactivity and thermostability. *Elife*, 13:RP98033, 2025b.

Yang Tan, Wenrui Gou, Bozitao Zhong, Liang Hong, Huiqun Yu, and Bingxin Zhou. VenusX: Unlocking fine-grained functional understanding of proteins. In *International Conference on Learning Representations (ICLR)*, 2026.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity.

Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. BERTology meets biology: Interpreting attention in protein language models. In *International Conference on Learning Representations (ICLR)*, 2021.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, 2024.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. DPLM-2: A multimodal diffusion protein language model. In *International Conference on Learning Representations (ICLR)*, 2025.

Xu Wang, Bingqing Jiang, Yu Wan, Baosong Yang, Lingpeng Kong, and Difan Zou. DLM-Scope: Mechanistic interpretability of diffusion language models via sparse autoencoders. *arXiv preprint arXiv:2602.05859*, 2026a.

Zhiyu Wang, Bingxin Zhou, Weishu Zhao, Yang Tan, Jing Wang, Pietro Liò, et al. Fast and interpretable protein substructure alignment via optimal transport. In *International Conference on Learning Representations*, volume 2026, pages 37232–37246, 2026b.

Markus Wenzel, Erik Grüner, and Nils Strodthoff. Insights into the inner workings of transformer models for protein function prediction. *Bioinformatics*, 40(3):btae031, 2024. doi: 10.1093/bioinformatics/btae031.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: A diffusion reasoning language model. *arXiv preprint arXiv:2508.15487*, 2025.

Bingxin Zhou, Yang Tan, Yutong Hu, Lirong Zheng, Bozitao Zhong, and Liang Hong. Protein engineering in the deep learning era. *mLife*, 3(4):477–491, 2024.

Appendix Overview

The supplementary material follows the main-text analysis: SAE training and biological validation (Appendices A–B), noise-conditioned feature dynamics (Appendix C), catalytic-first recovery and preactivation (Appendix D), the ESM-2 comparison (Appendix E), and independent ProteinGym validation (Appendix F).

A Training and Experimental Setup

DPLM model. We use the publicly released DPLM-650M [Wang et al., 2024] checkpoint without further fine-tuning. The model is a 33-layer transformer (d_{model}=1280, 20 attention heads) pretrained on UniRef50 with discrete masked diffusion.

SAE training. We train separate Top-K SAEs on residual-stream activations collected across six DPLM layers and five mask fractions. Table 2 reports the architecture, optimization settings, and activation collection protocol. All-SAE models are trained for 50,000 steps; Mask-SAE and Unmask-SAE use shorter, layer-dependent training runs, including 10,000–25,000 steps for some specialized models at L8, L22, and L32.

Table 2: SAE training configuration.

Parameter	**Value**
SAE type	Top-K
Input dimension (d)	1280
Dictionary size (D)	16,384
Expansion ratio	12.8×
Sparsity (K)	64
Optimizer	Adam (β_1=0.9, β_2=0.999)
Learning rate	3×10^{-4}
LR warmup	1,000 steps
Batch size	32,768
Training steps	50,000 (All-SAE) 10,000–30,000 (Mask/Unmask, layer-dependent)
Gradient clipping	1.0
Decoder normalization	Unit norm after each step
Activation collection	
Source data	UniRef50 (50K sequences, max length 512)
Noise levels	$t \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$
Target layers	$\ell \in \{3, 8, 15, 22, 28, 32\}$
Vectors per layer	∼95M (All-SAE, pooled across noise levels)
Position modes	All, Mask, Unmask

B Biological Feature Validation

B.1 SAE Reconstruction Fidelity

VenusX residue-level prediction. Table 3 gives the complete VenusX MF50 test results for active-site, binding-site, and conserved-site prediction. Reconstruction closely preserves AUPR at intermediate and deep layers, with larger gaps in shallow representations. The enrichment percentages peak at L15 for active-site and binding-site features, while the strongest raw predictive performance occurs at later layers.

Secondary structure prediction. We evaluate three-class secondary structure prediction (H/E/C) across six DPLM layers. SAE reconstruction remains within one percentage point of raw accuracy throughout; at L32, raw and reconstructed accuracies reach 0.851 and 0.849, respectively.

Thermostability prediction (Meltome). For Meltome thermostability regression [Jarzab et al., 2020], we use 5,054/639/1,336 proteins for training, validation, and testing. Residue representations

Table 3: Complete VenusX residue-level AUPR results. All values computed on VenusX MF50 test split.

Task	Layer	Raw AUPR	Recon AUPR	Enriched Features
Active Sites	3	0.606	0.504	53/1591 (3.3%)
	8	0.793	0.724	200/1454 (13.8%)
	15	0.892	0.888	284/1826 (15.6%)
	22	0.905	0.906	275/2194 (12.5%)
	28	0.909	0.905	174/1973 (8.8%)
	32	0.881	0.853	112/1959 (5.7%)
Binding Sites	3	0.804	0.738	110/1600 (6.9%)
	8	0.900	0.858	203/1515 (13.4%)
	15	0.942	0.929	292/1950 (15.0%)
	22	0.951	0.939	274/2319 (11.8%)
	28	0.936	0.934	149/2066 (7.2%)
	32	0.933	0.909	97/1998 (4.9%)
Conserved Sites	3	0.613	0.481	457/3314 (13.8%)
	8	0.812	0.724	890/3369 (26.4%)
	15	0.913	0.898	1052/6176 (17.0%)
	22	0.931	0.921	1225/8093 (15.1%)
	28	0.929	0.905	854/6797 (12.6%)
	32	0.909	0.848	476/5228 (9.1%)

are mean-pooled over non-special tokens and evaluated by ridge regression with validation-selected regularization. Table 4 reports test Spearman correlation for raw, reconstructed, and mean-pooled sparse representations.

