When Does Window Position Matter in Protein Language Models? A Controlled Benchmark and Cross-Architecture Taxonomy

Anonymous authors
Paper under double-blind review

Abstract

Protein language models (pLMs) are routinely applied to long proteins via fixed-length inference windows, so the *same* residue can be scored at the left edge, the center, or the right edge of its window. Using content-controlled probes, we ask whether and how this window position changes a pLM's behavior. We find a *real but modest* window-edge effect: bidirectional encoders (ESM-2, ESM C) are most faithful when a residue is centered and degrade toward the edges—an *inverted*-U, opposite the NLP "lost-in-the-middle" U and roughly an order of magnitude smaller. The effect is *architecture-specific*: causal models (ProGen2) instead depend monotonically on left context. We map it across **six backbones along three axes** (architecture, position encoding, structure-awareness). Mechanistically it is *one-sided context-loss*, not a terminal-token artifact; we find no decoder-style lost-in-the-middle and no truncation cliff—a controlled probe sliding a fixed-context residue across the training crop stays flat, consistent with rotary encodings extrapolating. Crucially, the effect is *largely metric-level*: it does **not** degrade three standard downstream tasks (variant-effect prediction, secondary structure, long-range contacts) *when the window retains task-relevant context*, the regime those tasks occupy by construction—but it does bite in the complementary regime, where cutting a long-range contact's partner to a different window makes the contact unrecoverable. We release a controlled benchmark, bias metrics, the taxonomy, the downstream scoping, and a *training-free* symmetric multi-crop remedy that recovers 68–80% of the edge loss without retraining.

1 Introduction

Protein language models (pLMs) have become the default tool for variant-effect prediction, structure and contact prediction, and residue-level functional annotation (Lin et al., 2023; EvolutionaryScale Team, 2024; Nijkamp et al., 2023). All of them operate on a finite context window—ESM-2 and SaProt on 1026 positions, ProtT5 on 512, ESM C on 2048, and ProGen2 on 1024 tokens (these are training/configuration context lengths, not hard limits—the rotary models extrapolate beyond them, Section 6; our ProGen2-base checkpoint allocates 2048). Real proteins routinely exceed these limits, so practitioners apply pLMs to long proteins by tiling, truncating, or pooling fixed-length windows. A direct consequence, usually left implicit, is that the *same* residue is presented to the model at different positions within its window: at the left edge in one tile, the center in another, the right edge in a third (Figure 1). Whether and how this window position changes the model's behavior is the question we study.

In natural language, decoder-only models exhibit a well-known *lost-in-the-middle* effect: evidence placed in the middle of a long context is used least effectively, producing a U-shaped accuracy curve (Liu et al., 2024), attributed to the interaction of causal masking with the long-range decay of rotary position encodings (Su et al., 2024a; Men et al., 2024). It is tempting to assume protein encoders inherit the same pathology, but the assumption is untested and the mechanisms differ: the dominant pLMs are *bidirectional* masked encoders without a causal mask, and they do not compress retrieved evidence into an autoregressive answer. We therefore ask, with content-controlled probes that hold a residue's local biology fixed while varying only its

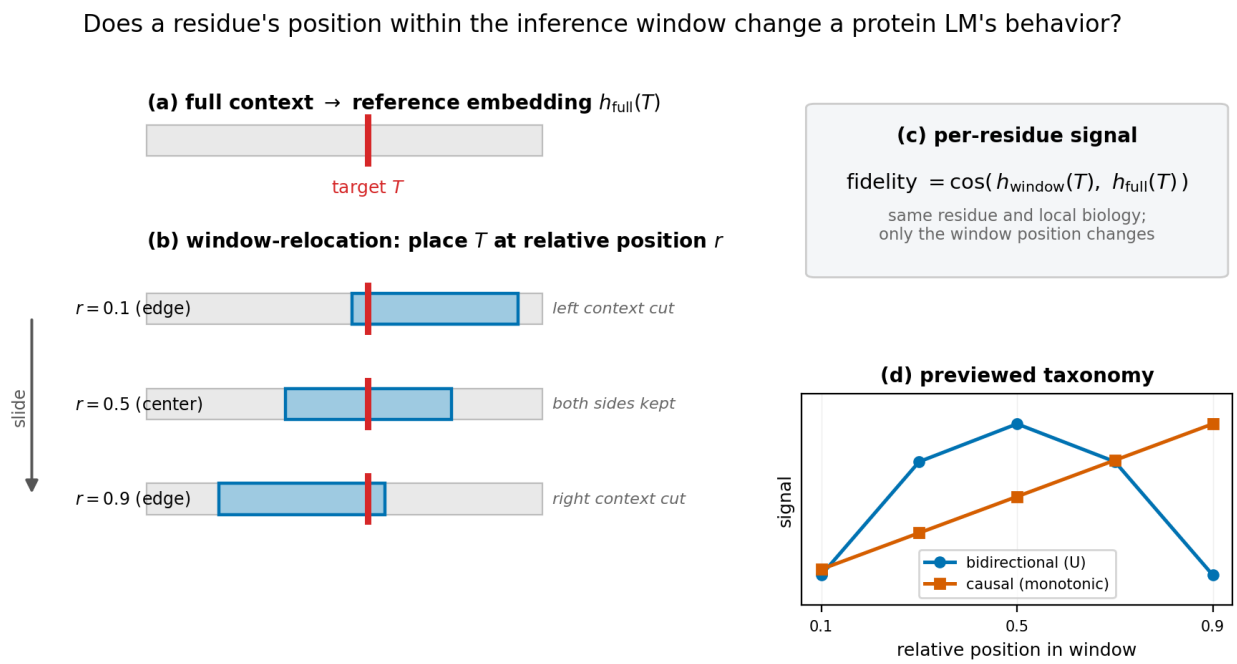

Figure 1: The crop-position probe. (a) A reference embedding for a non-terminal target residue T is computed with full context. (b) A fixed width-W inference window is slid so T lands at relative positions $\{0.1, 0.5, 0.9\}$; the near window edge truncates real context on one side, while T's own local neighborhood is unchanged. (c) The per-residue signal is the embedding fidelity—the cosine between the windowed and full-context embedding of T—so only window position varies. (d) Previewed taxonomy: bidirectional encoders show a center-favored inverted-U; causal models are monotonic.

window position: does a position effect exist, what shape does it take, what drives it, and—critically—does it change anything downstream?

We answer all four. **(C1)** A real but *modest* window-edge effect exists: bidirectional encoders (ESM-2, ESM C) are most faithful when a residue is centered and degrade toward the edges—an *inverted*-U, opposite the NLP U and roughly an order of magnitude smaller. **(C2)** The effect is *architecture-specific*, and mapping it is our centerpiece: across six backbones and three axes (architecture, position encoding, structure-awareness), causal models depend monotonically on left context whereas bidirectional encoders show the symmetric inverted-U. **(C3)** Two intuitions imported from NLP do not survive scrutiny: there is no decoder-style lost-in-the-middle, and no truncation cliff—a controlled probe sliding a fixed-context residue across the crop boundary stays flat, consistent with rotary encodings extrapolating. **(C4)** Mechanistically the effect is one-sided context-loss, and it is *largely metric-level*: across three standard downstream tasks (variant-effect prediction, secondary structure, long-range contacts) it does not degrade predictions when the window retains task-relevant context—the regime those tasks occupy by construction—while a matched probe that genuinely cuts the context (a long-range contact whose partner falls in a different window) shows the predicted degradation. Finally, we empirically show that the metric-level effect is cheaply mitigated by a training-free symmetric multi-crop that recovers 68–80% of the edge loss.

