TERTIARY STRUCTURE IS DIRECTLY ACCESSIBLE IN PROTEIN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein language models are trained on sequence alone, yet they learn enough three-dimensional structure that contact maps can be read out of them without structural supervision. The categorical Jacobian pioneered fully unsupervised readout, recovering contacts by substituting every amino acid at every position and measuring the response. We show the same contacts can be read directly from attention, across eight model families and fourteen models including encoder-decoder and causal architectures. Selecting a model's attention heads on a handful of labeled chains — or, to keep the pipeline unsupervised, bootstrapping off the Jacobian's own predictions, which select the same heads — and averaging the top few unweighted costs one forward pass per chain. On ESM-2 650M this exceeds the Jacobian on four evaluation sets by 2.9 to 27.0 percentage points of long-range precision (P@L/2), with the margin widening as training-leakage screening gets stricter. How accessible that structure is depends on the architecture: under both readouts every masked and encoder-decoder model above the smallest reads at least 0.47, while the three causal decoders reach at most 0.114. On the unsupervised path the Jacobian runs once per model, on ten chains, after which attention reads every protein 150× to 1,626× faster and, on designed repeat proteins, 27.9 to 45.9 points more precisely. It lets us ask how accessibly tertiary structure is encoded across frontier protein language models and architectures, and interrogate how the Jacobian conflates contact signal with additional information.

1. INTRODUCTION

A protein's function follows from how it folds in 3D. Predicting that structure from sequence is one of biology's long-standing computational problems (Jumper et al. 2021). Protein language models are never shown a structure; their objective is over sequence alone. They learn structure anyway, to such a degree that we can extract a chain's residue contacts directly (Rao, Meier, et al. 2021).

Earlier work located that signal in attention, demonstrating that attention maps align with the contact map of the sequence (Vig et al. 2021). To map that alignment into a readout, Rao, Meier, et al. (2021) employ a regression fit over every head while Bhattacharya et al. (2022) use only a small number of heads picked by their measured contact precision and averaged. Both choose or weight heads using known structures, and Rao, Meier, et al. (2021)'s regression also fits a weight per head.

The Categorical Jacobian (Zhang et al. 2024) removed the labels. It substitutes each residue with every standard amino acid and measures how the masked-LM logits move elsewhere in the chain; on 1,431 PDB chains with ESM-2-3B it reaches long-range P@L/2 of 0.80 against 0.67 for an inverse-covariance baseline. Unlike those attention readouts, it needs no structural label at any stage, head selection included. However, it has some limitations, including its robustness, such as its sensitivity to designed repeat proteins, and substantial cost; the perturbation costs $20L$ forward passes (where L is the length of the protein) and requires a per-position categorical head.

The frontier has also moved past the models these readouts were measured on, so whether attention exposes more or less contact signal on newer models with different architectures is an open question.

We adopt top-K attention fusion: rank a frozen model's layer-head maps by contact precision, then average the top K with equal weight. The ranking can come from ground truth contact labels or extends the Jacobian by bootstrapping labels off own predictions. On ESM-2-650M that selection recovers the identical ten heads that labels do, so the pipeline keeps the label-free property Zhang et al. (2024) introduced (Section 5.3).

Furthermore, for ESM-2-650M, fusion exceeds the Jacobian on all four evaluation sets, by 2.9 to 27.0 percentage points of long-range P@L/2, every paired 95% interval clear of zero (Section 5.1). End to end on one A100, its median per-protein wall-clock is 150× lower than the batched Jacobian's

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.

on ESM-2-35M and 1,626× lower on ESM-2-3B (Section 5.2). We run the readout on fourteen models from eight families, two of them encoder-decoders and three strictly causal decoders. The mean margin over each model's own Jacobian is positive on 10 of the 13 where both readouts run (Section 5.4). Accessibility turns on the architecture: every masked and encoder-decoder model above the smallest reads at least 0.47 under both readouts, while the three causal decoders reach at most 0.114 (Section 5.4). A pretraining-aware filter concentrates designed repeat proteins, on which the Jacobian's coupling signal tracks the repeat period rather than the fold, so filtered benchmarks should be reported stratified by repeat content (Section 5.1).

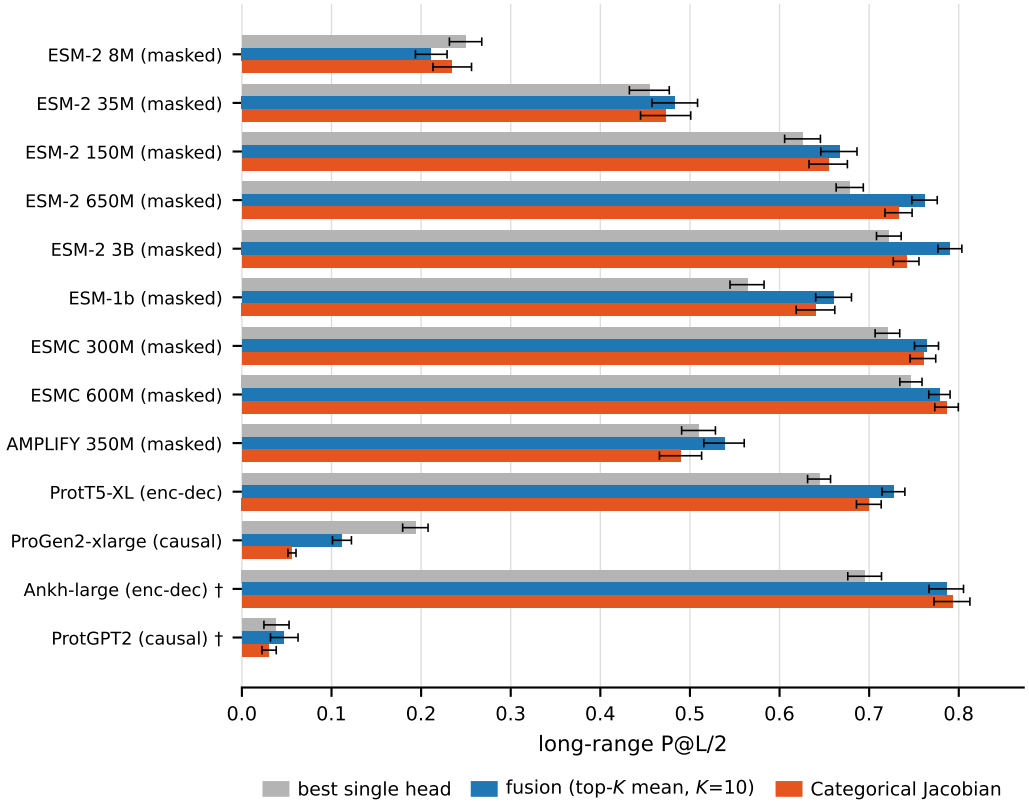

Figure 1: Long-range P@L/2 on Zhang eval-200 for the thirteen checkpoints with a Jacobian, per-row values in Appendix H, Table H1. Fusion is the unweighted mean of the top ten heads; the best single head and the model's own Categorical Jacobian are shown alongside. Bars carry 95% bootstrap intervals, the architecture follows each model name in parentheses, and † marks a 50-chain draw.