Subcellular localization (DeepLoc). For ten-class localization [Almagro Armenteros et al., 2017], the evaluation uses 6,622 training proteins after a 10% validation split and 1,842 test proteins. A separate multinomial logistic regression with validation-tuned ℓ_2 regularization is fitted to each representation. At L28, SAE reconstruction reaches 0.768 accuracy versus 0.773 for raw representations.

Table 4 provides the complete results for these three tasks. Reconstruction is close to the raw representations across all evaluated metrics, while the mean-pooled sparse codes show lower but nonzero predictive performance on Meltome and DeepLoc.

Table 4: Additional biological signal preservation results. SSP reports three-class accuracy; Meltome reports test Spearman ρ; DeepLoc reports ten-class accuracy and macro-F1. Gap is reconstruction minus raw. Sparse features are evaluated for Meltome and DeepLoc only.

Task / metric	Layer	Raw	SAE recon.	SAE sparse	Gap
SSP / Accuracy	3	0.724	0.715	—	−0.009
	8	0.743	0.733	—	−0.010
	15	0.776	0.766	—	−0.010
	22	0.800	0.795	—	−0.005
	28	0.838	0.838	—	0.000
	32	0.851	0.849	—	−0.002
Meltome / ρ	15	0.632	0.631	0.338	−0.001
	22	0.621	0.626	0.396	0.005
	28	0.629	0.637	0.363	0.008
DeepLoc / Accuracy	15	0.761	0.754	0.696	−0.007
	22	0.761	0.752	0.698	−0.009
	28	0.773	0.768	0.717	−0.005
DeepLoc / Macro-F1	15	0.631	0.622	0.570	−0.009
	22	0.625	0.629	0.563	0.004
	28	0.649	0.633	0.584	−0.016

B.2 Biological Feature Specificity

For each VenusX class $C \in \{\text{active-site}, \text{binding-biolip}, \text{conserved}\}$, we select the top-20 Fisher-enriched SAE features per layer at L15/L22/L28 using the corresponding VenusX training split. On held-out test proteins, we compute the mean per-position $\max_{f \in F_C}$ activation on residues annotated with each target class C'. Background denotes non-annotated positions in the respective task. Results are averaged across the three layers, using approximately 60 test proteins per class with at least two annotated positions.

Table 5: Mean per-position $\max_{f \in F_C}$ activation for each feature set (rows) and target residue class (columns). Background denotes non-annotated positions in the corresponding task.

	Active	Binding	Conserved	Background
F_{Active}	**105.5**	44.5	49.4	47.5
F_{Binding}	1.79	**7.47**	0.34	0.46
$F_{\text{Conserved}}$	7.56	2.95	**32.5**	2.30

Table 5 shows that each feature set activates most strongly on its corresponding class. Own-class activation exceeds the largest off-class value by 2.1×, 4.2×, and 4.3× for active-site, binding-site, and conserved-site feature sets, respectively. These comparisons are made within each feature-set row, avoiding comparisons between the absolute scales of different SAE features.

Representative catalytic-feature localizations. Figure 8 shows examples of catalytic-enriched feature activations near annotated catalytic residues on held-out VenusX proteins. The quantitative enrichment controls are reported below.

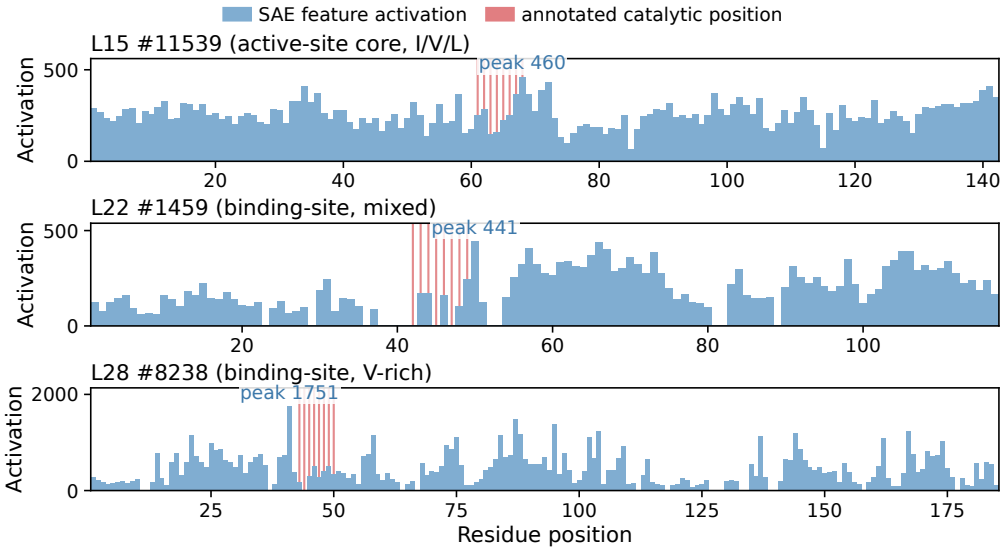

Figure 8: **Example catalytic feature localizations.** Three named SAE features at L15/L22/L28 on three test-split active-sites proteins; per-residue feature activation (blue bars) peaks within 0–2 residues of an annotated catalytic position (red lines) in each panel.