Contributions. (1) A controlled, backbone-agnostic benchmark for window-position behavior (sliding-cassette and window-relocation probes; a family-split snapshot; bias metrics). (2) A cross-architecture taxonomy along three axes (architecture, position encoding, structure-awareness; the latter two with one model per non-rotary class, so they are suggestive rather than controlled)—the centerpiece. (3) Evidence against two intuitions: no protein lost-in-the-middle and no truncation cliff (a controlled probe across the

crop is flat). (4) A downstream scoping result—the effect is largely metric-level, with a matched positive instance once the long-range context is actually cut. (5) A training-free remedy for the metric-level effect.

2 Related Work

Position bias in language models. The clearest precedent is in NLP, where decoder-only models exhibit *lost-in-the-middle*: evidence in the middle of a long context is used least effectively, yielding a U-shaped accuracy curve (Liu et al., 2024). Follow-up work traces this to a positional attention bias that can be partly corrected at inference by recalibrating it (Hsieh et al., 2024), and separates *having* the information from *using* it—models often encode the relevant evidence but fail to surface it (Lu et al., 2024). Mechanistically, rotary position encodings (Su et al., 2024a) impose a long-range decay that bounds the usable context length (Men et al., 2024), and recent analysis derives position bias in multi-layer attention (Wu et al., 2025). Crucially, this story rests on *causal* masking and an autoregressive readout; we treat it as a methodological template (control position; separate retrieval from utilization; separate architecture from training distribution) rather than as evidence that bidirectional protein encoders behave the same way.

Protein language models. Our backbones span the architectural axes this template asks about. ESM-2 (Lin et al., 2023) and ESM C (EvolutionaryScale Team, 2024) are sequence-only bidirectional masked encoders with rotary encodings; SaProt (Su et al., 2024b) adds a structure-aware vocabulary on the same rotary/bidirectional backbone; ProtT5 (Elnaggar et al., 2022) uses relative position encodings and a 512-token training crop (later extended to 1024); and ProGen2 (Nijkamp et al., 2023) is an autoregressive (causal) rotary model. This diversity in objective (masked vs. causal), position encoding (rotary, learned-absolute, relative), and structure-awareness is precisely what lets us ask whether a window-position effect is architecture-specific rather than a property of any single model.

What protein models attend to—and what they may miss. Interpretability work shows that pLM attention encodes biologically meaningful structure such as contacts, binding sites, and biochemical properties (Vig et al., 2021), with the broader landscape surveyed by Xiao et al. (2025). These studies establish *where* a model looks, but not whether the same evidence is used consistently as its window position changes. Relatedly, Zhang et al. (2024b) argue that pLM contact prediction reflects memorized co-evolutionary motif statistics rather than explicit pairwise reasoning; if models lean on such local statistics, cropping and window placement could plausibly alter how that evidence is invoked—motivating a controlled, content-matched position probe.

Benchmarks and the genomic analog. Protein-model evaluation is increasingly fine-grained and systematic—ProteinGym for zero-shot fitness (Notin et al., 2023), VenusX for residue/segment/domain-level functional understanding (Tan et al., 2025), PFMBench across foundation models (Gao et al., 2025), and residue-level resources such as BioLiP2 (Zhang et al., 2024a)—yet none isolates *window-position* behavior. The closest precedents are in genomics, where DNA models are probed for single-nucleotide-resolution annotation (de Almeida et al., 2025), long-context recall (Brixi et al., 2025), and—most pointedly—whether they learn positional regulatory logic rather than mere composition (Cheng & Zhang, 2026). These motivate the question but address a different modality and task family. To our knowledge, no prior work provides a controlled, content-matched benchmark of crop/window-position behavior for protein encoders, nor a cross-architecture taxonomy of it; that is the gap we fill.

3 Benchmark and Method

Our central methodological requirement is to vary a residue's *window position* while holding its biology fixed, so that any change in model behavior is attributable to position rather than to the residue's identity or local context (a confound we call the termini-biology confound: real protein ends are genuinely lower-information, so naively stratifying accuracy by position conflates a model effect with a data effect). We achieve this with two content-controlled probes and a backbone-agnostic signal, evaluated under pre-registered thresholds.

Probes. The *window-relocation* probe (in-distribution) slides a length-W inference window over a real protein so that a fixed, non-terminal target residue lands at relative positions rel-pos $\in \{0.1, 0.3, 0.5, 0.7, 0.9\}$; the residue keeps its native flanks, and only the amount of real context truncated by the near window edge changes. The *sliding-cassette* probe (controlled / out-of-distribution) transplants a recoverable site wrapped in a *constant* $\pm f$ native flank into a shuffled carrier and slides the whole cassette across window positions; because the residue's entire local context travels with it, this probe isolates any *pure* position/terminal effect from context-loss (used for the mechanism analysis in Section 5). Targets are kept at least W residues from the protein termini so the window edge is not a sequence terminus.

A uniform cross-backbone signal. To compare backbones that differ in vocabulary, training objective, and output head, we use *embedding fidelity*: for a target residue, the cosine similarity between its per-residue hidden state computed inside the window and the hidden state computed with full context. Fidelity requires only hidden states and so applies uniformly to sequence-only, structure-aware, learned-position, and causal models, avoiding masked-recovery idiosyncrasies (e.g. SaProt's sequence-only masked recovery is degenerate). For masked encoders we additionally report *masked recovery*—the softmax probability the model assigns to the true residue when it is masked—which is the quantity optimized during pre-training and exhibits the same effect at larger magnitude.

Bias metrics. For a target i with per-placement signal $s_i(r)$ over placements $r \in P$ (center c=0.5, edges E), we report three metrics, averaged over targets:

$$\text{Position Gap} = \frac{1}{N}\sum_i \Big(s_i(c) - \tfrac{1}{|E|}\textstyle\sum_{r\in E} s_i(r)\Big),$$
$$\text{Shift Sensitivity} = \frac{1}{N}\sum_i \mathrm{std}_{r\in P}\, s_i(r), \qquad \text{Robust Accuracy} = \frac{1}{N}\sum_i \min_{r\in P} s_i(r). \tag{1}$$

Position Gap (signed; positive = center-favored) is the primary effect size; Shift Sensitivity measures placement instability; Robust Accuracy is the worst-case over placements. Each gap is paired with a 95% confidence interval from a within-target/within-assay paired bootstrap.

Snapshot, splits, and pre-registration. The benchmark ships as a versioned snapshot of 200 proteins (one non-terminal target each) drawn from a pre-2024 reviewed-human set, with a manifest recording the seed, git revision, and provenance. Proteins span 480–1499 residues; every backbone processes the full length in one pass (rotary and relative encodings extrapolate beyond their training crops, and ProtBERT's learned-position table spans 40k), so each full-context reference embedding is genuine rather than itself windowed. To support fine-tuned baselines without leakage, we provide a family-aware split via MMseqs2 (`easy-cluster`, 30% identity, 80% coverage) into family-disjoint train/test partitions (140/60). The backbone panel spans the three taxonomy axes: ESM-2 (35M/150M/650M; RoPE, bidirectional), ESM C 300M (RoPE, bidirectional), SaProt 650M (RoPE, structure-aware), ProtBERT (learned absolute PE), and ProGen2-764M (RoPE, causal). All effect-size thresholds were *pre-registered* before running the main probes: a benchmark tier ($\geq$ 3 pp gap, or $\Delta\rho \geq 0.03$ for variant-effect ranking) and a method tier ($\geq$ 10 pp), both required at $p < 0.01$. Each center−edge gap is assessed with a within-target (within-assay) paired bootstrap (10,000 resamples, seed 42), reporting the two-sided bootstrap p-value and the 95% percentile confidence interval. Evaluating a new backbone requires only an `embed_fn`(seq)$\rightarrow(L, H)$ callback; everything else is shared.

