BOUNDARY SIGNALS BIAS REFERENCE COMPETITION IN PROTEIN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein language models can predict a masked amino acid by copying from a similar segment in the same sequence. When two matching segments suggest different residues, what determines their relative support? We study this choice in the 600-million-parameter ESM Cambrian (ESM-C) model (ESMC600M) by changing the boundary position encoding used by one attention head. We block attention to exterior sequence while keeping amino acids and physical positions fixed. Interventions identify a contribution from boundary values that affects later reference keys. We use this route to predict score changes under new boundary interventions. On unseen sequence sources and layouts, the predictor reduces mean absolute error from 0.115 with static input features to 0.093. Removing the boundary contribution attenuates the effect more than norm-matched random removal and preserves copying when the references agree. Natural-input tests show that reduced boundary sensitivity can coexist with worse residue prediction. Boundary stability therefore needs to be evaluated alongside prediction accuracy.

1 INTRODUCTION

Protein language models learn to predict missing amino acid residues from the surrounding sequence. These predictions also provide scores used to estimate mutation effects (Rives et al., 2021; Meier et al., 2021; Notin et al., 2023). A protein can contain similar sequence segments, so another segment can supply a candidate for a missing residue. Prior work shows that models retrieve these candidates and identifies components supporting exact and approximate repeat completion (Kantroo et al., 2025; Pomerants et al., 2026). This raises a further question when two matching segments provide different candidates. How does the model decide how much to favor each source?

Finding a reference does not explain why its candidate receives more weight. Similarity, location, and residue preferences offer different explanations for this choice. We study conflicts in which two references match the target's local context but disagree at the masked position. In Figure 1, one supplies T and the other supplies S. The same matched context supports two completions.

Sequence boundaries offer one possible influence on this competition. Previous work shows that the beginning-of-sequence token can supply a positional prior for protein prediction (Jedryszek & Crook, 2026). Can boundary information also change how the model compares two reference residues elsewhere in the sequence? Cropping an input cannot isolate this effect because it changes both the available sequence and its boundaries. We instead keep the amino acids and their physical positions fixed. We block attention between the repeated core and its exterior sequence to exclude information from the flanks. We then change only the boundary position encodings used by one attention head.

Attention normalization creates an identification problem. Changing attention to the boundaries also changes how the head reads other positions. We separate these contributions exactly at the head output. In ESMC600M (EvolutionaryScale, 2024), interventions identify a boundary-value contribution that affects later keys at the reference positions. Those keys determine how the target matches the two references.

We fit a predictor from this route on development inputs and test it on unseen sources and layouts. Targeted removal reduces the effect and preserves copying when references agree. Natural-input experiments test the route in ESMC600M. We separately compare exterior-context buffering in

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.

ESMC600M, ESMC300M, and ESM2-150M. Fixed value replacement reduces boundary sensitivity but harms residue prediction, so sensitivity alone cannot establish useful stabilization.

Our contributions are as follows.

- In ESMC600M, injecting the boundary-value contribution recovers the boundary-induced score change more accurately than injecting the non-boundary contribution. Later reference-key interventions recover this effect more accurately than target-query interventions.
- We construct a predictor from this route and test it on unseen inputs. Its advantage over static sequence and geometry features shows what the internal account adds. Targeted removal tests the same contribution while concordant references test whether copying is preserved.

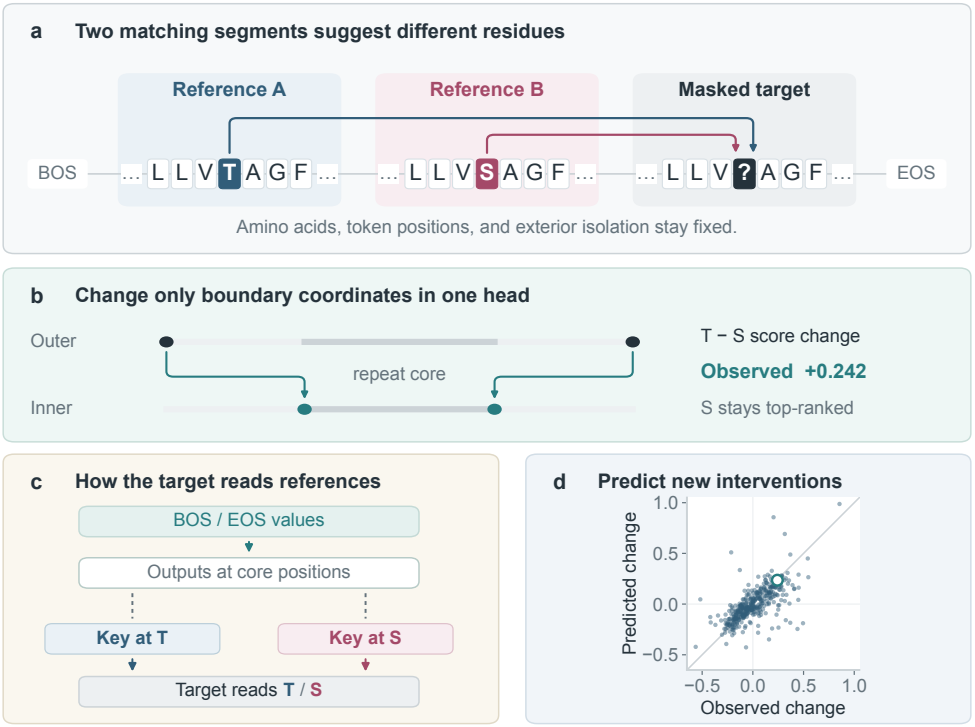

Figure 1: (a) Two matching references suggest different residues. (b) Changing boundary coordinates shifts their relative scores. (c) Interventions link boundary values to later reference keys. Dashed links mark unresolved intermediate computations. (d) Predicted and observed margin changes on held-out inputs. The circle marks the example and the diagonal indicates exact prediction.

2 Related Work

Protein representations and sequence context. Protein language models learn representations that support atomic structure prediction (Lin et al., 2023). Analyses of their attention reveal associations with residue contacts and functional sites (Vig et al., 2021; Rao et al., 2021a). These findings identify information present in model states. Explaining a particular prediction also requires tracing how the model uses that information. MSA Transformer explicitly receives aligned sequences and attends across them (Rao et al., 2021b). Our references are repeated segments within a single input sequence. Their competition occurs without a supplied multiple-sequence alignment.

From copying a residue to choosing a reference. In text models, induction heads connect repeated context with copying a previous continuation (Olsson et al., 2022). Kantroo et al. (2025)

show that repeated context can override sequence preferences and alter likelihood–fitness associations. Pomerants et al. (2026) identify positional heads, biochemical features, and induction heads supporting protein repeat completion. Their approximate-repeat task masks positions whose residue agrees with the aligned reference. We use their ESM-C induction-head cluster as a fixed reader set. Our question concerns disagreement between two aligned references after their surrounding context has matched. Evidence that a component copies an answer does not by itself explain how that answer acquired its relative weight.

Boundary values and positional computation. Rotary encoding makes attention scores depend on relative positions (Su et al., 2024). Initial tokens can attract attention despite limited semantic content (Xiao et al., 2024). In ESM2-8M, Jedryszek & Crook (2026) trace a BOS-derived methionine signal, test its values causally, and study upstream query composition and rotary geometry. They also separate norm and direction effects. Our contribution concerns a different computational role. We trace boundary-derived information to later keys at reference amino acids and predict its effect on competition between content sources.