B.3 AA-Identity-Matched Enrichment Controls

Several biologically enriched features also exhibit strong amino-acid preferences. To assess the contribution of amino-acid identity, we construct an AA-matched background for each feature's top-50 activating positions. For its amino-acid composition $\{p_a\}_a$, the expected annotation frequency is

$$\mathbb{E}[\text{freq}_{\text{null}}] = \sum_a p_a \Pr(\text{annotation} \mid \text{AA} = a).$$

The conditional rates are estimated from the pooled VenusX residue-level test splits ($n > 80{,}000$ residues). AA-matched enrichment divides the observed top-50 annotation frequency by this expectation. Empirical p-values use 10,000 random sets of 50 positions sampled with the same AA composition.

Table 6: AA-identity-matched enrichment for selected SAE features. Raw enrichment uses the unmatched annotation background; AA-matched enrichment accounts for the top-50 amino-acid composition. Empirical p-values use 10,000 matched permutations.

Layer	Feature	Annotation	Raw enr	AA-matched enr	Empirical p	Dom. AA
15	#11539	active	12.7	94.8	$< 10^{-4}$	I (32%)
15	#2620	active	13.0	79.2	$< 10^{-4}$	G (53%)
15	#13480	binding	9.2	4.8	0.071	G (45%)
22	#1459	binding	19.0	10.3	$< 10^{-3}$	F (37%)
22	#7071	active	10.2	66.5	$< 10^{-4}$	G (56%)
28	#8238	binding	12.3	9.6	0.003	V (78%)
28	#6054	active	3.8	37.2	$< 10^{-4}$	Y (52%)
28	#6054	binding	11.6	8.1	0.006	Y (52%)

Table 6 shows AA-matched enrichment of at least 4.8 across the listed feature–annotation pairs. All four active-site examples have empirical $p < 10^{-4}$; three of the four binding-site examples have $p < 0.01$, while feature #13480 yields $p = 0.071$. Across 38 pairs with raw enrichment ≥ 3, median AA-matched enrichment is 27.7 and 76% reach empirical $p < 0.05$. These results support feature selectivity beyond amino-acid composition for a substantial subset of enriched features.

C Feature Dynamics Across Noise Levels

C.1 Noise-Conditioned Feature Activity

The controlled mask-fraction sweep evaluates UniRef50 protein sequences at different masking levels, with a separate DPLM forward pass at each level. This analysis characterizes noise-conditioned representations; it is distinct from the iterative generation trajectories examined in Appendix D. Across 20 mask fractions from fully masked to unmasked, Figure 9 shows increasing consecutive-level feature-set similarity and decreasing activation entropy and turnover as sequence context becomes available.

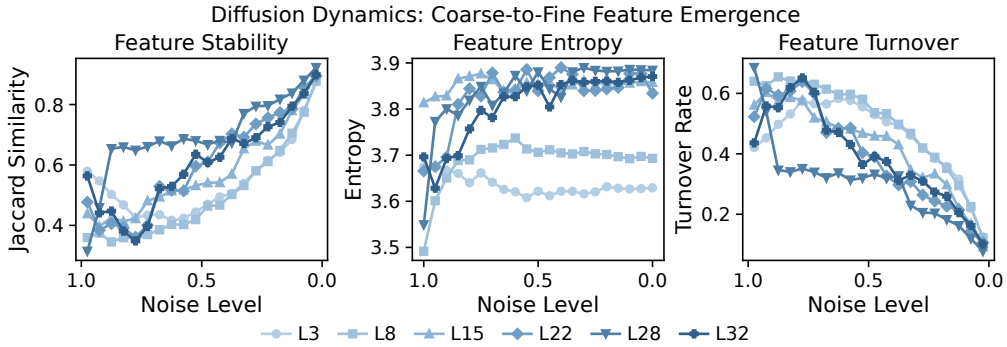

Figure 9: Noise-conditioned sparse-feature dynamics across 20 mask fractions ($t = 1.0$ to $t = 0.0$). Consecutive-level feature-set Jaccard similarity, activation entropy, and feature turnover are reported across layers. Each mask fraction is evaluated by a separate forward pass.

SAE substitution produces a smaller ΔLM loss at higher masking rates, with the intermediate layers approaching zero or slightly negative values at $t = 0.9$ (Figure 10). Negative values indicate a lower cross-entropy after residual-stream replacement under the specified evaluation setup.

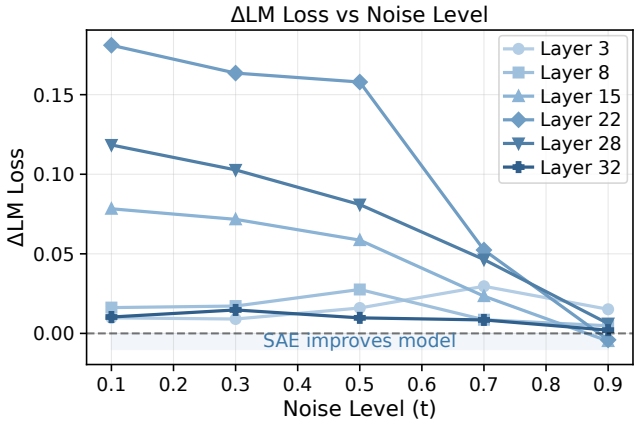

Figure 10: SAE substitution ΔLM loss across mask fractions and DPLM layers. At high masking rates the intermediate-layer reconstruction penalty approaches zero; L15 and L22 are slightly negative at $t = 0.9$.

