Structural Locality Differentiates Residue-Tokenised Bidirectional and Causal Protein Language Model Families

Wei Gao [1] **Alex Fedorec** [1]

Abstract

Protein language models (PLMs) trained under different bidirectional and causal regimes achieve strong downstream performance, yet how these architectural and objective differences jointly shape learned representations remains poorly understood. We apply TopK sparse autoencoders (SAEs) to three residue-tokenised PLMs (ESM-2, RITA, ProtT5; the ProtT5 encoder and decoder are probed separately) at nine matched relative depths. We find that ESM-2 (bidirectional) learns features with stronger structural locality than RITA (residue-tokenised causal) at every matched depth, robust across metric settings, held-out proteins, and SAE initialisations. On a BPE-tokenised PLM (ProtGPT2), we identify a tokenisation artefact in residue-pair locality metrics—a large naïve sequential-locality effect that reverses under inter-token control—motivating our restriction to residue-tokenised models in the main analyses. Within ProtT5, the nine-depth grid localises a structural-locality crossover between encoder and decoder to ≈42% relative depth, consistent with cross-attention propagating encoder-derived context that fades under the autoregressive objective. Within-model depth trajectories for ESM-2, RITA, ProtT5-enc, and ProtT5-dec are qualitatively distinct. The same family ordering appears on a downstream task—the two bidirectional models predict long-range contacts $\sim 2.7\times$ more accurately than the two causal ones under an unsupervised readout—though that readout is independent of the SAE, so these are convergent observations rather than a demonstrated link. These findings support a single robust bidirectional-vs-causal dissociation (on structural, not sequential, locality) and identify a methodological interaction between standard residue-projection and residue-pair locality metrics on BPE-tokenised PLMs. Code: https://github.com/Uninterpretable-Evolving-Blackbox/Structural-Locality-Differentiates-Residue-Tokenised-Bidirectional-and-Causal-Protein-Language-Model.

[1]University College London, London, United Kingdom. Correspondence to: Wei Gao <zcbtwga@ucl.ac.uk>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

1. Introduction

Proteins are linear chains of amino acids that fold into 3D structures, with residues that are distant in sequence often close in space—the long-range contacts that determine tertiary structure. Protein language models (PLMs) learn from millions of sequences and achieve strong performance on structure and function prediction (Lin et al., 2023; Elnaggar et al., 2021; Ferruz et al., 2022). Understanding what these domain-specific foundation models learn internally is a natural target for mechanistic interpretability beyond natural language.

PLMs differ in attention pattern and training regime: bidirectional masked language modelling (ESM-2; Lin et al. 2023), causal next-token prediction (RITA, Hesslow et al. 2022; ProtGPT2, Ferruz et al. 2022), and encoder–decoder denoising (ProtT5; Elnaggar et al. 2021). They also differ in tokenisation: most operate at residue (amino-acid) resolution, but ProtGPT2 uses byte-pair encoding (BPE), mapping multiple residues into a single token. RITA serves as our main residue-tokenised causal comparator against ESM-2 (H1); ProtGPT2 is included to examine whether residue-level locality metrics remain well-defined under BPE tokenisation. We treat the bidirectional-vs-causal family contrast as a coupled difference between architecture, training objective, training data, and other implementation details rather than a clean test of objective alone. Sparse autoencoders (SAEs) have emerged as a powerful tool for decomposing dense neural activations into interpretable sparse features (Bricken et al., 2023; Templeton et al., 2024; Gao et al., 2024). Prior SAE work on PLMs has focused on single architectures (Simon & Zou, 2025; Adams et al., 2025; Villegas Garcia & Ansuini, 2025); systematic cross-family comparison is

missing.

We ask how the bidirectional-vs-causal PLM-family contrast (which couples training objective with architecture and other implementation details) shapes the geometry of learned sparse features, testing two *a priori* hypotheses across matched relative depths plus an additional methodological observation. Our contributions are: (1) evidence that ESM-2 features have stronger structural locality than RITA features at every matched depth, supporting a bidirectional-vs-causal dissociation under residue tokenisation (H1; we do not isolate training objective from architectural and data-pipeline differences between the two families); (2) a depth-dependent reversal between ProtT5 encoder and decoder features consistent with cross-attention propagation (H2); and (3) a methodological observation that standard uniform residue-projection combined with residue-pair locality metrics produces a spurious sequential-locality effect on BPE-tokenised PLMs, which vanishes under an inter-token control. We further report, as a supporting observation rather than a fourth claim, that the same family ordering appears in unsupervised long-range contact prediction (§4.4)—a convergent measurement, not evidence that structural locality causes contact decodability.

2. Methods

2.1. Dataset and Models

We subsample 1,500 proteins from SCOPe 2.08 (Chandonia et al., 2022) (40% sequence-identity filter, fold-stratified across 432 folds, $\geq$ 80% DSSP coverage); length range 50–794, mean 197. A deterministic 90/10 protein-level split (seed 42) gives 1,350 SAE-training proteins and 150 held-out validation proteins. Feature-level analyses aggregate over all 1,500 proteins; we confirm main results hold on the 150 val-only subset (Table 2).

We compare three residue-tokenised PLMs (Table 1; the ProtT5 encoder and decoder are probed separately, giving four rows). ESM-2 vs. RITA is the main H1 contrast: both residue-tokenised and size-matched. ProtGPT2 is included for the BPE diagnostic (§4.2). The ProtT5 decoder is not strictly causal—each position cross-attends to the full bidirectional encoder output—making the encoder/decoder contrast (H2) a test of how autoregressive decoding interacts with a globally contextualised encoder state.

2.2. SAE Training

We train TopK SAEs (Gao et al., 2024) with expansion factor 8 (yielding 10,240 features for ESM-2/ProtGPT2, 12,288 for RITA, 8,192 for ProtT5). Ablations on ESM-2 layer 16 motivate k_{sparse}=256 (smallest train/val EV gap, Appendix A) and 8× (Appendix B): 8× has the highest val EV (0.717) and roughly doubles the absolute count of interpretable features over 4× (10,126 vs. 5,094) at near-identical per-feature interpretability; 16× and 32× degrade on both axes. Inputs are normalised to mean per-token L2 norm $\sqrt{D}$ (Gao et al., 2024) with unit-norm decoder columns (Bricken et al., 2023); AdamW (lr 5×10^{-5}, cosine), batch 4096. No latents met the dead criterion (zero activation across the last 10^6 training tokens) across any SAE; all train/val EV gaps < 0.10.

Table 1. PLMs compared. H1 is tested at nine matched relative depths (Table 2); H2 and the within-model depth analysis also use nine depths per model (see §4.3 and Appendix G).

Model	Objective	Token.	Params
ESM-2 t33	MLM (bidir.)	residue	650M
RITA-l	CLM (causal)	residue	680M
ProtGPT2	CLM (causal)	BPE	738M
ProtT5-enc	denoising (bidir.)	residue	1.2B
ProtT5-dec	denoising (causal)	residue	3.0B

2.3. Locality Score

For each SAE feature we compute structural (L_{struct}) and sequential (L_{seq}) locality scores via a shuffle-controlled standardised-contrast framework. Structural neighbours are residue pairs with C_α distance < 8 Å and sequence separation ≥ 12. The separation filter excludes α-helix periodicity ($\sim$3.6 residues/turn), isolating tertiary contacts. We use C_α throughout for a single consistent atom and ≥ 12 to retain medium- and long-range tertiary contacts; the standard long-range contact definition instead uses C_β (C_α for glycine) at separation ≥ 24 (Rao et al., 2019; Lin et al., 2023). Sequential neighbours are residues within ± 2 positions.

