WHAT GOES UP SHOULD COME DOWN: A TRAINING-DISTRIBUTION CONTROL FOR INTERPRETABILITY METRICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Interpretability metrics are often supported by showing that their results are reproducible across seeds, model conditions, and analytical choices. Such robustness does not establish that a metric is valid for the interpretation attached to it. We demonstrate this distinction using a relation-referenced metric for sparse autoencoder (SAE) features in protein language models. The metric, L_{struct}, measures whether feature activations cluster across residues in spatial contact. In a controlled comparison of masked and causal models, L_{struct} gives a plausible masked-model advantage at every sampled depth. The difference reproduces across model seeds, persists under prediction matching, and survives six diagnostic analyses spanning dictionary fitting, reconstruction quality, residue selectivity, and representation basis. However, after retraining the same models on sequences with native residue order destroyed, L_{struct} increases in 52 of 54 objective-by-depth-by-seed cells. Random-initialisation and no-model baselines reveal complementary problems: the trained masked-model score does not exceed the random-initialisation range, and a cysteine indicator exceeds every learned feature in 32 of the same 54 cells. These results do not rule out structural information in L_{struct}, but they show that its robust masked–causal difference is not specific evidence of structural organisation learned from native residue order. Applying the same framework to further published interpretability metrics reveals different patterns of failure across the training-distribution, random-initialisation, and no-model controls. Interpretability metrics therefore require controls matched to each link between a score and its intended interpretation; reproducibility within a measurement pipeline is not sufficient evidence of construct validity.

1 INTRODUCTION

Protein language models (pLMs) trained on amino-acid sequences learn representations from which structural and functional properties can be recovered, from residue contacts and secondary structure to binding sites and mutational effects (Rives et al., 2021; Rao et al., 2021; Vig et al., 2021). Interpreting relational information requires more than identifying the positions on which a model feature activates: it requires establishing that those activations capture the relationship between the participating residues.

Sparse autoencoders (SAEs) provide one approach to interpreting model representations by decomposing them into a large dictionary of sparsely activating features (Simon & Zou, 2025; Adams et al., 2025; Gujral et al., 2025; Villegas Garcia & Ansuini, 2025). Most evaluations of these features fall into two broad classes. *Annotation-referenced* evaluations compare feature activations with an external label, such as a Swiss-Prot annotation, an InterPro family or a secondary-structure class, using a metric such as F1 (Simon & Zou, 2025; Adams et al., 2025; Villegas Garcia & Ansuini, 2025). *Relation-referenced* evaluations instead ask whether the positions on which a feature activates stand in a specified relationship, such as proximity in the folded protein (Simon & Zou, 2025). The distinction matters because a residue-level property can correlate with a structural relationship without representing that relationship. A feature selective for cysteine may appear spatially clustered because disulphide bonds bring cysteine pairs into contact, even if the feature carries no information about which cysteines are paired.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Relation-referenced statistics are commonly assessed against a permutation null in which feature activations are reassigned among positions within each protein. Such a null can establish that the observed activations are more spatially organised than a random reassignment of the same values. It does not establish that this organisation was learned from structural relationships in the training data. Residues with similar physicochemistry preferentially contact one another in folded proteins (Sengupta & Kundu, 2012). High feature-level interpretability scores can also arise without a trained model (Heap et al., 2025) or without learned SAE directions (Korznikov et al., 2026). A statistic may therefore be reproducible across seeds, dictionaries and analysis settings while remaining invalid for the interpretation attached to it. Robustness asks whether a numerical result persists; construct validity asks whether the statistic measures the phenomenon it is intended to measure (Bean et al., 2025).

Related concerns have motivated deliberately uninformative controls elsewhere in interpretability. Control tasks, which assign each word type a random but fixed output, separate what a representation encodes from what an expressive probe can memorise (Hewitt & Liang, 2019). Weight-shuffled models (Simon & Zou, 2025), randomly initialised ones (Heap et al., 2025) and SAEs with random, frozen components (Korznikov et al., 2026) test whether interpretability scores depend on learned model or feature structure. These controls randomise the model or the dictionary but leave the training data intact; they do not test whether a relation-referenced statistic responds specifically to the relationship it names. A complementary strategy is to intervene on the training distribution, removing the target relationship while retaining other properties of the data. Sinha et al. (2021) used this logic to examine syntactic information in language models trained on word-order-permuted text. The effectiveness of such controls depends on how the shuffling is implemented, because residual order information can survive apparently order-destroying transformations (Abdou et al., 2022). One route they identify, shuffling before subword segmentation, cannot arise with single-residue tokens. The other, a dependence between sequence length and token frequencies, remains open by design. For protein sequences, residue order can be destroyed while preserving the exact length and amino-acid composition of every training example.

Masked and causal pLMs provide a useful setting in which to test such a control. A masked model can integrate information from residues on either side of a sequence position, whereas a causal model can access only earlier positions (Devlin et al., 2019). A structural contact can therefore influence the representations of both participating residues in a masked model, but in a causal model only that of the residue occurring later in the sequence. This motivates the hypothesis that masked training produces representations in which structural relationships are encoded more symmetrically. Existing comparisons of masked and causal pLMs confound the training objective with architecture, corpus, tokenisation or scale (Gao & Fedorec, 2026). Where those are held fixed, the models compared are still trained separately and matched only on size, compute and training tokens, with no reported shared initialisation, batch order or replicate seeds (Cheng et al., 2024). A pair of models sharing architecture, initialisation, corpus and batch order allows the objective contrast to be examined under substantially tighter control.

Here we use such a pair to evaluate L_{struct}, a relation-referenced statistic that measures co-activation of SAE features across contacting residues (Gao & Fedorec, 2026), which was adapted from the structural-clustering analysis of Simon & Zou (2025). We first ask whether the statistic produces the expected masked–causal difference and whether that difference survives conventional diagnostic checks. We then retrain the models on sequences in which residue order has been destroyed, and test whether L_{struct} responds specifically to structural organisation learned from native sequence order. Finally, we compare the training-distribution intervention with random-initialisation and no-model baselines, and apply all three to further published or adapted interpretability metrics. The metrics show different patterns across the controls, demonstrating that the controls test complementary rather than interchangeable requirements for interpretation.

2 METHODS

2.1 CONTROLLED MODEL COMPARISON

We trained paired masked and causal protein language models using a common 42.0M-parameter transformer architecture (41,981,504 parameters). The model comprised 30 pre-layer-normalised blocks with $d_{\mathrm{model}} = 320$, five attention heads, SwiGLU feed-forward layers and rotary position

embeddings, following the per-block design of ESM C (Candido et al., 2026). As in ESM C, the query and key projections are layer-normalised, linear layers are bias-free and the sub-layer outputs are residually scaled. ESM C is masked-only; the causal models add a lower-triangular attention mask. Input and output embeddings are separate. Within each seed, the two conditions used identical initial parameters, training sequences, batch order, architecture and optimisation procedure. The deliberately varied components were the attention mask and the associated prediction objective. The masked and causal conditions are referred to throughout as the two *objectives*. The models match ESM-2-8M (Lin et al., 2023), the model analysed by Simon & Zou (2025) and Villegas Garcia & Ansuini (2025), in width and exceed it in depth and size.

For masked training, 15% of eligible residue positions were selected as prediction targets at each step under the standard 80/10/10 replacement procedure (Devlin et al., 2019). The causal models used left-to-right attention and predicted the next residue at each eligible position. Both conditions were trained on the same three million UniRef50 sequences for 40,283 steps. We trained three paired model seeds, 42, 43 and 44. Full architecture, corpus and optimisation details are in Appendix A.

The primary comparison matched the number of input tokens. Because masked training receives supervision at fewer positions per sequence, we additionally trained prediction-matched masked models with enough optimisation steps to match the number of causal prediction targets (Appendix A). An extended-budget condition, a 32-fold increase in nominal tokens, was evaluated for seed 42 only.

We analysed representations from blocks $0, 4, 7, 11, 14, 18, 22, 26$ and 29, spanning the transformer stack. Block 0 denotes the output of the first transformer block, not the input embedding.

2.2 Training-distribution interventions

We constructed an order-destroyed training corpus by independently permuting the residues within each pretraining sequence. The intervention preserved the exact length and amino-acid composition of every sequence while disrupting native within-sequence residue order. Boundary tokens were added after permutation and never moved, and one order-destroyed corpus was shared across objectives and model seeds.

For each seed, masked and causal models were trained on this corpus using the same initial parameters, architecture, optimisation procedure, batch order and training budget as their native-corpus counterparts. All models were evaluated on native protein sequences and experimentally determined native structures. For the intervention-trained models these sequences are out of distribution. A metric of learned structural organisation should nonetheless not report more of it in models that never saw native residue order.

We also trained a block-shuffled intervention that permuted contiguous 16-residue blocks while preserving residue order within each block (Appendix A).

Our falsification criterion concerns the direction of the native-versus-intervention comparison. If a statistic specifically measures structural organisation learned from native residue order, its value should not systematically increase after that order has been disrupted during training. We do not require the statistic to fall to zero, because the intervened models can still learn sequence length, amino-acid composition, positional effects and other preserved regularities.

2.3 Structural evaluation data and sparse autoencoders

Structural evaluation used 1,500 protein domains from the SCOPe 2.08 ASTRAL subset filtered at 40% sequence identity, all with experimentally determined structures (Chandonia et al., 2022). Domains were truncated to their first 510 residues so that two boundary tokens fit within the 512-token context, giving 293,760 residue positions.

The domains were partitioned by complete SCOPe fold: 1,349 domains (272,525 residues) for dictionary fitting and 151 held-out domains (21,235 residues, 23 folds) for scoring. No held-out domain shares a fold with any fitting domain. Unless stated otherwise, every number below is computed on the held-out partition only, so that dictionary fitting and metric evaluation use disjoint domains. Analyses that did not use this partition are marked where they appear and listed in Appendix J. No structural classification guarantees unrelatedness, since all extant proteins share evolutionary history, but fold-disjointness removes the strongest source of overlap.

We fitted a separate TopK SAE (Gao et al., 2025) for each model condition, model seed and sampled block. Each dictionary contained 2,560 latent features (expansion factor 8) with $k = 256$ active features per residue. The primary dictionary seed was fixed across pLM seeds. SAE architecture, optimisation, normalisation and dictionary variants are in Appendices B and E.

2.4 STRUCTURAL CO-ACTIVATION

The construction, reported in Gao & Fedorec (2026), adapts the structural-clustering analysis of Simon & Zou (2025): the pairing of a sequential with a structural neighbourhood, the Cα contact criterion and the within-protein permutation null are theirs. Gao & Fedorec (2026) enlarge the contact radius from 6 to 8 Å, impose a minimum sequence separation, and score over an active set of residues rather than at each protein's maximum-activation residue.