2. Related work

The two supervised attention methods build on Vig et al. (2021)'s alignment finding. Rao, Meier, et al. (2021) fit an L1-sparse logistic regression over ESM-1b's 660 layer-head maps on 20 training structures, each map symmetrized and average-product-corrected on its own; its weights rank the heads. Bhattacharya et al. (2022) skip the fit: they take the six of ProtBERT-BFD's 480 heads with the highest individual contact precision over 500 randomly drawn families, average those six maps unweighted, symmetrize, and apply average-product correction. In Rao et al.'s own ablation, averaging the ten heads with the largest regression weights trails the regression on ESM-1b, 49.4 against 53.3 in long-range P@L/2, and they report that it leads on their other Transformer models (Rao, Meier, et al. 2021). Rao, Liu, et al. (2021) carry the regression to MSA Transformers. Reading representations rather than attention, Rives et al. (2021) fit a supervised low-rank bilinear projection of ESM-1b's last-layer representations on 15,297 SCOPe structures, cross-validated with fold-level holdout; it recovers long-range contacts ahead of coevolution (CCMpred) on their metric.

Zhang et al. (2024) read contacts by perturbation instead, post-processing the logit response with average-product correction (Dunn, Wahl, and Gloor 2008). Their method was first applied to protein models in the ESM family, and since to genomic language models (Nguyen et al. 2024; Cornman et al. 2024).

Supervision still buys accuracy over either unsupervised readout on the unfiltered sets (Section 5.1); when structure is the training target far more can be extracted from sequence and its alignments (Jumper et al. 2021). Our question is different: whether a language model learns structure as a by-product of modeling sequence and how sensitive that is to architecture and scale. Table 1 carries one supervised reference, Rao, Meier, et al. (2021)'s recipe refit on ESM-2 by its authors on the same 20 structures (Lin et al. 2023).

Hermann et al. (2024) show that test sets similar in sequence to the pretraining corpus inflate measured performance, and Rong et al. (2025) build a contamination-free benchmark by holding proteins out by date. The holdouts of TAPE (Rao et al. 2019) and FLIP (Dallago et al. 2021) control similarity within the downstream task and leave the pretraining corpus uncontrolled. We apply these ideas to develop our own evaluation sets to compare these methods fairly.

Interpretability work on these models has read attention (Vig et al. 2021) and now reads features: sparse autoencoders over their activations recover latents for binding sites, structural motifs and functional domains, scored against sequence annotations (Adams et al. 2025; Simon and Zou 2025). We read heads rather than features, score them against structure, and find the contact signal concentrated in a handful of late heads (Section 5.4).

3. Methods

Contacts and metric. A contact is a residue pair within 8 Å, Cβ to Cβ, with Cα for glycine; it is long-range when pair indices i and j have $|i - j| \geq 24$. Every readout produces one L-by-L score map per protein, scored by long-range P@L/2: the fraction of the L/2 highest-scoring long-range pairs that are true contacts. A 10,000-resample paired bootstrap over per-protein deltas tests every difference between readouts (Appendix A). The criterion differs from Zhang et al. (2024)'s 10 Å Cα, so the 0.733 their method scores on a 200-chain draw of their benchmark (Table 1) is not directly comparable to their reported values.

The Categorical Jacobian. For each position i and each substituted amino acid, one forward pass records how the predicted logits move at every position j. Mean-centering, a Frobenius norm over the amino-acid axes, average-product correction (Dunn, Wahl, and Gloor 2008) and symmetrization collapse that to an L-by-L map, at 20L forwards per protein. Architectures without a categorical head get a hidden-state form: the L2 displacement of the last-layer residual stream (the latent representation of the model before the output unembedding) at j when i is perturbed, post-processed identically. The two agree at per-protein Pearson r = 0.96 on ESM-2-650M (Appendix B).

Top-K attention fusion. For each layer-head pair we symmetrize the attention map, apply average-product correction and zero the diagonal. We score every head by long-range P@L/2 on a selection set of ten labeled chains, select the top K, and average them unweighted. We use K = 10 unless stated, and the same ten randomly drawn chains rank every model; they are disjoint from every evaluation set. Ranking heads by measured contact precision and averaging them unweighted, after symmetrization and average-product correction, is Bhattacharya et al. (2022)'s recipe; they averaged six heads. Appendix F shows the readout is insensitive to the number of labeled chains, and that the Jacobian's predictions can replace the labels. In Appendix G's sweep, K = 10 sits within 0.011 of each model's best K on seven of the ten swept; ESM-2 8M, AMPLIFY 350M and ESMC 600M do better at K = 3, by 0.063, 0.017 and 0.018. The ranking costs ten forward passes once per model; the per-protein cost is one forward pass.

Models. We test fourteen models from eight families. Nine are masked: ESM-2 at 8M, 35M, 150M, 650M and 3B (Lin et al. 2023), ESM-1b (Rives et al. 2021), AMPLIFY-350M (Fournier et al. 2024) and ESM Cambrian (ESMC) at 300M and 600M (ESM Team 2024). ProtT5-XL (Elnaggar et al. 2022) and Ankh-large (Elnaggar et al. 2023) are two encoder-decoder models, read at their encoder self-attention. ProGen2-large, ProGen2-xlarge (Nijkamp et al. 2023) and ProtGPT2 (Ferruz, Schmidt, and Höcker 2022) are three causal models from two families, read at the populated triangle (Appendix C).

4. Evaluation sets

A test protein with a close relative in the pretraining corpus lets both readouts score by recall, retrieving a memorized relative rather than applying structure the model has learned. We therefore report on four sets, two unfiltered and two filtered for pretraining overlap. Zhang eval-200 is a 200-chain length-stratified draw from the 1,431-chain benchmark of Zhang et al., in-distribution for every model tested. The Boltz-1 test split (Wohlwend et al. 2024) is that structure predictor's public partition, N=470 at L ≥ 75. Our pretraining-aware (PTA) set, N=83, is built in Appendix D: post-2022-09 PDB chains with no pre-cutoff homolog in UniRef90 (Suzek et al. 2007) or the PDB, screened on query coverage and for chimeras of pre-cutoff parents. The filtered Boltz-1 split applies the same screen to the Boltz partition, N=24.

A filter for sequence novelty keeps the chains with no pre-cutoff relative; those are disproportionately designed proteins. Zhang eval-200 has no chain with an internal repeat, while the Boltz split has 12 of 470. The two filtered sets have 13 of 83 and 12 of 24 internal-repeat chains, all designed or synthetic.

5. Results

5.1 Fusion and the Jacobian

On ESM-2-650M, fusion exceeds the categorical Jacobian on all four sets, by 2.9 to 27.0 percentage points, every paired interval clear of zero (Table 1). The margin widens as leakage is removed: fusion falls 25.8 points from Zhang eval-200 to PTA and the Jacobian falls 34.4.

Table 1: Long-range P@L/2 on ESM-2-650M. T1, best head. CJ, categorical Jacobian. F, fusion: unweighted mean of ten heads selected on ten labeled chains or, identically, on CJ's predictions. FH, fitted head: Rao, Meier, et al. (2021)'s regression refit on ESM-2 on 20 structures (Lin et al. 2023). Δ is F − CJ, paired, 10,000-resample bootstrap. Strata split sets with repeat chains. T1 pooled only.