Testing a mechanistic account. Activation patching tests whether selected components recover behavior (Wang et al., 2023; Conmy et al., 2023), but its interpretation depends on the metric and intervention (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024; Miller et al., 2024). Component overlap alone does not establish functional equivalence (Hanna et al., 2024). Self-repair and mediator interactions also limit additive accounts of patch effects (McGrath et al., 2023; Vaidyanathan et al., 2026). Gradient methods such as AtP* approximate intervention effects for efficient localization (Kramár et al., 2024). We combine a conditional derivative with a boundary-value decomposition to predict a specified perturbation on new inputs.

3 METHOD

3.1 SEPARATE BOUNDARY GEOMETRY FROM SEQUENCE EVIDENCE

A constructed core contains two reference windows and one target window. Each window has a central residue and eight residues on either side. Both references match the target context, but their centers contain different amino acids, a and b. The target center is masked. We measure their logit margin, $m = \ell_a - \ell_b$. A positive change increases support for the first candidate relative to the second. This measures competition even when the highest-scoring residue stays the same.

Exterior sequence flanks surround the core. To isolate position from information in those flanks, we place the core and BOS/EOS in one attention group. The flanks belong to another group. Attention between the groups is blocked in every layer. BOS and EOS are the start and end tokens. Core tokens can attend to both. Visible exterior context is examined separately in Section 5.4.

We compare the actual sequence boundaries with coordinates placed just outside the repeated core. The tokens themselves do not move. Let N be the number of amino acids. BOS and EOS occupy physical token indices 0 and $N + 1$, and the core occupies indices s through e. The two rotary-coordinate settings are

$$(p_{\mathrm{BOS}}, p_{\mathrm{EOS}}) = \begin{cases} (0, N+1), & \text{outer,} \\ (s-1, e+1), & \text{inner.} \end{cases} \tag{1}$$

The intervention changes only the BOS/EOS query and key coordinates in head 15.5. Layer and head indices are zero-based. Every other coordinate stays fixed. Queries and keys determine which positions attend to each other. Let q_t be the query vector at token t and k_j the key vector at token j. Their attention score is

$$s_{tj} = q_t^\top R(p_t)^\top R(p_j) k_j / \sqrt{d_h}, \qquad P_{tj} = \mathrm{softmax}_j(s_{tj}), \tag{2}$$

where R is the rotary transform and d_h is the head dimension (Vaswani et al., 2017; Su et al., 2024). Softmax normalizes these scores across all visible positions. Changing boundary coordinates can alter both the attention paid to the boundaries and the attention paid to other sequence positions. The next decomposition separates the resulting changes in the information read by the head.

3.2 TRACE BOUNDARY VALUES TO REFERENCE KEYS

To determine which part of the altered reading affects reference competition, we intervene on the two contributions separately. Head 15.5 was selected on development inputs and fixed before the later tests. We call it the writer because we trace the change introduced by its output.

Each replacement compares two runs of the same input under the two boundary coordinate settings. The donor supplies a component to the recipient run. We resume that run and compare its margin with the donor margin. Recovery error is their absolute difference. An outer-to-inner test uses the outer run as recipient and the inner run as donor. The reverse test exchanges their roles. Both runs have identical value vectors V entering the writer because the intervention changes only this head's boundary coordinates. Values carry the information read from each attended position. Its output change before projection has the exact decomposition

$$\Delta h_t = \underbrace{\sum_{j\in\mathcal{B}} \Delta P_{tj} V_j}_{\Delta h_{B,t}} + \underbrace{\sum_{j\notin\mathcal{B}} \Delta P_{tj} V_j}_{\Delta h_{\neg B,t}}, \qquad \mathcal{B} = \{\mathrm{BOS}, \mathrm{EOS}\}. \tag{3}$$

Here ΔP is donor attention minus recipient attention. The first term records changed reading of boundary values. The second records redistribution over other values. We inject a term into the recipient to test whether it reproduces the donor margin. Removing it from the donor tests its contribution to the geometry effect. Boundary values can be read at every query position, so the boundary term collects changes across the writer's output. Replacements restricted to the reference centers or remaining core distinguish the values' origin from their entry into later computation. The decomposition is exact at the writer. Later nonlinear computations can make the terms' margin effects nonadditive.

In a later head, the target query matches reference keys to determine how strongly to read each reference. We use the ESM-C induction-head cluster C1 from Pomerants et al. (2026) as the reader set. Replacing keys at the two reference centers tests the reference side of this comparison. Replacing the query at the masked position tests the target side. Replacements occur after Q/K normalization and before rotary encoding. Restoring recipient keys inside the donor run tests whether those keys carry the effect. Only readers later than the writer can mediate it. Earlier-reader clamps serve as identity controls. Appendix A lists all heads.

We then test which properties of the keys carry the score change. For recipient and donor keys k^r and k^d, the direction and norm replacements are

$$k^{\mathrm{dir}} = \|k^r\| \frac{k^d}{\|k^d\|}, \qquad k^{\mathrm{norm}} = \|k^d\| \frac{k^r}{\|k^r\|}. \tag{4}$$

The first retains recipient magnitude and the second retains recipient direction. Recombining direction and norm recovers the complete donor key.

To test whether the signal requires input-specific values, we replace V_j with development means μ_j. Each input retains its own attention change ΔP, so the contribution varies despite shared source vectors. A residual control uses $V_j - \mu_j$ to test information omitted by the mean. Means fitted on other source families test sensitivity to development-source choice. Norm-matched random vectors control for intervention size. These comparisons separate common values from input-specific routing and content.

3.3 PREDICT AN INTERVENTION FROM THE BASELINE NETWORK

To predict new interventions, we measure how the baseline margin responds to small writer-output changes. Let g be its gradient with respect to the entire writer output. We then fix the reference keys at their baseline values and obtain the gradient g_{K0}. The difference isolates the local response removed by fixing those keys. The clamp leaves the baseline prediction unchanged, so $g - g_{K0}$ compares responses around the same output.

Shared development values give $\Delta h_\mu = \sum_{j\in\mathcal{B}} \Delta P_j \mu_j$, with query positions collected into the same tensor as h. Baseline writer Q/K supplies attention probabilities for both geometries. We predict the margin change as

$$\widehat{\Delta m}_{K,\mu} = \alpha_K \, (g - g_{K0})^\top \Delta h_\mu. \tag{5}$$

We fit α_K through the origin by minimizing squared development error. The writer, reader set, two mean value vectors, and scalar are then fixed. For each new input, the baseline network supplies Q/K vectors and derivatives. Recomputing writer attention under the alternative coordinates supplies ΔP without running the perturbed downstream network.

The gradient contrast measures the sensitivity removed by the joint key clamp, including interactions among the clamped readers. This conditional response around the baseline supplies a numerical prediction for the finite coordinate intervention. Its accuracy tests how well the local derivative describes that intervention.

4 Experimental Setup

4.1 New sources test the proposed route

We scan layers 12–23 by changing boundary Q/K coordinates in one layer or all except one. On development inputs, layers 15, 16, and 18 rank highest in donor-margin recovery and its loss on omission. Both rankings average the two directions. Within these layers, recovery selects head 15.5.

Development sources come from the repeat dataset of Pomerants et al. (2026). The prediction pool excludes its components and those of the two 128-source Swiss-Prot natural-context cohorts. The sequence graph uses 30% identity and 80% coverage (Teufel et al., 2023). Table 1 summarizes the cohorts. Appendix A gives their layouts and source rules.