For feature f, let $a_f(i)$ denote its activation at residue i. The active set A_f contains residues whose activation strictly exceeds the 90th percentile of f's activation distribution across the held-out partition. The structural neighbourhood of residue i is

$$\mathcal{N}_{\text{struct}}(i) = \big\{\, j \;:\; \|\mathrm{C}\alpha_i - \mathrm{C}\alpha_j\|_2 < 8\,\text{Å},\;\; |i-j| \geq 12 \,\big\}. \tag{1}$$

The 8 Å threshold follows common contact-prediction evaluations (Rao et al., 2021; Schaarschmidt et al., 2018). Those evaluations define contacts between Cβ atoms (Cα for glycine); instead, the Cα distance is used, as in Simon & Zou (2025). The sequence-separation requirement excludes local backbone contacts and retains exactly their medium- and long-range classes. With $\bar{a}_f(i)$ the mean activation of f over $\mathcal{N}_{\text{struct}}(i)$, the observed contrast is computed as

$$C_f = \mathbb{E}_{i \in A_f}\big[\bar{a}_f(i)\big] - \mathbb{E}_i\big[\bar{a}_f(i)\big]. \tag{2}$$

Residues with no qualifying neighbour enter both expectations with $\bar{a}_f(i) = 0$; they are 33.4% of held-out residues, and Appendix C states the consequence. The feature activations are then permuted among residue positions independently within each protein, holding the contact graph fixed, and the contrast is recomputed as C_f^{π}. Averaging over five permutations gives

$$L_{\text{struct}}(f) = \frac{C_f - \mathbb{E}_{\pi}\big[C_f^{\pi}\big]}{\mathrm{SD}(a_f)}, \tag{3}$$

where $\mathrm{SD}(a_f)$ is the partition-wide standard deviation of the raw activations of f. The null preserves each protein's contact graph and its multiset of feature activations while disrupting their correspondence.

TopK activations are zero-inflated, so the 90th-percentile threshold is frequently exactly zero (52–58% of live features) and the active set then corresponds to the non-zero support, a median of 8.5–9.8% of residues. Features with fewer than five active residues are assigned $L_{\text{struct}} = 0$ and retained in dictionary means; excluding them instead leaves the order-destroyed result unchanged (Appendix C). The exact weighting, permutation, missing-neighbour and inactive-feature conventions are specified in Appendix C.

The statistic detects structural organisation only when it appears as co-activation of the same feature at contacting residues. It does not detect information distributed across features or encoded relationally across pairs of residue representations.

L_{seq} is defined using the same calculation with the structural neighbourhood replaced by residues at offsets $\{-2, -1, +1, +2\}$, the sequential neighbourhood of Simon & Zou (2025). L_{seq} is analysed separately and is never subtracted from L_{struct}.

2.5 DIAGNOSTIC ANALYSES AND VALIDITY BASELINES

We tested whether the masked–causal difference depended on the learned SAE basis using three dictionary seeds, three frozen random dictionaries (Korznikov et al., 2026) and a dictionary-free analysis in which the 320 residual-stream dimensions were treated as features. Residual dimensions are not interpreted as individually meaningful features. Two further diagnostics test the measurement rather than the dictionary. One compares the masked–causal difference with each objective's reconstruction quality. The other splits the structural neighbourhood into upstream and downstream

contacts, testing the prediction that follows from the causal models' lack of access to later positions. We also quantified feature selectivity for amino-acid identity from each feature's mean absolute activation per residue type.

We used three *reference readouts*: measures of structural information obtained without L_{struct}. First, linear and nonlinear probes predicted residue-level helix, strand and burial labels directly from pLM representations. Second, contact precision at $L/5$ ranked residue pairs without supervision. Third, pairwise probes predicted structural contacts from pairs of residue representations; they were run twice, on the raw residual stream and on the SAE latents, so that the two can be compared. We also scored a no-model layer of 29 deterministic indicators, among them the cysteine indicator of §1. Indicators are scored through the same L_{struct} implementation, but with the active set taken as the top 60% of residues, because a binary indicator is tied at the 90th-percentile threshold. Finally, we evaluated randomly initialised instances of the same transformer architecture. Each draw seeds both the network weights and the dictionary, so the three draws are independent complete pipelines. Full diagnostic, readout and baseline specifications are in Appendices B, E, I and G.

2.6 Statistical analysis

A *cell* is one objective-by-depth-by-seed combination, 54 in all from two objectives, nine depths and three model seeds; a *matched cell* is the same combination compared across two training corpora. The independently trained model seed was the experimental unit for inference about the training objective. Residues, proteins, residue pairs, SAE features, dictionary variants and representation depths were treated as nested or repeated observations within a trained model.

For each model seed s and depth b we calculated the raw masked-minus-causal difference Δ_{sb} in mean L_{struct} over the complete dictionary, and a standardised feature-distribution effect

$$d_{sb} = \frac{\overline{L_{\text{struct}}}_{sb}^{\text{masked}} - \overline{L_{\text{struct}}}_{sb}^{\text{causal}}}{\sqrt{\left(s^2_{\text{masked},sb} + s^2_{\text{causal},sb}\right)/2}}, \tag{4}$$

where the means and variances are taken over the 2,560 per-feature values of each objective. Depth-level effects were averaged within each paired model seed, and the three seed-level summaries were tested against zero with a two-sided one-sample t-test on two degrees of freedom.

Evaluation-set uncertainty was quantified with 1,000 fold-clustered bootstrap replicates: SCOPe folds were drawn with replacement and domains were then drawn with replacement within each drawn fold, with paired resampling indices for the masked and causal conditions. Each replicate recomputes the re-weighted contrasts, per-feature values and the effect size; per-feature standard deviations, percentile thresholds, the contact graph and the permutations are computed once on the held-out partition and held fixed. Bootstrap intervals are 95% percentile intervals, and the accompanying estimate is the bootstrap mean, which is shrunk toward zero relative to the original-sample estimate (Appendix D). Benjamini–Hochberg correction was applied across the nine analysed depths within each model seed.

For the order-destroyed control, native and intervention-trained models were paired by objective and seed. Each of the six objective-by-seed pairs contributed one summary, the mean of its nine depth-level differences, and an exact one-sided sign test was applied to the six model-level signs. Cell-level direction counts are reported descriptively and are not treated as independent model-level replicates.

3 Results

3.1 L_{struct} produces a robust masked–causal difference

Our hypothesis was that bidirectional access would mean that masked training would produce more symmetric structural representations than causal training. L_{struct} supported this prediction: the masked models scored higher at every sampled depth. Mean depth-averaged effects were 0.1387, 0.1425 and 0.1627 across the three paired seeds, giving $d = +0.1479$ (95% CI $[0.1159, 0.1800]$; $t = 19.88$, df $= 2$, $p = 0.0025$). Three, five and six of the nine depth-wise comparisons survived Benjamini–Hochberg correction within each seed (Appendix E).

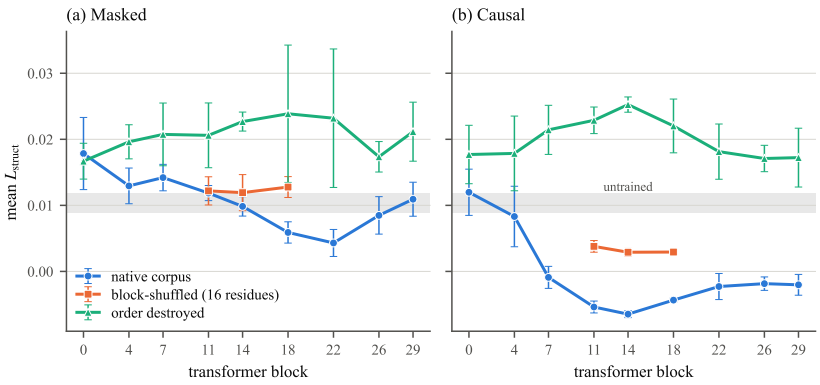


Figure 1: Mean L_{struct} across the nine sampled depths, averaged over three model seeds; error bars show the seed-to-seed standard deviation. The native and order-destroyed conditions were evaluated at all nine depths; the block-shuffled intervention at blocks 11, 14 and 18. The grey band spans the random-initialisation means — three draws each, both objectives pooled — computed at blocks 11, 14 and 18 and drawn across all depths for reference. It is not a confidence interval. On the causal models (b) the order-destroyed curve sits above native at all nine depths; in the masked models (a) it sits above native at every depth except block 0.

The direction persisted under prediction matching, with all nine comparisons at blocks 11, 14 and 18 supported across the three seeds, and 18 of 27 depth-by-seed comparisons supported across the full grid (Appendix E). It also remained positive across three dictionary seeds fitted to the seed-42 models, three frozen random dictionaries, and, at seed 42, a dictionary-free analysis of residual-stream dimensions. Reconstruction quality tracked the effect only at the two shallowest blocks; residue selectivity showed weak association with L_{struct} and strong association with L_{seq}; and the directional interaction had the opposite sign to that predicted by the causal models' lack of access to later sequence positions (Appendix E).

3.2 A TRAINING-DISTRIBUTION CONTROL BREAKS THE INTENDED INTERPRETATION

We tested the construct-level interpretation by retraining the same architecture after permuting residues within each training sequence. The intervention preserved exact length and amino-acid composition while removing native residue order and its correspondence with native protein structure. We expected L_{struct} scores to fall when the residue order is permuted, as protein structures depend on relative residue position.

L_{struct} increased in 52 of 54 objective-by-depth-by-seed cells (Figure 1); the two decreases occurred in the masked seed-43 and seed-44 models at block 0. All six objective-by-seed summaries increased (exact one-sided sign test, $p = 0.016$), and at the seed level the mean rise was $+0.01512$ ($t(2) = 31.76$, $p = 0.001$; causal $p = 0.004$, masked $p = 0.025$). Averaged over three seeds at blocks 11, 14 and 18, the causal mean moved from -0.00540 to $+0.02339$ and the masked mean from $+0.00920$ to $+0.02238$; all 18 differences at these depths are positive. The block-shuffled intervention of §2.2, which permutes 16-residue blocks but preserves residue order within each, increased L_{struct} in all nine causal cells and six of nine masked cells (Appendix F). The causal increase therefore reproduced across distinct order-disrupting interventions, while the masked response was intervention-dependent.

The masked–causal difference is reproducible, but it cannot support the interpretation that masked models learned more symmetric structural organisation from native residue order.

3.3 A SEQUENTIAL POSITIVE CONTROL RESPONDS AS EXPECTED IN ONLY ONE OBJECTIVE

The robustness analyses and the order-destroyed intervention test different claims. The sequential statistic illustrates the distinction. Under native training, L_{seq} was higher in the causal models at every sampled depth. At the five deepest masked-model blocks, order-destroyed training reduced L_{seq} to between 14% and 50% of its native value, below native in all 15 depth-by-seed cells. In the

Table 1: Mean L_{struct} for trained and randomly initialised networks at blocks 11, 14 and 18, on the held-out partition. Values are means over three model seeds (trained) or three independent random-initialisation draws, with the seed-to-seed standard deviation. A dash marks the causal rows, whose trained mean is below zero, so the ratio is not meaningful. Per-cell values are in Appendix G.

Objective	Block	Trained ($n = 3$)	Random initialisation ($n = 3$)	Trained / random (%)
Masked	11	$+0.01186 \pm 0.00114$	$+0.01177 \pm 0.00163$	100.8%
Masked	14	$+0.00985 \pm 0.00147$	$+0.01138 \pm 0.00177$	86.5%
Masked	18	$+0.00590 \pm 0.00161$	$+0.01125 \pm 0.00175$	52.4%
Causal	11	-0.00539 ± 0.00090	$+0.00908 \pm 0.00116$	—
Causal	14	-0.00647 ± 0.00052	$+0.00889 \pm 0.00152$	—
Causal	18	-0.00435 ± 0.00030	$+0.00962 \pm 0.00057$	—

causal models the same intervention increased L_{seq} to 1.48–1.92 times native, above native in all 15 cells (Appendix F). L_{struct} rose in all 30 of these cells.

If L_{seq} measures sequence-local organisation learned from native residue order, destroying that order should lower it. It did so in the masked models but not in the causal models, so L_{seq} is a partial rather than a general positive control: the co-activation construction can register loss of sequential organisation in one objective and depth regime, but the response depends on the representations to which the estimator is applied.

3.4 Random-initialisation and no-model controls diagnose complementary failures

Random-initialisation and no-model baselines test two further requirements: whether the absolute score range requires model training, and whether it requires a learned representation at all.