4 A Cross-Architecture Taxonomy of Window-Position Behavior

We first establish that window position systematically affects pLM behavior, and that *how* it does is determined by architecture, using the backbone-agnostic *embedding-fidelity* signal of Section 3. For each non-terminal target residue we slide the window so the residue lands at relative positions rel-pos $\in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and measure the cosine between its windowed and full-context embedding at each placement; the residue's identity and immediate biology are held fixed, so any change is attributable to position.

Table 1: Window-position behavior across six backbones spanning three axes (architecture, position encoding, structure-awareness), measured with the uniform embedding-fidelity probe. "center−edge" is the mean within-residue fidelity gap (percentage points) between the centered placement and the edge placements; positive values indicate a center-favored inverted-U (n=120 non-terminal targets). Bidirectional encoders show a modest inverted-U regardless of family, position encoding, or structure-awareness; the causal model is monotonic in relative position.

Backbone	**Family / position enc.**	**Shape**	**center−edge**	p
ESM-2 650M	bidirectional, RoPE	inverted-U	+3.29 pp	< 0.001
ESM C 300M	bidirectional, RoPE	inverted-U	+3.04 pp	< 0.001
SaProt 650M	bidirectional, RoPE, structure	inverted-U	+2.82 pp	< 0.001
ProtT5	bidirectional, *relative* PE	inverted-U	+5.49 pp	< 0.001
ProtBERT	bidirectional, *learned* PE	inverted-U	+8.68 pp	< 0.001
ProGen2-764M	**causal**, RoPE	monotonic	— (ρ_S=+1.0)	< 0.001

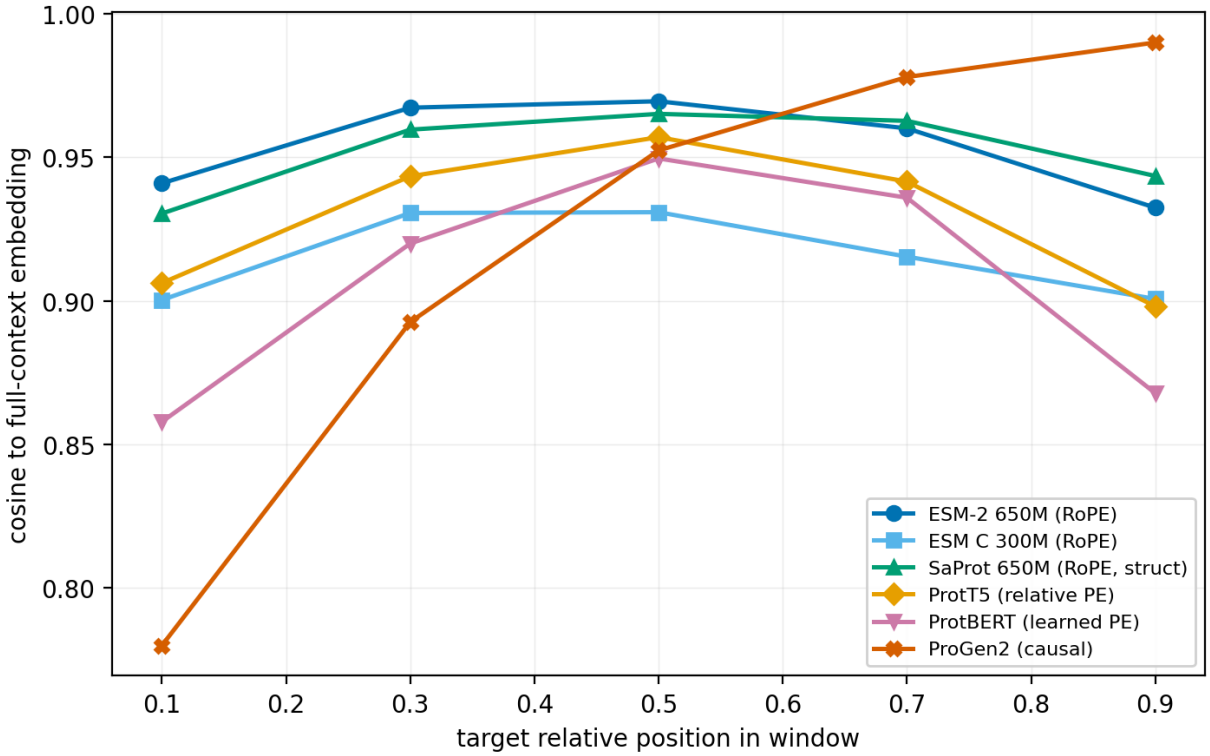

Figure 2: Cross-architecture taxonomy. Embedding fidelity versus the target residue's relative position in the window. Bidirectional encoders (ESM-2, ESM C, SaProt) are most faithful at the center and degrade symmetrically toward both edges (a modest inverted-U); the effect grows along the position-encoding axis from rotary (RoPE) to T5 relative-bias (ProtT5) to learned-absolute (ProtBERT), all with the same inverted-U shape; the causal model ProGen2 instead rises monotonically with relative position (Spearman ρ=+1.0), reflecting its dependence on left context only.

Bidirectional encoders show a modest, center-favored inverted-U. Table 1 and Figure 2 show that ESM-2, ESM C, and SaProt are each most faithful when the target is centered and lose fidelity toward the window edges, with a center−edge gap of 2.8–3.3 pp (all $p < 0.001$); these per-backbone fidelity gaps sit near

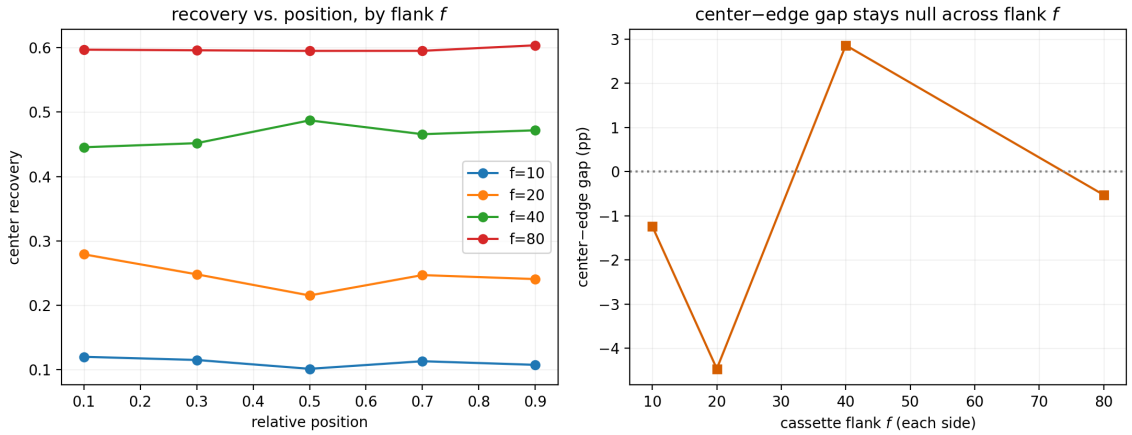

Figure 3: Mechanism. When a recoverable site is wrapped in a *constant* $\pm f$ native flank and slid across window positions, recovery-versus-position is flat at every flank size f, and the center−edge gap is null at the pre-registered bar. With the residue's own context held fixed, window position has no intrinsic effect; the window-edge effect of Section 4 is therefore one-sided truncation of real context, not a terminal-token or pure-position bias.

the pre-registered 3 pp benchmark-tier threshold, while a larger signal—whose point estimate reaches the method tier—appears on the masked-recovery metric (below). This is the *opposite* of the NLP "lost-in-the-middle" U-shape and roughly an order of magnitude smaller. The same qualitative effect appears at larger magnitude on the masked-recovery metric used during pre-training (e.g. a center−edge gap of +10.7 pp for ESM-2 at W=256), and it is present and grows across ESM-2 scales (35M–650M; Appendix A).