Table 1: Source cohorts and their roles in the constructed-input experiments.

Cohort	Sources and conditions	Role
Development	8 inspected families, 32 conditions	Fit slopes and ridge. Estimate shared boundary means from these families.
Writer-path tests	32 families, two directions and two flank lengths	Test value injection and Q/K routes.
Held-out test	48 Swiss-Prot components. Four layouts vary order, flanks, and gaps. Both residue orientations.	Test prediction, removal, and concordant copying.

4.2 Prediction, attenuation, and copying have separate tests

Prediction error compares the predicted and measured donor-minus-recipient margin changes. We compare the main predictor with zero change and a static ridge model. Ridge uses geometry, reference similarity, and amino-acid contrasts. We standardize features on development data and fix the ridge penalty at one without hyperparameter search. Unrestricted-gradient predictors test what the selected route omits. They use shared boundary values, input-specific boundary values, or the complete writer change, each with its own development-fitted scalar. These controls separate the restriction to reference keys from the shared-value approximation.

For attenuation, we remove the boundary contribution or restore baseline reference keys. The residual effect is the absolute margin difference from the unmodified recipient. Random removal tensors match the boundary contribution's norm at each writer-output position. Concordant-reference controls give both references the same central residue and test whether copying is preserved. Four primary comparisons jointly test prediction against static and zero baselines, attenuation against random removal, and newly broken copying below 10%. Inference averages repeated conditions within each source and applies Holm correction. Figures 2–4 show means and pointwise 95% source-bootstrap intervals. Appendix A gives source selection, layouts, and statistical procedures.

4.3 Natural context tests the scope of the mechanism

Natural conflicts use non-overlapping 17-residue windows whose references provide different central residues. Both match at least half the target context and have comparable similarity. Exterior residues remain visible. One cohort tests a development-selected eight-head set and the single writer.

A source-separated cohort tests exterior-context buffering in ESMC600M, ESMC300M, and ESM2-150M. Boundary sensitivity is the mean absolute change in the native-versus-competing-residue logit margin between the outer and inner boundary coordinates. Buffering means lower sensitivity. A separate comparison evaluates native versus shuffled exterior order using native-residue log probability.

Background-value replacement changes exterior values at every layer and head. Keys and attention remain input dependent. The common template averages development values. The affine offset $b = W_V\beta + b_V$ combines the LayerNorm offset β with value-projection parameters. Native accuracy is the fraction of targets whose top-ranked prediction matches the original residue, measured at the outer coordinates. Figure 5 reports replacement results with exterior order shuffled. Its intervals use 20,000 paired bootstrap draws over the 128 independent source components.

We also test whether exterior references affect scores used for functional evaluation. ProteinGym provides the assay sequences and mutation measurements (Notin et al., 2023). The GB1 assay measures binding to IgG (Olson et al., 2014). At each target position, we mask the residue and score a substitution by its log probability minus that of the wild type. Full-input, domain-only, and selected-reference conditions use the same assayed target positions. Position-bootstrap draws keep variants at the same position together. Additional proteins test how the context effect varies across sequences.

5 Results

5.1 Boundary values change reference competition

Boundary coordinates change relative scores while the competing residues and their positions stay fixed. Figure 1 illustrates increased support for T while S remains the top prediction.

For outer-to-inner transfer with 32-residue flanks, boundary-term injection reduces mean donor-margin error from 0.244 to 0.052 logits. Non-boundary injection leaves 0.200 (Figure 2). Reference-key injection gives 0.073, whereas target-query injection gives 0.274. The same ordering holds with 96-residue flanks and in the reverse direction. Boundary-derived information thus reaches the reference comparison through later keys.

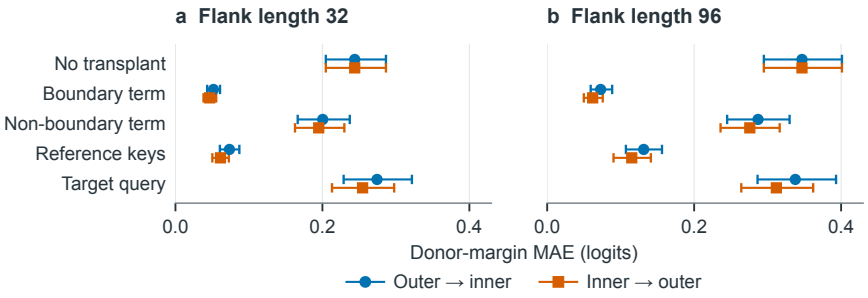

Figure 2: Donor-margin recovery with flank lengths of (a) 32 and (b) 96 residues. Boundary-value and reference-key interventions outperform the non-boundary and target-query alternatives. Bars show pointwise 95% source-bootstrap intervals.

In the same outer-to-inner condition, direction-only key replacement gives 0.101-logit reconstruction error versus 0.262 for norm-only replacement. Shared development-fitted boundary values give 0.062, close to the actual boundary term's 0.052, whereas their input-specific residual gives 0.237 (Figure 3). The reverse direction has the same ordering. A common boundary signal can therefore alter competition through changes in reference-key direction.

The writer change acts beyond the reference centers. For short-flank outer-to-inner transfer, replacing writer outputs only at those centers leaves 0.243-logit error, against a 0.244 baseline. Replacing other core positions reduces it to 0.093 even after excluding both reference centers and the masked target. Longer flanks and reverse transfer give the same ordering (Appendix D.1).

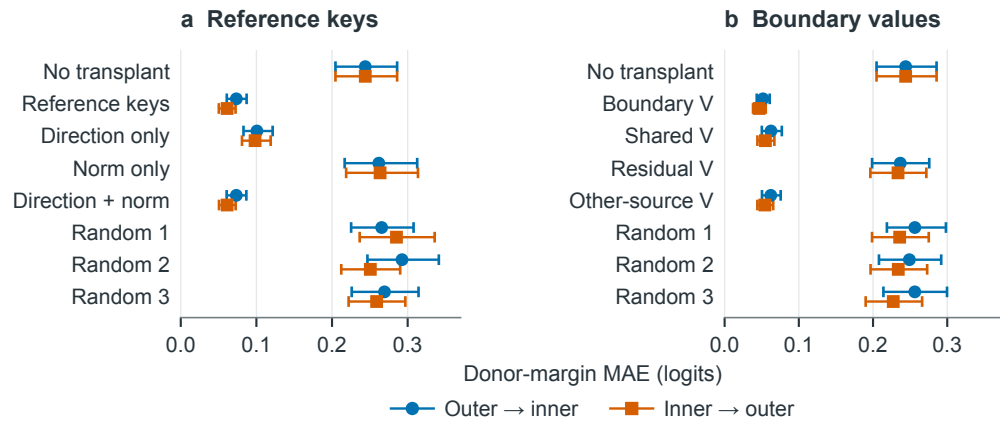

Figure 3: (a) Reference-key direction recovers the donor margin more accurately than key magnitude. (b) Shared boundary values outperform residual values. Both panels compare the two transfer directions. Bars show pointwise 95% source-bootstrap intervals.

5.2 The route predicts new boundary interventions

On held-out inputs, the pathway predictor has mean absolute error 0.093 logits, compared with 0.115 for static features and 0.145 for zero change (Figure 4). Its paired advantage over static features is 0.022 logits (95% source-bootstrap interval [0.010, 0.035]). Both primary comparisons survive Holm correction. The internal response adds predictive information beyond sequence identity and geometry.