At blocks 11, 14 and 18, randomly initialised networks obtained mean L_{struct} between $+0.009$ and $+0.012$ (Table 1, Figure 4). The trained masked means occupied the same range at block 11, fell to 86.5% of it at block 14 and to 52.4% at block 18, while the trained causal means were below zero throughout. The masked–causal difference was therefore produced by the trained causal score falling below the random-initialisation range, not by the trained masked score rising above it.

The no-model analysis showed that even high individual feature scores did not require a learned representation. The cysteine indicator obtained $L_{\mathrm{struct}} = +0.4067$, exceeding every learned feature in four of six tested objective-by-depth conditions at seed 42, and in 32 of the 54 objective-by-depth-by-seed cells overall. On the causal models it was 3.3–4.7 times the best learned feature. Several hydrophobic and aromatic indicators also reached the learned-feature range (Appendix G). These high scores are consistent with known non-random organisation of residue identities in folded proteins, including disulphide pairing and hydrophobic core packing.

3.5 Metric construction determines the apparent model comparison

The interpretation of L_{struct} also depended on how the target relationship was defined. At a fixed 6 Å contact radius we varied the minimum sequence separation used to define structural neighbours. At block 11 the masked-minus-causal effect changed from $d = -1.9322$ at a minimum separation of one residue to $d = +0.1444$ at a separation of 12 residues, crossing zero at five residues; blocks 14 and 18 crossed at five and four residues (Figure 2). The ranking of the two training objectives therefore depended on whether immediately local sequence neighbours were admitted.

This reversal arose because a structural neighbourhood without a separation floor is dominated by sequentially local edges: at 6 Å all adjacent residue pairs qualify as structural neighbours, and pairs with $|i - j| \leq 3$ account for approximately 74% of admitted edges. At block 11, where the reversed effect was largest, the resulting masked–causal comparison closely matched the one obtained with L_{seq} (Appendix H). The nominal structural statistic was therefore reporting primarily sequential locality under that definition.

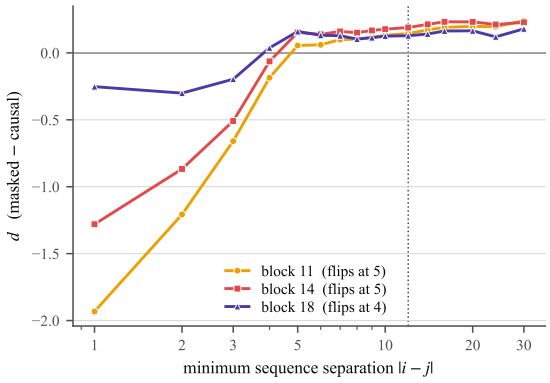


Figure 2: Masked-minus-causal standardised effect as a function of minimum sequence separation at a fixed 6 Å contact radius, averaged over three model seeds. Lines show blocks 11, 14 and 18, with each block's first positive separation in the legend. The dotted line marks the primary minimum separation of 12 residues. The comparison reverses when immediately local sequence neighbours are included.

The selected threshold of 12 residues was not the value that maximised the masked-model advantage; the effect grew more slowly after approximately six residues and was larger at more restrictive thresholds. Nevertheless, the sign reversal demonstrates that a plausible analytical choice can determine the apparent model ranking.

3.6 Reference readouts do not reproduce the same depth profile

The reference readouts were not intended to identify a single correct ranking of masked and causal representations. They tested whether the depth profile obtained with L_{struct} was supported by other measures of related structural information. The probes read the residual stream rather than the SAE latents. This does not explain the disagreement: at seed 42, L_{struct} computed directly on the 320 residual dimensions reproduces the masked advantage at blocks 11, 14 and 18 (Appendix E).

At blocks 11 and 14, linear probes for helix and strand favoured the causal models in all 12 label-by-seed comparisons, with every seed-level interval excluding zero. At block 18 the helix probe was indistinguishable from zero ($+0.0044$; two of three seeds favouring masked, no interval excluding zero) while the strand probe continued to favour the causal models (-0.0305; three of three seeds). Contact precision at $L/5$ gave a third pattern, favouring the masked model at block 11 and the causal model at blocks 14 and 18. Of the four helix and strand comparisons at blocks 11 and 14, a nonlinear probe strengthened one and weakened three, and all four still favoured the causal models.

The reference readouts therefore did not reproduce the L_{struct} profile: where the masked advantage was largest the residue-level probes favoured the causal models, and at block 18 they were mixed. These differences do not determine which model contains more structural information, because the metrics evaluate distinct properties. They show that the conclusion drawn from the model comparison depends on the operational measure used.

Pairwise contact probes provided an additional test of the SAE representation. For the seed-42 models, separation-matched contact AUROC was 0.570–0.594 for the raw residual stream against 0.525–0.567 for the SAE representation, higher for the raw stream in all six tested comparisons by 0.020–0.045, with shuffled-label controls at 0.489–0.507. Under this probe, contact information was less linearly recoverable after SAE encoding. Complete probe architectures and values are in Appendix I.

3.7 Published metrics fail different validity requirements

To examine whether the control logic was specific to L_{struct}, we applied the three checks to further interpretability metrics for which the corresponding intervention is defined. These were evaluated on the full 1,500-domain evaluation set, which includes the dictionaries' fitting domains. Their

protocols require more qualifying proteins per feature or per label than 151 domains in 23 folds can supply. The co-activation–decoder-geometry correlation is also reported on the held-out partition. Appendix K gives the protocol and partition for each.

Every metric failed at least one applicable check, but the pattern of outcomes differed across metrics (Table 16). Two failures echo L_{struct}. The InterPLM structural-clustering statistic d_{struct} (Simon & Zou, 2025) rose in the causal models after order-destroyed training, while falling in the masked models. Single-latent F1, adapted from Adams et al. (2025), was exceeded by the no-model indicator layer in both objectives, just as the cysteine indicator exceeds learned features under L_{struct}. The other three fail differently. The co-activation–decoder-geometry correlation ρ, adapted from Li et al. (2025), responded to the intervention in both objectives, but scored randomly initialised networks *above* trained ones in the masked models. k-sparse probing, adapted from SAEBench (Karvonen et al., 2025), stayed above both baselines, significantly in three of the four comparisons, and showed no detectable response to the intervention. Concept-F1 (Simon & Zou, 2025) responded in the causal models only.

These are seed-level comparisons over three model seeds, the unit used in §2.6. Under the order-destroyed intervention, directions reverse between objectives for three of the five metrics, so objective-pooled summaries are misleading.

4 Discussion

A metric can be robust and still support the wrong interpretation. L_{struct} produced a plausible masked–causal difference that reproduced across model seeds, prediction-matching protocols, dictionary constructions and representation bases. These checks established that the numerical result was stable within the measurement pipeline. They did not establish that the result measured structural organisation learned from native residue order.

The training-distribution control tested that interpretation directly. After native residue order was destroyed, L_{struct} increased in 52 of 54 cells. The original masked–causal difference remains reproducible, but it cannot support the conclusion attached to it. This does not mean that L_{struct} contains no structural information: the evaluation proteins retain real three-dimensional structure, residue physicochemistry and non-random contact patterns. It means that a higher score is not specific evidence of structural organisation learned from native sequences.

The additional controls expose different gaps in the evidential chain. Random initialisation asks whether the score requires model training; no-model indicators ask whether it requires a learned representation; the training-distribution intervention asks whether it requires the target relation during learning. The sequence-separation analysis identifies a further requirement: the operational definition of the construct must itself be tested, since plausible neighbourhood definitions produced opposite model rankings.

These conclusions are limited to one architecture, model scale, masking rate and structural dataset, with a single-seed extended-budget analysis. The held-out partition is small and unevenly composed: 151 domains in 23 folds, four of which hold 77% of the set, with a mean domain length below that of the full set, so absolute levels and effect sizes reflect that sample. The statistic assigns zero to features with insufficient active residues, uses five permutations per feature, and enters residues without qualifying neighbours into both expectations with a neighbour average of zero. Two further limitations we have not resolved: order-destroyed dictionaries reconstruct their inputs better than native ones in all 54 held-out cells, so reconstruction quality and the effect are not independent (Appendix E), a confound that matching reconstruction quality across corpora would remove; and the mechanism generating the score remains unresolved.

Relation-referenced metrics can still identify reproducible organisation and generate useful hypotheses. But reproducibility shows that a measurement is stable, not that its interpretation is correct. When a metric is claimed to measure a learned relationship, validation should include an intervention that disrupts that relationship while preserving plausible alternatives. L_{struct} provides a cautionary example of why that stronger test is necessary.

Reproducibility statement

The controlled pair is 30 pre-LN blocks, $d_{\text{model}} = 320$, 5 heads, SwiGLU and RoPE, trained with AdamW ($\beta = 0.9, 0.95$), decoupled weight decay 0.1 with zero decay on normalisation weights, gradients clipped to global norm 1.0, learning rate 6×10^{-4} with 500 warm-up steps and cosine decay to one tenth of peak, bfloat16 autocast, no gradient accumulation, batch size 32, maximum sequence length 512. Within each seed the two objectives share initial parameter values, training corpus and batch order; only the attention mask and the loss built from it differ. Dictionaries are TopK with $k = 256$ and expansion 8, trained with AdamW at 5×10^{-5} with cosine annealing to 5×10^{-6}, batch size 8,192 residues in bfloat16, 60 epochs, final-epoch weights, no checkpoint selection, and an auxiliary loss over 64 latents at coefficient $1/32$. Appendices A and B give the full configuration; Appendix K gives reproduction notes for the reimplemented statistics. The supplementary material contains the code and the per-cell results. A single command regenerates every table and figure from them and checks each table cell against the paper.

AI Usage Statement

We used Claude (Anthropic; Opus 4.7, 5.0 and 5.5) to help with implementing the training, evaluation and analysis code; with reviewing the methodology and finding implementation faults; with drafting and revising parts of the manuscript, figures and tables; with suggesting related papers for the authors to read; and with double-checking citations against their sources. Google Paper Assistant Tool (PAT) provided automated pre-submission feedback. No AI tool generated data, proposed hypotheses or formulated mathematical claims. The authors reviewed all AI-assisted code and text, tested the code, and take responsibility for the content.

References

Mostafa Abdou, Vinit Ravishankar, Artur Kulmizev, and Anders Søgaard. Word order does matter and shuffled language models know it. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6907–6919. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.476. URL https://aclanthology.org/2022.acl-long.476/.

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*, volume 267 of *PMLR*, pp. 460–476, 2025.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Kaili-May Liu, Lennart Luettgau, Jabez Magomere, Jonathan Rystrøm, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Nicolaus Foerster, Yarin Gal, Scott A. Hale, Inioluwa Deborah Raji, Christopher Summerfield, Philip H. S. Torr, Cozmin Ududec, Luc Rocher, and Adam Mahdi. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=mdA5lVvNcU.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, et al. Language modeling materializes a world model of protein biology. *bioRxiv preprint*, 2026. doi: 10.64898/2026.06.03.729735.

John-Marc Chandonia, Lindsey Guan, Shiangyi Lin, Changhua Yu, Naomi K. Fox, and Steven E. Brenner. 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. doi: 10.1093/nar/gkab1054.

Xingyi Cheng, Bo Chen, Pan Li, Jing Gong, Jie Tang, and Le Song. Training compute-optimal protein language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

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. In *International Conference on Learning Representations (ICLR)*, 2025.

Wei Gao and Alex Fedorec. Structural locality differentiates residue-tokenised bidirectional and causal protein language model families. In *Mechanistic Interpretability Workshop at ICML*, 2026.