C.2 Masked and Unmasked Position-Specialized SAEs

Figure 4 summarizes the representative-layer comparisons in the main text. Table 7 provides the full ΔLM results for Mask-SAE and Unmask-SAE at six layers, and for All-SAE at three representative layers. Mask-SAE replacement cost generally decreases as masking increases, whereas Unmask-SAE cost increases most clearly at L15/L22/L28. The specialized models have layer-dependent training budgets, as reported in Appendix A.

Table 7: ΔLM loss across mask fractions. Mask-SAE and Unmask-SAE results cover six layers; All-SAE results cover L3, L15, and L28. Specialized training duration varies by layer.

Mode	**Layer**	t=0.1	t=0.3	t=0.5	t=0.7	t=0.9
Mask	3	0.529	0.498	0.416	0.325	0.127
	8	0.490	0.436	0.341	0.217	0.072
	15	0.366	0.311	0.262	0.158	0.045
	22	0.459	0.462	0.322	0.173	0.027
	28	0.342	0.335	0.286	0.152	0.032
	32	0.019	0.014	0.010	0.006	0.001
Unmask	3	0.007	0.018	0.026	0.056	0.025
	8	0.018	0.027	0.026	0.022	0.004
	15	0.068	0.082	0.127	0.201	0.295
	22	0.199	0.137	0.174	0.176	0.344
	28	0.123	0.134	0.179	0.293	0.572
	32	0.036	0.029	0.013	0.019	0.005
All	3	0.006	0.008	0.018	0.025	0.020
	15	0.074	0.081	0.056	0.025	−0.002
	28	0.115	0.118	0.091	0.047	0.006

Decoder-direction overlap. For each Mask-SAE decoder direction, we record its highest cosine similarity with the Unmask-SAE decoder columns at the same layer (16,384 directions per layer). The mean/median best-match cosine similarities are 0.40/0.32 at L3, 0.71/0.76 at L15, and 0.68/0.72 at L28. The fractions with cosine above 0.9 are 3.0%, 17.4%, and 7.4%, respectively; the fractions above 0.7 are 11.9%, 60.8%, and 54.2%, and those below 0.3 are 37.8%, 5.1%, and 4.6%. Decoder directions therefore overlap more strongly at L15 and L28, despite the differences in reconstruction loss between masked and unmasked positions.

D Catalytic Recovery and SAE Preactivation

D.1 Decoding Protocol and Statistical Specification

We use 148 proteins from the held-out VenusX active-sites test split, each 30–320 residues long with at least two annotated catalytic sites. An initial random mask hides 50% of positions, with the remaining native residues retained as context. DPLM's reparameterized decoding uses Gumbel-argmax sampling and deterministic linear re-masking for 100 steps. For each masked position, we record the first recovery step and prediction-distribution entropy during decoding.

For masked residue i of protein p, we normalize the observed recovery step within that protein and standardize its residue-level entropy statistic:

$$t_i^{\text{norm}} = \frac{\text{step}_i - \min_{j \in p} \text{step}_j}{\max_{j \in p} \text{step}_j - \min_{j \in p} \text{step}_j}, \qquad H_i^{(z)} = \frac{H_i - \overline{H}_p}{\text{sd}(H_p)}.$$

The entropy-adjusted recovery analysis fits

$$t_i^{\text{norm}} = \alpha + \beta_H H_i^{(z)} + \beta_{\text{active}} \mathbf{1}_{\text{cat}}(i) + \varepsilon_i.$$

A categorical amino-acid fixed-effect term is added as a separate robustness model in Appendix D.4. Standard errors are CR1 cluster-robust by protein ($G = 148$; $G = 120$ for the cross-task pool), and reported p-values use the cluster-t reference distribution.

D.2 Sequence Perturbations and Recovery Robustness

We use the same initial random-mask draw across the baseline and three perturbation conditions: (i) permutation of non-catalytic residues while retaining annotated catalytic residues; (ii) permutation of the full sequence, with labels carried along by the permutation; and (iii) additional masking of residues within ± 3 positions of annotated catalytic sites. Table 8 reports the entropy-adjusted recovery coefficients. Under the 50% mask, the catalytic coefficient decreases in magnitude from -0.038 to -0.007 or -0.009 after sequence shuffling, while it remains -0.041 when local neighbours are masked.

Table 8: Catalytic-first recovery under sequence perturbations. Cluster-robust OLS adjusts for within-protein standardized prediction entropy (148 protein clusters). Negative β_{active} indicates earlier recovery of annotated catalytic positions.

Condition	β_{active}	SE	p (cluster)	β_{entropy}	Effect
Baseline (native sequence)	$\mathbf{-0.038}$	0.007	$\mathbf{2.9 \times 10^{-8}}$	0.18	**present**
(i) Non-catalytic residues shuffled	-0.007	0.011	0.56	0.07	absent
(ii) Full sequence shuffled	-0.009	0.007	0.24	0.05	absent
(iii) ± 3 motif neighbours masked	$\mathbf{-0.041}$	0.007	$\mathbf{9.4 \times 10^{-10}}$	0.17	**present**

Mask-fraction and decoding-step robustness. We repeat the sequence perturbation experiments across initial masking fractions of 0.3, 0.5, and 0.7 and with 100, 250, or 500 decoding steps (Table 9). Under native context, the catalytic association remains negative across both sweeps. It also persists with the ± 3 neighbours masked. Full-sequence shuffling yields coefficients near zero throughout, while non-catalytic shuffling gives a borderline negative coefficient at 30% masking ($p = 0.052$). The decoding-step sweep holds the initial mask fraction at 0.5; the mask-fraction sweep uses 100 steps.