Causal models are monotonic, not U-shaped. ProGen2, an autoregressive decoder with rotary position encodings, shows no inverted-U: its fidelity increases monotonically with relative position (Spearman ρ=+1.0; Figure 2), because a later placement grants the residue more visible left context. Window position therefore acts on causal and bidirectional pLMs through qualitatively different mechanisms—left-context hunger versus a symmetric window-edge effect—and this distinction is the core of our taxonomy.

The effect is not ESM-specific, not RoPE-specific, and not removed by structure. The taxonomy answers four questions a reader will ask. *Is it an ESM-2 artifact?* No: ESM C (a different family) and SaProt reproduce it. *Is it specific to rotary position encodings?* No: it appears across the position-encoding axis and in fact *grows* as the encoding becomes more absolute: RoPE (relative; 2.8–3.3 pp) < T5 relative-bias (ProtT5; +5.49 pp) < learned absolute (ProtBERT; +8.68 pp). This gradient is consistent with RoPE's relative encoding being comparatively crop-robust, rather than rotary encodings being the cause of the effect; we note that each non-rotary position-encoding class is represented by a single model, so the axis is suggestive rather than a controlled comparison. *Does structure-awareness help?* No: SaProt's gap (+2.82 pp) matches sequence-only ESM-2. *Do causal and bidirectional models behave the same?* No, as above. The window-edge effect is thus a general property of bidirectional pLM encoders, modulated by the position-encoding scheme.

5 Mechanism: One-Sided Context-Loss

The window-edge effect admits two qualitatively different explanations. It could be an *intrinsic* position artifact—a terminal-token or attention-sink effect, or a pure dependence on the absolute index within the window—that would persist even if the residue's information were unchanged. Alternatively, it could simply be *one-sided context-loss*: placing a residue near a window edge truncates the real context on the near side, removing evidence the model would otherwise use. These have opposite implications for downstream behavior, so we disentangle them with the sliding-cassette probe.

The cassette holds the residue's entire local context ($\pm f$ native flank) constant while moving the whole block to different window positions in a shuffled carrier. If the effect were intrinsic, we would still see a center−edge gap; if it is context-loss, the curve should flatten because no real context is removed. Figure 3 shows the latter: across flank sizes $f \in \{10, 20, 40, 80\}$ the recovery-versus-position curve is flat, and the center−edge gap is not significant at the pre-registered $p < 0.01$ bar ($|\text{gap}| \leq 4.5$ pp, min $p = 0.019$ at f=40, with inconsistent sign; at f=80 the gap is -0.53 pp, $p = 0.57$). There is thus no intrinsic terminal-token, attention-sink, or pure-position bias: the window-edge effect is one-sided context-loss. This mechanism is exactly what makes the effect benign downstream (Section 7)—when the window still contains the task-relevant context, nothing the prediction needs has been removed.

6 Refuting the Naive Stories

Two intuitions imported from NLP and from finite-context training would, if true, predict a large effect. We find neither holds for protein encoders.

No decoder-style lost-in-the-middle. The canonical NLP probe inserts a "needle" into a long context and tests whether the model recovers it. Transplanted to a protein masked language model—insert a short motif into a carrier and recover it—this paradigm fails as an instrument: masked recovery of an inserted motif sits at chance ($0.049 \approx 1/20$, the uniform amino-acid prior), with position amplitude below 2 pp everywhere. The reasons are protein-specific: an inserted motif creates an out-of-distribution junction with no informative neighboring context, and a generic consensus motif is low-information by design. The needle paradigm therefore does not transfer, which is precisely why we adopt content-controlled probes (window-relocation and sliding-cassette, Section 3) that never insert foreign content. This is a statement about the *instrument*, not evidence for or against an effect. The positive evidence against a decoder-style U is the center-favored inverted-U of Section 4, whose direction is opposite to lost-in-the-middle.

No training-crop truncation cliff. ESM-2 is trained on crops of up to 1022 residues, which suggests a sharp competence drop for positions beyond the crop. We test this by scoring variant effects with full-sequence wt-marginals, which lets the rotary encoding extrapolate past 1022. On BRCA1—the only ProteinGym assay whose mutated positions straddle the 1022 boundary—the in-crop and out-of-crop Spearman correlations are $\rho_{\text{in}} = 0.475$ and $\rho_{\text{out}} = 0.537$, a gap of $\Delta\rho_{\text{in}-\text{out}} = -0.062$ (95% CI $[-0.14, 0.01]$, $p = 0.11$)—the *wrong sign* and not significant; the ρ-versus-position curve even peaks past 1022 (the BRCT domain). A pooled cross-assay comparison can look like a cliff, but that is a confound: the out-of-crop pool is dominated by intrinsically low-ρ proteins, not by position. We therefore observe no truncation cliff (Figure 4). BRCA1 is in fact the *only* ProteinGym substitution assay whose single mutants straddle the crop (of the nine proteins longer than 1022 residues, only BRCA1 has variants on both sides), so this direct DMS evidence cannot be powered up with a second assay—a limitation of the benchmark, not of the test.

We therefore complement it with a *controlled* cliff test that holds local context constant. We implant a recoverable site, wrapped in a constant ± 40 native flank, at different *absolute* positions in a fixed-length (2200) shuffled carrier and measure masked recovery as the target crosses the 1022 boundary; because only the absolute index changes, this isolates a training-crop effect from one-sided context-loss and from sequence length. Recovery is flat (Figure 5): in-crop ($p \leq 1022$) 0.348 versus out-of-crop 0.345, a gap of +0.28 pp (95% CI $[-1.17, 1.64]$, $p = 0.68$; n=60 cassettes), and the step across the boundary ($900 \rightarrow 1100$) is +0.91 pp ($p = 0.35$)—no discontinuity. With content held fixed, the rotary model extrapolates past its training crop, confirming the single-assay BRCA1 result with a powered, confound-free probe.

7 Downstream Scoping: When It Does (Not) Matter

A metric-level position effect is only consequential if it changes downstream predictions. We test this directly: keeping the window-relocation protocol, we replace the per-residue signal with the output of a standard downstream task and ask whether the prediction degrades when the target is placed at the window edge rather than the center. We use the same paired-bootstrap statistics and pre-registered thresholds throughout, and analyze each task per-assay / per-protein without pooling.

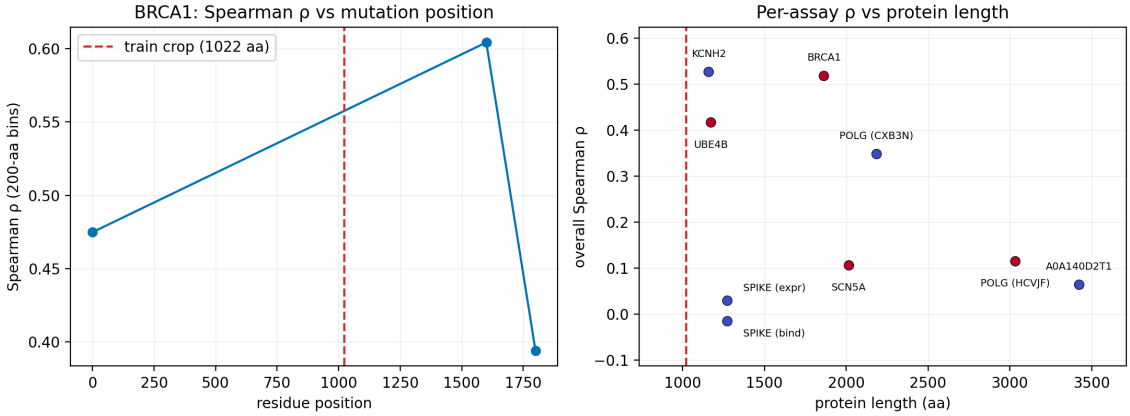

Figure 4: No training-crop truncation cliff. For BRCA1 (the only assay straddling the 1022 training crop), the Spearman correlation does not drop beyond the crop boundary—if anything it is higher out-of-crop ($\Delta\rho_{\text{in-out}} = -0.062$, $p = 0.11$)—consistent with rotary position encodings extrapolating past the training crop length.