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.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. Automated interpretability metrics do not distinguish trained and random transformers. *arXiv preprint arXiv:2501.17727*, 2025.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *International Conference on Machine Learning (ICML)*, 2025.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Y. Rogov, Ivan Oseledets, and Elena Tutubalina. Sanity checks for sparse autoencoders: Do SAEs beat random baselines? *arXiv preprint arXiv:2602.14111*, 2026.

Yuxiao Li, Eric J. Michaud, David D. Baek, Joshua Engels, Xiaoqing Sun, and Max Tegmark. The geometry of concepts: Sparse autoencoder feature structure. *Entropy*, 27(4):344, 2025. doi: 10.3390/e27040344.

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. *Science*, 379(6637):1123–1130, 2023.

Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. In *International Conference on Learning Representations (ICLR)*, 2021.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118.

Joerg Schaarschmidt, Bohdan Monastyrskyy, Andriy Kryshtafovych, and Alexandre M. J. J. Bonvin. Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins: Structure, Function, and Bioinformatics*, 86(S1):51–66, 2018. doi: 10.1002/prot.25407.

Dhriti Sengupta and Sudip Kundu. Role of long- and short-range hydrophobic, hydrophilic and charged residues contact network in protein's structural organization. *BMC Bioinformatics*, 13 (1):142, 2012. doi: 10.1186/1471-2105-13-142.

Elana Simon and James Zou. InterPLM: Discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, 22(10):2107–2117, 2025. doi: 10.1038/s41592-025-02836-7.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2888–2913, 2021. doi: 10.18653/v1/2021.emnlp-main.230.

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. URL `https://openreview.net/forum?id=YWtLZvLmud7`.

Edith Natalia Villegas Garcia and Alessio Ansuini. Interpreting and steering protein language models through sparse autoencoders. *arXiv preprint arXiv:2502.09135*, 2025. Presented at the LMRL workshop, ICLR 2025.

A Model training and training-distribution interventions

Architecture. 30 pre-layer-normalised transformer blocks, $d_{\text{model}} = 320$, five attention heads, SwiGLU feed-forward layers and rotary position embeddings; 41,981,504 parameters. The input embedding and the output head are separate modules. Residues are tokenised at single-residue resolution over the standard 20 amino acids plus start, end and mask tokens, with inputs limited to 512 tokens including two boundary tokens. For comparison, ESM-2-8M has the same width, $d_{\text{model}} = 320$, with 6 blocks and about 8M parameters (Lin et al., 2023; Simon & Zou, 2025).

Objectives. Masked objective: 15% of eligible positions are selected as targets at each step, of which 80% are replaced by the mask token, 10% by a random amino acid and 10% left unchanged; cross-entropy is computed at selected positions only. Causal objective: left-to-right attention with next-residue prediction at every eligible position. Start, end and padding tokens are never masked targets; the end token is a causal target.

Optimiser. AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$ and decoupled weight decay 0.1, with no decay on normalisation weights or biases. Gradients are clipped to global norm 1.0. The learning rate rises linearly to 6×10^{-4} over 500 warm-up steps and then follows a cosine schedule to one tenth of peak. Training uses bfloat16 autocast, batch size 32 and no gradient accumulation.

Corpus. `ConvergeBio/uniref50`, snapshot through 2026-01-28, streamed with a 50,000-sequence shuffle buffer seeded with 42. Sequences are truncated to 510 residues; sequences shorter than 16 residues, sequences containing non-standard residues, and exact matches to the SCOPe 40% evaluation sequences are removed, which retains 3,000,000 of 3,024,561 drawn sequences, of which 2,000 are held out for validation. The evaluation domains are therefore separate from the pretraining corpus, but removal is by exact match only, so homologues of evaluation domains may remain in it.

Budget. Three model seeds (42, 43, 44), 40,283 optimisation steps each. The nominal token count, $40{,}283 \times 32 \times 512 \approx 6.6 \times 10^8$, counts padded grid positions; the corresponding residue count is approximately 4.5×10^8, or 10.8 tokens per parameter, drawn from about 43% of the corpus.

Prediction-matched training. The prediction-matched masked models run $\lfloor 40{,}283/0.15 \rfloor = 268{,}553$ steps at the same batch size, which matches the expected number of supervised targets to the causal models within 0.3% (approximately 451.5M against 452.8M). They therefore see more input tokens (approximately 3.0×10^9 residues, 2.9 passes over the corpus) and take more optimisation steps than the causal models; the protocol matches supervision, not compute or data repetition.

Extended budget. The extended-budget condition runs 1,281,738 steps, nominally 2.1×10^{10} tokens (500 per parameter) and approximately 1.4×10^{10} residues, at seed 42 only.

Order-destroyed intervention. Residues within each pretraining sequence are permuted independently at corpus-construction time, from a single random stream, after filtering and truncation; boundary tokens are added afterwards. One corpus is shared across objectives and seeds, and each intervention-trained model reuses the architecture, initial parameters, optimisation procedure, budget and batch order of its native counterpart.

Block-shuffled intervention. Each sequence is divided into contiguous 16-residue blocks from the N-terminus, with a shorter terminal block when the length is not divisible by 16, and the blocks are permuted while residue order within each block is preserved.

Representation depths. Blocks $0, 4, 7, 11, 14, 18, 22, 26, 29$, corresponding to relative depths $0, 13, 25, 38, 50, 63, 75, 88$ and 100%. Blocks 11, 14 and 18 are used as diagnostic depths in the analyses that do not span the full grid. Block 0 is the output of the first transformer block; block 29 is taken before the final layer normalisation.

B Evaluation data and sparse autoencoder training

Structural evaluation dataset. 1,500 SCOPe 2.08 ASTRAL domains at 40% sequence identity with experimentally determined structures (Chandonia et al., 2022), truncated to 510 residues, giving 293,760 residue positions. Structural calculations use $\mathrm{C}\alpha$ coordinates; residues without a resolved $\mathrm{C}\alpha$ (66 of the 21,235 held-out residues) remain rows of the feature matrix and lose only their contacts. Contacts crossing the truncation boundary are excluded. The held-out contact graph has 55,886 edges.

Partition. Complete SCOPe folds are assigned to the fitting or held-out side, giving 1,349 fitting domains (272,525 residues) and 151 held-out domains (21,235 residues, 23 folds). Dictionaries are fitted on the fitting partition only and L_{struct} is scored on the held-out partition only; thresholds, standard deviations, contact graphs and permutations are all computed within whichever partition is being scored.

Architecture. TopK sparse autoencoders with $k = 256$ active latents per residue out of 2,560 (expansion factor 8). No L1 penalty or activation shrinkage. An auxiliary loss over $k_{\mathrm{aux}} = 64$ latents at coefficient $1/32$ revives inactive features, following Gao et al. (2025) with four differences. The residual it reconstructs is not normalised. It is not detached, whereas their released implementation detaches it. k_{aux} is 64 rather than of order $d_{\mathrm{model}}/2$ (they use 512). A feature counts as dead after 10^6 consecutive training residues without activation, rather than after their 10^7 tokens.

Normalisation. Activations are scaled by one global scalar, $\sqrt{320}$ divided by the mean row norm of the fitting rows, and a fixed per-dimension mean of the scaled fitting rows is subtracted before encoding and added back after decoding. Per-token normalisation is not used in any reported pipeline except the frozen random dictionary comparison below, where those variants and the fitted baseline they are compared against both use it.

Optimisation. AdamW at learning rate 5×10^{-5} with cosine annealing to 5×10^{-6} over 60 epochs, batch size 8,192 residue representations in bfloat16, decoder columns renormalised after each step, final-epoch weights, and no checkpoint selection on validation loss, L_{struct} or any downstream evaluation.

Dictionary-seed sensitivity. Three independently initialised dictionaries were fitted for each of six objective-by-depth conditions; see Appendix E.

Frozen random dictionaries. Three variants fix one SAE component at random values and train the rest, following Korznikov et al. (2026): a *frozen decoder* of random directions drawn with the activation covariance and held fixed; a *frozen encoder* of random isotropic directions held fixed;

and a *soft-frozen decoder* initialised at random isotropic directions and projected after each step to keep every column within cosine 0.8 of its initialisation. These variants were trained with per-token normalisation at batch size 4,096 in float32, at blocks 7, 14 and 22 of the seed-42 models. They were fitted on 1,350 domains of an earlier random split and scored on all 1,500 domains. They test whether the masked-minus-causal difference requires *learned* SAE directions; they do not force the two objectives onto one fitted dictionary.

C Implementation of L_{struct}

Active set. For each feature, condition, seed and depth, A_f contains residues whose activation strictly exceeds the 90th percentile of that feature's activations over every residue of the held-out partition, zeros included and with linear interpolation. Because TopK activations are zero-inflated the threshold is frequently exactly zero, in which case A_f is the non-zero support.

Neighbourhood and weighting. Neighbourhoods follow Eq. 1 with Cα distances. Residues without a qualifying neighbour receive a neighbour average of zero and are retained in both expectations of Eq. 2; they are 7,100 of the 21,235 held-out residues. Both expectations pool residues across domains, so domains are weighted by their number of residues.

This zero-fill is not neutral: the zeros enter the all-residue expectation, so a feature that merely avoids isolated residues could in principle score above zero without any spatial clustering. We tested it directly by recomputing L_{struct} with those residues excluded from both expectations and from active-set membership, holding the contact graph, the active-set threshold, the permutation draw and the denominator fixed, so that the two versions differ only in the zero-fill. Across all 126 cells L_{seq}, whose neighbourhood is $\{\pm1, \pm2\}$ and which therefore has almost no isolated residues, moved by exactly zero, which is the control on the procedure. On the native-corpus models mean L_{struct} fell slightly, from $+0.00519$ to $+0.00452$; on the order-destroyed models it rose from $+0.02031$ to $+0.02386$; and across the 18 random-initialisation cells it rose from $+0.01033$ to $+0.01209$, in 18 of 18. Across the 28 indicators it moved from $+0.0309$ to $+0.0322$, and the cysteine indicator rose from $+0.4067$ to $+0.4669$, remaining the highest-scoring indicator under both conventions. The gap between order-destroyed and native training therefore widened from $+0.01512$ to $+0.01934$ once the convention was removed. The rise reported in §3.2 is not an artefact of the missing-neighbour convention, and neither are the baselines these scores are compared against.

Permutation. Activations are permuted among residue positions independently within each protein, with one permutation shared by all features, drawn once per scoring run from a fixed seed. The contact graph, the standard deviations and the active-set size are unchanged by the permutation; the active set moves with the activations, so each active value is evaluated at a random position's neighbourhood within the same protein. Five permutations are averaged. The same permutations are used for both objectives, all seeds, all depths and both corpora within a partition.

Normalisation. The denominator is the standard deviation of the unpermuted raw activations of the feature over every residue of the held-out partition, zeros included, computed once and reused for the permuted contrasts.

Inactive features. Features with fewer than five active residues are assigned $L_{\mathrm{struct}} = 0$ and retained in dictionary means. Per held-out cell this affects at most 18 features under native training and at most 97 under order-destroyed training. Excluding them instead of assigning zero leaves the order-destroyed intervention at 52 of 54 cells with the same two exceptions, the model-level count at 6 of 6, and the seed-level rises at $+0.01013$ (masked) and $+0.02031$ (causal).

Sequential statistic. L_{seq} uses the same active set, contrast, permutation and normalisation with the neighbourhood $\{i-2, i-1, i+1, i+2\}$, dropping partners outside the protein.

D Statistical analysis

Effect size. Eq. 4 is computed over the 2,560 per-feature values of each objective, with the sample standard deviations of those two distributions. The bootstrap implementation rebuilds per-feature

values from per-protein contributions and does not apply the fewer-than-five rule, so its estimates differ from the dictionary-mean version at the two shallowest depths.