Set	N	T1	CJ	F	FH	Δ [95% CI]
Zhang eval-200	200	0.678	0.733	0.762	0.784	+0.029 [+0.024, +0.035]
Boltz-1 test split	470	0.504	0.520	0.583	0.593	+0.063 [+0.054, +0.073]
non-repeat	458		0.532	0.585	0.598	+0.053 [+0.045, +0.060]
repeat	12		0.030	0.489	0.399	+0.459 [+0.369, +0.561]
PTA	83	0.434	0.389	0.504	0.500	+0.115 [+0.086, +0.147]
non-repeat	70		0.451	0.535	0.546	+0.085 [+0.063, +0.110]
repeat	13		0.055	0.334	0.251	+0.279 [+0.168, +0.396]
Boltz-1, PTA	24	0.365	0.195	0.465	0.421	+0.270 [+0.179, +0.367]
non-repeat	12		0.360	0.441	0.442	+0.081 [+0.026, +0.145]
repeat	12		0.030	0.489	0.399	+0.459 [+0.369, +0.561]

The width of that range, 2.9 to 27.0 points across the four sets, is a composition effect disentangled by the strata rows of Table 1. On chains without an internal repeat the margin is 2.9 to 8.5 points, with all intervals still significant. On repeat chains it is 27.9 to 45.9, because the Jacobian returns a long-range precision of 0.00 on 5 chains on PTA and 6 on the filtered Boltz split. We hypothesize this is due to the model relying disproportionately on the repeat sequences to predict structure — residues within repeat sequences receive high signal from the whole sequence — rather than contacts. Perturbing i moves the distribution at i + d and its harmonics whether or not the residues touch. The resulting diagonal band holds about L pairs per harmonic and consumes the top-L/2 budget before a single true contact is reached: all 8 zero-scoring chains we inspected place every pick on the chain's own period of maximum sequence self-identity or its multiples (Appendix E). Periodic bands in coevolution maps of repeat proteins are a known artifact (Espada et al. 2015; Anishchenko et al. 2017). Appendix A.11 of Rao, Meier, et al. (2021) shows a repeat protein on which a coupling method picks up extra bands and their fitted head keeps only the band nearest the diagonal. The artifact carries over to the Jacobian, where it is total. Averaging ten contact-selective heads dilutes a periodic contribution instead of being dominated by it, and the heads that respond to periodicity are a separate population (Appendix E). The Jacobian's zero rate runs 0% on Zhang eval-200, 2.6% on the full Boltz split and 25% on the filtered one, so a pretraining-aware benchmark must be reported stratified to avoid this limitation.

The fitted-head column quantifies what supervision buys: one learned weight per layer-head map, fit by Rao, Meier, et al. (2021)'s recipe on 20 structures and refit on ESM-2 by its authors (Lin et al.

2023). It leads fusion by 2.2 points on Zhang eval-200; on the two pretraining-aware sets the two sit within 0.4 and 4.4 points.

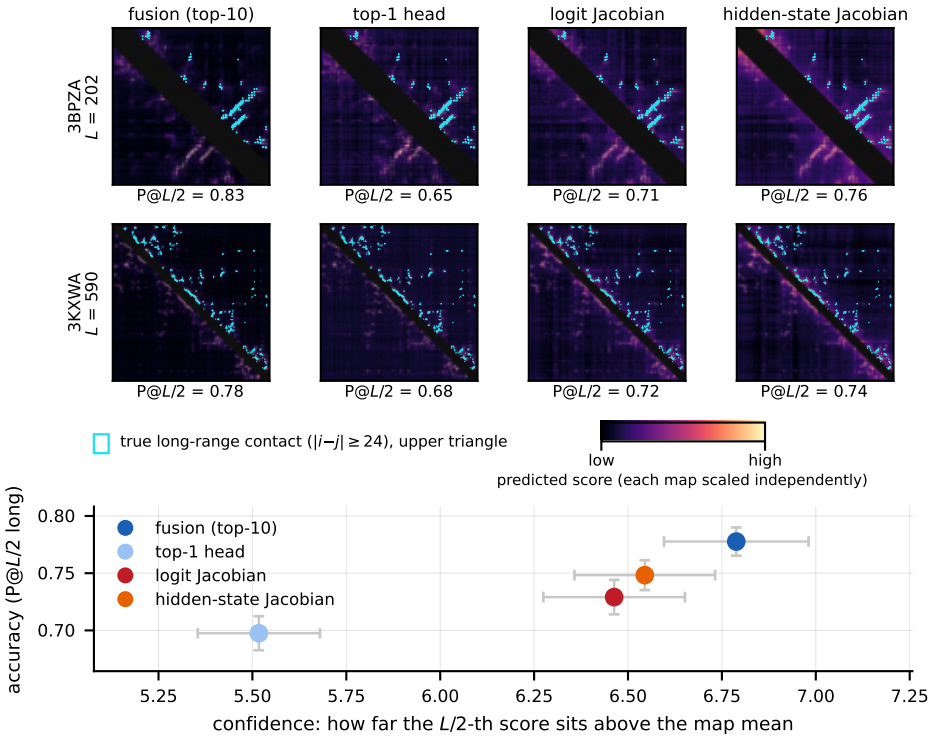

Figure 2: Readouts on ESM-2-650M. Top, predicted long-range maps for the shortest and longest of the 40 proteins, the short-range band masked. Cyan outlines mark true long-range contacts above the diagonal; below it, the bare prediction. Bottom, accuracy against confidence, the height of the L/2-th score above the map mean, mean ± sem over those 40 proteins.

5.2 Compute Efficiency

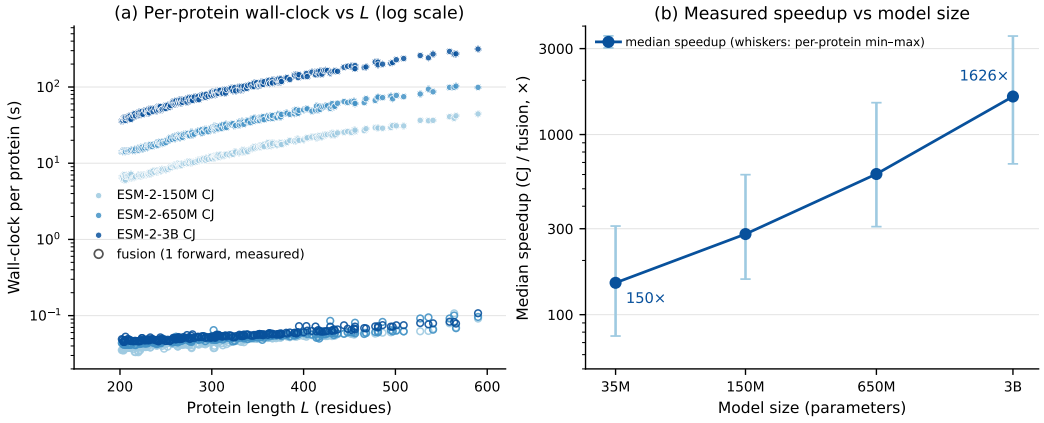

Figure 3: Compute economy on ESM-2. Left, per-protein wall-clock against length at 150M, 650M and 3B; the Jacobian scatters along the 20L trend and fusion sits two to three orders of magnitude below it on the same proteins. Right, median per-protein speedup against model size on Zhang eval-200, 150× at 35M to 1,626× at 3B.

Batching the Jacobian's 20L passes cuts wall-clock but not compute: each perturbed sequence is a distinct input that needs its own full forward pass, while fusion reads pair structure from the one pass it already makes. On one A100-80GB, over the 200 Zhang eval-200 chains of 202 to 590 residues, the Jacobian batched at 32 perturbations per forward takes a median 4.7 s per protein on ESM-2-35M and 87 s on ESM-2-3B, against 0.03 to 0.05 s for fusion. Median speedups end-to-end are 150×, 280×, 604× and 1,626× at 35M, 150M, 650M and 3B, and 466× with 3B on the 14 CASP14 (Kryshtafovych et al. 2021) free-modeling (FM) targets inside ESM-2's 1,026-position context.

5.3 The Jacobian selects the same heads

To keep fusion unsupervised, we replace the true contact labels with the Jacobian's own predictions on the same ten chains, taking its top-scoring pairs per chain as pseudo-contacts at three budgets, L/2, L and 2L for a chain of L residues. ESM-2-650M's Jacobian and ESM-2-150M's select the identical ten heads as label selection at all three budgets, so every fusion cell in Table 1 is unchanged on every protein. ESM-2-35M's recovers 8, 7 and 6 of the ten at L/2, L and 2L. At budget L its ten score 0.739 on Zhang eval-200 against label selection's 0.762; over ESM-2-650M's own Jacobian, at 0.733, the margin is insignificant at +0.006 [−0.002, +0.014]. Rao, Meier, et al. (2021), Rao, Liu, et al. (2021) and Lin et al. (2023) fit contact regressions over attention, and Bhattacharya et al. (2022) select heads, all against contacts from experimental structures. To our knowledge, none ranks heads against a label-free readout's predictions.