A residue is *active* for feature j if its activation exceeds feature j's 90th-percentile activation across all 295,240 residues. Let z_{ij} denote feature j's activation at residue i, and $A \in \{0,1\}^{N\times N}$ be the adjacency matrix (structural or sequential) with $\deg_i = \sum_k A_{ik}$; the neighbour co-activation of feature j at residue i is $\bar{z}_{ij} = \deg_i^{-1} \sum_k A_{ik}\, z_{kj}$, and the per-feature standardised *locality score* is

$$L_j = \frac{\mathbb{E}_{i\in\mathcal{A}_j}[\bar{z}_{ij}] - \mathbb{E}_i[z_{ij}]}{\sigma_j} - \overline{L_j^{\text{shuf}}}, \qquad (1)$$

where $\mathcal{A}_j$ is the active set, σ_j the global activation SD of feature j, and $\overline{L_j^{\text{shuf}}}$ the mean of the same contrast across 5 within-protein permutations of $z_{\cdot j}$. Structural neighbours yield L_{struct}; sequential neighbours yield L_{seq}. We reserve d for cross-model effect sizes between two L distributions (Cohen's-d standardised mean difference).

Residue projection for BPE. ProtGPT2's BPE tokeniser produces multi-residue tokens (mean $\sim$3 residues/token).

To compare features at residue resolution, we assign each residue the activation of its containing BPE token scaled by $1/L$ (token length), preserving aggregate per-token magnitude. A direct consequence is that residues within the same BPE token receive identical SAE activations; this is the confound examined in §4.2.

Feature interpretability. We also report an auxiliary *interpretability* metric: the fraction of SAE features whose per-residue activation correlates significantly (Benjamini–Hochberg $q < 0.05$) with at least one of three structural annotations: DSSP secondary structure (helix, strand) and a C_α coordination number computed from PDB coordinates as the count of other C_α atoms within 10 Å (a contact-density proxy for residue burial). It is the %Interp column in Appendix B and Appendix C, and is used only for SAE hyperparameter calibration; the hypothesis tests in §4 use the locality score.

2.4. Probing Protocol and Robustness Checks

We probe each residue-tokenised PLM in Table 1 at *nine matched relative depths* (layer index / total transformer layers; layer pairs in Table 2; the ProtT5 encoder and decoder are probed separately). ProtGPT2 is probed separately at five depths $\{0, 25, 50, 75, 100\}\%$ for the BPE diagnostic (§4.2). Main settings: C_α cutoff 8 Å, separation ≥ 12, quantile 10%, window ± 2. We declare *a priori* a robustness sweep over four metric axes ($C_\alpha \in \{6, 8, 10\}$ Å; separation $\in \{8, 12, 24\}$; quantile $\in \{5, 10, 20\}\%$; window $\in \{\pm 1, \pm 2, \pm 4\}$; Appendix I), three SAE seeds at $k_{\text{sparse}} = 256$, an alternative $k = 128$ and 99/1 split (Appendix F), a fresh 1,500-protein SCOPe subsample, and a smaller PLM pair (ESM-2 t12 vs. RITA-s; Appendix J). To test whether the H1 layer-0 effect inherits from raw input-embedding geometry, we additionally compute an embedding-only cosine-cohesion baseline at layer 0 (Appendix L).

3. Hypotheses

H1 (structural locality). ESM-2 features exhibit higher structural locality (L_{struct}) than RITA features at every matched relative depth.

H2 (ProtT5 depth-dependence of structural locality). Within ProtT5, early decoder layers retain encoder-like structural locality (via cross-attention propagation), and this signal fades with decoder depth so that late decoder layers show weaker structural locality than encoder layers at matched relative depth.

We do not state an *a priori* sequential-locality hypothesis: the residue-resolution RITA-vs-ESM-2 contrast shows no consistent bidirectional-vs-causal sequential effect (§4.2) and ProtT5 sequential data is non-monotonic with three sign flips across depth (§4.3). H2 is therefore a purely structural claim.

Table 2. H1 effect sizes d (ESM−RITA, positive = ESM wins) on L_{struct} and L_{seq} at nine matched depths, on the 150-protein held-out validation subset. Each cell is the bootstrap mean over 1000 protein-level resamples with 95% percentile CI in brackets. All L_{struct} cells exclude 0. The L_{seq} pattern is asymmetric across depth (§4); full-set table and ProtT5 H2 in Appendix K. Layer pairs ESM-2/RITA.

Depth	Layers	d on L_{struct}	d on L_{seq}
0%	L0/L0	$+0.58\,[.50, .66]$	$+1.24\,[1.17, 1.30]$
13%	L4/L3	$+0.05\,[.03, .09]$	$-0.05\,[-.08, -.02]$
25%	L8/L6	$+0.18\,[.07, .28]$	$+0.30\,[.22, .37]$
38%	L12/L9	$+0.23\,[.14, .32]$	$+1.17\,[1.03, 1.31]$
50%	L16/L12	$+0.26\,[.16, .37]$	$+1.88\,[1.67, 2.09]$
63%	L20/L15	$+0.22\,[.12, .32]$	$+2.09\,[1.86, 2.33]$
75%	L24/L18	$+0.13\,[.04, .21]$	$+1.97\,[1.73, 2.19]$
88%	L28/L21	$+0.17\,[.12, .22]$	$+0.49\,[.33, .64]$
100%	L32/L23	$+0.34\,[.29, .40]$	$-0.89\,[-1.02, -.76]$

4. Results

4.1. H1: ESM-2 vs. RITA Structural Locality

H1 is supported at all nine matched depths on both the full 1,500-protein set and the 150-protein held-out validation subset (Table 2 reports val with bootstrap CIs; full-set values and ProtT5 H2 in Appendix K). A protein-level cluster bootstrap (B=1000, percentile interval) excludes 0 from the 95% CI at every L_{struct} cell on both sets; the residue-level Mann–Whitney $p < 10^{-6}$ throughout but conflates feature-set size with effective sample size, so we treat the bootstrap as primary. Validation-set effect sizes range from $d = +0.05$ at 13% (coinciding with ESM-2's own layer-4 within-model trough, Appendix G) to $d = +0.58$ at the input embedding; full-set values are larger (Appendix K). We show in §5 that the layer-0 effect is not inherited from input-embedding geometry.

Robustness. Across the a priori metric sweep, H1 holds at 27/27 $C_\alpha \times$depth cells and 36/45 separation$\times$quantile cells (the 9 non-significant cells all at sequence separation ≥ 24 at intermediate depths, where the stricter filter leaves too few residue pairs). Cross-model d is seed-deterministic (cross-seed SD ≤ 0.044); direction is preserved under k=128, the 99/1 split, the fresh protein subsample (max $|\Delta d| = 0.08$), and the smaller ESM-2 t12/RITA-s pair (5/5 significant). Tables in Appendices F, I, J.

4.2. Sequential Locality at Residue Resolution

On the native-residue ESM-2-vs-RITA L_{seq} contrast (Table 2, no a priori hypothesis), ESM-2 leads at 7 of 9 depths

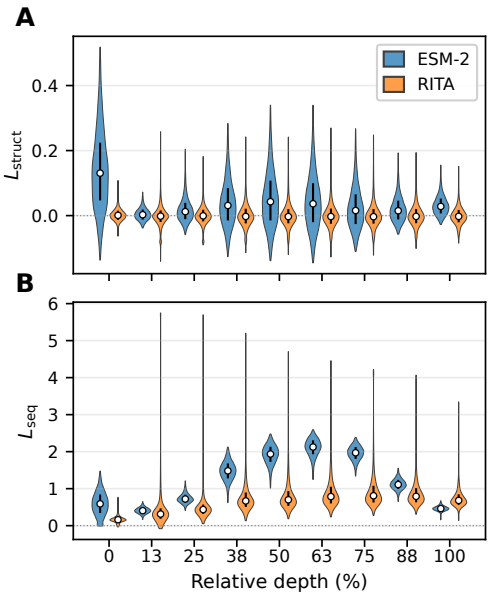

Figure 1. Distribution of L_{struct} (A) and L_{seq} (B) across SAE features for ESM-2 (blue) and RITA (orange) at nine matched relative depths. Violins clipped to 1st–99th percentile; whiskers show IQR, white dots the median. ESM-2 features concentrate at positive L_{struct} at every depth (H1); on L_{seq} ESM-2 leads at 7/9 depths (often $|d| > 1$ at mid-depth) with RITA leading decisively only at the final layer—a post-hoc observation discussed in §4.

with RITA leading only at 100% ($d = -0.97$); a post-hoc reading is that MLM may distribute sequential motif features throughout the network whereas CLM concentrates them at the prediction-head layer. Applying the same metric to a BPE-tokenised PLM (ProtGPT2) reveals a tokenisation artefact: multi-residue BPE tokens combined with our standard residue projection assign within-token residues bit-identical SAE activations, inflating naïve L_{seq} as a measurement of token-span identity rather than feature-level structure. A naïve ProtGPT2-vs-ESM-2 contrast shows a large effect favouring the causal model (up to $+5.9$); restricting to inter-token residue pairs reverses the effect at every depth (27/27 cells across the $\{\pm1, \pm2, \pm4\}$ window sweep; Appendix E). At residue resolution, causal PLMs therefore have no sequential-locality advantage over bidirectional ones. Structural locality (H1) requires separation ≥ 12, far beyond any BPE span, and is unaffected.