Bootstrap. 1,000 replicates. Each replicate draws the 23 held-out folds with replacement and then draws domains with replacement within each drawn fold, using one set of resampling indices shared by the masked and causal conditions and by all depths. Each replicate recomputes the re-weighted active-set and all-residue contrasts, the per-feature values, the per-objective means and standard deviations, and the effect size. Percentile thresholds, per-feature standard deviations, the contact graph and the permutations are computed once on the held-out partition and held fixed, so active-set membership is re-weighted rather than re-thresholded. Holding the thresholds and standard deviations fixed ignores their sampling variability, so the per-depth intervals are likely somewhat narrow. The model-level conclusions do not depend on them: the depth-averaged effect is tested with a t-test over seed summaries, and the order-destroyed intervention uses an exact sign test over models.

Reported estimates. Tables and figures report the bootstrap mean beside the 95% percentile interval. The bootstrap mean is closer to zero than the original-sample estimate in all 27 held-out depth-by-seed cells (0.71–0.82 times at blocks 7 and deeper), because resampling noise inflates the across-feature standard deviation in the denominator more than it moves the numerator. The two estimators are reported consistently rather than mixed.

Significance. Two-sided bootstrap p-values are $p = \mathrm{clip}\big(2\min(f, 1-f),\, 1/B,\, 1\big)$, where f is the fraction of replicates in which the effect is positive, with Benjamini–Hochberg correction across the nine depths within each seed. Seed-level tests use the three seed summaries with a two-sided one-sample t-test; the interval for the depth-averaged effect quoted in §3.1 is such a t-interval, not a bootstrap interval. With three seeds an exact sign-flip test cannot fall below $p = 0.25$, which is why the seed-level t-test is reported.

Corpus-control sign test. Each cell value is the mean L_{struct} over all 2,560 features; the rise is order-destroyed minus native; each of the six models is summarised by the mean rise over its nine depths, and the exact one-sided sign test on six positive summaries gives $p = 1/64 = 0.016$.

E Robustness of the masked–causal comparison

Depth-wise effects. Table 2 and Figure 3 report seed 42. BH-survival counts are 3, 5 and 6 of 9 for seeds 42, 43 and 44.

Table 2: Depth-wise masked-minus-causal effects on the held-out partition, seed 42. Estimates are bootstrap means; intervals are fold-clustered percentile intervals; q values are Benjamini–Hochberg adjusted across the nine depths.

Relative depth	d (bootstrap mean)	95% CI	q_{BH}	Supported
0%	$+0.0177$	$[-0.033, +0.064]$	0.454	no
13%	$+0.0460$	$[-0.029, +0.113]$	0.2649	no
25%	$+0.1211$	$[+0.058, +0.183]$	0.006	yes
38%	$+0.1806$	$[+0.099, +0.262]$	0.0045	yes
50%	$+0.1936$	$[+0.110, +0.273]$	0.0045	yes
63%	$+0.1438$	$[-0.010, +0.271]$	0.1395	no
75%	$+0.1177$	$[-0.139, +0.322]$	0.454	no
88%	$+0.1901$	$[-0.096, +0.424]$	0.2649	no
100%	$+0.2376$	$[-0.034, +0.464]$	0.144	no

Prediction-matched analysis. All nine comparisons at blocks 11, 14 and 18 have intervals excluding zero and survive correction; across the full grid 18 of 27 do (Table 3). Per-seed depth-averaged effects are $+0.1726$, $+0.1329$ and $+0.1407$.

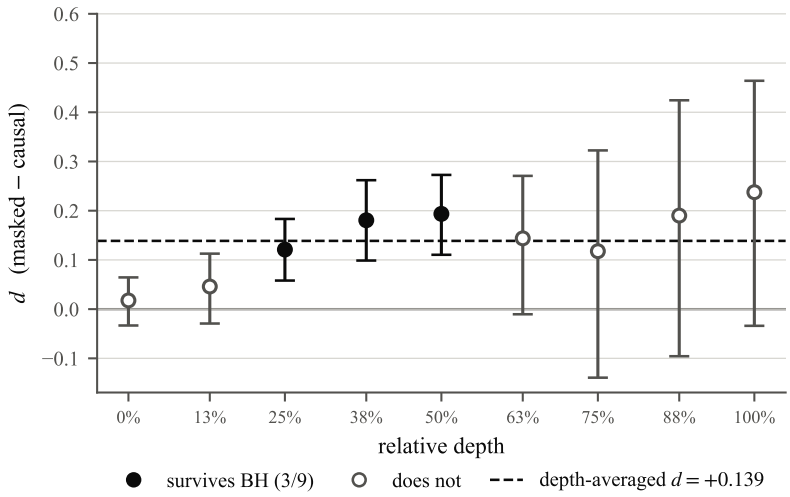


Figure 3: Depth-wise masked-minus-causal effect d at seed 42 on the held-out partition. Points are bootstrap means with fold-clustered 95% percentile intervals; filled points survive Benjamini–Hochberg correction across the nine depths (3 of 9). The dashed line is the seed-42 depth-averaged effect ($d = +0.139$). Values are in Table 2.

Dictionary seeds. Three dictionaries per condition at six objective-by-depth cells (Table 4): the median dictionary-seed standard deviation is 0.00060 against 0.00102 across model seeds, and both are far below $\Delta = +0.01460$ at blocks 11, 14 and 18.

Frozen random dictionaries. With the three frozen random dictionaries of Appendix B, the raw masked-minus-causal difference remains positive in all nine variant-by-block cells on the full evaluation set: $+0.0210$, $+0.0205$, $+0.0057$ (frozen decoder), $+0.0218$, $+0.0241$, $+0.0063$ (frozen encoder) and $+0.0196$, $+0.0205$, $+0.0037$ (soft-frozen) at blocks 7, 14 and 22, against $+0.0146$, $+0.0166$ and $+0.0073$ with independently fitted dictionaries.

Dictionary-free residual dimensions. Treating each of the 320 residual-stream coordinates as a feature leaves the difference positive at all three control depths on the held-out partition at seed 42: $+0.01182$, $+0.01174$ and $+0.00688$ at blocks 11, 14 and 18. Under the order-destroyed intervention the same analysis rises in three of three causal cells and two of three masked cells, so the SAE is not necessary for the effect to appear.

Reconstruction quality. Across the 27 depth-by-seed cells, the Spearman correlation between the masked-minus-causal gap in validation explained variance and the masked-minus-causal L_{struct} effect is $\rho = +0.347$ ($p = 0.076$); excluding blocks 0 and 4 it is $\rho = -0.096$ ($p = 0.68$). The association is therefore carried by the two shallowest blocks. Separately, order-destroyed dictionaries reconstruct their inputs *better* than native ones in all 54 held-out cells, and the size of that gap correlates with the rise ($\rho = 0.63$), which is discussed as a limitation in §4.

Amino-acid selectivity. For feature f and amino acid a, let $r_{f,a}$ be the mean absolute activation of f over all evaluation residues of type a, and $p_{f,a} = r_{f,a} / \sum_b r_{f,b}$. The selectivity score is $S(f) = 1 - H(p_f)/\log 20$. It uses all residues, not the active set, and the division by the per-type mean is its only background correction. On native models the median Spearman correlation between $S(f)$ and L_{struct}(f) over the nine held-out cells is $+0.047$ in the masked models (range $[-0.019, +0.099]$) and $+0.080$ in the causal models (range $[+0.027, +0.091]$); with L_{seq} it is -0.745 and -0.396. On order-destroyed models the corresponding medians are -0.011 and -0.063 against L_{struct}, and -0.476 and -0.414 against L_{seq}. Composition is therefore strongly associated with L_{seq} and weakly with L_{struct}.

Table 3: Prediction-matched masked models against the causal models, held-out partition. A comparison is supported when the percentile interval excludes zero and the BH-adjusted q is at most 0.05.

Seed	Relative depth	d (bootstrap mean)	95% CI	q_{BH}	Supported
42	0%	$+0.0163$	$[-0.032, +0.063]$	0.5783	no
42	13%	$+0.0382$	$[-0.016, +0.106]$	0.27	no
42	25%	$+0.1318$	$[+0.069, +0.188]$	0.0045	yes
42	38%	$+0.2324$	$[+0.111, +0.335]$	0.0045	yes
42	50%	$+0.3254$	$[+0.229, +0.414]$	0.0045	yes
42	63%	$+0.1929$	$[+0.046, +0.318]$	0.0108	yes
42	75%	$+0.0716$	$[-0.209, +0.312]$	0.676	no
42	88%	$+0.2325$	$[+0.037, +0.431]$	0.021	yes
42	100%	$+0.3119$	$[+0.134, +0.463]$	0.0045	yes
43	0%	$+0.0387$	$[+0.007, +0.064]$	0.03	yes
43	13%	$+0.0190$	$[-0.022, +0.062]$	0.404	no
43	25%	$+0.1213$	$[+0.078, +0.170]$	0.00225	yes
43	38%	$+0.1935$	$[+0.116, +0.272]$	0.00225	yes
43	50%	$+0.3155$	$[+0.196, +0.414]$	0.00225	yes
43	63%	$+0.1193$	$[+0.035, +0.203]$	0.0144	yes
43	75%	$+0.0622$	$[-0.061, +0.169]$	0.3983	no
43	88%	$+0.0886$	$[-0.004, +0.168]$	0.07971	no
43	100%	$+0.2379$	$[+0.136, +0.339]$	0.00225	yes
44	0%	-0.0028	$[-0.062, +0.038]$	0.94	no
44	13%	$+0.0087$	$[-0.044, +0.052]$	0.8167	no
44	25%	$+0.0669$	$[+0.025, +0.104]$	0.003	yes
44	38%	$+0.2193$	$[+0.102, +0.314]$	0.0018	yes
44	50%	$+0.2074$	$[+0.106, +0.290]$	0.0018	yes
44	63%	$+0.1481$	$[+0.051, +0.239]$	0.0018	yes
44	75%	$+0.1149$	$[-0.103, +0.293]$	0.4449	no
44	88%	$+0.2160$	$[+0.090, +0.316]$	0.0018	yes
44	100%	$+0.2882$	$[+0.193, +0.368]$	0.0018	yes

Table 4: Mean L_{struct} under three dictionary seeds, seed-42 models, held-out partition.

Objective	Block	Dict. 42	Dict. 43	Dict. 44	Dict.-seed SD	Model-seed SD
masked	11	$+0.01153$	$+0.00963$	$+0.01061$	0.00095	0.00114
masked	14	$+0.00820$	$+0.00889$	$+0.00992$	0.00087	0.00147
masked	18	$+0.00450$	$+0.00553$	$+0.00500$	0.00051	0.00161
causal	11	-0.00489	-0.00353	-0.00398	0.00069	0.00090
causal	14	-0.00613	-0.00543	-0.00564	0.00036	0.00052
causal	18	-0.00408	-0.00394	-0.00399	0.00007	0.00030

Directional interaction. Partitioning the neighbourhood into downstream ($j > i$) and upstream ($j < i$) contacts and taking the masked-minus-causal difference in differences gives a mean of -0.0855 (SD 0.0387, range $[-0.1544, -0.0511]$), negative in all nine held-out cells, with $t(2) = -6.26$, $p = 0.025$ over seed means. A sequential-graph control gives $+0.0241$, positive in five of nine cells. A positive interaction was predicted if the masked advantage arose from access to later-in-sequence contact partners; the observed sign is the opposite.

F TRAINING-DISTRIBUTION CONTROLS

Complete order-destroyed grid. Tables 5 and 6 give all 54 matched cells.