D.3 Latent Catalytic-Feature Preactivation

Feature selection and timing. The catalytic feature set F_{cat} is selected on 1,142 qualifying proteins from the VenusX active-sites training split, disjoint from the 148-protein test set under the MF50 split. At L15, L22, and L28, the top-50 features are ranked by Fisher exact enrichment on their top-50 activating positions, with Benjamini–Hochberg correction. We evaluate top-10, top-20, and top-50 prefixes and training-pool activation thresholds q_{90}, q_{95}, and q_{99}. The null set F_{rand} contains equally

Table 9: Robustness of the entropy-adjusted catalytic recovery coefficient to initial mask fraction (first three numeric columns, 100 denoising steps) and decoding iterations (last two columns, 50% initial mask). Each entry reports β_{active} and its protein-clustered p-value. The 50% / 100-step condition is shared by both sweeps.

Condition	Mask 30% ($T = 100$)	Mask 50% ($T = 100$)	Mask 70% ($T = 100$)	$T = 250$ (mask 50%)	$T = 500$ (mask 50%)
Baseline	-0.026 (0.011)	-0.038 (2.9×10^{-8})	-0.056 (3.1×10^{-14})	-0.043 (1.5×10^{-8})	-0.047 (2.3×10^{-8})
Non-catalytic shuffle	-0.025 (0.052)	-0.007 (0.56)	0.008 (0.48)	-0.012 (0.34)	-0.014 (0.23)
Full-sequence shuffle	-0.004 (0.68)	-0.009 (0.24)	0.002 (0.79)	-0.009 (0.13)	0.006 (0.39)
± 3 neighbours masked	-0.021 (6.0×10^{-4})	-0.041 (9.4×10^{-10})	-0.045 (5.6×10^{-10})	-0.040 (7.2×10^{-11})	-0.047 (1.1×10^{-8})

many non-enriched features outside the top-200 active-site ranking, each with at least 50 nonzero training activations.

During baseline, full-shuffle, and local-neighbour-masking runs, SAE feature activations are recorded at masked positions at every decoding step. The first threshold crossing across the selected features and layers defines $t_{\text{feat}}(i)$; the preactivation indicator is $P_i = \mathbf{1}[t_{\text{feat}}(i) < t_{\text{unmask}}(i)]$. We evaluate an unadjusted linear probability model (M0), an AA-fixed-effect model (M1), and a model additionally adjusting for within-protein prediction entropy (M2), with standard errors clustered by protein.

Table 10: Catalytic-feature preactivation at q_{95} across feature-set sizes and perturbations. M0 is unadjusted; M2 includes amino-acid fixed effects and within-protein entropy. SEs are clustered by protein ($n_p = 148$).

Condition	top-K	spec	P_{cat}	P_{non}	β_{active}	CR SE	p
Baseline	10	M0	0.138	0.058	0.079	0.023	8.5×10^{-4}
	10	M2	0.138	0.058	0.063	0.024	9.1×10^{-3}
	20	M0	0.166	0.058	**0.107**	0.024	$\mathbf{1.1\times10^{-5}}$
	20	M2	0.166	0.058	**0.091**	0.024	$\mathbf{2.5\times10^{-4}}$
	50	M0	0.236	0.061	0.174	0.027	2.0×10^{-9}
	50	M2	0.236	0.061	0.157	0.028	6.8×10^{-8}
(ii) full shuffle	10	M2	0.153	0.157	-0.002	0.013	0.86
	20	M2	0.160	0.164	-0.003	0.014	0.84
	50	M2	0.200	0.202	-0.000	0.015	0.98
(iii) ± 3 masked	10	M2	0.177	0.066	0.095	0.023	4.8×10^{-5}
	20	M2	0.210	0.067	**0.127**	0.023	$\mathbf{2.0\times10^{-7}}$
	50	M2	0.298	0.073	0.206	0.026	7.9×10^{-13}

Under the top-20 q_{95} setting, 16.6% of catalytic positions and 5.8% of non-catalytic positions exhibit preactivation. The entropy- and AA-adjusted coefficient is 0.091 ($p = 2.5 \times 10^{-4}$). Table 10 shows the same direction across feature-set sizes and preservation under local-neighbour masking, with the effect disappearing after full-sequence shuffling. The random non-enriched feature control exhibits a negative association in the native-sequence condition (Table 11).

Table 11: Preactivation regression for the random non-enriched feature control at q_{95} (top-20 features per layer); M2 includes AA identity and within-protein entropy.

Condition	spec	P_{cat}	P_{non}	β_{active}	CR SE	p
Baseline	M0	0.041	0.110	-0.070	0.0095	1.6×10^{-11}
Baseline	M2	0.041	0.110	$\mathbf{-0.033}$	0.0083	$\mathbf{1.0\times10^{-4}}$
(ii) full shuffle	M0	0.615	0.602	0.014	0.0165	0.41
(ii) full shuffle	M2	0.615	0.602	0.017	0.0165	0.30
(iii) ± 3 masked	M0	0.108	0.133	-0.025	0.0142	0.078
(iii) ± 3 masked	M2	0.108	0.133	0.007	0.0133	0.61

Layer-subset sensitivity. At the q_{95} threshold, the unadjusted M0 preactivation coefficients are 0.023 at L15 ($p = 1.2 \times 10^{-4}$), 0.074 at L22 ($p = 1.1 \times 10^{-3}$), and 0.035 at L28 ($p = 9.4 \times 10^{-4}$).