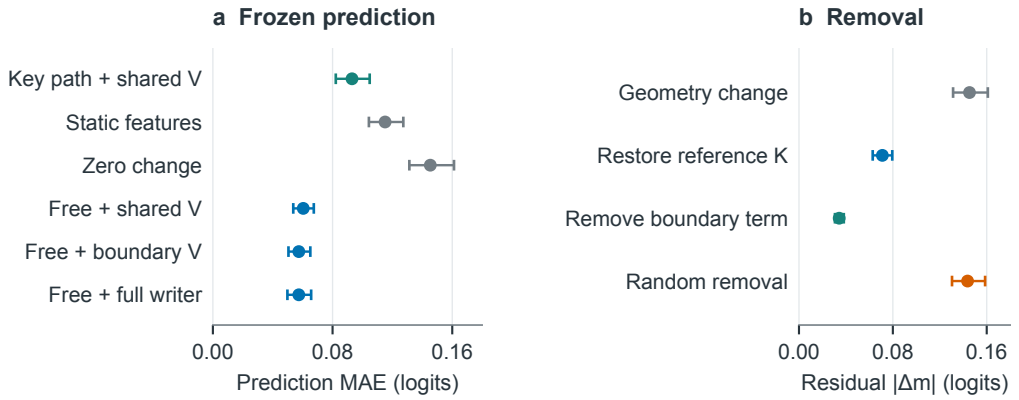

Figure 4: (a) Frozen prediction on held-out sources. (b) Residual margin change after restoring reference keys or removing the boundary contribution. Bars show pointwise 95% source-bootstrap intervals.

Removing the key-path restriction lowers prediction MAE from 0.093 to 0.060 with the same shared values. Input-specific boundary values lower it further to 0.057. Predictive information therefore remains outside the selected keys and in each input's boundary values. A quarter-step full-writer perturbation also lowers Taylor error, indicating curvature over the full intervention (Appendix E).

The pathway's largest error is 0.163 logits when the target lies between references with long flanks. With short flanks and the target after both references, static features slightly outperform the pathway (0.097 versus 0.099). Geometry therefore affects the predictive value of the route (Figure 8).

5.3 Removing the contribution reduces sensitivity and preserves copying

Boundary removal reduces the mean absolute perturbation from 0.145 to 0.034 logits, compared with 0.144 after matched random removal (Figure 4). The paired advantage over random removal is 0.109 logits (95% interval [0.099, 0.121]). Reference-key restoration leaves 0.071. Both interventions attenuate the identified contribution, with greater reduction at the upstream boundary term.

When both references support the same residue, copying remains correct under the coordinate change and boundary removal. With no new errors in 48 independent source controls, the exact one-sided 95% binomial upper bound is $1 - 0.05^{1/48} = 6.05\%$. It lies below the prespecified 10% threshold (Appendix A). Attenuation also holds in groups split by a development-frozen threshold on predicted effect strength.

5.4 Natural context changes the route's influence

The eight-head set reduces outer-to-inner donor-margin error from 0.741 to 0.262 logits on constructed inputs, but only from 0.179 to 0.147 on natural conflicts. Random head sets reach 0.151–0.173 on the natural inputs (Figure 5). The selected route accounts for much less of the natural-input response.

In the source-separated cohort, shuffled exterior context buffers boundary perturbations in all three models. A separate test finds no significant native-order advantage in native-residue log probability after joint correction. Buffering does not require the native exterior order.

Lower sensitivity alone does not establish useful stabilization. In ESMC600M, replacing background values by the affine offset lowers sensitivity from 0.180 to 0.160 logits relative to the common template, but native-residue accuracy falls from 87/128 to 55/128 (Figure 5). The offset and the remaining template contribution can oppose each other. Their balance matters for residue prediction as well as boundary sensitivity (Appendix F).

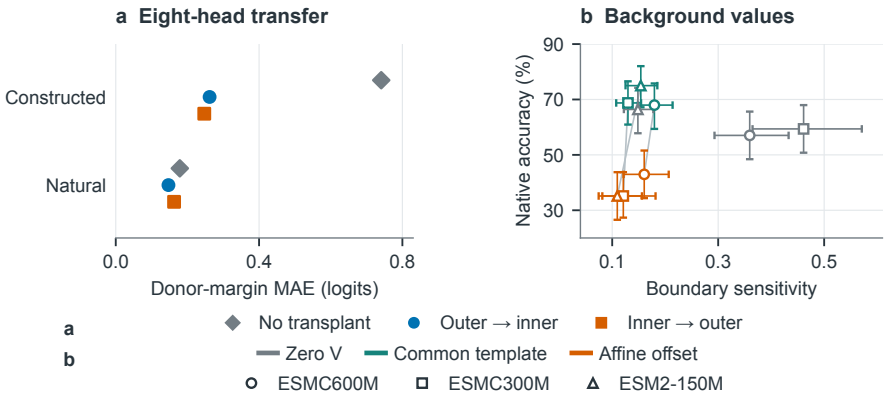

Figure 5: (a) Eight-head transfer on constructed and natural inputs. (b) Background-value replacement with shuffled exterior context. Affine offsets reduce sensitivity but harm native-residue accuracy. Horizontal and vertical bars give pointwise 95% source-bootstrap intervals for sensitivity and accuracy.

5.5 Reference context changes functional scores

For ESMC600M, the Spearman correlation with GB1 single-mutant binding falls from 0.275 with the full input to 0.014 with the domain alone (Figure 6). Retaining the closest sequence-matched exterior reference preserves a correlation of 0.276 even after shuffling the remaining exterior. Scrambling that reference gives -0.070; retaining a random window gives -0.053. Reference selection uses sequence identity alone. These controls connect the binding association to a specific exterior repeat.

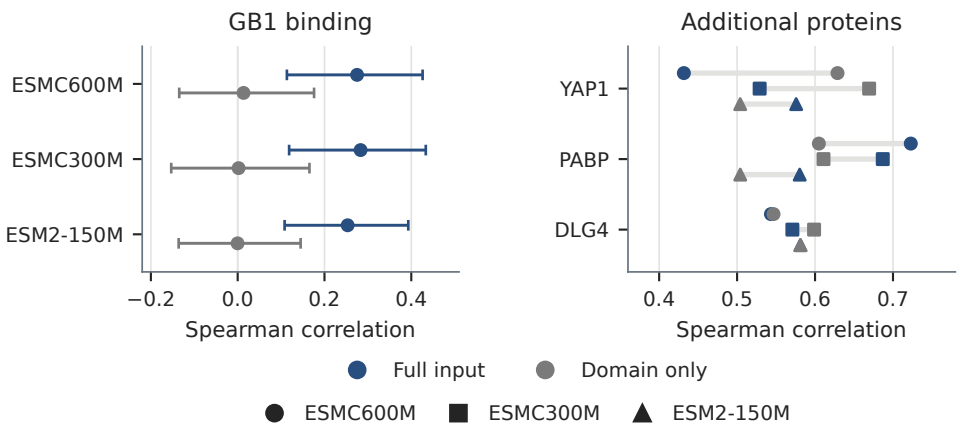

Figure 6: Exterior references support the GB1 binding association and have different effects across additional proteins.