Language-model performance on the order-destroyed corpus. Relative to their native-corpus counterparts, causal held-out perplexity increased by 10.6% and masked pseudo-perplexity by 11.1%. These are objective-specific quantities and their absolute values are not comparable, but

Table 5: Causal objective: mean L_{struct} under native and order-destroyed training, held-out partition.

Seed	Block	Native	Order-destroyed	Difference
42	0	$+0.00918$	$+0.01308$	$+0.00390$
42	4	$+0.00373$	$+0.01416$	$+0.01043$
42	7	-0.00278	$+0.01806$	$+0.02084$
42	11	-0.00489	$+0.02055$	$+0.02544$
42	14	-0.00613	$+0.02454$	$+0.03067$
42	18	-0.00408	$+0.02604$	$+0.03012$
42	22	-0.00067	$+0.02036$	$+0.02103$
42	26	-0.00240	$+0.01798$	$+0.02038$
42	29	-0.00282	$+0.01437$	$+0.01719$
43	0	$+0.01087$	$+0.01813$	$+0.00726$
43	4	$+0.00832$	$+0.02437$	$+0.01606$
43	7	-0.00033	$+0.02541$	$+0.02574$
43	11	-0.00485	$+0.02398$	$+0.02882$
43	14	-0.00621	$+0.02659$	$+0.03280$
43	18	-0.00467	$+0.02212$	$+0.02679$
43	22	-0.00449	$+0.02073$	$+0.02522$
43	26	-0.00068	$+0.01849$	$+0.01917$
43	29	-0.00021	$+0.02236$	$+0.02257$
44	0	$+0.01590$	$+0.02189$	$+0.00599$
44	4	$+0.01289$	$+0.01506$	$+0.00217$
44	7	$+0.00040$	$+0.02082$	$+0.02042$
44	11	-0.00643	$+0.02410$	$+0.03052$
44	14	-0.00707	$+0.02464$	$+0.03171$
44	18	-0.00429	$+0.01791$	$+0.02219$
44	22	-0.00170	$+0.01331$	$+0.01501$
44	26	-0.00248	$+0.01481$	$+0.01729$
44	29	-0.00305	$+0.01496$	$+0.01801$

the similar relative degradation gives no indication that one objective adapted substantially better to the order-destroyed corpus.

Block-shuffled intervention. The block-shuffled models were trained at the same three seeds and both objectives, under the same initial-parameter, batch-order and budget matching as the order-destroyed models, and scored at blocks 11, 14 and 18, giving 18 cells (Table 7): L_{struct} increased in all nine causal cells and six of nine masked cells.

Extended-budget comparison. Table 8 reports seed 42 at both budgets on the held-out partition. Order-destroyed cells change by -20.3% to $+71.3\%$ and native cells by -97.8% to $+178.7\%$; order-destroyed remains above native in all six cells at both budgets. The causal native values are close to zero, so their ratios are unstable, and the condition has one model seed, so this is a descriptive sensitivity analysis rather than a replicated estimate.

Sequential-locality comparator. Table 9 gives L_{seq} at the five deepest blocks. On the masked models the order-destroyed value is below native in all 15 cells; in the causal models it is above native in all 15.

Fixed-size active sets. Because the active set is defined by a percentile of a zero-inflated distribution, its size varies between conditions. Repeating the order-destroyed intervention with active sets of fixed size (the top 1%, 2.5% and 5% of residues per feature) leaves the direction unchanged (Table 10). The fixed-size rules exclude features that cannot reach the required count, and exclude more of them under order-destroyed training, so they are not a size-only intervention.

Table 6: Masked objective: mean L_{struct} under native and order-destroyed training, held-out partition. The two negative differences are the only exceptions among the 54 cells.

Seed	Block	Native	Order-destroyed	Difference
42	0	+0.01175	+0.01737	+0.00562
42	4	+0.01037	+0.01996	+0.00959
42	7	+0.01235	+0.02488	+0.01253
42	11	+0.01153	+0.02244	+0.01090
42	14	+0.00820	+0.02135	+0.01315
42	18	+0.00450	+0.02434	+0.01984
42	22	+0.00547	+0.01844	+0.01297
42	26	+0.00811	+0.01958	+0.01147
42	29	+0.01043	+0.02342	+0.01299
43	0	+0.01952	+0.01900	−0.00052
43	4	+0.01576	+0.02202	+0.00626
43	7	+0.01394	+0.01556	+0.00162
43	11	+0.01313	+0.01505	+0.00191
43	14	+0.01032	+0.02421	+0.01389
43	18	+0.00766	+0.01324	+0.00558
43	22	+0.00195	+0.01593	+0.01398
43	26	+0.00585	+0.01750	+0.01165
43	29	+0.00864	+0.01601	+0.00737
44	0	+0.02228	+0.01370	−0.00858
44	4	+0.01272	+0.01689	+0.00417
44	7	+0.01632	+0.02179	+0.00547
44	11	+0.01092	+0.02432	+0.01340
44	14	+0.01102	+0.02250	+0.01147
44	18	+0.00554	+0.03402	+0.02848
44	22	+0.00551	+0.03521	+0.02971
44	26	+0.01148	+0.01498	+0.00349
44	29	+0.01372	+0.02404	+0.01032

Table 7: Native against 16-residue block-shuffled training, held-out partition.

Objective	Seed	Block	Native	Block-shuffled	Difference
causal	42	11	−0.00489	+0.00291	+0.00780
causal	42	14	−0.00613	+0.00236	+0.00849
causal	42	18	−0.00408	+0.00285	+0.00693
causal	43	11	−0.00485	+0.00468	+0.00953
causal	43	14	−0.00621	+0.00285	+0.00906
causal	43	18	−0.00467	+0.00339	+0.00806
causal	44	11	−0.00643	+0.00378	+0.01021
causal	44	14	−0.00707	+0.00350	+0.01057
causal	44	18	−0.00429	+0.00255	+0.00684
masked	42	11	+0.01153	+0.01135	−0.00018
masked	42	14	+0.00820	+0.01078	+0.00258
masked	42	18	+0.00450	+0.01312	+0.00862
masked	43	11	+0.01313	+0.01462	+0.00149
masked	43	14	+0.01032	+0.01504	+0.00473
masked	43	18	+0.00766	+0.01415	+0.00649
masked	44	11	+0.01092	+0.01061	−0.00031
masked	44	14	+0.01102	+0.00995	−0.00108
masked	44	18	+0.00554	+0.01105	+0.00551

G RANDOM-INITIALISATION AND NO-MODEL BASELINES

Random-initialisation baseline. Each draw seeds the network weights before construction and the dictionary with the same seed, so the three draws are independent complete pipelines; within

Table 8: L_{struct} at the primary and extended budgets, seed 42, held-out partition.

Corpus	Objective	Block	Primary budget	Extended budget	Change
Native	Masked	11	+0.01153	+0.02068	+79.3%
Native	Masked	14	+0.00820	+0.02286	+178.7%
Native	Masked	18	+0.00450	+0.00576	+28.0%
Native	Causal	11	−0.00489	−0.00480	+1.7%
Native	Causal	14	−0.00613	−0.00602	+1.8%
Native	Causal	18	−0.00408	−0.00807	−97.8%
Order-destroyed	Masked	11	+0.02244	+0.03843	+71.3%
Order-destroyed	Masked	14	+0.02135	+0.03189	+49.4%
Order-destroyed	Masked	18	+0.02434	+0.03408	+40.0%
Order-destroyed	Causal	11	+0.02055	+0.01638	−20.3%
Order-destroyed	Causal	14	+0.02454	+0.02316	−5.6%
Order-destroyed	Causal	18	+0.02604	+0.02915	+11.9%

Table 9: Native/order-destroyed L_{seq} per cell, held-out partition. The final column is the range of order-destroyed to native ratios across the three seeds.

Objective	Block	Seed 42	Seed 43	Seed 44	Ratio range
Masked	14	0.2369/0.0641	0.1141/0.0569	0.1676/0.0583	27–50%
Masked	18	0.3482/0.0549	0.2512/0.0518	0.3673/0.0544	15–21%
Masked	22	0.2994/0.0446	0.3557/0.0570	0.3214/0.0448	14–16%
Masked	26	0.2756/0.0435	0.2803/0.0402	0.3096/0.0443	14–16%
Masked	29	0.2641/0.0607	0.2554/0.0451	0.2781/0.0482	17–23%
Causal	14	0.3576/0.6088	0.3845/0.5692	0.3986/0.5959	148–170%
Causal	18	0.3638/0.6975	0.3678/0.6762	0.3878/0.6943	179–192%
Causal	22	0.3898/0.7213	0.3684/0.7019	0.3988/0.7296	183–191%
Causal	26	0.4020/0.7402	0.3821/0.7311	0.4101/0.7272	177–191%
Causal	29	0.4145/0.7529	0.4010/0.7386	0.4213/0.7460	177–184%

a draw the masked and causal networks share initial weights and differ only in the attention mask, mirroring the trained pairs. Table 11 gives per-cell values.

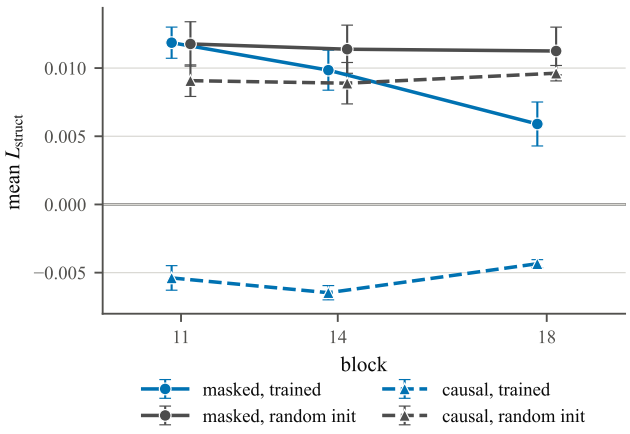


Figure 4: Mean L_{struct} at blocks 11, 14 and 18 for trained models (blue) and randomly initialised networks (grey), masked (solid, circles) and causal (dashed, triangles). Error bars are standard deviations over three model seeds (trained) or three random-initialisation draws. Values are in Tables 1 and 11.

Table 10: Corpus control under the default and fixed-size active-set rules, held-out partition. Cells are objective-by-depth-by-seed combinations in which order-destroyed exceeds native; models are per-model means over depths.

Rule	All cells	Causal	Masked	Models	Features excluded (max)
default	52/54	27/27	25/27	6/6	0
top0.01	49/54	24/27	25/27	6/6	461
top0.025	49/54	27/27	22/27	6/6	674
top0.05	48/54	27/27	21/27	6/6	904

Table 11: Mean L_{struct} for the three random-initialisation draws, held-out partition.

Objective	Block	Draw 42	Draw 43	Draw 44	Mean $\pm$ SD
Masked	11	$+0.00990$	$+0.01250$	$+0.01290$	$+0.01177 \pm 0.00163$
Masked	14	$+0.00940$	$+0.01279$	$+0.01195$	$+0.01138 \pm 0.00177$
Masked	18	$+0.00987$	$+0.01322$	$+0.01066$	$+0.01125 \pm 0.00175$
Causal	11	$+0.00775$	$+0.00962$	$+0.00987$	$+0.00908 \pm 0.00116$
Causal	14	$+0.00716$	$+0.00950$	$+0.01001$	$+0.00889 \pm 0.00152$
Causal	18	$+0.00904$	$+0.01017$	$+0.00965$	$+0.00962 \pm 0.00057$

No-model indicator baseline. The no-model indicator layer contains 29 indicators: one per amino acid, eight biochemical groupings and one all-residue column, each scored through the same neighbourhood, permutation and normalisation procedure as the learned features. The active-set rule differs: with a threshold at the 90th percentile, any indicator covering more than 10% of residues would have an empty active set, so indicators are scored with the active set taken as the top 60% of residues, a threshold at the 40th percentile. No scored indicator covers more than 47.6% of residues, so for every scored indicator the threshold is zero and the active set is exactly the residues where the indicator equals one. The all-residue column cannot be scored, because its active set cannot be formed, leaving 28 scored indicators (Table 12). The cysteine indicator reaches $L_{\mathrm{struct}} = +0.4067$. At seed 42 and blocks 11, 14 and 18 the best learned feature, rescored under the same top-60% rule, scores 0.1178, 0.0864 and 0.1241 on the causal models and 1.0166, 0.5905 and 0.2206 in the masked models, so the indicator exceeds every learned feature in four of those six conditions, and in 32 of the 54 objective-by-depth-by-seed cells. The number of indicators exceeding all 2,560 learned features at seed 42 is 4, 6 and 4 of 28 in the causal models at blocks 11, 14 and 18.