Layer pairs L15/L22, L15/L28, and L22/L28 yield coefficients of 0.087, 0.057, and 0.095; combining all three reaches 0.107 ($p = 1.1 \times 10^{-5}$). These values use M0 and are distinct from the entropy- and AA-adjusted M2 coefficient in Section 5.

D.4 Additional Recovery Controls

Amino-acid identity. Adding 19 amino-acid fixed effects to the entropy-adjusted recovery model changes the baseline catalytic coefficient from -0.0384 to -0.0378. The unadjusted coefficient is -0.2100, and the adjusted coefficient under local-neighbour masking is -0.0413 (Table 12).

Structural and clean-input entropy controls. ESMFold-predicted structures provide Cα contact counts at 8 Å and neighbour counts at 12 Å; DPLM clean-input entropy is evaluated as an additional sequence-context proxy. The three covariates are standardized within each protein. Table 12 reports their inclusion in the entropy-adjusted recovery model. The catalytic coefficient stays negative, including when all three are entered together (-0.0424). These structural/context specifications do not jointly include the amino-acid fixed effects used in the separate models above.

Table 12: Additional controls for the catalytic-first recovery association. Panel A reports amino-acid identity adjustments (M0: catalytic indicator only; M1: + within-protein entropy; M2: + 19 AA fixed effects). Panel B adds structural/context covariates to M1, without AA fixed effects. All estimates use CR1 standard errors clustered by protein.

Specification	n_{res}	β_{active}	SE	p
A. Amino-acid identity controls				
Native, M0 (unadjusted)	19,568	-0.2100	0.0124	$< 10^{-50}$
Native, M1 (+ entropy)	19,568	-0.0384	0.0066	2.9×10^{-8}
Native, M2 (+ AA FE)	19,568	-0.0378	0.0063	1.4×10^{-8}
Non-catalytic shuffle, M2	19,568	-0.0064	0.0111	0.56
Full-sequence shuffle, M2	19,568	-0.0079	0.0072	0.28
± 3 neighbours masked, M2	20,977	-0.0413	0.0060	1.5×10^{-10}
B. Structural and clean-input entropy controls (all $n = 19{,}568$)				
M1 baseline	19,568	-0.0384	0.0066	2.9×10^{-8}
+ contact (8 Å)	19,568	-0.0400	0.0066	9.4×10^{-9}
+ neighbour count (12 Å)	19,568	-0.0424	0.0066	1.7×10^{-9}
+ clean entropy	19,568	-0.0376	0.0065	4.6×10^{-8}
+ contact + neighbour count	19,568	-0.0426	0.0066	1.5×10^{-9}
+ contact + clean entropy	19,568	-0.0397	0.0065	1.0×10^{-8}
+ contact + neighbour count + clean entropy	19,568	-0.0424	0.0066	1.4×10^{-9}

Selectivity across VenusX annotation classes. A separate pool of 120 sequence-deduplicated VenusX test proteins (40 each from active-site, binding-site, and conserved-site tasks; truncated to 384 residues) is used for class-wise recovery regressions. Table 13 shows earlier unadjusted recovery for all three annotations. After adjusting for prediction entropy, the active-site coefficient remains negative (-0.037, $p = 0.032$), while binding and conservation coefficients are near zero; these results pertain to separate class-wise models.

Table 13: Per-class breakdown of the catalytic-residue-first finding on the 120-protein cross-task pool. Univariate columns report Mann–Whitney one-sided p-values; cluster-robust columns control for within-protein logit entropy.

Class	n pos.	uni. mean step (pos)	uni. mean step (neg)	uni. MWU p	β_{class}	cluster p
Active-site	197	0.512	0.700	7.5×10^{-17}	-0.037	0.032
Binding (BioLiP)	348	0.589	0.705	1.9×10^{-14}	-0.005	0.69
Conserved	364	0.602	0.710	4.6×10^{-12}	-0.007	0.64

E Comparison with ESM-2

E.1 SAE Feature Comparison

We train Top-K SAEs on ESM-2 650M [Lin et al., 2023] at L15 and L22 with the same dictionary size ($D = 16{,}384$) and activation budget ($K = 64$) as the DPLM SAEs. The ESM-2 SAEs use 2.5 million activation vectors sampled at 15% masking and are trained for 15,000 steps; the DPLM SAEs use the training configuration in Appendix A. Thus architecture and sparsity are aligned, while activation data and training budgets differ.

Table 14 reports both the learned decoder-direction overlap and active-site feature discrimination. The median best-match cosine is 0.60 at L15 and 0.56 at L22, while the maximum individual-feature active-site AUPRs are almost identical for the two models.

Table 14: ESM-2 versus DPLM SAE feature comparison at matched layers. Top: best-match decoder-direction cosine from each ESM-2 SAE feature to the DPLM SAE dictionary. Bottom: single-feature active-site AUPR on VenusX, including best-feature and top-20 mean values (random-guess AUPR $= 0.039$). Dictionary size and sparsity are matched; activation data and training budgets differ.