Bootstrapping works because of how the model learns to allocate the function of its attention heads. One head's top-L/2 list covers 0.4% of the long-range candidate space, so a uniformly placed false label rarely touches the list being scored. Corrupting the true labels to a known precision p confirms it: scored with a rank-fusion proxy on 50 held-out chains (Appendix F), selection stays within 0.5 points of the fully labeled one until p falls to 0.1 (when nine of ten labels are wrong). Meanwhile the Jacobian's errors are not uniform but register-shifted: 66.9% of its false pairs lie within two residues of a true contact, against 8.8% for a uniform draw. Where labels are used, one structure is enough: it recovers 8.9 of the ten heads and, under the same proxy, scores 0.745 against 0.743 with fifty (Appendix F).

The unsupervised pipeline inherits one requirement that label selection does not. On a designed repeat chain the Jacobian's picks are its periodic band, so the pseudo-labels are coherent and wrong: a selection set of ten with four such chains still scores 0.702, and with five it collapses to 0.346.

5.4 Performance across architectures

Figure 1 sets fusion, the best head and the model's own Jacobian side by side on Zhang eval-200, with per-row values and intervals in Appendix H, Table H1. Fusion's mean margin over the model's own Jacobian is positive on 10 of the 13 models both readouts run on, and 7 of the 13 survive Holm correction (Appendix A).

How accessible the structure is depends strongly on the architecture. Every masked and encoder-decoder model above the smallest reads at least 0.47 under both readouts, while the three causal decoders reach at most 0.114 and their best single head at most 0.193. Masked and causal models differ in objective as well as architecture, the causal side is three models from two families, and their Jacobian is the hidden-state form, validated as functionally corresponding only within masked models (Appendix B).

The number of heads that carry contact signal tracks how precise a model is. We count the heads whose excess precision over the median head is at least half the best head's excess. Over the first ten models that count correlates with best-head precision at Spearman ρ = +0.81, p = 0.004. It holds on Ankh-large, measured afterwards, and fails only on ProtGPT2, whose best head sits at about 4 times the chance contact rate; the median is then within a factor of two of the best and the count measures noise (Appendix H).

Where those heads sit depends on the evaluation set as much as on the model. A model's late share is the fraction of its excess head-precision mass over the median head that lies in its last $\lceil n/5 \rceil$ of n layers. On Zhang eval-200 the top contact head lies at relative depth 0.91 to 1.00 across the nine non-causal models. The three causal models place far less of it late, and so does ESM-2-3B (one of the largest models we tested) on CASP14-FM, a set of hard free-modeling targets that exposes less contact signal. Its share there is 0.347, inside the causal models' range of 0.029 to 0.365 (Appendix H). A low late share therefore reflects weak contact signal as much as architecture.

Ankh-large, a second encoder-decoder measured after the analysis above, follows the same pattern. Its top head sits at relative depth 0.979 of 48 encoder layers and its late share is 0.686. It is the strongest non-masked model we measure: its best head reads 0.675 against ProtT5-XL's 0.644, and fusion 0.774 against 0.727 on all 200 chains.

Every attention readout collapses on the causal models. On ProGen2-xlarge the top head reaches 0.193, fusion 0.111 and the hidden-state Jacobian 0.056. ProtGPT2 is lower still: its best head reads 4.3 times the corpus contact rate at one residue per token and 5.1 times under its own tokenization, against 51 times for ESM-2-650M. At about 4 times chance there is little contact signal to concentrate.

6. Discussion and limitations

A masked-language objective does not merely leave contact information decodable somewhere in the network; it concentrates it in a handful of late attention maps, in a form that a fixed unweighted average recovers. It is perhaps surprising that a larger signal — perturbations over the whole model — is more sensitive than a much more focused signal: a small number of attention heads. Perturbing one position moves the output through every pathway the model has, so the Jacobian sums every coupling the model learned, contact or not; on Table 1's repeat strata the repeat period dominates, 0.030 and 0.055 against fusion's 0.489 and 0.334. Ten heads selected for contact isolate the part of that signal that is contact. How accessibly a model holds that information depends on its architecture: masked and encoder-decoder models expose it to both readouts, while the three causal decoders expose little to either.

Compared with the Jacobian, fusion is faster everywhere we timed it, 150× to 1,626× across ESM-2 sizes. It is more precise on ESM-2-650M on all four sets, with a positive mean margin on 10 of 13 models, and more robust on designed repeats, by 27.9 to 45.9 points. Pretraining-aware evaluation matters: it costs fusion about a third of its precision and the Jacobian nearly half, 25.8 and 34.4 points. How much tertiary structure an unsupervised readout recovers depends deeply on the architecture it reads: those architectures with inductive biases that allow for two-way interaction between residues across a sequence outperform causal models that enforce one-way coupling. However, this ordering does not carry over to downstream quality: causal models such as ProGen2 roughly match masked models on zero-shot fitness prediction (Notin et al. 2023).

Some limitations bound the claims. The bootstrap removes the dependence on labels and on the student's own Jacobian, but not on a teacher of sufficient capacity, which Section 5.3 locates between ESM-2-35M and ESM-2-150M. The hidden-state Jacobian is validated only within masked models, at r = 0.96 on ESM-2-650M. The pretraining-aware filter is sequence-level and leaves fold-level overlap uncontrolled. The causal result rests on only two families of autoregressive decoders, and the encoder-decoder result on two models. Every result in the main text uses one contact criterion, 8 Å Cβ. Head selection draws its ten chains from one benchmark's distribution, and the pseudo-label pipeline needs the repeat screen of Appendix F on that set.

Reproducibility statement

Code, the pretraining-aware filter pipeline, the head-selection artifacts and the data for the tables and figures in this paper are in an anonymized repository at [anonymized URL]; the archive will be linked on acceptance.

AI use statement

Large language models were used for code, data analysis, figures, drafting and review, under the authors' direction. The authors carefully reviewed all figures, text and citations, and take responsibility for all claims, results and content.

References

Adams, Etowah, Liam Bai, Minji Lee, Yiyang Yu, and Mohammed AlQuraishi. 2025. "From Mechanistic Interpretability to Mechanistic Biology: Training, Evaluating, and Interpreting Sparse Autoencoders on Protein Language Models." In *Proceedings of the 42nd International Conference on Machine Learning*, 267:460–76. Proceedings of Machine Learning Research. https://proceedings.mlr.press/v267/adams25a.html.

Anishchenko, Ivan, Sergey Ovchinnikov, Hetunandan Kamisetty, and David Baker. 2017. "Origins of Coevolution Between Residues Distant in Protein 3D Structures." *Proceedings of the National Academy of Sciences* 114 (34): 9122–27. https://doi.org/10.1073/pnas.1702664114.

Bhattacharya, Nicholas, Neil Thomas, Roshan Rao, Justas Dauparas, Peter K. Koo, David Baker, Yun S. Song, and Sergey Ovchinnikov. 2022. "Interpreting Potts and Transformer Protein Models Through the Lens of Simplified Attention." In *Biocomputing 2022: Proceedings of the Pacific Symposium*, 27:34–45. Pacific Symposium on Biocomputing. https://doi.org/10.1142/9789811250477_0004.