4.3. H2: ProtT5 Encoder vs. Decoder

H2 predicts that within ProtT5, late decoder layers show weaker structural locality than encoder layers at matched relative depth. With nine depths per side (full table in Appendix H), the structural crossover localises to a narrow window: the decoder leads at 0–39% ($d \in [-0.38, -0.12]$), the encoder leads at 52–100% ($d \in [+0.37, +0.66]$), with the sign flip between layers 9 and 12 (≈42% relative depth).

Table 3. Structural locality vs. contact prediction (P@$L/5$, Cα < 8 Å, $|i-j| \geq 12$) across nine matched depths per model. Contact P@$L/5$ is the mean over 500 proteins (per-layer means stable to ±0.01 under a 1000× protein bootstrap). ρ is the Spearman correlation between L_{struct} and contact P@$L/5$ across the nine layers (p from n=9). Bidirectional models predict contacts ∼ 2.7× better than causal ones, the same direction as their L_{struct} ordering; the readout is independent of the SAE (§4.4), so this is a convergent observation, not a demonstrated link.

Model	Dir.	$\overline{L_{\text{struct}}}$	$\overline{\text{P@}L/5}$	ρ (p)
ESM-2	bi.	0.043	0.098	−0.12 (0.77)
ProtT5-enc.	bi.	0.023	0.099	+0.92 (<0.001)
RITA	ca.	0.003	0.042	+0.37 (0.33)
ProtT5-dec.	ca.	0.013	0.030	+0.87 (0.003)
Bidirectional			**0.098**	
Causal			**0.036**	

The autoregressive objective appears to progressively override encoder-derived structural context past the first third of the decoder, though direct validation would require cross-attention-specific analyses.

4.4. Structural Locality and Downstream Contact Prediction

A natural question is whether the structural-locality gap reflects a difference in *downstream usefulness*, not merely feature-space geometry. We test this on the same 500 SCOPe domains using unsupervised, APC-corrected contact prediction from raw per-residue representations, scoring *exactly* the contacts that define L_{struct} (Cα < 8 Å, $|i-j| \geq 12$, top-$L/5$ precision). We adopt the unsupervised, APC-corrected top-$L/5$ metric of Rao et al. (2021) but predict from an embedding outer-product rather than their attention-head regression, and score at $|i-j| \geq 12$ rather than their long-range threshold of ≥ 24; L_{struct} is averaged over SAE seeds $\{42, 43, 44\}$.

Two findings emerge (Table 3). First, at the *family* level the two orderings agree: bidirectional models (ESM-2, ProtT5 encoder) predict long-range contacts ∼ 2.7× more accurately than causal ones (RITA, ProtT5 decoder; 0.098 vs. 0.036 mean P@$L/5$), the same direction as their L_{struct} ordering. We stress that this is an agreement between two *independent* measurements over n=4 models, not a demonstrated link: the readout runs on the raw representations and does not involve the SAE, so it does not establish that structural locality is what confers contact decodability. Second, the *within*-model, across-depth relationship is model-dependent: for ProtT5 (encoder and decoder) more structurally-local layers are strongly better contact predictors ($\rho \approx 0.9$), whereas for ESM-2 the two are uncorrelated ($\rho = -0.12$, p=0.77) as contact accuracy rises with depth independently of locality. Notably, the H2 encoder–decoder L_{struct} crossover does *not* produce a

contact-accuracy crossover—the encoder is more accurate at every depth even where the decoder is more structurally local—indicating L_{struct} captures local co-activation geometry related to, but distinct from, raw contact decodability. We use an unsupervised readout for consistency with L_{struct}; a supervised contact probe is left to future work, and the family-level comparison (n=4 models) inherits the confounds discussed below.

5. Discussion

The H1 dissociation is qualitative: under residue tokenisation, ESM-2 features' co-activating residues cluster in 3D space while RITA features show markedly weaker structural locality at every matched depth. We attribute this to the bidirectional-vs-causal family contrast as a whole—architecture, objective, training data, and pipeline—not to training objective in isolation, since ESM-2 and RITA differ on multiple coupled axes (Lin et al., 2023; Hesslow et al., 2022). The peak layer-0 effect ($d = +1.22$) is not visible to whole-vector cosine on raw input embeddings: ESM-2's layer-0 cosine landscape is near-saturated and cannot resolve contact-vs-non-contact at H1 magnitude (Appendix L). The SAE basis transformation exposes structural information that whole-vector averaging cannot detect; we cannot distinguish whether this information was always present in specific decompositional directions of the input embedding or emerges through SAE training.

The BPE finding (§4.2) recommends inter-token filtering for future work applying residue-pair metrics to multi-residue tokenisation. H2's encoder-decoder crossover shows the contrast is not binary when cross-attention is present.

Future work. Causal necessity and objective isolation remain the two clearest open directions. Feature ablation with a supervised contact-probe readout would test whether the structural-locality features are causally load-bearing rather than merely present, and a gradient-attribution readout could quantify their output-relevance—a causal complement to the representational metrics used here. A contact readout built directly from the L_{struct}-selected SAE features, rather than the raw representations of §4.4, would test convergent validity of the metric on held-out proteins. Finally, a matched-objective controlled experiment (identical architecture, data, and compute, varying only MLM vs. CLM) would isolate the training objective from the family-level confounds, provided the sparse-feature comparison controls for any difference in the intrinsic dimensionality of the two objectives' representations.

Limitations. (1) The locality score is correlational and reflects one specific operationalisation (shuffle-corrected standardised neighbour co-activation); alternative metrics—concept-annotation correlation (Simon & Zou, 2025), linear probing (Adams et al., 2025), targeted steering (Villegas Garcia & Ansuini, 2025)—ask different questions and may surface different features. Causal validation via feature-clamping is a natural next step. The protein-level bootstrap addresses residue-within-protein non-independence but not feature-level correlation across SAE features (which share decoder weights), so reported CIs are conditional on the trained feature set. (2) Parameter counts span 650M–3B; ESM-2 vs. RITA is size-matched at ∼650M, and the smaller-PLM-pair test (ESM-2 t12/RITA-s, Appendix J) rules out scale-driven H1, but ProtT5 enc/dec differ (1.2B/3B) so H2 conflates capacity with attention pattern. (3) H1 rests on a single residue-tokenised causal comparator (RITA); ProtGPT2 is excluded from H1 due to BPE-tokenisation artefacts (§4.2). A second residue-tokenised causal PLM would strengthen the family-level claim. (4) H1 effect-size magnitudes (ESM-2 vs. RITA on L_{struct}) at extreme depths (layer 0, layer $L_{\max}$) depend on the active-residue threshold; mid-depth (38–63%) is the most stable regime, suggesting extreme-depth peaks are partly driven by many weakly-firing residues rather than a few strongly-firing ones (Appendix M). Direction is preserved at every threshold. (5) SAE hyperparameters (k_{sparse}=256, expansion $8\times$) were selected via ablation on ESM-2 layer 16 only (Appendices A, B); we use these as conservative defaults rather than asserting cross-model optimality. (6) Scope: cosine baseline is layer-0 only (Appendix L); only TopK SAEs tested (no JumpReLU or GatedSAE comparison). (7) ESM-2 and RITA differ markedly in activation geometry (participation ratio ≈260 vs. ≈1; SAE val EV 0.72 vs. 0.98 at the headline layer), so L_{struct} is compared across SAE bases of different reconstruction character; direction survives k=128 (Appendix F) but controlling for this intrinsic-dimensionality gap is left to future work.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning interpretability, specifically for domain-specific foundation models of proteins. We do not foresee direct societal risks from the findings.