Changing the aligned reference residue redirects most GB1 masked predictions. For mutation $w \to a$, the parameter-free score $\mathbf{1}[a = r_t] - \mathbf{1}[w = r_t]$ uses the aligned reference residue r_t. Its single-mutant binding correlation is 0.324. A paired comparison with the full-input model does not establish an ordering. Reference agreement thus captures much of this association. Full-input versus cropped effects differ across YAP1, PABP, and DLG4, so the near-deterministic copying in GB1 does not extend to every protein. Appendix B.4 gives the controls and assay mapping.

6 Discussion

Explaining copying requires accounting for the relative weight of competing references. Here, a common boundary signal changes that weight through later reference keys. Input-dependent attention routes the shared values, and key direction carries more of the response than key magnitude.

The same boundary contribution supports prediction and causal removal tests. Unrestricted gradients reveal predictive information outside the selected keys.

Visible exterior context buffers positional changes. Fixed value replacement can also reduce these changes while harming prediction. Evaluating residue accuracy distinguishes useful stabilization from lost sequence information.

Scope. The identified route concerns a single-head intervention in ESMC600M with exterior attention blocked. Its predictor is local, and key-direction replacement changes complete reference keys. Natural-context experiments test its extension beyond isolated repeats.

7 Conclusion

Boundary-derived information changes reference competition through later keys in ESMC600M. A local predictor forecasts unseen interventions. Targeted removal attenuates the effect and preserves concordant copying. Natural-input results show why reducing boundary sensitivity must be assessed together with residue accuracy.

AI USE STATEMENT

Generative AI tools assisted manuscript writing and translation.

REPRODUCIBILITY STATEMENT

Appendix A specifies input construction, interventions, and numerical checks. Appendix C describes the model and software environment.

REFERENCES

Carlos L. Araya, Douglas M. Fowler, Wentao Chen, Ike Muniez, Jeffery W. Kelly, and Stanley Fields. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proceedings of the National Academy of Sciences*, 109 (42):16858–16863, 2012. doi: 10.1073/pnas.1209751109.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards Automated Circuit Discovery for Mechanistic Interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352, 2023. doi: 10.52202/075280-0719.

EvolutionaryScale. ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning. EvolutionaryScale blog, December 4, 2024.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have Faith in Faithfulness: Going Beyond Circuit Overlap When Finding Model Mechanisms. In *Conference on Language Modeling*, 2024.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. arXiv:2404.15255, 2024.

Piotr Jedryszek and Oliver M. Crook. Retrieval and competition: how a protein foundation model starts a protein. arXiv:2605.16331v2, 2026.

Pranav Kantroo, Günter P. Wagner, and Benjamin B. Machta. In-Context Learning can distort the relationship between sequence likelihoods and biological fitness. arXiv:2504.17068, 2025.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components. arXiv:2403.00745, 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The Hydra Effect: Emergent Self-repair in Language Model Computations. arXiv:2307.15771, 2023.

Richard N. McLaughlin, Jr., Frank J. Poelwijk, Arjun Raman, Walraj S. Gosal, and Rama Ranganathan. The spatial architecture of protein function and adaptation. *Nature*, 491(7422):138–142, 2012. doi: 10.1038/nature11500.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29287–29303, 2021.

Daniel Melamed, David L. Young, Caitlin E. Gamble, Christina R. Miller, and Stanley Fields. Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA*, 19(11):1537–1551, 2013. doi: 10.1261/rna.040709.113.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer Circuit Faithfulness Metrics are not Robust. In *Conference on Language Modeling*, 2024.

Pascal Notin, Aaron W. Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, et al. ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64331–64379, 2023. doi: 10.52202/075280-2810.

C. Anders Olson, Nicholas C. Wu, and Ren Sun. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Current Biology*, 24(22):2643–2651, 2014. doi: 10.1016/j.cub.2014.09.072.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, et al. In-context Learning and Induction Heads. arXiv:2209.11895, 2022.

Gal Pomerants, Yaniv Nikankin, Anja Reusch, Tomer Tsaban, Ora Schueler-Furman, and Yonatan Belinkov. Induction Meets Biology: Mechanisms of Repeat Detection in Protein Language Models. In *International Conference on Machine Learning*, 2026.

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*, 2021a.

Roshan M. Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA Transformer. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8844–8856. PMLR, 2021b.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, et al. 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.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Felix Teufel, Magnús Halldór Gíslason, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Ole Winther, and Henrik Nielsen. GraphPart: homology partitioning for biological sequence analysis. *NAR Genomics and Bioinformatics*, 5(4):lqad088, 2023. doi: 10.1093/nargab/lqad088.

Sankaran Vaidyanathan, David Arbour, Aaron Mueller, Scott Niekum, and David Jensen. The Curse of Multiple Mediators: Hidden Interaction Effects in Activation Patching. arXiv:2606.27510, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small. In *International Conference on Learning Representations*, 2023.

Nicholas C. Wu, Lei Dai, C. Anders Olson, James O. Lloyd-Smith, and Ren Sun. Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife*, 5:e16965, 2016. doi: 10.7554/eLife.16965.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient Streaming Language Models with Attention Sinks. In *International Conference on Learning Representations*, pp. 21875–21895, 2024.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *International Conference on Learning Representations*, pp. 1651–1678, 2024.

A PREDICTION TASK AND STATISTICAL ANALYSIS

A.1 SOURCE SELECTION AND CONSTRUCTION

The held-out source pool uses the sequence-only Swiss-Prot candidates. Components connected to the original-study data or to the development and natural-context cohorts are excluded using the 30%-identity/80%-coverage graph. The first 48 remaining components in frozen source-rank order with sequence length at least 224 are retained. Each source contributes one target-derived 17-residue pattern. Its existing candidate supplies the two conflicting reference letters. No model response participates in the selection. Constructed cores are unique across the selected components.

The four fixed layouts are shown below. A and B denote reference windows. T denotes the masked target. Both orientations of the two conflicting central letters are evaluated. The exterior flanks come from the source ends but are isolated from the core through sequence-ID attention masking. BOS/EOS remain in the visible core group. Only rotary boundary coordinates change in writer head 15.5.

Table 2: Held-out configuration blocks. Each has 48 components and two letter orientations.

Configuration	Window order	Flank length	Inter-window gap
1	ABT	48	7
2	TAB	48	29
3	ATB	112	7
4	BTA	112	29

The development set contains 32 conditions from eight source families with flank length 64 and gap 17, using ABT/TAB layouts and both letter orientations. It determines the four through-origin slopes and the static baseline. The shared boundary-value means are estimated separately from these eight development families. Test sources and their model responses enter neither fit. Source selection and fitted quantities are fixed before test evaluation.

A.2 EXAMPLE SEQUENCE AND COORDINATE MAPPING

The sequence below is the example in Figure 1. The central residues in the two references are T and S. The target is replaced by the model's mask token. Positions in the table use one-based amino-acid indexing, equivalently model-token indexing after the leading BOS token. The core attention group retains BOS and EOS while isolating the exterior flanks. Figure 1 shows the actual seven-residue windows centered at these positions.

```
MKHKFLVSFLLGLTISFAGAQVIAKTPSVNNAPVARSVNQSPIQLEVISQDGKLLVTAGFDGTARGSGGSGGSQDGKLLVSAGFDGTARGSGG⌋
↪  SGGSQDGKLLV<mask>AGFDGTARLRVWNLQGQLLGQYNPNLERLGSVSFHPNGKNLAIAAYHGTIILLELQ
```

Table 3: Predicted and observed margin changes for the Figure 1 example.