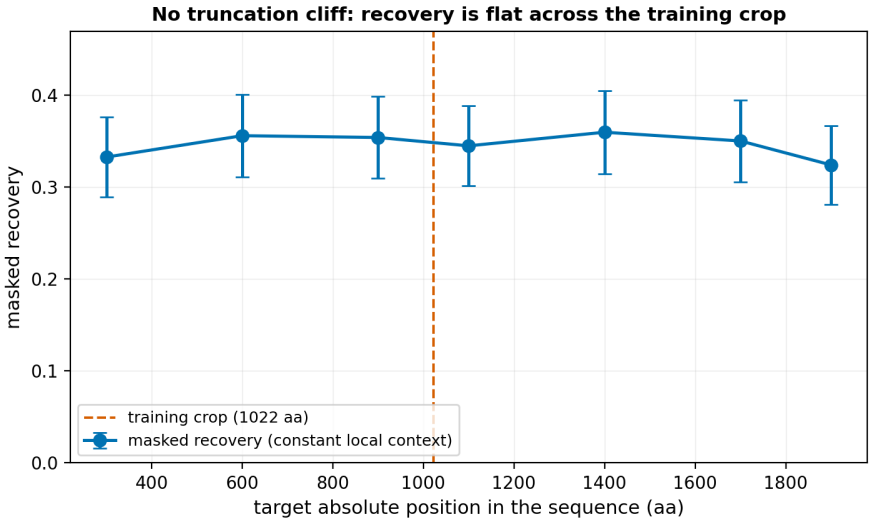

Figure 5: A controlled truncation-cliff test. A recoverable site with a *constant* ±40 native flank is placed at different absolute positions in a fixed-length (2200) shuffled carrier, so only the target's absolute index changes. Masked recovery is flat across the 1022 training crop (in-crop − out-of-crop = +0.28 pp, $p = 0.68$; n=60 cassettes), confirming that rotary encodings extrapolate past the training crop without a competence cliff.

Standard downstream tasks are position-robust. Table 2 and Figure 6 show the contrast cleanly. The two controlled probes clear the pre-registered bar (+10.70 and +3.58 pp, $p < 0.01$), but all three downstream tasks fail to reach significance: variant-effect prediction (VEP) has a Spearman gap of +0.025 ($p = 0.16$), secondary-structure Q3 accuracy a gap of −0.58 pp ($p = 0.58$), and long-range contact probability a gap of −0.34 pp ($p = 0.51$). The window-edge effect is therefore largely a property of the masked-recovery / representation metric and does not, by itself, degrade these standard predictions.

A unifying principle. Because the mechanism is one-sided context-loss (Section 5), a task inherits the effect only if its prediction is *both* (a) a per-position output *and* (b) dependent on the long-range context that

Table 2: Downstream scoping. The center−edge gap, with 95% paired-bootstrap confidence intervals, under one common statistical idiom. The two controlled probes (metric level) clear the pre-registered detection threshold; all three standard downstream tasks straddle zero and are not significant. "pp" = percentage points; for variant-effect prediction the effect is a Spearman gap $\Delta\rho$ with threshold 0.03. The embedding-fidelity gap shown here (+3.58 pp, n=60) is measured on the held-out benchmark test split; the taxonomy in Table 1 reports +3.29 pp on the full n=120 target set. Masked recovery's point estimate (+10.70 pp) reaches the method tier ($\geq$ 10 pp), but its CI lower bound (6.45) does not exclude smaller values, so we report it as a robust benchmark-tier clearance.

Probe / task	**Level**	n	**center−edge**	95% **CI**	p	**Verdict**
Masked recovery	metric	62	+10.70 pp	$[6.45, 15.31]$	< 0.01	clears benchmark tier
Embedding fidelity	metric	60	+3.58 pp	$[2.77, 4.47]$	< 0.01	clears (benchmark tier)
ProteinGym VEP	downstream	27	+0.025 ($\Delta\rho$)	$[-0.009, 0.062]$	0.16	n.s. (CI crosses 0)
Secondary structure	downstream	600	−0.58 pp	$[-2.50, 1.33]$	0.58	n.s. (CI crosses 0)
Long-range contacts	downstream	444	−0.34 pp	$[-1.37, 0.71]$	0.51	n.s. (CI crosses 0)

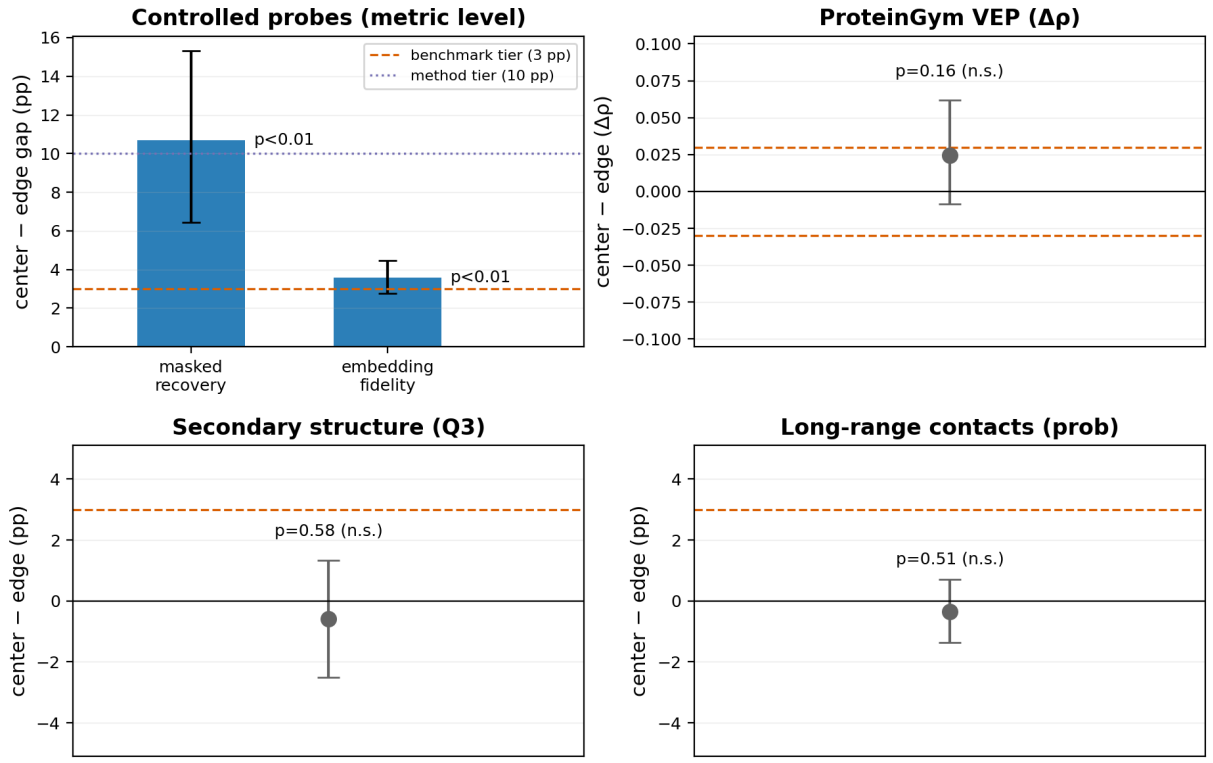

Figure 6: The window-edge effect is real at the metric level but does not propagate to standard downstream tasks. Top-left: the controlled probes (masked recovery, embedding fidelity) clear the pre-registered benchmark- and method-tier thresholds ($p < 0.01$). The other three panels: variant-effect prediction, secondary structure, and long-range contact prediction each show a center−edge gap whose confidence interval crosses zero (not significant), given that the window retains the task-relevant context.