References

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

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language

models with dictionary learning. *Anthropic Transformer Circuits Thread*, 2023.

Chandonia, J.-M., Guan, L., Lin, S., Yu, C., Fox, N. K., and Brenner, S. E. SCOPe: improvements to the structural classification of proteins – extended database to facilitate variant interpretation and machine learning. *Nucleic Acids Research*, 50(D1):D553–D559, 2022.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 7112–7127, 2021.

Ferruz, N., Schmidt, S., and Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13:4348, 2022.

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

Hesslow, D., Zanichelli, N., Notin, P., Poli, I., and Marks, D. RITA: a study on scaling up generative protein sequence models. *arXiv preprint arXiv:2205.05789*, 2022.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., and Song, Y. Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems*, 2019.

Rao, R. M., Meier, J., Sercu, T., Ovchinnikov, S., and Rives, A. Transformer protein language models are unsupervised structure learners. In *International Conference on Learning Representations*, 2021.

Simon, E. and Zou, J. InterPLM: Discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, 2025. doi: 10.1038/s41592-025-02836-7.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jermyn, A., et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Anthropic Transformer Circuits Thread*, 2024.

Villegas Garcia, E. N. and Ansuini, A. Interpreting and steering protein language models through sparse autoencoders. *arXiv preprint arXiv:2502.09135*, 2025.

A. Sparsity Ablation

k	Train EV	Val EV	Gap
64	0.746	0.612	0.134
128	0.775	0.658	0.117
256	0.809	0.717	0.092

Table 4. Sparsity ablation on ESM-2 layer 16 at fixed expansion $8\times$ (10,240 features). k=256 chosen for smallest val/train gap.

B. Expansion Factor Ablation

Exp.	Val EV	Gap	%Interp	n_{interp}	%Dead
4×	0.710	0.083	99.5	5,094	0.0
8×	0.717	0.092	98.9	10,126	0.0
16×	0.711	0.105	96.8	19,821	0.0
32×	0.685	0.125	89.7	36,724	0.0

Table 5. Expansion sweep on ESM-2 layer 16 at fixed k=256. n_{interp} is the absolute count of interpretable features (%Interp $\times$ dictionary size). $8\times$ is selected because it has the highest validation EV and roughly doubles the absolute interpretable-feature count over $4\times$ at near-identical per-feature interpretability; further expansion degrades both metrics.

At matched sparsity density (2.5%, scaling k with expansion), interpretability still drops ($99.5\% \rightarrow 95.1\%$) and $32\times$ overfits (gap 0.137), indicating that extra dictionary atoms fragment existing biological signals rather than capturing new ones.

C. Feature Interpretability

We report %Interp (the fraction of SAE features whose per-residue activation correlates with at least one of helix / strand / C_α coordination at Benjamini–Hochberg $q < \tau$; definition in §2.3) per model and depth at two thresholds: a lenient $\tau = 0.05$ and a strict $\tau = 10^{-6}$.

Table 6. %Interp at lenient threshold $\tau = 0.05$ (sign convention: ESM−RITA, ESM−ProtGPT2). ESM-2 wins 5/5 cells against both comparators; all three models approach ceiling.

Depth	ESM-2	RITA	ProtGPT2	ESM−RITA	ESM−ProtGPT2
0%	95.60	86.23	92.86	+9.4 pp	+2.7 pp
25%	98.32	89.16	93.30	+9.2 pp	+5.0 pp
50%	98.92	91.89	94.06	+7.0 pp	+4.9 pp
75%	99.35	95.17	97.94	+4.2 pp	+1.4 pp
100%	99.28	93.33	95.58	+6.0 pp	+3.7 pp

At $\tau = 0.05$ the ESM-2 lead is 1–9 pp, dismissable as ceiling effects. At $\tau = 10^{-6}$ the gap widens to 11–32 pp and is monotonically positive at every cell: the bidirectional advantage is not "ESM-2 has slightly more features above an arbitrary cutoff" but "ESM-2 has many more strongly-interpretable features, by 1–2 orders of magnitude in q". ESM-2's largest lead is at depth 100% under the strict threshold (+32 pp vs. RITA, +25 pp vs. ProtGPT2): ESM-2's deepest layer retains 92% strongly-interpretable features whereas RITA's drops to 59% (its own minimum). RITA and ProtGPT2 essentially tie at $\tau = 10^{-6}$, depth 50% (61.4% vs. 61.7%); neither causal model develops the strong-signal feature population that ESM-2 acquires at mid-depth.

D. SAE Validation Explained Variance

Table 8 reports SAE validation explained variance (val EV) at each of the nine matched H1 depths for ESM-2 and RITA-l. RITA-l saturates near 1.0 at depths $\geq 13\%$ (val EV $\in [0.957, 0.999]$), indicating activations that admit very low-effective-dimensional reconstruction; ESM-2 stays well below 1.0 throughout (val EV $\in [0.654, 0.942]$). At a saturated EV plateau the SAE solution can become under-determined (different initialisations find different equally-valid feature bases), but RITA-l's H1 cross-model d remains seed-deterministic at our hyperparameters (cross-seed SD ≤ 0.04, Table 10).

Table 7. %Interp at strict threshold $\tau = 10^{-6}$ (sign convention: ESM−RITA, ESM−ProtGPT2). ESM-2 wins 5/5 against both comparators; the bidirectional advantage widens to 11–32 pp, separating ESM-2's strongly-interpretable feature population from the two causal models.

Depth	ESM-2	RITA	ProtGPT2	ESM−RITA	ESM−ProtGPT2
0%	71.53	49.66	59.52	+21.9 pp	+12.0 pp
25%	83.57	52.80	61.82	+30.8 pp	+21.8 pp
50%	88.73	61.43	61.69	+27.3 pp	+27.0 pp
75%	91.65	69.86	80.56	+21.8 pp	+11.1 pp
100%	91.74	59.45	66.94	+32.3 pp	+24.8 pp

Table 8. SAE validation explained variance (val EV) at each of the nine matched H1 depths for ESM-2 and RITA-l. RITA-l saturates near 1.0 at depths ≥13%, indicating low-effective-dimensional activations; ESM-2 stays well below 1.0 throughout. Despite the high-EV regime, RITA-l locality scores remain seed-deterministic (cross-seed SD ≤ 0.04, Table 10).

Rel. depth	ESM-2 layer	ESM-2 val EV	RITA layer	RITA val EV
0%	L0	0.942	L0	0.957
13%	L4	0.718	L3	0.999
25%	L8	0.708	L6	0.996
38%	L12	0.748	L9	0.993
50%	L16	0.717	L12	0.989
63%	L20	0.692	L15	0.985
75%	L24	0.666	L18	0.981
88%	L28	0.654	L21	0.976
100%	L32	0.733	L23	0.967

E. BPE Artifact Additional Detail

The residue projection used throughout maps each BPE token's activation uniformly to its constituent residues with weight $1/L$, preserving aggregate magnitude. As a direct consequence, residues within the same BPE token receive bit-identical SAE activations. For ProtGPT2 (mean ∼3 residues/token), this means a large fraction of ± 1 and ± 2 neighbour pairs used by the sequential locality metric have activation differences of identically zero, inflating the measured L_{seq}. Restricting to inter-token residue pairs (60,718 edges on the validation set, 49.9% of 121,648 directed ± 2 pairs) yields the "inter-tok." column in Table 9.

Table 9. Sequential locality L_{seq} at five matched depths, sign oriented causal−bidirectional. Positive values indicate the causal model wins. "Naïve" ProtGPT2 uses the standard residue projection. "Inter-token" restricts to residue pairs that cross BPE token boundaries. For the RITA column, $^{\text{n.s.}}$ marks contrasts not significant at $p < 0.05$; all ProtGPT2 rows are significant at $p < 10^{-6}$.