Quantity	Value
Source component	P73595
Reference positions / target	57 / 81 / 105
Reference letters	T / S
Predicted margin change	0.23549
Observed margin change	0.24159
Baseline / changed-geometry top residue	S / S

The margin $\ell_T - \ell_S$ increases by 0.242 logits under the boundary change, while S remains the top prediction. The sequence and physical token positions are unchanged. Only the BOS/EOS rotary coordinates in writer head 15.5 move.

A.3 Exact predictor and intervention implementation

The reference-reader set follows ESM-C induction-head cluster C1 from Pomerants et al. (2026), fixed at heads 12.2, 13.5, 17.4, 17.12, 19.8, 19.14, 19.16, 21.0, 28.3, 29.13, 30.12 and 35.16. Only heads later than writer 15.5 can mediate its perturbation. Earlier baseline clamps are identity controls. The gradient difference in Eq. 5 is conditional on this jointly clamped network. Interactions among the clamped readers are included in this derivative contrast.

Attention coefficients for both boundary geometries are computed from baseline writer Q/K, and all four predictor deltas use baseline V or its frozen mean. Predictions are computed before running the perturbed network and use derivatives of the baseline model. Its comparison with static geometry measures the added predictive information from that internal response.

The static ridge model uses 11 geometry/identity features plus 20 amino-acid contrasts. Features are standardized by development means and standard deviations. Constant scales are set to one. Ridge minimizes mean squared error plus unit squared coefficient norm. The intercept is fitted separately. Mechanistic slopes minimize squared error through the origin. No test-set tuning or model selection occurs. A development median of absolute main predictions defines strong/weak groups for descriptive diagnostics only.

Actual conditions include unmodified baseline, single-writer geometry change, geometry with its boundary term removed, geometry with reference keys restored, baseline with the shared term added, and three random removal controls. Each random tensor matches the boundary-term norm per writer-output position. A quarter-step full-writer perturbation diagnoses local linearity and does not replace the primary full-step forecast. One concordant-reference control per source uses configuration 1 and the first orientation.

A.4 Prediction error and statistical comparisons

Table 4: Prediction and preservation comparisons with Holm-adjusted p-values.

Hypothesis	Effect / pointwise 95% interval	Adjusted p
Pathway MAE below static	0.02211 [0.01002, 0.03490]	7.80×10^{-4}
Pathway MAE below zero	0.05233 [0.03767, 0.06838]	2.00×10^{-5}
Boundary removal exceeds random	0.10940 [0.09859, 0.12128]	2.00×10^{-5}
New copy errors below 10%	0/48; upper bound 6.05%	0.00636

The first three intervals are pointwise source-component bootstrap intervals for mean paired differences. The copying bound is one-sided and exact. All four primary tests use one Holm family. Two-sided sign-flip tests require the predicted effect direction and assume symmetry. Exact binomial inference evaluates the copying bound. Layouts and residue orientations remain grouped within each source component.

Table 5: Prediction error depends on the reference layout and exterior geometry.

Predictor	1	2	3	4
Reference-key pathway, shared V	.0991	.0447	.1627	.0654
Static ridge	.0974	.0848	.1690	.1093
Zero	.1728	.1405	.1670	.1010
Free gradient, shared V	.0643	.0329	.1137	.0308
Free gradient, boundary V	.0564	.0310	.1140	.0283
Free gradient, full writer	.0609	.0265	.1180	.0243

The pathway MAE is highest for configuration 3, where the target lies between the references with long flanks. Configuration 1 slightly favors the static baseline. Development slopes are 1.52887 for the key pathway, 1.22141 for unrestricted/shared values, 1.16037 for unrestricted/boundary values, and 0.94904 for unrestricted/full-writer changes.

A development-derived threshold separates strong and weak predicted effects. Their mean absolute perturbations are 0.231 and 0.117 logits; boundary removal leaves 0.056 and 0.027. Primary inference weights sources equally and retains all layouts and orientations within each source.

B NATURAL CONTEXT AND FUNCTIONAL EVALUATION

B.1 EIGHT-HEAD TRANSFER AND ITS RANDOM CONTROLS

The eight-head set comprises 15.5, 16.11, 16.3, 18.13, 18.14, 18.6, 18.17, and 18.1, with zero-based layer and head indices. This transfer test has eight constructed source families and 32 conditions per direction, plus 126 natural sources and 126 conditions per direction. These are distinct from the 32 frozen families used for the writer-path experiment in Figure 2. The matched random-head sets compare the selected route with equally sized alternatives.

Table 6: Donor-margin errors for eight-head transfer and every random-head control.

Cohort	Direction	Original	Eight heads	Random 1	Random 2	Random 3
Constructed	Outer to inner	0.7410	0.2619	0.7452	0.7465	0.7253
Constructed	Inner to outer	0.7410	0.2471	0.7232	0.7544	0.6858
Natural	Outer to inner	0.1788	0.1471	0.1525	0.1735	0.1509
Natural	Inner to outer	0.1788	0.1628	0.1735	0.1807	0.1711

B.2 NATURAL-SEQUENCE SELECTION

From reviewed, repeat-annotated Swiss-Prot sequences of length 150–1022, fixed accession-hash ordering selects 4,096 for repeat-window search. Seventeen-residue windows must have at least 50% similarity excluding the central residue, comparable match quality to both references, no overlap, sufficient sequence entropy and removable exterior context. The original cohort uses a 50%-identity graph, and the more strictly separated cohort uses a 30%-identity graph. These thresholds define the source components used for separation and uncertainty estimation.

The natural test cohort excludes components touching development or original-study sources and retains one representative per selected component. Component bootstrap resampling keeps related sequences together. Native target labels are the residues in the observed sequences.

B.3 COMMON BACKGROUND VALUES

The common background prototype averages development-cohort values with equal weights over its 81 strict components. It has no held-out natural inputs or assay labels in fitting. At every layer and head, replacement leaves current-sequence K and other computation input dependent. Zero, norm-matched random, sequence-mean, shuffled-value and layer-range replacements provide comparison controls.

The prototype decomposes into $b + (\mu - b)$, where $b = W_V\beta + b_V$. In ESMC, the projection can have no explicit bias while LayerNorm beta still contributes an affine offset. Bias/prototype median norm ratios are 2.69/3.26/1.37 across the three models, with negative median cosines. Removing one of two opposing terms need not be a benign information deletion. The affine offset explains how a low-sensitivity intervention can remove a contribution needed for native-residue prediction.

B.4 GB1 REFERENCE CONTROLS AND ASSAY MAPPING

ProteinGym-v1 supplies the fixed assay data and input sequences (Notin et al., 2023). The GB1 metadata input has 448 residues. The Olson assay measures binding to IgG over positions 228–282 of that input (Olson et al., 2014). This 55-residue region excludes the initiator methionine. A 56-residue crop with added M tests sensitivity to that input choice. Figure 6 compares full-input and domain-only scoring.

We select the GB1 reference as the highest-identity non-overlapping ungapped window of the assay-domain length. Ties are resolved by the earliest position. This exterior repeat matches 48 of the 55 aligned domain residues. Three random windows have the same length and are disjoint from the assay core. Selection uses only the sequence. Retaining the reference while shuffling other exterior content preserves much of the binding correlation. Scrambling the reference or retaining a random window does not reproduce it (Figure 7).