Biochemical groupings. Hydrophobic (A, V, L, I, M, F, W, C); charged (D, E, K, R); polar (S, T, N, Q, Y, H); aromatic (F, W, Y, H); small (A, G, S, T, C, V, P, N, D); tiny (A, G, S); negative (D, E); positive (K, R).

H Sequence-separation and normalisation analyses

Separation sweep. The minimum sequence separation was varied over 16 values at blocks 11, 14 and 18 with a fixed 6 Å contact radius, at three model seeds (Table 13). The seed-mean effect crosses zero at separations of five, five and four residues, and the sweep maximum lies at separations of 16 or more at every block rather than at the paper's choice of 12.

Across three blocks, 16 separations and two objectives, order-destroyed L_{struct} exceeds native L_{struct} in 79 of 96 cells: 42 of 48 in the causal models, and 16, 11 and 10 of 16 in the masked models at blocks 11, 14 and 18. The 11 masked exceptions lie at separations of six residues or fewer, at blocks 14 and 18. The six causal exceptions lie at the longest separations tested, 24 and 30 (Figure 5).

Alternative normalisations. Four denominators were compared: the evaluation-wide standard deviation of raw activations (the primary statistic), a fixed denominator of one, the interquartile range

Table 12: All 28 scored indicator features, held-out partition. Occupancy is the fraction of residues at which the indicator equals one. Indicator scores do not depend on the model.

Indicator	Occupancy	L_{struct}
`aa:C`	0.0139	$+0.4067$
`class:hydrophobic`	0.3972	$+0.1925$
`aa:I`	0.0564	$+0.1332$
`aa:L`	0.1050	$+0.1311$
`aa:V`	0.0692	$+0.1140$
`aa:F`	0.0393	$+0.0994$
`class:aromatic`	0.1107	$+0.0810$
`aa:W`	0.0141	$+0.0374$
`aa:M`	0.0216	$+0.0372$
`aa:T`	0.0526	$+0.0252$
`aa:G`	0.0654	$+0.0235$
`aa:N`	0.0384	$+0.0191$
`aa:A`	0.0777	$+0.0117$
`aa:Y`	0.0335	$+0.0109$
`class:small`	0.4756	$+0.0094$
`class:polar`	0.2478	-0.0004
`class:tiny`	0.2048	-0.0047
`aa:H`	0.0238	-0.0120
`aa:D`	0.0524	-0.0174
`aa:P`	0.0444	-0.0220
`aa:S`	0.0617	-0.0263
`class:charged`	0.2452	-0.0263
`aa:E`	0.0736	-0.0460
`aa:R`	0.0562	-0.0484
`aa:Q`	0.0378	-0.0526
`class:positive`	0.1193	-0.0622
`class:negative`	0.1260	-0.0635
`aa:K`	0.0631	-0.0848

Table 13: Masked-minus-causal standardised effect against minimum sequence separation, seed means, held-out partition, 6 Å contact radius.

Minimum separation	Block 11	Block 14	Block 18
1	-1.9322	-1.2803	-0.2523
2	-1.2075	-0.8675	-0.3006
3	-0.6593	-0.5092	-0.1966
4	-0.1855	-0.0633	$+0.0375$
5	$+0.0552$	$+0.1530$	$+0.1599$
6	$+0.0607$	$+0.1388$	$+0.1316$
7	$+0.0989$	$+0.1601$	$+0.1282$
8	$+0.1032$	$+0.1512$	$+0.1032$
9	$+0.1160$	$+0.1676$	$+0.1137$
10	$+0.1318$	$+0.1772$	$+0.1254$
12	$+0.1444$	$+0.1895$	$+0.1282$
14	$+0.1718$	$+0.2133$	$+0.1411$
16	$+0.1908$	$+0.2325$	$+0.1636$
20	$+0.1983$	$+0.2319$	$+0.1649$
24	$+0.1955$	$+0.2120$	$+0.1198$
30	$+0.2412$	$+0.2283$	$+0.1797$

of the strictly positive activations, and a rank transformation in which each feature's activations are ranked over all residues with average ties and scaled to $[0, 1]$ before the estimator is re-run. The native-to-order-destroyed increase holds in all 18 cells under the standard deviation, the fixed denominator and the interquartile range, and in 17 of 18 under the rank transformation (Figure 6).

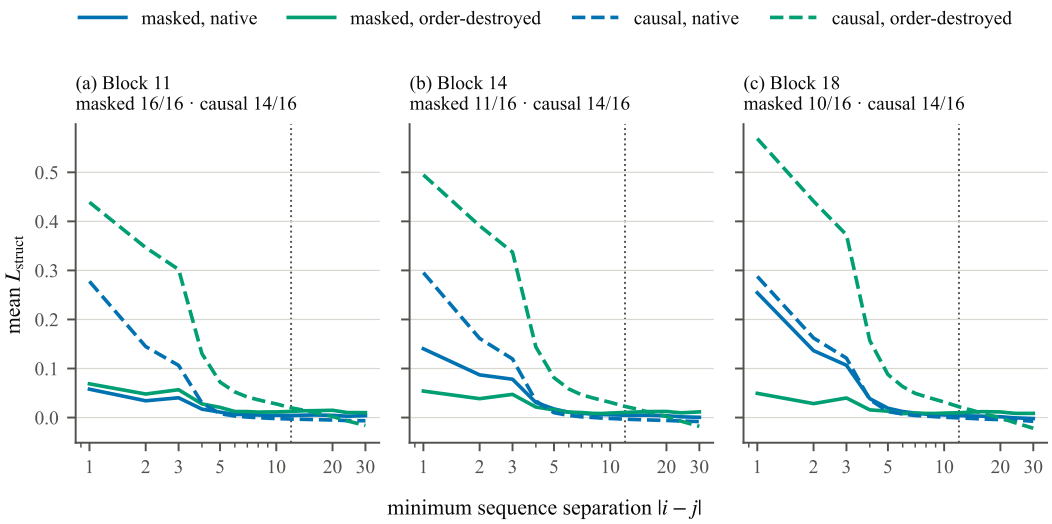


Figure 5: Mean L_{struct} against minimum sequence separation at a 6 Å contact radius, seed means on the held-out partition: native (blue) and order-destroyed (green) training, masked (solid) and causal (dashed). The dotted line marks the primary separation of 12. Panel titles give the number of the 16 separations at which order-destroyed exceeds native.

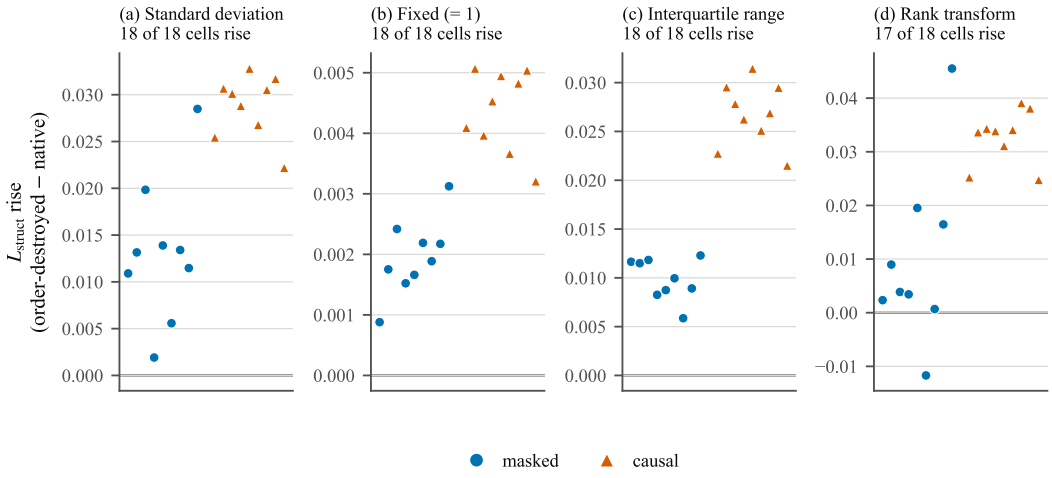


Figure 6: Rise in L_{struct} from native to order-destroyed training under four denominators, 18 objective-by-depth-by-seed cells per panel (blue circles masked, orange triangles causal): (a) the standard deviation of raw activations, used throughout; (b) a fixed denominator of one; (c) the interquartile range of the positive activations; (d) a rank transformation. Every cell rises except one masked cell under the rank transformation.

Relationship to sequential locality. At 6 Å without a separation floor, all adjacent residue pairs qualify as structural neighbours and pairs with $|i - j| \leq 3$ account for 73.8% of admitted edges, so the neighbourhood is dominated by sequentially local pairs. That composition is a property of protein backbones rather than of the sample: it is 41,837 of 56,669 edges on the held-out partition and 596,357 of 807,547 on the full evaluation set, both 73.8%, with 21,005 of 21,006 adjacent held-out pairs inside 6 Å.

Without the separation floor the nominally structural comparison closely tracks the sequential one. On the held-out partition the masked-minus-causal effects at blocks 11, 14 and 18 are -1.9322, -1.2803 and -0.2523 for the no-floor structural statistic against -2.0586, -1.3385 and -0.2929 for L_{seq} (seed means, native models). The corresponding full-set pair at block 11 is -2.2096 against -2.2663.

I Reference structural readouts

Labels. Helix labels are DSSP states H, G and I; strand labels are E and B. Residues whose DSSP state is polyproline II are excluded from the secondary-structure probes (4,482 fitting and 421 held-out residues). Burial is defined by a residue neighbour count above the median of the fitting partition. The neighbour count of a residue is the number of residues whose $\mathrm{C}\alpha$ lies within 10 Å of its own, and the fitting-partition median is 17. Recomputing the count from the deposited structures reproduces the cached labels for 99.6% of the 292,533 residues and gives the same median; the residues that differ are ones whose $\mathrm{C}\alpha$ the sequence-to-structure alignment does not place.

Probe architectures. Representations enter the probes unstandardised. The linear probe is L2-regularised logistic regression ($C = 1.0$, class-balanced). The nonlinear probe is a single-hidden-layer network of 256 ReLU units trained with Adam and early stopping on a random 10% of fitting residues, with the majority class downsampled; it uses no dropout. Contact precision at $L/5$ is not a trained probe: it ranks residue pairs by the average-product-corrected cosine similarity of their representations and scores the top $0.2L$ pairs with $|i - j| \geq 12$.

Partitions. Probes are fitted once on the 1,349 fitting domains and scored once on the 151 held-out domains; there is no cross-validation. Intervals are 1,000 two-stage fold-clustered bootstrap resamples of the held-out proteins with the fitted probe held fixed, paired across objectives.

Residue-level results. Table 14 gives masked-minus-causal differences averaged over three model seeds. At blocks 11 and 14 the helix and strand probes favour the causal models in all 12 label-by-seed comparisons, and every seed-level interval excludes zero.