Cornman, Andre, Jacob West-Roberts, Antonio Pedro Camargo, Simon Roux, Martin Beracochea, Milot Mirdita, Sergey Ovchinnikov, and Yunha Hwang. 2024. "The OMG Dataset: An Open MetaGenomic Corpus for Mixed-Modality Genomic Language Modeling." *bioRxiv*. https://doi.org/10.1101/2024.08.14.607850.

Dallago, Christian, Jody Mou, Kadina E. Johnston, Bruce J. Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K. Yang. 2021. "FLIP: Benchmark Tasks in Fitness Landscape Inference for Proteins." In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Vol. 1. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/2b44928ae11fb9384c4cf38708677c48-Abstract-round2.html.

Dunn, Stanley D., Lindi M. Wahl, and Gregory B. Gloor. 2008. "Mutual Information Without the Influence of Phylogeny or Entropy Dramatically Improves Residue Contact Prediction." *Bioinformatics* 24 (3): 333–40. https://doi.org/10.1093/bioinformatics/btm604.

Elnaggar, Ahmed, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, and Burkhard Rost. 2023. "Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling." *arXiv Preprint arXiv:2301.06568*. https://arxiv.org/abs/2301.06568.

Elnaggar, Ahmed, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, et al. 2022. "ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (10): 7112–27. https://doi.org/10.1109/TPAMI.2021.3095381.

ESM Team. 2024. "ESM Cambrian: Revealing the Mysteries of Proteins with Unsupervised Learning." https://evolutionaryscale.ai/blog/esm-cambrian. https://evolutionaryscale.ai/blog/esm-cambrian.

Espada, Rocío, R. Gonzalo Parra, Thierry Mora, Aleksandra M. Walczak, and Diego U. Ferreiro. 2015. "Capturing Coevolutionary Signals in Repeat Proteins." *BMC Bioinformatics* 16: 207. https://doi.org/10.1186/s12859-015-0648-3.

Ferruz, Noelia, Steffen Schmidt, and Birte Höcker. 2022. "ProtGPT2 Is a Deep Unsupervised Language Model for Protein Design." *Nature Communications* 13 (1): 4348. https://doi.org/10.1038/s41467-022-32007-7.

Fournier, Quentin, Robert M. Vernon, Almer van der Sloot, Benjamin Schulz, Sarath Chandar, and Christopher J. Langmead. 2024. "Protein Language Models: Is Scaling Necessary?" *bioRxiv*. https://doi.org/10.1101/2024.09.23.614603.

Hermann, Leon, Tobias Fiedler, Hoang An Nguyen, Melania Nowicka, and Jakub M. Bartoszewicz. 2024. "Beware of Data Leakage from Protein LLM Pretraining." In *Proceedings of the 19th Machine Learning in Computational Biology Meeting*, 261:106–16. Proceedings of Machine Learning Research. https://proceedings.mlr.press/v261/hermann24a.html.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature* 596 (7873): 583–89. https://doi.org/10.1038/s41586-021-03819-2.

Kryshtafovych, Andriy, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. 2021. "Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round XIV." *Proteins: Structure, Function, and Bioinformatics* 89 (12): 1607–17. https://doi.org/10.1002/prot.26237.

Lin, Zeming, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, et al. 2023. "Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model." *Science* 379 (6637): 1123–30. https://doi.org/10.1126/science.ade2574.

Nguyen, Eric, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David B. Li, Liam J. Bartie, et al. 2024. "Sequence Modeling and Design from Molecular to Genome Scale with Evo." *Science* 386 (6723): eado9336. https://doi.org/10.1126/science.ado9336.

Nijkamp, Erik, Jeffrey A. Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. 2023. "ProGen2: Exploring the Boundaries of Protein Language Models." *Cell Systems* 14 (11): 968–978.e3. https://doi.org/10.1016/j.cels.2023.10.002.

Notin, Pascal, Aaron W. Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Hansen Spinner, Nathan Rollins, et al. 2023. "ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction

and Design." In *Advances in Neural Information Processing Systems*, 36:64331–79. https://doi.org/10.52202/075280-2810.

Rao, Roshan, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. 2019. "Evaluating Protein Transfer Learning with TAPE." In *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 32. https://papers.nips.cc/paper_files/paper/2019/hash/37f65c068b7723cd7809ee2d31d7861c-Abstract.html.

Rao, Roshan, Jason Liu, Robert Verkuil, Joshua Meier, John F. Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. 2021. "MSA Transformer." In *Proceedings of the 38th International Conference on Machine Learning*, 139:8844–56. Proceedings of Machine Learning Research. https://proceedings.mlr.press/v139/rao21a.html.

Rao, Roshan, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. 2021. "Transformer Protein Language Models Are Unsupervised Structure Learners." In *International Conference on Learning Representations (ICLR)*. https://doi.org/10.1101/2020.12.15.422761.

Rives, Alexander, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, et al. 2021. "Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences." *Proceedings of the National Academy of Sciences* 118 (15). https://doi.org/10.1073/pnas.2016239118.

Rong, Dingyi, Zijian Chen, Qi Jia, Kaiwei Zhang, Haotian Lu, Guangtao Zhai, and Ning Liu. 2025. "LiveProteinBench: A Contamination-Free Benchmark for Assessing Models' Specialized Capabilities in Protein Science." *arXiv Preprint arXiv:2512.22257*. https://arxiv.org/abs/2512.22257.

Simon, Elana, and James Zou. 2025. "InterPLM: Discovering Interpretable Features in Protein Language Models via Sparse Autoencoders." *Nature Methods* 22 (10): 2107–17. https://doi.org/10.1038/s41592-025-02836-7.

Steinegger, Martin, and Johannes Söding. 2017. "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets." *Nature Biotechnology* 35 (11): 1026–28. https://doi.org/10.1038/nbt.3988.

Suzek, Baris E., Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H. Wu. 2007. "UniRef: Comprehensive and Non-Redundant UniProt Reference Clusters." *Bioinformatics* 23 (10): 1282–88. https://doi.org/10.1093/bioinformatics/btm098.

Vig, Jesse, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. "BERTology Meets Biology: Interpreting Attention in Protein Language Models." In *International Conference on Learning Representations (ICLR)*. https://doi.org/10.1101/2020.06.26.174417.

Wohlwend, Jeremy, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, et al. 2024. "Boltz-1 Democratizing Biomolecular Interaction Modeling." *bioRxiv*. https://doi.org/10.1101/2024.11.19.624167.

Zhang, Zhidian, Hannah K. Wayment-Steele, Garyk Brixi, Haobo Wang, Dorothee Kern, and Sergey Ovchinnikov. 2024. "Protein Language Models Learn Evolutionary Statistics of Interacting Sequence Motifs." *Proceedings of the National Academy of Sciences* 121 (45). https://doi.org/10.1073/pnas.2406285121.

Appendix

A. Significance procedure

Every comparison between two readouts is paired over the proteins present in both result files. We resample the per-protein deltas 10,000 times with replacement at seed 0 and take the 2.5 and 97.5 percentiles as the 95% interval. On the same deltas we run a two-sided Wilcoxon signed-rank test. An interval clear of zero whose Wilcoxon does not reject marks a margin carried by a few proteins with large gaps.

Table A1 applies Holm's correction at $\alpha = 0.05$ to the thirteen checkpoints on which both readouts run, and every Holm value in the paper is over these thirteen tests. The mean delta favors fusion on 10 of 13, the per-protein majority on 9 of 13, and the Holm-adjusted test on 7 of 13. In either direction, 8 of 13 are significant after correction. On the eight checkpoints whose Jacobian is the validated logit form, the mean delta favors fusion on 6 of 8.