Depth	ProtGPT2 (naïve)	ProtGPT2 (inter-tok.)	RITA (residue-level)
0%	+3.24	−2.28	$-1.64^{\text{n.s.}}$
25%	+2.55	−3.01	$-0.26^{\text{n.s.}}$
50%	+1.69	−6.10	$-2.03^{\text{n.s.}}$
75%	+2.37	−2.12	$-2.19^{\text{n.s.}}$
100%	+5.91	−1.47	+0.98

F. Multi-Seed Reproducibility

Across seeds 42/43/44 varying SAE initialisation only, the cross-model effect size d is essentially deterministic (max SD 0.044 across five depths, occurring at 100% depth). Under k=128 (half the TopK sparsity of our main setting) the direction of H1 is preserved at all depths with modestly reduced effect sizes; under a 99/1 train/validation split (replacing the main 90/10 split with a smaller validation set and larger training set), results are quantitatively similar to the main seed-42 run. This addresses dependence on SAE initialisation and on the two dataset/hyperparameter axes separately.

Depth	seed 42	seed 43	seed 44	mean	SD
0%	+1.915	+1.904	+1.897	+1.906	0.0091
25%	+1.248	+1.225	+1.252	+1.242	0.0148
50%	+1.398	+1.386	+1.378	+1.387	0.0101
75%	+0.632	+0.619	+0.614	+0.622	0.0090
100%	+1.680	+1.766	+1.712	+1.720	0.0436

Table 10. H1 cross-model effect size d across three SAE initialisation seeds for ESM-2 vs. RITA at five matched depths $\{0, 25, 50, 75, 100\}\%$. (These values are the seeds' raw paired-distribution Cohen's-d estimates and differ from Table 2's shuffle-corrected protein-bootstrap point estimates—the seeds-only diagnostic uses an alternative scaling, retained here because it is what was originally computed.) Per-depth cross-seed SD is ≤ 0.044 at every depth and ≤ 0.015 at four of five; H1 is essentially deterministic across SAE initialisation. Under an alternative sparsity (k=128, half the main setting) and an alternative 99/1 train/validation split, H1 direction is preserved at all five depths with quantitatively similar effect sizes; numerical values omitted for space.

G. Within-Model Depth Signatures

We probed nine layers per model for ESM-2, RITA, ProtT5-enc, and ProtT5-dec and computed mean L_{struct} per layer across SAE features with bootstrap 95% confidence intervals over features (1000 resamples). Each model exhibits a qualitatively distinct depth trajectory (Figure 2, Table 11).

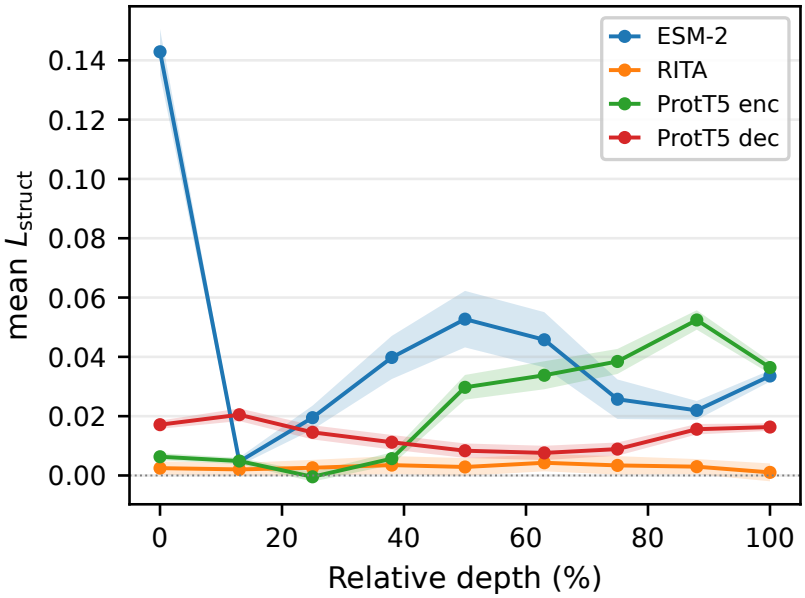

Figure 2. Within-model trajectories of mean L_{struct} vs. relative depth for the four model variants (ESM-2, RITA, ProtT5-enc, ProtT5-dec) at nine matched relative depths each. Shaded bands are 95% bootstrap CIs over proteins (B=1000). The four models display qualitatively distinct depth signatures summarised in Table 11.

Four signatures emerge. **ESM-2** spikes at the input embedding, dips sharply at layer 4, rises to a mid-depth plateau around layer 16, and rebounds at layer 32; the trajectory is non-monotonic. **ProtT5-enc** shows a clean monotonic accumulation of structural locality with depth, peaking at layer 21. **ProtT5-dec** is a shallow U: front-loaded at layer 3, trough at layer 15, mild recovery at the output. **RITA** is flat: every layer's bootstrap CI overlaps every other layer's except the layer-23 dip, and early-vs-late means are indistinguishable. RITA's mean L_{struct} is substantially smaller than ESM-2's at most depths ($\sim$57$\times$ smaller at the ESM-2 layer-0 peak, where ESM-2 reaches $+0.143$ vs. RITA's $+0.0025$ at the same depth; the ratio collapses near ESM-2's own layer-4 trough), consistent with H1.

These trajectories are descriptive rather than hypothesis-tested: we do not perform a per-model depth-trend hypothesis test at feature level, as stable per-feature depth trends would require more layers than 9 points can resolve. What the 9-point grid does support is aggregate comparisons across models, which is what we report here.

Model	Peak layer	Peak d	Trough layer	Trough d
ESM-2	0	$+0.143$	4	$+0.005$
ProtT5-enc	21	$+0.052$	6	-0.000
ProtT5-dec	3	$+0.021$	15	$+0.008$
RITA	15	$+0.004$	23	$+0.001$

Table 11. Within-model depth trajectories of mean L_{struct} across 9 probed layers: peak and trough layer indices and effect sizes. We do not report a Pearson r as ESM-2 and ProtT5-dec are non-monotonic across depth (Figure 2).

H. Full H2 Table

Table 12. Full per-layer ProtT5 encoder vs. decoder bootstrap mean $\overline{d}$ with 95% percentile bootstrap CIs (B=1000, protein-level resampling), on the full 1,500-protein set and the 150-protein val subset. Positive = encoder greater. The structural sign flip between layers 9 and 12 is preserved on val. Sequential locality is non-monotonic on both sets (three sign flips at L9→L12, L18→L21, L21→L23) and is not interpreted as a depth-direction trend.

		d on L_{struct}		d on L_{seq}	
Layer	Rel. depth	full	val	full	val
0	0.0%	-0.27 $[-0.31, -0.24]$	-0.15 $[-0.20, -0.09]$	-0.48 $[-0.52, -0.42]$	-0.45 $[-0.53, -0.38]$
3	13.0%	-0.38 $[-0.42, -0.33]$	-0.23 $[-0.29, -0.16]$	-2.57 $[-2.62, -2.51]$	-2.10 $[-2.22, -1.98]$
6	26.1%	-0.35 $[-0.39, -0.32]$	-0.20 $[-0.26, -0.14]$	-1.60 $[-1.67, -1.53]$	-1.25 $[-1.40, -1.10]$
9	39.1%	-0.12 $[-0.15, -0.09]$	-0.06 $[-0.10, -0.01]$	-1.53 $[-1.60, -1.47]$	-1.06 $[-1.20, -0.91]$
12	52.2%	$+0.37$ $[+0.32, +0.42]$	$+0.22$ $[+0.16, +0.29]$	-0.08 $[-0.18, +0.03]$	$+0.15$ $[-0.04, +0.33]$
15	65.2%	$+0.39$ $[+0.35, +0.44]$	$+0.23$ $[+0.16, +0.31]$	$+1.65$ $[+1.56, +1.74]$	$+1.41$ $[+1.22, +1.59]$
18	78.3%	$+0.47$ $[+0.43, +0.51]$	$+0.28$ $[+0.23, +0.34]$	$+1.27$ $[+1.18, +1.37]$	$+0.99$ $[+0.84, +1.15]$
21	91.3%	$+0.66$ $[+0.62, +0.69]$	$+0.40$ $[+0.36, +0.44]$	-0.46 $[-0.56, -0.27]$	-0.43 $[-0.55, -0.30]$
23	100.0%	$+0.46$ $[+0.43, +0.50]$	$+0.24$ $[+0.19, +0.29]$	$+2.39$ $[+2.34, +2.44]$	$+1.60$ $[+1.47, +1.72]$

The L12 (52.2%) sequential-locality cell spans zero on both full and val (the only cells that do): full $[-0.18, +0.03]$; val $[-0.04, +0.33]$. All 18 structural cells and the remaining 16 sequential cells have CIs that exclude 0. The structural crossover is therefore tight and unambiguous; the sequential axis is genuinely non-monotonic and contains the L12 transition cell as a near-null contrast.