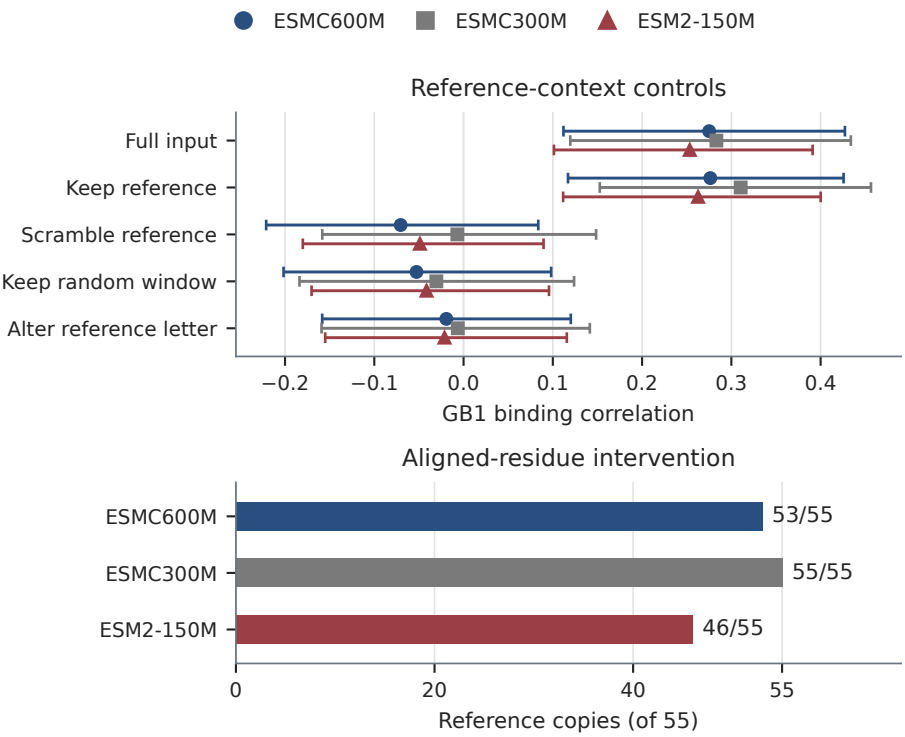

Figure 7: Sequence-selected reference controls connect GB1 assay association to exterior context and show strong copying after changing the aligned reference residue.

At each assayed target position, the target residue is masked. The score for substituting a for wild-type w is $\log p(a \mid x_{\setminus t}) - \log p(w \mid x_{\setminus t})$. All mutants at that position share the masked context. Position-bootstrap draws resample positions with all their variants together. The parameter-free reference indicator is $\mathbf{1}[a = r_t] - \mathbf{1}[w = r_t]$. Here r_t is the aligned exterior residue. Its Olson single-mutant correlation is .324 with a pointwise 95% interval [.121,.480]. The paired interval comparing it with the full-input model includes either ordering. Paired comparisons use the same 4,995 bootstrap draws with nonconstant indicators.

Multi-mutant scores sum constituent single-site contributions. The reconstructed full Olson population has 536,962 variants and the measured Wu population has 149,360 (Wu et al., 2016). Within Wu mutation-count strata, the indicator is constant. Its pooled correlation captures between-stratum association. These additive scores do not model epistasis, and the two GB1 assays do not provide independent protein replication. The measured Wu population is also distinct from the full combinatorial possibility space.

The additional sequence-selected cases use YAP1 WW (Araya et al., 2012), PABP (Melamed et al., 2013), and DLG4 (McLaughlin et al., 2012). Their assay regions cover 33, 75, and 83 positions with 362, 1,187, and 1,576 measured variants. Continuous reference effects survive composition controls. Near-deterministic copying of an altered reference residue generally does not. Full-input/domain-only single-mutant Spearman correlations are shown below.

Protein	ESMC600M	ESMC300M	ESM2-150M
YAP1	.432/.629	.529/.669	.576/.504
PABP	.723/.605	.687/.611	.580/.504
DLG4	.544/.547	.571/.599	.581/.581

Position-bootstrap intervals measure variation among assayed positions within each protein. The differing full-input and cropped correlations show that reference context affects functional evaluation differently across proteins.

Physical swaps of artificial references preserve whole-sequence composition and the target domain but also move other content. A virtual shift of only the reference window can duplicate or skip rotary coordinates. A complete coordinate permutation reproduces physical permutation up to numerical error. Distinguishing these operators separates reference relocation from a change in position encoding alone.

C Model and numerical settings

We use `ESMC_600M_202412` in evaluation mode and float32. TF32 is disabled. Gradients are taken with respect to the selected writer output while model parameters remain fixed. The environment uses PyTorch 2.8.0+cu128, Transformers 4.46.3, and `esm` 3.1.1 on an NVIDIA A800 with 80 GB memory.

D Decomposition, key geometry, and shared values

D.1 The boundary decomposition is exact at one writer

Section 3.2 defines the boundary decomposition and distinguishes source-value positions from output positions. The table below reports the complete positional comparison in the frozen writer-path cohort. Lower error means closer donor reconstruction. The other-core condition excludes both reference centers and the masked target. Complete writer replacement at all positions exactly replays the donor in these runs.

Table 7: Donor-margin error after replacing the complete writer output at selected positions.

Flank	Direction	Baseline	Reference centers	Other core
32	Outer to inner	0.244	0.243	0.093
32	Inner to outer	0.244	0.233	0.093
96	Outer to inner	0.347	0.357	0.111
96	Inner to outer	0.347	0.351	0.105

D.2 The shared boundary value retains input-dependent routing

The shared-value and residual terms sum to the boundary contribution at the writer. Subsequent nonlinear computation can make their separately injected effects nonadditive.

E Prediction error across sources and interventions

E.1 A decomposition of prediction error

Write $g_K = g - g_{K0}$ and let R be the finite-step Taylor remainder of the full network around the baseline writer output. Then the main prediction error admits the algebraic expansion

$$\begin{aligned}\Delta m - \alpha_K g_K^\top \Delta h_\mu = g_K^\top(\Delta h_B - \Delta h_\mu) + g_{K0}^\top \Delta h_B + g^\top \Delta h_{\neg B} \\ + (1-\alpha_K) g_K^\top \Delta h_\mu + R. \end{aligned} \tag{6}$$

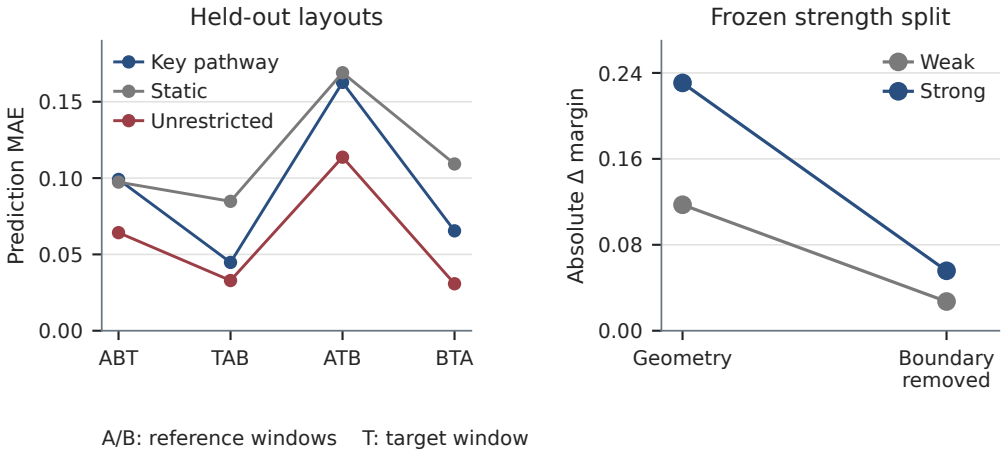

Figure 8: Configuration-specific errors and a development-frozen strength split reveal heterogeneity behind the aggregate prediction and removal results.