Table A1: The thirteen-checkpoint family. Δ is fusion − Jacobian. W/L counts per-protein wins and losses. p (W) is the raw two-sided Wilcoxon p, and Holm p corrects it over all thirteen tests. † marks a 50-chain draw.

Checkpoint	N	Δ	W/L	p (W)	Holm p
ESM-2 8M	200	−0.0232	72/121	6.05e-05	0.000423
ESM-2 35M	200	+0.0103	92/100	0.400	0.616
ESM-2 150M	200	+0.0119	108/74	2.09e-04	0.00125
ESM-2 650M	200	+0.0290	147/40	3.57e-18	4.28e-17
ESM-2 3B	200	+0.0486	176/14	8.87e-29	1.15e-27
ESM-1b	200	+0.0200	123/63	6.93e-09	5.54e-08
ESMC 300M	200	+0.0040	98/90	0.205	0.616
ESMC 600M	200	−0.0078	80/113	0.018	0.090
AMPLIFY 350M	200	+0.0489	150/44	1.38e-17	1.38e-16
ProtT5-XL	200	+0.0275	150/40	1.16e-17	1.27e-16
ProGen2-xlarge	200	+0.0557	145/50	1.98e-17	1.78e-16
Ankh-large †	50	−0.0063	19/28	0.210	0.616
ProtGPT2 †	50	+0.0168	25/22	0.122	0.489

Interval endpoints vary over bootstrap seeds with a standard deviation of about 0.0001 at N = 200 and 0.0012 at N = 24. The seed decides the sign of one endpoint, ESM-2-35M's lower one in Table H1, which clears zero by 0.000125 against a seed standard deviation of 0.000119. That checkpoint's Wilcoxon p is 0.400 and it loses on 100 of 200 proteins.

B. The hidden-state Jacobian

Architectures without a per-position categorical head give no logit response, so we read the perturbation from the residual stream. For each position i and each of the 20 substituted amino acids, one forward pass records the last-layer residual stream at every position j. The score at (i, j) is the L2 displacement of that vector, summed over substitutions. Average-product correction and symmetrization follow, as in the logit form. The cost is the same 20L forward passes, batched 32 to a forward.

On ESM-2-650M both forms run. Over the 200 Zhang eval-200 chains they agree at per-protein Pearson r = 0.955, with set means of 0.733 for the logit form and 0.732 for the hidden-state form (Figure 4). ESM-2-3B gives 0.742 against 0.746, and ESM-1b 0.641 against 0.637.

The validation covers only masked models, because no checkpoint tested has both a non-masked objective and a categorical head. The five rows of Table H1 read with the hidden-state form inherit that limit.

C. The fourteen checkpoints

Table C1 lists the checkpoints and where fusion reads them.

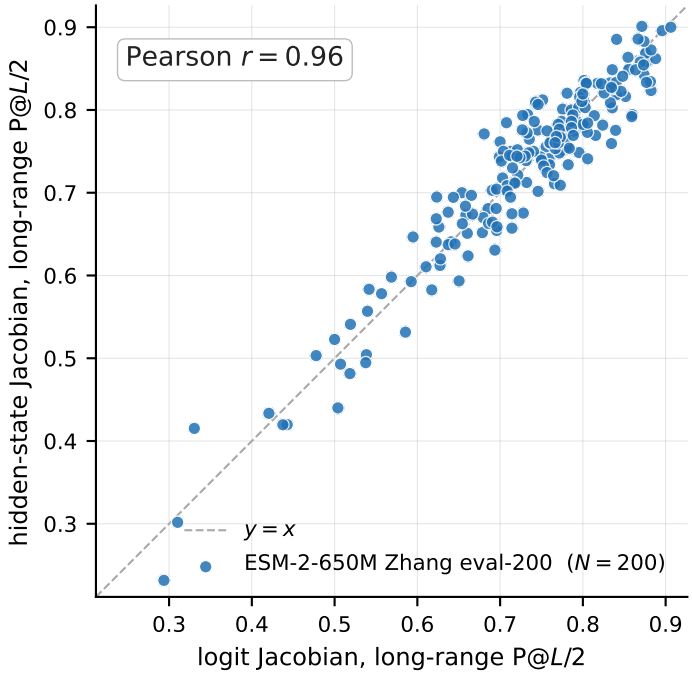

Figure 4: Per-protein agreement of the two Jacobian forms on ESM-2-650M, Zhang eval-200, N = 200. The dashed line is y = x.

Table C1: The checkpoints. Form is the Jacobian form, logit where a categorical head exists and hidden-state otherwise. Read point names the attention tensor fusion reads. Geometry is layers × heads of that tensor.

Checkpoint	Objective	Form	Read point	Geometry
ESM-2 8M	masked	logit	self-attention	6×20
ESM-2 35M	masked	logit	self-attention	12×20
ESM-2 150M	masked	logit	self-attention	30×20
ESM-2 650M	masked	logit	self-attention	33×20
ESM-2 3B	masked	logit	self-attention	36×40
ESM-1b	masked	logit	self-attention	33×20
ESMC 300M	masked	logit	self-attention	30×15
ESMC 600M	masked	logit	self-attention	36×18
AMPLIFY 350M	masked	hidden	self-attention	32×15
ProtT5-XL	span corruption	hidden	encoder self-attn	24×32
Ankh-large	span corruption	hidden	encoder self-attn	48×16
ProGen2-large	causal	hidden	populated triangle	32×32
ProGen2-xlarge	causal	hidden	populated triangle	32×16
ProtGPT2	causal	hidden	populated triangle	36×20

ProGen2-large has 2.7B parameters, ProGen2-xlarge 6.4B, and ProtGPT2 a 0.708B transformer body. Other parameter counts follow the checkpoint names, with ESM-1b at 650M. For the encoder-decoders only the encoder is loaded, 1.151B parameters for Ankh-large. The ESMC results store only each model's top 20 heads, so the two ESMC checkpoints appear in Table G1 but not in Tables H2 and H3, which need every head.

A causal checkpoint attends only backward, so half its attention map is empty by construction and the readout uses the populated triangle. Each forward checks this per chain, with an upper-triangle non-zero fraction of 0.000 on a causal checkpoint against 1.000 on Ankh-large.

ProtGPT2's byte-pair vocabulary averages 2.98 residues per token on this corpus, so a map in token coordinates is constant over blocks of residues. Table H1 reports a per-residue condition that feeds one residue per token, with L tokens and no special tokens. A native condition keeps the model's own tokenization and appears only in the depth share of Table H2, which survives the blocking.

Ankh-large's tokenizer has no normalizer or pre-tokenizer, so a spaced input emits a space token per residue and a raw string is prefixed with an unknown token. We build its ids from the vocabulary directly and check for L + 1 tokens with no unknown token inside the sequence.

D. The pretraining-aware filter

The filter keeps a chain only if no close relative of it was public before the model's pretraining cutoff. It runs in six stages.

1. **Pool.** Single-chain PDB entries released on or after 2022-09-01, at $L \geq 75$, so intra-chain contacts are the whole prediction task.
2. **Sequence search.** Each chain is searched against current UniRef90 (Suzek et al. 2007) with MMseqs2 (Steinegger and Söding 2017) at a 50% identity floor, which only generates candidates. Coverage is 80% of the query alone, so a chain excerpted from a longer pre-cutoff protein is still caught. The reference is UniRef90 because ESM-2 samples UniRef90 members within UniRef50 clusters. The prefilter cap is set explicitly, since its default truncates long hit lists.
3. **Chimera pass.** A second search at 50% coverage catches fusion constructs whose partners pull query coverage below 80%. It moves 10 chains from clean to contaminated and leaves 6 whose hits cannot be dated; all 16 are excluded.
4. **Dating.** A hit counts against a chain only if its representative accession predates ESM-2's pretraining corpus, UniRef 2021_04, dated 2021-10-27.
5. **Pre-cutoff PDB search.** A pre-cutoff homolog can have a PDB structure and no adequate UniRef90 representative, which hides it from stage 2. Candidates are therefore searched against all PDB seqres at 80% identity, below the 90% bar because this identity mode counts gap columns. Dates are RCSB release dates, since deposition dates run earlier and would over-flag.
6. **Acceptance gate.** The gate checks the built set itself. It rejects nucleic-acid records, duplicate sequences, contact maps whose size disagrees with the sequence length, and any surviving pre-cutoff hit above 90% identity at 80% query coverage.