Layer	Mean best-cos	Median	$\% > 0.9$	$\% < 0.3$
L15	0.59	0.60	6.6	10.7
L22	0.55	0.56	2.2	10.4

Layer	ESM-2 max	DPLM max	ESM-2 top-20 mean	DPLM top-20 mean
L15	0.121	0.120	0.085	0.088
L22	0.152	0.152	0.084	0.086

E.2 ESM-2 Iterative-Fill Baseline

Starting from the same 148 proteins and initial 50% mask draw as DPLM, we use ESM-2 in greedy iterative-fill mode: at each iteration, the lowest-entropy masked position is fixed to its highest-probability amino acid. DPLM uses its native Gumbel-based, linear re-masking policy (Appendix D.1). These procedures share the test proteins and initial masking but differ in decoding policy; ESM-2 also encounters a 50% initial masking rate despite training at 15%.

Table 15 shows earlier recovery of catalytic positions in both models. ESM-2 has a raw mean-step separation of 0.231 (0.511 versus 0.280), compared with 0.210 for DPLM (0.711 versus 0.501). After within-protein prediction-entropy adjustment, the catalytic coefficient is -0.0159 for ESM-2 and -0.0384 for DPLM. The larger adjusted magnitude for DPLM is observed under these distinct decoding policies.

Table 15: ESM-2 iterative-fill vs. DPLM iterative denoising on the full 148-protein test split, 50% initial random mask.

Quantity	ESM-2 iterfill	DPLM (Table 8 baseline)
n residues	19,568	19,568
n proteins	148	148
univariate mean step (catalytic)	0.280	0.501
univariate mean step (non-catalytic)	0.511	0.711
univariate Mann–Whitney p (one-sided)	3.4×10^{-127}	$< 10^{-100}$
cluster-robust β_{active}	-0.0159	-0.0384
cluster-robust SE	0.0077	0.0066
cluster-robust p	0.040	2.9×10^{-8}
β_{entropy}	0.259	0.18

F ProteinGym External Validation

F.1 Data and Evaluation Protocol

We begin with 217 ProteinGym substitution DMS assays [Notin et al., 2023], retaining assays with a single wild-type sequence of at most 1,022 residues and at least one single-substitution variant; 61 assays meet these criteria. Within each assay, we rank-normalize substitution scores to $[0, 1]$ so that higher values indicate greater deleteriousness and average the normalized ranks across variants at each wild-type position. For the ESM-2 baseline, we mask one position at a time and evaluate the wild-type residue's masked-marginal log probability.

We also audit sequence overlap between the 61 ProteinGym wild-type sequences and the VenusX feature-selection pool ($\sim 4{,}372$ sequences) using maximum 5-mer Jaccard similarity. Only one assay has a maximum similarity above 0.3 (CALM1, 0.51). Excluding this assay leaves the pooled SAE preactivation coefficient at 0.046, compared with 0.046 in the full sample; this audit concerns the SAE association, not a separate recovery-time rerun.

F.2 Recovery Timing and Sequence-Prior Baselines

Table 16 reports per-assay Spearman correlation and top-decile retrieval for recovery order, ESM-2 masked marginal, DPLM clean-input prediction entropy, and SAE-derived timing measures. ESM-2 has a higher median correlation (0.471) than recovery order (0.287), while $-t_{\text{unmask}}$ is positively associated with experimental mutation sensitivity in 82% of assays. The SAE timing measures show smaller standalone associations.

Table 16: Per-assay Spearman ρ between each predictor and within-assay rank-normalized DMS sensitivity, and per-assay top-10% retrieval on the same 61 assays. Columns: median ρ, fraction of assays with $\rho{>}0$, one-sided Wilcoxon p-value, mean enrichment of top-10% sensitive residues among the predictor's top decile, mean AUROC. The $-t_{\text{unmask}}$ row corresponds to the headline result of Section 6; the rows above and below are appendix-only comparisons.

Predictor	median ρ	frac $\rho{>}0$	W p	enrich@10%	AUROC
ESM-2 $\log p$	0.471	80%	$6{\times}10^{-8}$	3.22×	0.699
$-$DPLM clean entropy	0.353	80%	$3{\times}10^{-7}$	3.03×	0.669
$-t_{\text{unmask}}$	**0.287**	**82%**	$\mathbf{7{\times}10^{-7}}$	**2.91**×	0.666
Lead time L^{q95}	0.081	76%	$1{\times}10^{-5}$	1.11×	0.512
Early activation $1{-}t^{q95}_{\text{feat}}/T$	0.122	76%	$2{\times}10^{-5}$	1.26×	0.518
P^{q95}_{func} (binary)	0.112	75%	$3{\times}10^{-5}$	1.18×	0.515

SAE-derived timing measures. The binary preactivation indicator P^{q95}_{func}, lead time $L^{q95} = \max(t_{\text{unmask}} - t^{q95}_{\text{feat}}, 0)$, and early-activation score $1 - t^{q95}_{\text{feat}}/T$ have median per-assay correlations of 0.081–0.122 and AUROC values near 0.51 as standalone rankers (Table 16). They are reported as complementary descriptive analyses rather than competitive retrieval predictors.

F.3 Leave-One-Assay-Out Retrieval

For each held-out ProteinGym assay, we fit a ridge model on the other 60 assays and evaluate retrieval of the top 10% most mutation-sensitive positions by AUROC. Table 17 compares the reported sequence-prior baseline with additions of recovery time and SAE-derived timing variables. Adding $-t_{\text{unmask}}$ changes AUROC from 0.654 to 0.687 across all assays and from 0.673 to 0.713 in the enzyme-activity subset.

F.4 SAE Preactivation Association and Robustness

A pooled regression with assay-clustered standard errors includes assay effects, amino-acid covariates, ESM-2 masked marginal, and DPLM clean-input entropy. Table 18 reports a negative coefficient for recovery time and a positive additional association for the SAE preactivation indicator; the incremental R^2 from adding preactivation is 0.003.