the window edge actually removes. Each downstream task fails one prong. *VEP* is a *ranking* of mutations: crop placement shifts the absolute score roughly monotonically (shift-sensitivity $\approx$ 0.76 nats), and Spearman ranking absorbs this drift, so it fails (a). *Secondary structure* depends on a ±10–20-residue neighborhood, which an edge-placed target retains at W=256, so it fails (b). For *long-range contacts* we slid only pairs

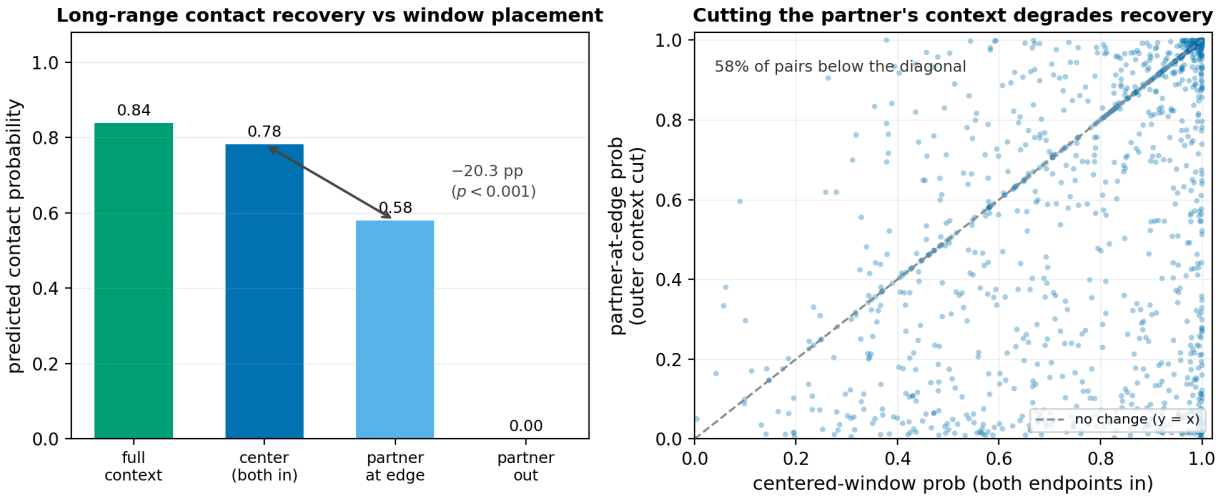

Figure 7: A positive downstream instance: long-range contact prediction in the regime where the window edge cuts the defining context. *Left:* predicted contact probability under four window placements—full context; the pair centered (both endpoints in-window, ≈ the null); the partner flush at the window edge (its outer context cut, partner still in-window); and the partner pushed outside the window (contact unrecoverable, probability 0). *Right:* per-pair centered vs. partner-at-edge probability (n=1178); most pairs fall below the diagonal. Contact prediction is position-robust only while the window retains both endpoints' context.

whose *both* endpoints remain inside the window, so the defining context is never cut, and it fails (b). The effect surfaces only when both conditions hold simultaneously. The three nulls above each keep the task-relevant context inside the window by construction; we therefore add a matched probe in the complementary regime—where the long-range context is genuinely cut—to test the principle directly rather than infer it.

A positive downstream instance. Long-range contact prediction supplies the test. We take confident full-context contacts (predicted probability > 0.5, sequence separation 30–200; n=1178 pairs) and, instead of keeping both endpoints in-window, slide the width-W window so the contact sits against its edge (Figure 7). With the pair *centered* (both endpoints balanced) the predicted contact probability is 0.78, matching the full-context value 0.84 and reproducing the null. Sliding the window so the partner residue sits flush at the edge—truncating the real context beyond it while the partner is still in-window—drops the predicted probability to 0.58, a center−edge gap of +20.3 pp (95% CI $[18.3, 22.2]$, $p < 0.001$). Pushing the window one residue further places the partner *outside* the window, so the contact spans two tiles and is unrecoverable; a fixed non-overlapping W=256 tiling does this to 19.1% of these confident long-range contacts. The *same* task that is position-robust when the window retains context (−0.34 pp, n.s.) thus degrades sharply once the window removes it—a direct downstream confirmation of the unifying principle, and a concrete consequence for practitioners who tile long proteins.

Power. The secondary-structure and contact nulls are well powered ($n = 600$ and 444 within-target pairs). The VEP result now rests on 27 long assays—within the 22–56 range our power analysis indicated for detecting $\Delta\rho \approx 0.03$—and remains a null ($\Delta\rho = +0.025$, $p = 0.16$, with the point estimate below the 0.03 threshold). The practical message is unchanged: with retained context, windowing is safe for local residue tasks and for ranking tasks that keep the relevant residues in-window.

8 A Training-Free Remedy

Although the window-edge effect is largely metric-level, it still matters when a model is used in a way that is sensitive to that metric—masked-recovery or pseudo-perplexity scoring, or per-residue embeddings

Table 3: Training-free multi-crop remedy at the recovery level, on two bidirectional families. Each strategy's recovered fraction of the center−edge gap (%), with the absolute gain over the edge baseline. Centrality-weighted multi-crop recovers most of the loss with no retraining; centering is the 100% upper bound. All gains over edge are significant at $p < 0.001$ (paired bootstrap).

Backbone	W	**uniform multi-crop**	**centrality-weighted**	**centering**
ESM-2 650M	256	46% (+4.9 pp)	**68**% (+7.3 pp)	100% (upper bound)
ESM-2 650M	384	56% (+5.1 pp)	**80**% (+7.3 pp)	100% (upper bound)
ESM C 300M	256	55% (+6.5 pp)	**79**% (+9.3 pp)	100% (upper bound)
ESM C 300M	384	54% (+4.8 pp)	**78**% (+7.0 pp)	100% (upper bound)