I. Metric Sweep

We re-ran the locality analysis over the full metric configuration space. Tables 13 and 14 report H1 and H2 effect sizes under three C_α distance cutoffs. Table 15 reports the sequential-locality contrast under three window sizes for the three causal-vs-bidirectional comparisons (RITA residue-level, ProtGPT2 naïve, ProtGPT2 inter-token).

Table 13. H1 bootstrap mean $\overline{d}$ (ESM-2 > RITA, on L_{struct}, full set) under C_α cutoff sweep at nine matched depths. 27/27 cells positive (95% percentile bootstrap CI excludes 0).

Rel. depth	6 Å	8 Å(main)	10 Å
0%	$+1.00$	$+1.22$	$+1.25$
13%	$+0.03$	$+0.04$	$+0.05$
25%	$+0.27$	$+0.30$	$+0.27$
38%	$+0.33$	$+0.44$	$+0.42$
50%	$+0.39$	$+0.54$	$+0.54$
63%	$+0.33$	$+0.45$	$+0.43$
75%	$+0.28$	$+0.29$	$+0.22$
88%	$+0.34$	$+0.33$	$+0.28$
100%	$+0.58$	$+0.67$	$+0.70$

J. Additional Robustness: Scale and Protein Resampling

We re-ran the main H1 and H2 comparisons under two additional axes not covered by the metric sweep in Appendix I: (i) a smaller PLM pair at comparable size to each other, and (ii) a fresh protein subsample.

Table 14. H2 effect size d (ProtT5 encoder − decoder on L_{struct}, positive = encoder greater) under C_α cutoff sweep. The sign flip between layers 9 and 12 is preserved at every cutoff. † marks cells where $d > 0$ *and* Mann–Whitney $p < 0.05$ (encoder wins, significant): 15/27 cells. The remaining 12 cells (early depths 0–39%) have $d < 0$ (decoder leads); these are also individually significant ($p < 10^{-6}$ in all but the 39% cells, which are still $p < 0.05$), so the H2 sign reversal across depth is supported by significant effects on both sides of the crossover, not by a non-significant negative side.

Rel. depth	6 Å	8 Å(main)	10 Å
0%	−0.26	−0.30	−0.50
13%	−0.34	−0.40	−0.51
26%	−0.29	−0.38	−0.51
39%	−0.11	−0.13	−0.19
52%	+0.29†	+0.40†	+0.41†
65%	+0.32†	+0.43†	+0.42†
78%	+0.40†	+0.50†	+0.51†
91%	+0.52†	+0.69†	+0.80†
100%	+0.45†	+0.50†	+0.57†

Table 15. Sequential-locality cross-model effect size d on L_{seq} under window sweep, sign-oriented causal−bidirectional. "RITA" = RITA residue-level vs. ESM-2: $d > 0$ (RITA wins) at 3/27 cells, all at 100% depth. The remaining 24 cells are split: 22 cells with $d < 0$ (ESM-2 wins, mostly at depths 25–88% with $|d| > 1$); 2 cells with $d > 0$ at 13% (± 2 and a near-null ± 1). The pattern is non-monotonic across depth and not interpretable as a coherent architectural trend. "PG2 raw" = ProtGPT2 naïve BPE projection vs. ESM-2: 27/27 cells positive (causal wins, BPE artifact). "PG2 inter" = ProtGPT2 inter-token only vs. ESM-2: 27/27 cells negative (artifact reverses completely under inter-token control).

	RITA (residue)			ProtGPT2 raw			ProtGPT2 inter-tok.		
Depth	±1	±2	±4	±1	±2	±4	±1	±2	±4
0%	−1.37	−1.64	−1.92	+3.46	+3.24	+2.64	−2.37	−2.28	−2.36
13%	+0.26	+0.18	+0.03	+2.57	+2.51	+2.44	−2.97	−1.96	−1.83
26%	−0.16	−0.26	−0.61	+2.67	+2.58	+2.37	−4.46	−3.01	−2.90
39%	−0.91	−1.10	−1.69	+1.96	+1.82	+1.45	−6.01	−4.91	−4.87
52%	−1.84	−2.03	−2.79	+1.95	+1.69	+0.96	−7.66	−6.10	−6.15
65%	−2.14	−2.31	−3.00	+1.96	+1.75	+1.17	−5.76	−3.44	−3.56
78%	−2.17	−2.19	−2.66	+2.61	+2.38	+1.70	−5.55	−2.12	−1.95
91%	−0.55	−0.46	−0.54	+4.24	+4.15	+3.73	−4.82	−1.60	−1.15
100%	+0.96	+0.98	+1.16	+6.50	+5.92	+5.26	−3.99	−1.47	−1.16

PLM scale ablation. To test whether the H1 pattern depends on the 650M/680M scale of ESM-2/RITA, we repeated H1 on smaller residue-tokenised checkpoints: ESM-2 t12 (35M, 12 transformer layers) vs. RITA-s (85M, 12 transformer layers), probed at relative depths $\{0, 27, 55, 82, 100\}\%$. Table 16 reports H1 (structural) and the auxiliary sequential contrast at these depths.

Table 16. H1 structural locality (ESM-2-small > RITA-small) and sequential contrast at five matched depths on the smaller PLM pair (35M/85M). ✓ denotes $p < 0.05$ by one-sided Mann–Whitney. H1 is significant at 5/5 depths; the sequential contrast is significant only at 100% depth, matching the paper's 650M/680M pattern.

Rel. depth	Layers (E/R)	H1 d	aux d_{seq} (causal−bidir.)
0%	0/0	+1.19 ✓	−2.83
27%	3/3	+0.50 ✓	+0.30
55%	6/6	+1.04 ✓	−0.35
82%	9/9	+0.72 ✓	−0.26
100%	11/11	+0.76 ✓	+0.60 ✓

H1 is preserved at the smaller scale and, if anything, Cohen's-d is amplified (e.g. $+1.04$ vs. $+0.54$ at 50%)—contrary to what a pure scale confound would predict. RITA-s SAE validation EV saturates at ≥ 0.99 from layer 3 onward, matching RITA-l (Appendix D); this suggests the SAE saturation on causal PLMs is architectural, not scale-dependent.

Fresh protein subsample. To check whether H1 depends on the particular 1,500-protein subsample used throughout, we ran a replication on a second 1,500-protein SCOPe subsample drawn with a different random seed (same 40% sequence-

identity filter, same fold-stratified scheme). H1 replicates at 5/5 matched depths (Table 17); the maximum $|d|$ deviation from the main run is 0.08, mean 0.06—comparable to cross-seed SAE-initialisation variability and far below the H1 signal. The ProtGPT2 BPE artefact and the ProtT5 encoder-decoder crossover also reproduce (5/5 significant for naïve L_{seq}; sign flip preserved between layers 6 and 12).

Table 17. H1 cross-model effect size d (original-sample, full set) for ESM-2 vs. RITA on the fresh 1,500-protein subsample, compared against the main paper subsample at the same five depths. $|\Delta|$ = absolute difference between the two runs. Both columns are original-sample d values (pre-bootstrap) to enable a direct subsample-to-subsample comparison; the bootstrap mean $\overline{d}$ headline numbers (Appendix K) attenuate by a similar magnitude on both runs.