Table 17: Leave-one-assay-out AUROC for retrieval of the top 10% mutation-sensitive residues. Additions are evaluated against the reference sequence-prior model.

Feature set	global (n=61)	Activity (n=11)	OrgFitness (n=43)	Binding (n=3)	Expression (n=3)
Sequence-prior reference	0.654	0.673	0.640	0.732	0.823
$+(-t_{\text{unmask}})$	**0.687**	**0.713**	**0.663**	0.780	0.828
$+L^{q95}$ lead time	0.655	0.655	0.637	0.734	0.819
$+P_{\text{func}}^{q95}$	0.656	0.655	0.640	0.721	0.822
$+(-t_{\text{unmask}})+L^{q95}$	0.689	0.709	0.667	0.777	0.828

Table 18: Pooled ProteinGym regression across 61 assays (n = 15,103 positions), with assay-clustered standard errors and the indicated sequence-prior and amino-acid covariates.

Spec	$\beta_{P_{\text{func}}}$ (p)	$\beta_{t_{\text{unmask}}}$ (p)	$\beta_{\text{ESM-2}}$ (p)	R^2
M0 (no diffusion variables)	—	—	0.025 (4×10^{-5})	0.127
M1 ($+\,t_{\text{unmask}}$)	—	-0.123 (1×10^{-4})	0.026	0.139
M2 ($+\,P_{\text{func}}$)	0.046 (6×10^{-4})	-0.122 (1×10^{-4})	0.026	0.142

Raising the preactivation threshold from q_{90} to q_{99} increases the estimated preactivation coefficient from 0.037 ($p = 2.0 \times 10^{-3}$) to 0.065 ($p = 7.3 \times 10^{-3}$), with 0.046 ($p = 6.3 \times 10^{-4}$) at q_{95}, using four-seed pooled data. In a single-seed feature-class analysis, the catalytic-enriched set retains a positive coefficient, the binding-enriched set is not statistically distinguishable from zero, and the conservation-enriched set has a negative coefficient (Table 19).

Table 19: ProteinGym per-task dissection on F_{func} (single-seed, n=8,102). M2 cluster-robust OLS with 61 assay clusters; predictor is $\mathbf{1}[t_{\text{feat}}^{q95} < t_{\text{unmask}}]$ over the indicated feature subset.

Predictor source	$\beta_{P_{\text{func}}}$	SE	p	per-assay $\overline{\rho}$
F_{cat} (active sites)	0.040	0.013	2.0×10^{-3}	0.083
F_{bind} (binding)	-0.057	0.040	0.16	0.072
F_{cons} (conserved)	-0.270	0.018	$< 10^{-15}$	-0.116
F_{func} (union, single seed)	0.032	0.014	0.030	0.080
F_{func} (union, 4-seed mean)	0.046	0.013	6.3×10^{-4}	0.084

Mask-fraction sensitivity. A single-seed mask-fraction sweep gives SAE preactivation coefficients of 0.058 (SE 0.015, $p = 1.7 \times 10^{-4}$, n = 4,852 residues) at 30% masking, 0.032 (SE 0.014, $p = 0.030$, n = 8,102) at 50%, and 0.029 (SE 0.013, $p = 0.026$, n = 11,303) at 70%. At full masking, the association is close to zero (0.001; SE 0.005, $p = 0.82$, n = 16,123). This full-mask condition serves as a negative control for the SAE-derived association in the specified pipeline.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction (§1) state the central claim and §4–§6 present the supporting evidence with controls.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Discussion (§7) and the Limitations paragraph explicitly discussed the limitation.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper presents an empirical mechanistic analysis with no theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: §3 states the SAE training recipe (Top-K, D=16,384, K=64, six layers and five noise levels of DPLM-650M, $\sim$ 95M activation vectors, 50,000 optimisation steps, batch 32,768, learning rate 3×10^{-4}); Appendix A documents activation collection, hyperparameters, and seeds; Appendix D.1 specifies the protein sample, masking and decoding protocol, and cluster-robust regression used for the catalytic-first analyses; Appendix D.3 specifies catalytic-feature selection, activation thresholds, and the preactivation indicator; Appendix F.1 specifies the ProteinGym assay filtering and the residue-level sensitivity score.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The code and checkpoints will be released on acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specify all the training and test details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper report error bars suitably and correctly.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: SAE training runs on a single RTX4090 GPU; each layer-noise SAE takes $\sim$2 hours at 50,000 steps; the ProteinGym timing pipeline runs in $\sim$ 10 min on one GPU under length-bucketed batched forward.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The work is an interpretability analysis of a publicly released protein language model; it involves no human subjects, no scraped or sensitive data, and no models with high misuse risk.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Discussion (§7) notes that interpretability of generative protein models is dual-use — the same internal-structure understanding that aids beneficial protein design could in principle aid biosecurity-relevant misuse; we release only analysis tooling (SAEs and scripts), not a steering-controlled generation pipeline.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases SAEs as interpretability tooling for an existing public model. We do not release a new generative model, dataset of generated proteins, or steered design pipeline that would warrant gated-release safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets are cited where used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New assets introduced in the paper are properly documented and provided alongside the assets. The code and checkpoint will later be released on acceptance.

    Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or research with human subjects was conducted.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects were involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Large language models were used only as a writing aid (drafting, polishing, and reformatting) and were not part of the core methodology, evaluation, or scientific claims.

    Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.