A 95% identity bar would exceed UniRef90's clustering radius, so a chain sharing a cluster with a pre-cutoff member could pass. A 50% bar would leave a set enriched for synthetic and designed constructs. The filter leaves 83 chains, and all 13 with an internal repeat are designed or synthetic.

E. Repeat proteins

A chain is a repeat protein when its maximum cyclic sequence self-identity, searched over periods of 10 to 400 residues, reaches 0.40. On 12 inspected chains the statistic separates cleanly, 0.119 to 0.255 on 3 non-repeat controls against 0.579 to 0.969 on 9 repeats. The Jacobian scores 0.0000 on all 9 repeats, although its maps are numerically healthy, with no NaN or infinity and a variance like that of chains that score well. We read the top-L/2 picks of eight zero-scoring chains, and all eight land on the chain's own period and its integer multiples (Table E1).

Table E1: Four of the eight inspected zero-scoring chains. Pick separations are the sequence separations of the Jacobian's top-L/2 picks. Period is the chain's period of maximum self-identity, and self-id. that identity as a fraction. The control row is chains from the same inspection scoring 0.50 to 0.71.

Chain	Jacobian pick separations	Period	Self-id.
8EOV_1	every pick at 35	35	0.793
8UGC_1	64, 128, 192	64	0.866
7UDL_1	47, 94, 188	47	0.906
8GA6_1	123, 369	123	0.789
controls	smooth spread	—	0.07 to 0.12

Table E2: Zero-scoring chains by set. Repeats counts chains with an internal repeat, zeros counts chains on which the Jacobian scores 0.0000, and the last column is how many of those zeros are repeat chains. Per-stratum precisions are in Table 1.

Set	N	repeats	zeros	zeros that are repeats
Zhang eval-200	200	0	0	0 of 0
Boltz-1 test split	470	12	12	6 of 12
PTA	83	13	5	5 of 5
Boltz-1, PTA	24	12	6	6 of 6

On both filtered sets every zero-scoring chain is a repeat chain (Table E2). On the full Boltz-1 split six of the twelve are not, and we have not established what those six share.

The heads that respond to periodicity are separate from the contact heads. We ranked ESM-2-650M's 660 heads against synthetic pseudo-labels on a single off-diagonal band, over 23 configurations covering 11 periods, 4 harmonic stacks and 8 widths. No configuration's top ten shares a head with the contact ten. These band heads are shallow, at mean layer 4.6 to 20.1 against 29.8 for the contact ten, and fused they score 0.031 to 0.078 on a held-out 50, below the 0.103 of ten random heads. Selected on the Jacobian's output for nine real repeat chains, ten heads place 0.694 of their top-L/2 picks at the modal sequence separation, against 0.192 over all heads and 0.118 for the contact ten.

F. Head selection

Contact signal in ESM-2-650M sits in a few heads. Of its 660 heads, 18 exceed a long-range P@L/2 of 0.3 and 6 exceed 0.5, while the median head scores 0.046, the random-ranking floor. The ten selected heads average 0.553, which is 10.4 standard deviations above the mean of the other 650. Ranking heads on all 200 evaluation proteins returns the same ten as the ten selection chains.

Tables F1 and F2 score the fused ten under truncated reciprocal-rank fusion on a held-out 50 chains, the only operator that can be rebuilt from the stored top-L/2 pair lists. This proxy runs 0.011 below the unweighted mean. Under it the paper's ten heads score 0.747 on all 200 Zhang eval-200 chains and 0.743 on the held-out 50.

Table F1: The label-budget curve under the rank-fusion proxy on the held-out 50. P is long-range P@L/2 of the fused ten and sd its standard deviation, over resamples of the labeled set. Shared is the mean number of heads in common with the paper's ten.

n labeled proteins	1	2	3	5	10	20	50	150
P	0.745	0.746	0.745	0.745	0.744	0.743	0.743	0.743
sd	0.009	0.007	0.005	0.004	0.003	0.002	0.000	0.000
shared	8.88	9.34	9.54	9.68	9.80	9.94	10.00	10.00

Table F2: The teacher-quality curve, true labels degraded to precision p at pseudo-label budget L, under the rank-fusion proxy on the held-out 50. Random places false labels uniformly; near-miss places them at ±2 register shifts of a true contact. p = 0.73 is the Jacobian's own precision on these chains.

target p	1	0.9	0.73	0.5	0.3	0.1	0.05	0.02	0
random P	0.745	0.745	0.745	0.744	0.746	0.748	0.739	0.271	0.061
near-miss P	0.745	0.745	0.745	0.746	0.747	0.733	0.659	0.481	0.359

Selection tolerates uniform false labels down to p = 0.1 because one head's top-L/2 list covers 0.0039 of the long-range candidate space on the mean chain, so a false label rarely touches it. At p = 0 the readout falls to 0.061, below the 0.103 of ten random heads, so the measurement does register a worthless teacher. Placement matters as well as precision. A label set of the same size and precision moved onto a length-matched different chain scores 0.086 and recovers none of the ten.

The Jacobian's own false pairs land on a selected head's top-L/2 list 49.1 times more often than uniform placement would, against 45.0 times for true contacts. With every true label removed, they alone separate the selected ten from the other 650 heads by 13.46 standard deviations, against 8.19 for the true contacts. Against label selection, the ten heads chosen by ESM-2-35M's Jacobian give a paired Δ of -0.023 [-0.029, -0.018] at Wilcoxon p = 2.78e-14. Their +0.006 margin over ESM-2-650M's own Jacobian has p = 0.187. Uniform false labels and a sequence-reversed teacher recover 0 of 10 heads, and an empty teacher 1 of 10, by tie-breaking.

With zero, three, four and five designed repeat chains in a selection set of ten, fused precision is 0.747, 0.741, 0.702 and 0.346, as the band pulls in Appendix E's periodicity heads. One, two and three non-repeat chains from the same foreign pool give 0.744, 0.743 and 0.741, so the slip up to three repeat chains is the cost of any foreign chain. The screen measures the share of the teacher's picks at one sequence separation. It reads 0.031 to 0.073 on the ten selection chains and 0.050 to 0.213 on three non-repeat probes, against 0.349 to 0.616 on nine repeat chains. Any threshold in the gap classifies all 22 chains, which fixes an operating point and does not measure an error rate.

G. The K sweep and the fusion operator

Table G1: Best K per checkpoint, unweighted mean on Zhang eval-200 over K in 1, 2, 3, 5, 7, 10, 20. Best is the best head on the evaluation set itself, unlike Table H1's T1. ProGen2-xlarge's K = 10 value is its standalone run, not its sweep. § swept over K ≤ 10 only; its best-head value is not stored.