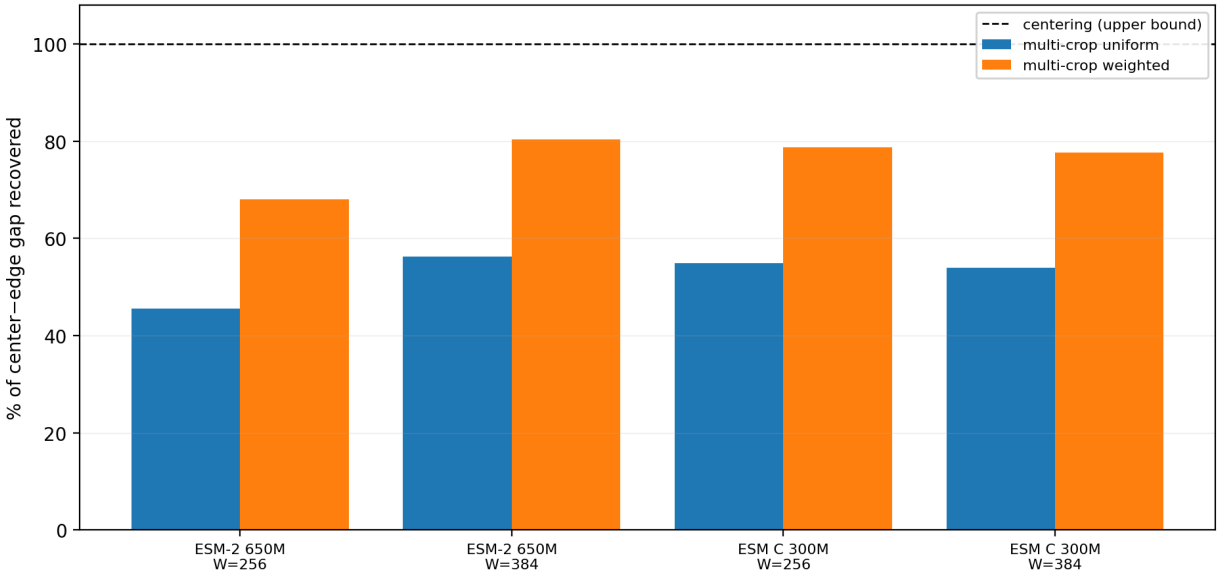

Figure 8: Training-free multi-crop recovers the window-edge loss without retraining, consistently across two bidirectional families. Uniform symmetric multi-crop recovers ∼46–56% of the center−edge gap; centrality-weighted multi-crop recovers ∼68–80%; per-residue centering is the 100% upper bound (dashed line).

consumed before any pooling. Because the effect is one-sided context-loss (Section 5), it can be mitigated *without retraining* whenever a residue can be scored from several overlapping windows, as is natural when tiling a long protein. We compare four placements/strategies: scoring from a window *edge* (the bad case); from the window where the residue is *centered* (the upper bound, but requiring a re-tile per residue); a *uniform* average of the per-placement distributions; and a *centrality-weighted* average that weights each window by the residue's centrality in it, $w(r) \propto \min(r, 1 - r)$ (its balanced-context budget). We quantify each strategy by its *recovered fraction* of the center−edge gap, 0 at the edge and 1 at the center.

Centrality-weighted multi-crop recovers most of the loss. Table 3 and Figure 8 show that uniform multi-crop already recovers ∼46–56% of the edge loss for free, and the centrality-weighted variant recovers ∼68–80%—a clear, significant improvement whose only added ingredient is the placement weight—while per-residue centering attains the 100% upper bound at the cost of more forward passes. The pattern is consistent across ESM-2 and ESM C, matching the bidirectional inverted-U of the taxonomy (Section 4).

Scope. This remedy operates at the recovery/scoring metric level. Per Section 7, it does *not* translate into a downstream-accuracy gain on tasks that are already position-robust; it is intended for metric-level usage,

and we state this explicitly to avoid over-claiming. It is also a *bidirectional-encoder* tool: causal models are monotonic rather than U-shaped (Section 4), so their analogous "fix" is simply to grant more left context.

9 Discussion and Limitations

What the small effect size means. The window-edge effect is modest, and we report it as such. The contribution is not a downstream accuracy gain but a *measurement*: a controlled benchmark, a cross-architecture taxonomy, an honest scoping of when the effect matters, and a deployable tool for the case where it does. The work is deliberately *orthogonal* to backbone fitness—structure- and MSA-based methods dominate sequence-only models on ProteinGym (Notin et al., 2023), and we make no claim to improve fitness prediction. A reader should take away *where windowing is safe* (local residue tasks; ranking tasks that keep the relevant residues in-window) and *where to be careful* (per-position outputs that depend on long-range context), plus a free mitigation for the metric-level case.

Anticipated objections. We address the questions a reader is most likely to raise. *Is this just termini biology?* No: targets are non-terminal, and the sliding-cassette control (Section 5) holds local biology fixed, isolating a model-side position effect. *Is it specific to ESM-2 or to rotary encodings?* No: ESM C, SaProt, ProtT5, and ProtBERT all reproduce it, and it in fact grows as the position encoding becomes more absolute (Section 4; one model per non-rotary class, so that axis is suggestive). *Given the small size, why care?* Because the value is the taxonomy and the scoping corrective: practitioners currently lack any principled account of window-position behavior across pLM architectures. *Is the result obvious?* The opposite—the three most natural hypotheses (a decoder-style U, a truncation cliff, and *pervasive* downstream degradation) do not hold as stated: the first two are absent on the probes we can test, and downstream degradation appears only where the window actually removes the needed context, which is itself the finding.

Limitations. The variant-effect-prediction null is now powered with 27 long assays (within the indicated 22–56 range) and remains non-significant with a point estimate below the detection threshold; while this is a well-powered null rather than a sharp zero (the upper CI admits a small effect), it no longer rests on an underpowered handful of assays. We leave SaProt's structure-token mode (we run it sequence-only) to future work. Embedding fidelity measures how much a representation *shifts*, not whether it worsens; that a shifted representation is not necessarily a degraded one is precisely what the downstream nulls (Section 7) establish, while the positive contact instance marks where the shift does become a degradation. Finally, absolute embedding fidelity may be influenced by pre-training exposure, but every reported effect is a *within-residue* center-versus-edge gap, which is internally controlled and unaffected by this.

10 Conclusion

We presented a controlled study of how a residue's position within the inference window affects protein language models. Using content-controlled probes and a backbone-agnostic fidelity signal, we established that bidirectional encoders exhibit a real but modest center-favored inverted-U, that this behavior is architecture-specific (causal models are instead monotonic in left context), and that it is neither a decoder-style lost-in-the-middle nor a training-crop truncation cliff. Mechanistically the effect is one-sided context-loss, and—our central practical message—it is largely metric-level: it does not degrade standard downstream tasks when the window retains task-relevant context, and degrades them only when it does not. We release the benchmark, bias metrics, the taxonomy, the downstream-scoping battery, and a training-free multi-crop remedy, giving practitioners a principled account of when windowing is safe and a cheap fix for when it is not.

Broader Impact Statement

This work is a measurement and benchmarking study of existing protein language models; it introduces no new generative capability. By clarifying when window-position effects do and do not matter, it aims to prevent both over- and under-correction in deployed pLM pipelines.

References

Garyk Brixi, Matthew G. Durrant, Jerome Ku, Michael Poli, Greg Brockman, Patrick D. Hsu, Brian L. Hie, et al. Genome modeling and design across all domains of life with Evo 2. *bioRxiv*, 2025. doi: 10.1101/2025.02.18.638918.

Bryan Cheng and Jasper Zhang. The mechanistic invariance test: Genomic language models fail to learn positional regulatory logic, 2026.

Bernardo P. de Almeida, Hugo Dalla-Torre, Guillaume Richard, Christopher Blum, Lorenz Hexemer, Maxence Gélard, Javier Mendoza-Revilla, Ziqi Tang, Frederikke I. Marin, David M. Emms, Priyanka Pandey, Stefan Laurent, Marie Lopez, Alexandre Laterre, Maren Lang, Uğur Şahin, Karim Beguir, and Thomas Pierrot. Annotating the genome at single-nucleotide resolution with DNA foundation models. *Nature Methods*, 22(11):2301–2315, 2025. doi: 10.1038/s41592-025-02881-2.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

EvolutionaryScale Team. ESM Cambrian (esmc): Revealing the mysteries of proteins with unsupervised learning. EvolutionaryScale technical release, 2024. URL `https://github.com/evolutionaryscale/esm`.

Zhangyang Gao, Hao Wang, Cheng Tan, Chenrui Xu, Mengdi Liu, Bozhen Hu, Linlin Chao, Xiaoming Zhang, and Stan Z. Li. PFMBench: Protein foundation model benchmark. *arXiv preprint arXiv:2506.14796*, 2025.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14982–14995, 2024. doi: 10.18653/v1/2024.findings-acl.890. Preprint arXiv:2406.16008.

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

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. Preprint arXiv:2307.03172.