The five terms correspond to source-specific boundary values, sensitivity under the key clamp, the non-boundary writer change, calibration, and finite-step curvature. They can cancel, so their magnitudes are not additive causal fractions.

The unrestricted/shared predictor removes the key-path restriction but retains shared source values. The unrestricted/boundary predictor also restores source-specific boundary values. The full-writer predictor includes the non-boundary change. Each uses its own development-fitted scalar. Differences in final MAE compare complete predictors.

E.2 Finite-amplitude error and difficult sources

The quarter-step control injects one quarter of the complete writer delta into the baseline network. It compares the actual resulting margin with one quarter of the raw full-writer derivative. The full-step control uses the complete delta. Figure 9 reports errors averaged within each of the 48 source components and retains every source. Its distribution separates a broad residual from an aggregate driven only by a few sources.

E.3 Source-level uncertainty

For a predictor f, the inference unit is $e_s(f) = \frac{1}{8}\sum_{c,o} |f(x_{sco}) - \Delta m_{sco}|$. Differences between two predictors are computed within the same source and then averaged over sources. Bootstrap samples draw 48 source IDs with replacement and keep all their conditions together. Each source receives equal weight in the point estimate. The strength split in Figure 8 uses a development-derived cutoff. The strata are descriptive because sources do not appear in both with equal counts.

F Natural background and functional outcomes

F.1 Visible background and sequence order

The source-separated natural cohort separates two questions. Native-order benefit compares the native sequence to shuffled exterior content while retaining the local core. Buffering compares boundary sensitivity when shuffled exterior context is available or isolated. Figure 10 shows both endpoints for all three models. The order comparison uses native-residue log probability. The buffering comparison uses the reduction in absolute boundary-induced margin change. Their scales are not interchangeable.

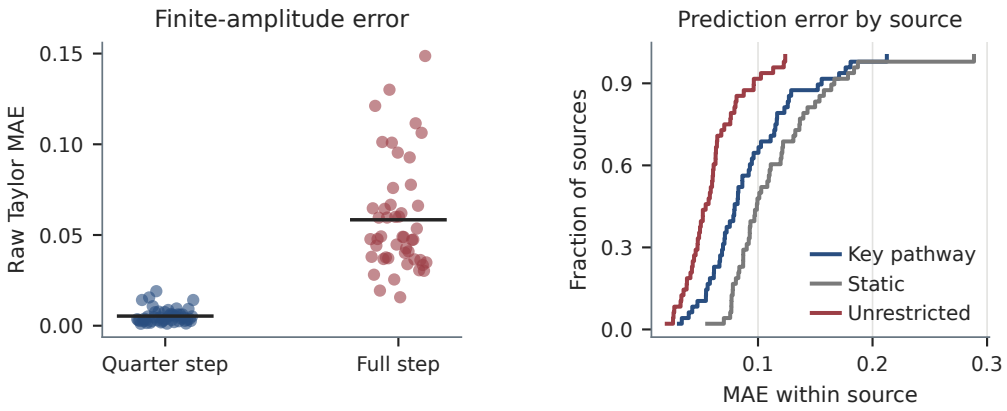

Figure 9: Finite-step Taylor errors and the distribution of source-level prediction errors expose approximation limits beyond average MAE.

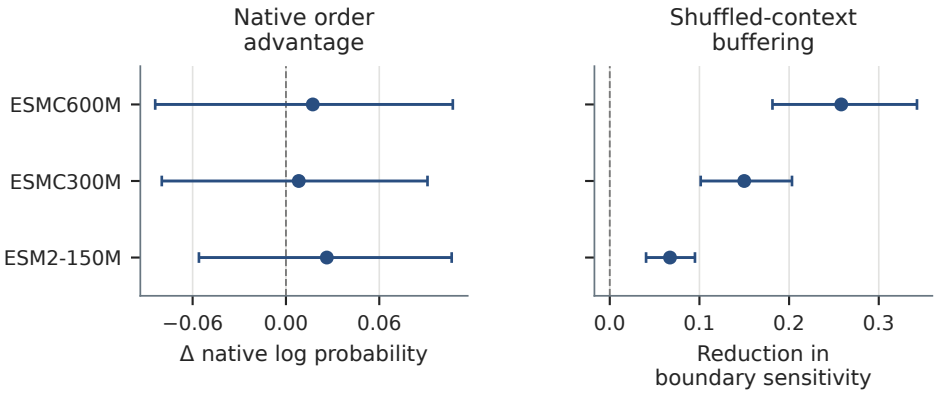

Figure 10: Strict-cohort intervals support buffering from visible shuffled context without confirming a uniform benefit from its native order.

F.2 AFFINE VALUE OFFSETS CHANGE THE MEANING OF REMOVAL

The zero, common-template, affine-offset, and residual interventions are evaluated on the same 128 source components. Figure 11 retains both boundary sensitivity and native accuracy for all four. The residual is the common template minus its parameter-derived offset. The two pieces can point in opposing directions. Either part alone can produce a qualitatively different model.

For a normalized attention row, let P_{bg} be total background attention and let b be the fixed replacement value. Then

$$P_{\mathrm{core}}V_{\mathrm{core}} + P_{\mathrm{bg}}b = P_{\mathrm{core}}(V_{\mathrm{core}} - b) + b. \tag{7}$$

The identity holds for the current attention probabilities and values. The fixed offset changes the origin of the core values before later projections and nonlinearities. Its lower boundary sensitivity can therefore accompany worse residue prediction.

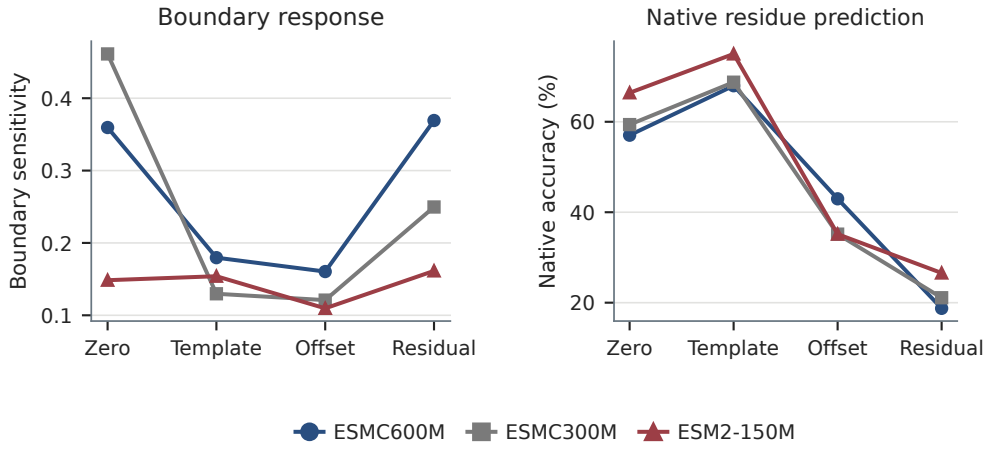

Figure 11: All four background-value replacements reveal that stability and native-residue accuracy can move in different directions.