Checkpoint	Best	Best K	P at best K	P at K=10
ESM-2 8M	0.250	3	0.275	0.211
ESM-2 35M	0.455	7	0.486	0.483
ESM-2 150M	0.626	20	0.677	0.667
ESM-2 650M	0.682	20	0.767	0.762
ESM-2 3B	0.722	10	0.790	0.790
ESM-1b	0.564	7	0.669	0.661
AMPLIFY 350M	0.510	3	0.556	0.539
ProtT5-XL	0.644	10	0.727	0.727
ESMC 300M §	—	7	0.765	0.764
ESMC 600M §	—	3	0.797	0.779
ProGen2-xlarge	0.193	withheld	—	0.111

In Table G1, K = 10 sits within 0.011 of the best K on seven of the ten checkpoints swept. ESM-2-8M and AMPLIFY-350M peak at K = 3, by 0.063 and 0.017, and have the fewest strong heads, 3 and 5 by the guarded count of Table H2. ESMC 600M also peaks at K = 3, by 0.018; its stored results keep only the top 20 heads, so it has no Table H2 row. Best K is chosen on the evaluation data, so these gaps bound what tuning K per checkpoint could buy. ProGen2-xlarge's sweep reads 0.014 at K = 10 against 0.111 in its standalone run, with the same ranking set, K and head list in both files, so we withhold its best K.

We also compared the unweighted mean with nine other fusion operators. They are elementwise max, geometric mean, z-score mean, reciprocal rank fusion, positional-filter mean, entropy-gated

mean, graph closure, Hopfield iteration and spectral consensus. Each of 25 checkpoint-and-set cells searches operator and K jointly, at Bonferroni $\alpha = 0.05/63 \approx 0.0008$, or $\alpha \approx 0.0056$ for the K = 10-only CASP14-FM cells. Three cells clear their own threshold, all with spectral consensus, by 0.018, 0.013 and 0.006. At K = 10 with the paper's selection, the best alternative leads the unweighted mean by at most 0.003 on any of the four evaluation sets.

H. The per-checkpoint table, depth and redundancy

Table H1: Long-range P@L/2 on Zhang eval-200 at K = 10, heads ranked on ten chains. T1, the head ranked first on those ten chains, scored on the evaluation set. CJ, the model's own Jacobian. F, fusion. Δ is paired over proteins, 10,000 resamples, and the interval is marginal.

Checkpoint	Objective, form	T1	CJ	F	Δ F − CJ [95% CI]
ESM-2 8M	masked, logit	0.249	0.234	0.211	−0.023 [−0.034, −0.013]
ESM-2 35M	masked, logit	0.455	0.473	0.483	+0.010 [+0.000, +0.021] ‡
ESM-2 150M	masked, logit	0.626	0.655	0.667	+0.012 [+0.006, +0.018]
ESM-2 650M	masked, logit	0.678	0.733	0.762	+0.029 [+0.024, +0.035]
ESM-2 3B	masked, logit	0.722	0.742	0.790	+0.049 [+0.043, +0.054]
ESM-1b	masked, logit	0.564	0.641	0.661	+0.020 [+0.014, +0.027]
ESMC 300M	masked, logit	0.720	0.760	0.764	+0.004 [−0.002, +0.010] ‡
ESMC 600M	masked, logit	0.747	0.787	0.779	−0.008 [−0.015, −0.001] ‡
AMPLIFY 350M	masked, hidden	0.510	0.490	0.539	+0.049 [+0.039, +0.059]
ProtT5-XL	span corr., hidden	0.644	0.700	0.727	+0.027 [+0.022, +0.033]
ProGen2-xlarge	causal, hidden	0.193	0.056	0.111	+0.056 [+0.046, +0.066]
ProGen2-large	causal, hidden	0.109	—	0.114	—
Ankh-large †	span corr., hidden	0.695	0.793	0.786	−0.006 [−0.015, +0.002] ‡
ProtGPT2 †	causal, hidden	0.038	0.030	0.046	+0.017 [−0.001, +0.035] ‡

Column 2 gives objective and Jacobian form, ‡ marks a Holm p above 0.05 over the thirteen tests of Table A1, and † a 50-chain draw. ProtGPT2 is read at one residue per token. ProGen2-large has no Δ because its stored hidden-state Jacobian does not reproduce, at pooled Pearson r = −0.263 against a fresh run of the same chains. The same code reproduces the sibling checkpoint's stored result unchanged. The fusion cell, 0.114, is unaffected. The checkpoint's 50-chain head grid is also withdrawn, so its rows in Tables H2 and H3 come from a fresh grid on the ten selection chains.

Table H2: Depth and redundancy per checkpoint on Zhang eval-200, N = 200. Best is the best head on this set, Depth its relative depth. Share is the fraction of excess head-precision mass in the last $\lceil n/5 \rceil$ of n layers. Raw counts heads within a factor of two of the best; guarded asks the same above the median head.

Checkpoint	Best	Depth	Share [95% CI]	Raw	Guarded
ESM-2 8M	0.250	1.000	0.660 [0.639, 0.676]	3	3
ESM-2 35M	0.455	1.000	0.745 [0.733, 0.764]	4	3
ESM-2 150M	0.626	0.967	0.740 [0.724, 0.749]	11	9
ESM-2 650M	0.682	1.000	0.678 [0.668, 0.692]	13	13
ESM-2 3B	0.722	0.917	0.595 [0.584, 0.607]	16	16
ESM-1b	0.564	0.909	0.698 [0.686, 0.714]	14	14
AMPLIFY 350M	0.510	1.000	0.629 [0.614, 0.646]	5	5
ProtT5-XL	0.644	0.958	0.705 [0.692, 0.719]	15	15
Ankh-large	0.675	0.979	0.686 [0.665, 0.703]	10	9
ProGen2-xlarge	0.193	0.844	0.365 [0.349, 0.387]	2	2
ProGen2-large	0.153	0.531	0.221 [0.163, 0.282]	16	6
ProtGPT2	0.057	0.111	0.029 [0.024, 0.043]	469	102
ProtGPT2, native	0.068	0.861	0.081 [0.067, 0.112]	136	13

Share intervals are per-protein bootstraps, and a depth of 1.000 is the last layer. The nine non-causal checkpoints run 0.595 to 0.745 on the share and 0.909 to 1.000 on depth, and the four causal rows 0.029 to 0.365 on the share. The share also depends on the evaluation set. Over non-causal checkpoints it runs 0.593 to 0.757 on the 50-chain draw and 0.535 to 0.700 on the ten selection chains. It is 0.347

to 0.497 on CASP14-FM (Kryshtafovych et al. 2021). ESM-2-3B's 0.347 on CASP14-FM is below ProGen2-xlarge's 0.365 on Zhang eval-200.

The raw near-best count fails at the floor. When a checkpoint's best head barely clears its median head, the median is itself within a factor of two of the best, and almost every head qualifies. ProtGPT2 counts 469 of 720 heads this way, against ESM-2-3B's 16 of 1440 at about thirteen times the precision. Table H3 therefore uses the guarded count; over the eight non-causal checkpoints the raw count gives Spearman ρ = +0.881, permutation p = 0.006.

Table H3: Spearman ρ between best-head precision and the guarded near-best count, with the Spearman test's p, as checkpoints are added to the ten the relation was stated on. Ankh-large and the two ProtGPT2 conditions were scored after it was stated.

Subset	n	ρ	p
ten reference checkpoints	10	+0.815	0.0041
+ Ankh-large	11	+0.799	0.0032
+ Ankh-large + ProtGPT2, per-residue	12	+0.382	0.2198
+ Ankh-large + ProtGPT2, native	12	+0.594	0.0417
eight non-causal, both ProGen2 removed	8	+0.874	0.0045
reference minus ProGen2-large	9	+0.912	0.0006
reference minus ProGen2-large, + Ankh	10	+0.854	0.0017

ProtGPT2 lowers the relation because its best head, at 0.057 and about 4 times chance, sits near its median, so the guarded count of 102 measures noise more than redundancy. Dropping ProGen2-large raises ρ to +0.912, so the causal rows do not carry the ten-checkpoint value.