Table 14: Masked-minus-causal probe differences, held-out partition, mean of three model seeds. Contact precision at $L/5$ is an unsupervised readout rather than a probe; it is listed in the linear rows only for compactness and has no nonlinear counterpart.

Probe	Block	Helix F1	Strand F1	Burial F1	Contact P@$L/5$
linear	11	-0.0370	-0.0769	-0.0355	$+0.0057$
linear	14	-0.0363	-0.0742	-0.0274	-0.0029
linear	18	$+0.0044$	-0.0305	-0.0076	-0.0137
mlp	11	-0.0493	-0.0569	-0.0206	—
mlp	14	-0.0280	-0.0726	-0.0091	—
mlp	18	$+0.0168$	-0.0185	$+0.0182$	—

Probes on trained and randomly initialised networks. Across objectives and blocks, AUROC for randomly initialised networks runs 0.654–0.693 (helix), 0.650–0.695 (strand) and 0.704–0.723 (burial), against trained ranges of 0.694–0.794, 0.689–0.809 and 0.725–0.768. The causal models clear the random-initialisation range at every block; the masked models at block 11 do not (strand F1 0.4624 trained against 0.4655 randomly initialised).

Pairwise contact probes. Positive pairs are contacts with $|i - j| \geq 12$ and Cα distance below 8 Å; each is matched with one negative pair from the same protein at the same sequence separation. Features are the elementwise absolute difference and product of the two residue representations, standardised, with L2-regularised logistic regression. The probe is fitted on 80% of the held-out domains (31,553 pairs) and tested on the remaining 20% (8,445 pairs), and the control permutes the labels. This split is by protein but not by fold, so unlike every other comparison in this paper the probe's fitting and test proteins can share a fold; the shuffled-label control sits at chance, but the absolute values should be read as optimistic. Both objectives use the same split, so the comparison between them is unaffected. Table 15 gives all six comparisons.

Table 15: Separation-matched pairwise contact AUROC on the held-out partition, seed 42.

Objective	Block	Raw stream	SAE latents	Raw − SAE	Shuffled labels (raw / SAE)
causal	11	0.5841	0.5571	+0.0270	0.497 / 0.501
causal	14	0.5786	0.5533	+0.0253	0.495 / 0.507
causal	18	0.5858	0.5655	+0.0203	0.501 / 0.489
masked	11	0.5702	0.5251	+0.0451	0.500 / 0.501
masked	14	0.5772	0.5458	+0.0314	0.507 / 0.502
masked	18	0.5940	0.5673	+0.0267	0.498 / 0.497

J Evaluation-set design

The dictionary-fitting and scoring partitions are disjoint at the level of complete SCOPe folds (§2.3). The held-out partition is small and unevenly composed: 151 domains in 23 folds, of which four folds hold 116 domains, with a mean domain length of 141 residues against 196 for the full set. Comparisons made on the same held-out domains are unaffected, but absolute levels and effect sizes reflect that sample.

The following analyses are reported on the full 1,500-domain evaluation set, because their protocols require more domains or more dictionary refits than the held-out partition allows: the frozen random dictionaries (Appendix B), and the metric audits of Appendix K other than the co-activation–decoder-geometry correlation, which is computed on the held-out partition like L_{struct} itself.

K Additional metric audits

The training-distribution, random-initialisation and no-model controls test different requirements, and their applicability depends on each metric's definition. We therefore describe the observed response to each control rather than imposing a universal pass-or-fail label.

Table 16 gives every metric-by-check comparison. Each entry is the mean difference between the intervened condition and its matched native cell, over three model seeds at blocks 11, 14 and 18, tested with the seed-level t-test of §2.6. A negative difference is the direction the check intends: the intervention lowers the score, the randomly initialised network scores below the trained one, or the no-model layer scores below both. A positive difference indicates that this prespecified directional criterion was not met. An exact two-sided sign test on three seeds cannot fall below $p = 0.25$ (Appendix D), so a sign-test summary of this grid returns no significant cell for any metric; that is a property of the test, not of the data.

d_{struct}. d_{struct} is our name for InterPLM's structural Cohen's d, which identifies each protein's maximum-activation residue for a feature, averages activation within 6 Å of it with no separation floor, and compares the result with five within-protein permutations, requiring at least 25 qualifying proteins and sampling at most 100. The released repository does not implement it, so we reimplemented it from the published Methods (Simon & Zou, 2025). Simon & Zou (2025) use d_{struct} to select features for subsequent interpretation rather than to compare models, and apply their weight-shuffled model baseline to their concept-F1 analysis rather than to structural clustering, which uses a within-protein permutation null. What follows therefore tests the interpretation of the statistic

Table 16: Every metric-by-check comparison on the full evaluation set, blocks 11, 14 and 18, three model seeds. Δ is the mean difference from the matched native cell; negative is the prespecified direction for the check, positive indicates that the directional criterion was not met. Bold marks $p < 0.05$ from a two-sided seed-level t-test on three seed summaries. The no-model cell for ρ does not exist because the indicator layer has no decoder.

Metric	Check	Masked		Causal	
		Δ	p	Δ	p
d_{struct}	order-destroyed	-0.212	$\mathbf{0.008}$	$+0.158$	$\mathbf{0.008}$
	no-model	-0.354	$\mathbf{0.004}$	-0.774	$< \mathbf{0.001}$
	random init	-0.191	$\mathbf{0.016}$	-0.310	$< \mathbf{0.001}$
Concept-F1	order-destroyed	$+0.010$	0.08	-0.057	$\mathbf{0.003}$
	no-model	-0.069	$\mathbf{0.006}$	-0.006	0.24
	random init	-0.001	0.65	-0.090	$\mathbf{0.002}$
Single-latent F1	order-destroyed	$+0.009$	0.14	$+0.052$	$\mathbf{0.007}$
	no-model	$+0.048$	$\mathbf{0.015}$	$+0.105$	$\mathbf{0.001}$
	random init	-0.023	$\mathbf{0.027}$	$+0.012$	0.08
ρ	order-destroyed	-0.062	$\mathbf{0.002}$	-0.240	$\mathbf{0.001}$
	no-model	—	—	—	—
	random init	$+0.062$	$\mathbf{0.004}$	-0.087	$\mathbf{0.023}$
SAEBench	order-destroyed	-0.013	0.17	$+0.012$	0.26
	no-model	-0.058	$\mathbf{0.009}$	-0.041	$\mathbf{0.029}$
	random init	-0.020	0.11	-0.028	$\mathbf{0.037}$

under a model comparison its authors did not propose, and the random-initialisation result below has no counterpart in the original. As in InterPLM, the Bonferroni correction is applied across the features tested. The procedure gates features at 0.6 of a maximum activation. InterPLM fixes that maximum as each feature's maximum over a scan of 50,000 Swiss-Prot proteins, and the released normalisation code computes the same dataset-wide per-feature maximum; we use the maximum over our evaluation set. Under the dataset-wide gate, with three model seeds, mean d_{struct} rises in the causal models from 0.572 under native training to 0.731 under order-destroyed training, and falls in the masked models from 0.153 to -0.060. The no-model indicator layer scores -0.202, below both objectives in all 18 cells (Table 16); five of its 28 scored features are significant two-sided at $p < 0.05/28$, and all five are negative.

Concept-F1. Concept-F1 adapts the domain-adjusted F1 of Simon & Zou (2025), with SCOPe taxonomy and residue-level labels in place of their Swiss-Prot annotations; as in their protocol the feature is selected on a validation half and scored on a test half, and we additionally fix the activation threshold on the validation half. The concept set contains 74 SCOPe taxonomy labels and six residue-level labels, each with at least ten domains, and the aggregate is the mean test F1 over concepts. On the audit grid of blocks 11, 14 and 18 at three model seeds, the aggregate falls under order-destroyed training in the causal models (-0.057, $p = 0.003$) but not in the masked models ($+0.010$, $p = 0.08$), so the response is objective-dependent rather than consistent (Table 16). The no-model comparison for this metric is scored over the 12 concepts the indicator layer supports, against 80 for the other two checks, so its column is not on a common denominator with them. Its decomposition shows where it departs from the residue-level probes. Its six residue-level concepts favour the masked models at every usable depth at seed 42, whereas the probes of Appendix I favour the causal models. The 74 taxonomy concepts favour the causal models at the deeper blocks and dominate the unweighted mean. The two analyses weight different classes of structural information and are not direct replications of one another.

Co-activation and decoder geometry. This statistic is the Spearman correlation between pairwise feature co-activation, quantified by the Jaccard overlap of the binarised active sets, and the cosine similarity of the corresponding decoder vectors, over 400,000 sampled feature pairs. Li et al. (2025) compare co-occurrence and decoder-cosine clusterings by adjusted mutual information and prefer

the phi coefficient to the Jaccard overlap; the per-pair Spearman correlation used here is our simplification. Native exceeds order-destroyed in all 18 matched held-out cells, and in all six objective-by-seed summaries (exact one-sided sign test, $p = 0.016$). Against the random-initialisation baseline the metric splits by objective: trained exceeds randomly initialised in all nine causal cells and in none of the nine masked cells (Table 17). Both directions replicate on the full evaluation set (Table 16), which is a separate population from the held-out cells reported here and is not merged with them.

Table 17: Co-activation/decoder-geometry correlation under native training, order-destroyed training and random initialisation, held-out partition.

Objective	Block	Seed	Native ρ	Order-destroyed ρ	Random-init ρ
masked	11	42	0.0713	0.0373	0.1397
masked	11	43	0.0764	0.0363	0.1459
masked	11	44	0.0726	0.0361	0.1460
masked	14	42	0.0941	0.0421	0.1562
masked	14	43	0.0927	0.0410	0.1639
masked	14	44	0.0977	0.0390	0.1587
masked	18	42	0.1500	0.0482	0.1717
masked	18	43	0.1377	0.0478	0.1812
masked	18	44	0.1498	0.0444	0.1781
causal	11	42	0.2816	0.1098	0.2163
causal	11	43	0.2649	0.1077	0.2257
causal	11	44	0.2613	0.1058	0.2284
causal	14	42	0.3583	0.1152	0.2400
causal	14	43	0.3274	0.1151	0.2500
causal	14	44	0.3365	0.1157	0.2539
causal	18	42	0.4180	0.1218	0.2639
causal	18	43	0.4092	0.1232	0.2746
causal	18	44	0.3969	0.1248	0.2841

Single-latent F1 and SAEBench probing. Both were computed through our own pipeline on the controlled pair rather than through released code. Both use the same concepts: SCOPe class, fold, superfamily and family labels, and residue-level secondary-structure and solvent-accessibility labels, each present in at least ten domains. The domains are split by SCOPe fold into two halves. Single-latent F1, following Adams et al. (2025), selects for each concept the one latent and activation threshold with the highest per-residue F1 on the first half, and reports that latent's per-residue F1 on the second half. The candidate thresholds are any positive activation and the 90th to 99.9th percentiles of the latent's positive activations. The k-sparse probe, following Karvonen et al. (2025), ranks latents by the difference in their mean activation between positive and negative residues and fits a logistic-regression probe on the top k latents, for $k \in \{1, 2, 4, 8, 16, 32\}$. The value of k is chosen by AUROC on a fold-disjoint selection subset of the first half, and the probe's AUROC on the second half is reported. In a paired comparison, each metric is averaged over the concepts scored in both cells. Single-latent F1 is averaged over the 12 concepts shared by every paired cell, of 47 scored. Single-latent F1 rises under order-destroyed training in the causal models and is exceeded by the no-model indicator layer on both objectives. k-sparse probing shows no detectable response to the intervention on either objective and stays above both baselines on average (Table 16).