Taiming Lu, Muhan Gao, Kuai Yu, Adam Byerly, and Daniel Khashabi. Insights into LLM long-context failures: When transformers know but don't tell. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. Preprint arXiv:2406.14673.

Xin Men, Mingyu Xu, Bingning Wang, Qingyu Zhang, Hongyu Lin, Xianpei Han, and Weipeng Chen. Base of RoPE bounds context length. In *Advances in Neural Information Processing Systems 37*, 2024. Preprint arXiv:2405.14591.

Erik Nijkamp, Jeffrey A. Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978.e3, 2023. doi: 10.1016/j.cels.2023.10.002.

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

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024a. doi: 10.1016/j.neucom.2023.127063. Preprint arXiv:2104.09864 (2021).

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SaProt: Protein language modeling with structure-aware vocabulary. In *International Conference on Learning Representations*, 2024b. Spotlight.

Yang Tan, Wenrui Gou, Bozitao Zhong, Liang Hong, Huiqun Yu, and Bingxin Zhou. VenusX: Unlocking fine-grained functional understanding of proteins. *arXiv preprint arXiv:2505.11812*, 2025.

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*, 2021.

Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the emergence of position bias in transformers. In *International Conference on Machine Learning*, 2025. arXiv:2502.01951.

Yijia Xiao, Wanjia Zhao, Junkai Zhang, Yiqiao Jin, Han Zhang, Zhicheng Ren, Renliang Sun, Haixin Wang, Guancheng Wan, Pan Lu, Xiao Luo, Yu Zhang, James Zou, Yizhou Sun, and Wei Wang. Protein large language models: A comprehensive survey. *arXiv preprint arXiv:2502.17504*, 2025.

Chengxin Zhang, Xi Zhang, Peter L. Freddolino, and Yang Zhang. BioLiP2: an updated structure database for biologically relevant ligand-protein interactions. *Nucleic Acids Research*, 52(D1):D404–D412, 2024a. doi: 10.1093/nar/gkad630.

Zhidian Zhang, Hannah K. Wayment-Steele, Garyk Brixi, Haobo Wang, Matteo Dal Peraro, Dorothee Kern, and Sergey Ovchinnikov. Protein language models learn evolutionary statistics of interacting sequence motifs. *Proceedings of the National Academy of Sciences*, 121(45):e2406285121, 2024b. doi: 10.1073/pnas.2406285121.

A Additional Results

Masked-recovery magnitude and scale-robustness. On the masked-recovery metric, the window-relocation center−edge gap reaches the method-tier point estimate ($\geq$ 10 pp; the W=256 ESM-2 CI is $[6.45, 15.31]$, so it clears the benchmark tier robustly while the method threshold sits inside the interval): +10.7 pp (W=256) and +9.1 pp (W=384) for ESM-2, and +11.8 and +9.0 pp for ESM C (all $p < 0.01$). The effect is present at every ESM-2 scale and grows in absolute magnitude rather than vanishing—35M +4.65, 150M +5.95, 650M +10.94 pp (Figure 9)—ruling out a single-checkpoint artifact and showing it does not disappear with scale.

All three bias metrics. Table 4 reports the full bias-metric triple of Section 3 for every backbone. The three metrics are broadly consistent: across position-encoding classes the Position Gap grows (rotary < relative < learned-absolute, with the causal model apart), and higher Shift Sensitivity generally tracks lower Robust Accuracy. ProtBERT (learned-PE) and ProGen2 (causal) are the most placement-sensitive (Shift Sensitivity 0.078 and 0.083; Robust Accuracy 0.776 and 0.779); ESM-2 and SaProt are the most placement-robust (Robust Accuracy $\approx$ 0.91), with ESM C somewhat less so (0.85).

Seed stability. The main-text gaps carry within-seed bootstrap confidence intervals; re-sampling the n=120 target set under five seeds $\{42, 43, 44, 45, 46\}$ confirms between-seed stability for the ESM-2 anchor: Position Gap 3.51 ± 0.46 pp, Shift Sensitivity 0.028 ± 0.001, and Robust Accuracy 0.910 ± 0.004 (mean $\pm$ std over seeds). The effect's sign and magnitude are stable under target re-sampling. We sweep seeds on the ESM-2 anchor rather than on all six backbones because every per-backbone gap in Table 1 already carries a within-target paired-bootstrap CI excluding zero at $p < 0.001$, which controls sampling noise at a fixed target set; the seed sweep additionally controls the choice of target set, and the anchor shows that source of variance is small ($\pm$0.46 pp).

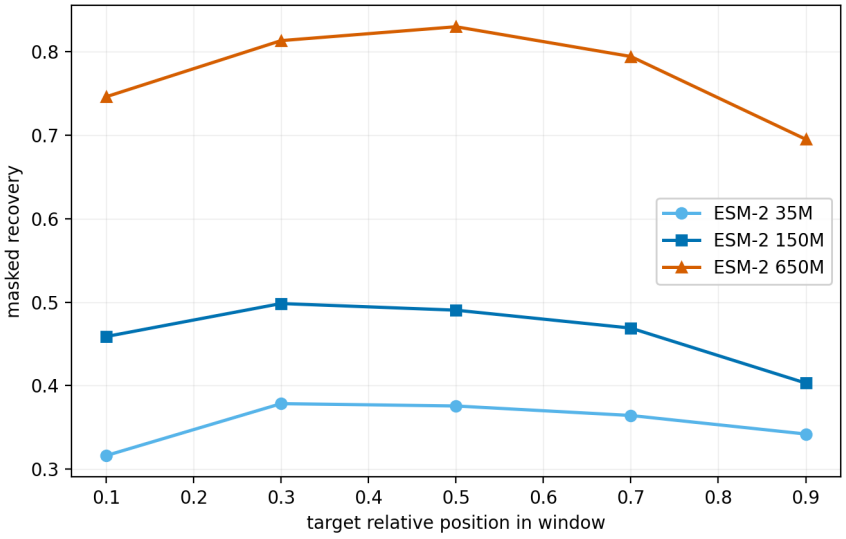

Figure 9: Masked-recovery center−edge gap across ESM-2 scales (35M/150M/650M); the window-edge inverted-U holds at every scale and grows in absolute magnitude.

Table 4: All three bias metrics (Section 3) per backbone on the embedding-fidelity probe (n=120 targets). Position Gap is in percentage points; higher Shift Sensitivity and lower Robust Accuracy both indicate greater placement sensitivity. ProGen2 is monotonic (Table 1); its Position Gap (†) reflects the concavity of a monotone fidelity curve, not a center-favored inverted-U, so it is not comparable to the bidirectional rows.

Backbone	n	**Position Gap**	**Shift Sensitivity**	**Robust Acc.**
ESM-2 650M	120	+3.29 pp	0.028	0.909
ESM C 300M	120	+3.04 pp	0.041	0.853
SaProt 650M	120	+2.82 pp	0.027	0.907
ProtT5	120	+5.49 pp	0.036	0.870
ProtBERT	120	+8.68 pp	0.078	0.776
ProGen2-764M	117	+6.75 pp†	0.083	0.779

Per-task scoping and per-backbone remedy. The per-task downstream curves (secondary-structure Q3 and long-range contact probability versus window position) are flat, and the per-assay ProteinGym $\Delta\rho$ distribution is centered near zero, consistent with the consolidated scoping of Section 7. The training-free remedy of Section 8 reproduces on each backbone separately (ESM-2 and ESM C), with the centrality-weighted strategy recovering 68–80% of the gap at all tested window sizes.

Reproducibility. The benchmark ships as a fixed snapshot (seed 42, git revision, and package versions recorded in a manifest), with a family-disjoint MMseqs2 split and a single-callback evaluation harness; a new backbone is evaluated by supplying an `embed_fn`(seq)$\rightarrow(L, H)$ function.