Rel. depth	Main run	Fresh subsample	$\|\Delta\|$
0%	$+1.44$	$+1.51$	0.07
25%	$+0.34$	$+0.26$	0.08
50%	$+0.62$	$+0.66$	0.04
75%	$+0.32$	$+0.40$	0.08
100%	$+0.73$	$+0.77$	0.04

K. Protein-level Bootstrap CIs for H1

We replace the residue-level Mann–Whitney test with a protein-level cluster bootstrap (B=1000) that resamples the 1,500 proteins (or 150 val proteins) with replacement and recomputes the cross-model effect size d on each resample. Per-feature shuffle thresholds and global activation SDs are held fixed across resamples (recomputing them per iteration is several orders of magnitude more expensive and is not standard practice); reported CIs are conditional on those nuisance parameters. We report the bootstrap mean $\overline{d}$ across the 1000 resamples as the point estimate, with the 95% percentile bootstrap interval $[P_{2.5}, P_{97.5}]$. Under cluster resampling the bootstrap mean attenuates 0.05–0.22 below the original-sample d on L_{struct}—a known artefact of ratio statistics under cluster bootstrap—and we report the bootstrap-derived estimate as the more conservative summary. Original-sample d values (computed on the unsampled data) are larger but follow the same depth pattern.

Table 18. Per-depth bootstrap mean $\overline{d}$ with 95% percentile CIs (B=1000, protein-level resampling), on the full 1,500-protein set and the 150-protein held-out validation subset. `frac_pos` is the fraction of bootstrap resamples in which $d > 0$. 9/9 cells exclude 0 from the 95% CI on both sets.

Rel. depth	$\overline{d}$ (full)	95% CI (full)	frac_pos	$\overline{d}$ (val)	95% CI (val)	frac_pos
0%	$+1.224$	$[+1.156, +1.284]$	1.000	$+0.577$	$[+0.496, +0.656]$	1.000
13%	$+0.040$	$[+0.013, +0.068]$	0.996	$+0.055$	$[+0.025, +0.094]$	1.000
25%	$+0.303$	$[+0.237, +0.363]$	1.000	$+0.175$	$[+0.071, +0.283]$	0.999
38%	$+0.436$	$[+0.358, +0.515]$	1.000	$+0.226$	$[+0.139, +0.320]$	1.000
50%	$+0.544$	$[+0.455, +0.632]$	1.000	$+0.257$	$[+0.157, +0.367]$	1.000
63%	$+0.447$	$[+0.363, +0.535]$	1.000	$+0.218$	$[+0.119, +0.322]$	1.000
75%	$+0.288$	$[+0.214, +0.363]$	1.000	$+0.129$	$[+0.044, +0.209]$	0.997
88%	$+0.331$	$[+0.281, +0.383]$	1.000	$+0.169$	$[+0.118, +0.221]$	1.000
100%	$+0.671$	$[+0.615, +0.721]$	1.000	$+0.343$	$[+0.286, +0.398]$	1.000

L. Embedding-only Cosine Baseline at Layer 0

The peak ESM-2 effect at layer 0 ($d = +1.22$) raises the question of whether structural locality is already present in raw input-embedding geometry, before any sparse decomposition. The SAE-feature locality metric (L_j) is ill-defined on dense rotation-invariant hidden states: per-dimension statistics measure coordinate-system happenstance rather than intrinsic geometry. We use neighbour cosine similarity instead, which is rotation-invariant and tests the underlying question directly.

Metric. For each protein p and each model, compute the mean cosine similarity of layer-0 hidden states across structural-contact pairs ($C_\alpha < 8$ Å, sequence separation ≥ 12) minus the mean cosine across sequence-distance-matched non-contact pairs (sampled at the same separation $s = |j - i|$ from the same protein). This isolates structural contact from generic sequence proximity. Per-protein scalar L_p^{model}, with cross-model effect size d computed on the resulting paired distributions of 1,499 per-protein values, plus 1,000-iteration protein-level bootstrap for 95% CI.

Table 19. Bootstrap mean $\overline{d}$ on L_{seq} (ESM−RITA, positive = ESM wins) with 95% percentile bootstrap CIs (B=1000, protein-level), on the full 1,500-protein set and the 150-protein val subset. ESM-2 wins at 7 of 9 depths on both sets; RITA wins at 13% (small inversion) and 100% (decisive). The 100% reversal is the cell first flagged in the Appendix I window sweep (Table 15 uses the opposite sign orientation, causal−bidirectional, to group RITA with the ProtGPT2 BPE diagnostic; the underlying finding is the same).

Rel. depth	$\overline{d}$ (full)	95% CI (full)	frac_pos	$\overline{d}$ (val)	95% CI (val)	frac_pos
0%	$+1.568$	$[+1.475, +1.622]$	1.000	$+1.239$	$[+1.170, +1.302]$	1.000
13%	-0.169	$[-0.184, -0.151]$	0.000	-0.048	$[-0.077, -0.016]$	0.000
25%	$+0.260$	$[+0.229, +0.288]$	1.000	$+0.301$	$[+0.224, +0.371]$	1.000
38%	$+1.080$	$[+1.029, +1.133]$	1.000	$+1.173$	$[+1.027, +1.315]$	1.000
50%	$+1.986$	$[+1.907, +2.063]$	1.000	$+1.880$	$[+1.668, +2.086]$	1.000
63%	$+2.252$	$[+2.155, +2.349]$	1.000	$+2.093$	$[+1.860, +2.329]$	1.000
75%	$+2.143$	$[+2.048, +2.242]$	1.000	$+1.969$	$[+1.732, +2.195]$	1.000
88%	$+0.459$	$[+0.394, +0.522]$	1.000	$+0.488$	$[+0.334, +0.638]$	1.000
100%	-0.968	$[-1.022, -0.912]$	0.000	-0.894	$[-1.021, -0.759]$	0.000

Result. ESM-2 mean per-protein cohesion $+0.0009$ (sd 0.0008); RITA $+0.0090$ (sd 0.0121). Cross-model $d = -0.94$, 95% bootstrap CI $[-1.06, -0.83]$, $0/1000$ resamples favoured ESM-2.

Saturation caveat. The raw cohesion magnitudes are not directly comparable. ESM-2's layer-0 hidden states live on a near-saturated angular shell: across 500 randomly sampled residue pairs the pairwise cosine has mean $+0.975$, sd 0.008, range $[+0.90, +1.00]$. RITA's are more dispersed: mean $+0.803$, sd 0.049, range $[+0.44, +0.98]$. RITA therefore has roughly $8\times$ the angular dynamic range that ESM-2 has, and ESM-2's contact-vs-non-contact difference is mechanically capped at $\sim$0.025 by saturation regardless of how much structural information the embeddings carry. The cohesion comparison is thus qualitative: it shows that raw whole-vector cosine, in ESM-2's saturated regime, cannot resolve the H1 layer-0 advantage that the SAE feature basis recovers ($d = +1.22$). Sparse coding therefore does substantive work at layer 0 even though no transformer computation has yet occurred: the basis change exposes structural information that whole-vector averaging on a saturated shell cannot detect.

M. Active-residue Threshold Sensitivity

The paper defines a residue as "active" for feature j if its activation exceeds the 90th-percentile of feature j's activations. For TopK SAEs at $k = 256/10240$ this collapses to "any non-zero activation" for many features, giving an unstable operational definition across sparsity regimes. We re-ran H1 with three fixed-magnitude thresholds based on the median non-zero activation s (per model, per layer): $0.5\times s$, $1.0\times s$, $2.0\times s$, with bootstrap CIs computed by the same procedure as Table 18.

Table 20. H1 cross-model bootstrap mean $\overline{d}$ at nine depths under three fixed-magnitude active-residue thresholds, with 95% percentile bootstrap CIs (B=1000). Direction is preserved at 52/54 cells (CI excludes 0); the two non-significant cells are at 13% relative depth at $1.0\times s$, where the quantile-based $\overline{d}$ is also near-null. Magnitudes vary substantially across thresholds at the extreme-depth peaks: the 0% full-set $\overline{d}$ falls from $+1.22$ (quantile) to $+0.20$ at $2.0\times s$, and the 100% full-set $\overline{d}$ falls from $+0.67$ to $+0.23$. Mid-depth values (38−63%) are stable across thresholds. The extreme-depth peaks are partly driven by many weakly-firing residues; mid-depth effects reflect stable strong feature firing.

	$0.5\times s$		$1.0\times s$		$2.0\times s$	
Rel. depth	full	val	full	val	full	val
0%	$+1.18$ $[+1.11, +1.24]$	$+0.56$ $[+0.49, +0.64]$	$+0.59$ $[+0.54, +0.64]$	$+0.30$ $[+0.23, +0.37]$	$+0.20$ $[+0.15, +0.24]$	$+0.10$ $[+0.05, +0.15]$
13%	$+0.05$ $[+0.02, +0.08]$	$+0.06$ $[+0.03, +0.10]$	$+0.01$ $[-0.03, +0.04]$	$+0.02$ $[-0.01, +0.06]$	$+0.06$ $[+0.03, +0.10]$	$+0.05$ $[+0.03, +0.07]$
25%	$+0.31$ $[+0.24, +0.37]$	$+0.18$ $[+0.07, +0.29]$	$+0.30$ $[+0.24, +0.36]$	$+0.17$ $[+0.07, +0.28]$	$+0.21$ $[+0.14, +0.28]$	$+0.06$ $[+0.01, +0.13]$
38%	$+0.44$ $[+0.36, +0.52]$	$+0.23$ $[+0.14, +0.32]$	$+0.51$ $[+0.43, +0.59]$	$+0.25$ $[+0.16, +0.34]$	$+0.34$ $[+0.27, +0.43]$	$+0.22$ $[+0.17, +0.26]$
50%	$+0.55$ $[+0.46, +0.64]$	$+0.26$ $[+0.16, +0.37]$	$+0.66$ $[+0.58, +0.74]$	$+0.30$ $[+0.20, +0.41]$	$+0.47$ $[+0.41, +0.53]$	$+0.27$ $[+0.23, +0.32]$
63%	$+0.45$ $[+0.37, +0.54]$	$+0.22$ $[+0.12, +0.33]$	$+0.61$ $[+0.52, +0.69]$	$+0.31$ $[+0.21, +0.40]$	$+0.43$ $[+0.37, +0.48]$	$+0.29$ $[+0.24, +0.33]$
75%	$+0.29$ $[+0.22, +0.37]$	$+0.13$ $[+0.05, +0.21]$	$+0.48$ $[+0.41, +0.55]$	$+0.27$ $[+0.19, +0.35]$	$+0.31$ $[+0.26, +0.35]$	$+0.19$ $[+0.15, +0.23]$
88%	$+0.33$ $[+0.28, +0.39]$	$+0.17$ $[+0.12, +0.22]$	$+0.57$ $[+0.51, +0.63]$	$+0.36$ $[+0.30, +0.42]$	$+0.22$ $[+0.18, +0.27]$	$+0.16$ $[+0.13, +0.19]$
100%	$+0.71$ $[+0.64, +0.76]$	$+0.35$ $[+0.29, +0.41]$	$+0.85$ $[+0.76, +0.93]$	$+0.48$ $[+0.41, +0.56]$	$+0.23$ $[+0.18, +0.27]$	$+0.15$ $[+0.13, +0.18]$

N. Val-only Robustness Tables

For completeness we report the main-text robustness sweeps recomputed on the 150-protein held-out validation subset alone (i.e. excluding all proteins the SAEs were trained on). Direction is preserved at every cell of every sweep on val. The full-set + val H2 ProtT5 encoder–decoder table appears in Appendix H (Table 12); we do not duplicate it here.

Table 21. H1 cross-model bootstrap mean $\overline{d}$ (ESM−RITA) under C_α cutoff sweep at nine matched depths, with 95% percentile bootstrap CIs (B=1000, protein-level resampling). Both full and val sets shown for $C_\alpha \in \{6, 10\}$ Å; the $C_\alpha = 8$ Å (main) cells appear in Table 18. 35/36 cells exclude 0 (the one near-null cell is at 75% under $C_\alpha = 10$ Å on val).

	$C_\alpha = 6$ Å		$C_\alpha = 10$ Å	
Rel. depth	full	val	full	val
0%	$+1.00$ $[+0.93, +1.07]$	$+0.44$ $[+0.38, +0.51]$	$+1.25$ $[+1.17, +1.32]$	$+0.58$ $[+0.49, +0.67]$
13%	$+0.03$ $[+0.00, +0.07]$	$+0.05$ $[+0.02, +0.08]$	$+0.05$ $[+0.01, +0.08]$	$+0.06$ $[+0.02, +0.10]$
25%	$+0.27$ $[+0.21, +0.33]$	$+0.13$ $[+0.04, +0.23]$	$+0.27$ $[+0.20, +0.34]$	$+0.18$ $[+0.06, +0.30]$
38%	$+0.33$ $[+0.27, +0.40]$	$+0.17$ $[+0.10, +0.24]$	$+0.42$ $[+0.34, +0.51]$	$+0.19$ $[+0.08, +0.31]$
50%	$+0.39$ $[+0.32, +0.46]$	$+0.17$ $[+0.09, +0.25]$	$+0.54$ $[+0.44, +0.64]$	$+0.25$ $[+0.13, +0.38]$
63%	$+0.33$ $[+0.26, +0.39]$	$+0.16$ $[+0.08, +0.24]$	$+0.43$ $[+0.34, +0.53]$	$+0.20$ $[+0.08, +0.32]$
75%	$+0.28$ $[+0.22, +0.34]$	$+0.12$ $[+0.06, +0.19]$	$+0.22$ $[+0.14, +0.30]$	$+0.08$ $[-0.02, +0.18]$
88%	$+0.34$ $[+0.30, +0.39]$	$+0.16$ $[+0.11, +0.20]$	$+0.28$ $[+0.22, +0.34]$	$+0.16$ $[+0.10, +0.22]$
100%	$+0.58$ $[+0.53, +0.62]$	$+0.29$ $[+0.24, +0.34]$	$+0.70$ $[+0.62, +0.77]$	$+0.40$ $[+0.33, +0.47]$

Table 22. Sequential-locality bootstrap mean $\overline{d}$ on L_{seq} under window sweep, RITA (residue) vs. ESM-2, with 95% percentile bootstrap CIs (B=1000). Sign-oriented *causal − bidirectional* (consistent with Table 15), so positive = RITA (causal) wins (note: opposite sign to Table 19, which uses ESM−RITA). Window-±2 cells (main setting) on val are in Table 19. The pattern is the same as the full set: a trough at ∼50–75% where ESM-2 strongly wins, then a flip at 100% where RITA wins.

	Window ±1		Window ±4	
Depth	full	val	full	val
0%	-1.31 $[-1.36, -1.22]$	-1.06 $[-1.12, -0.99]$	-1.84 $[-1.89, -1.73]$	-1.42 $[-1.49, -1.35]$
13%	$+0.25$ $[+0.23, +0.26]$	$+0.11$ $[+0.08, +0.13]$	$+0.02$ $[+0.00, +0.04]$	-0.10 $[-0.16, -0.06]$
25%	-0.16 $[-0.19, -0.13]$	-0.20 $[-0.26, -0.13]$	-0.60 $[-0.64, -0.56]$	-0.61 $[-0.72, -0.49]$
38%	-0.89 $[-0.94, -0.85]$	-1.00 $[-1.13, -0.86]$	-1.64 $[-1.72, -1.57]$	-1.61 $[-1.79, -1.43]$
50%	-1.80 $[-1.87, -1.73]$	-1.74 $[-1.92, -1.53]$	-2.69 $[-2.80, -2.58]$	-2.31 $[-2.57, -2.06]$
63%	-2.10 $[-2.18, -2.01]$	-1.99 $[-2.21, -1.76]$	-2.89 $[-3.02, -2.75]$	-2.42 $[-2.70, -2.14]$
75%	-2.13 $[-2.22, -2.04]$	-1.99 $[-2.21, -1.76]$	-2.57 $[-2.70, -2.44]$	-2.15 $[-2.42, -1.87]$
88%	-0.55 $[-0.61, -0.48]$	-0.57 $[-0.71, -0.40]$	-0.54 $[-0.62, -0.45]$	-0.54 $[-0.73, -0.34]$
100%	$+0.95$ $[+0.89, +1.01]$	$+0.87$ $[+0.74, +1.00]$	$+1.13$ $[+1.07, +1.19]$	$+1.01$ $[+0.86, +1.15]$