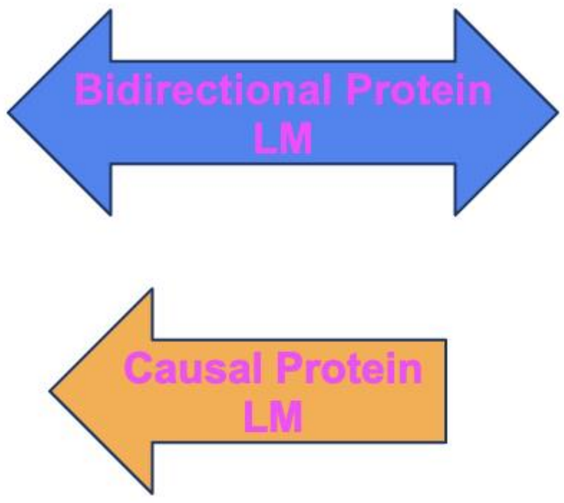
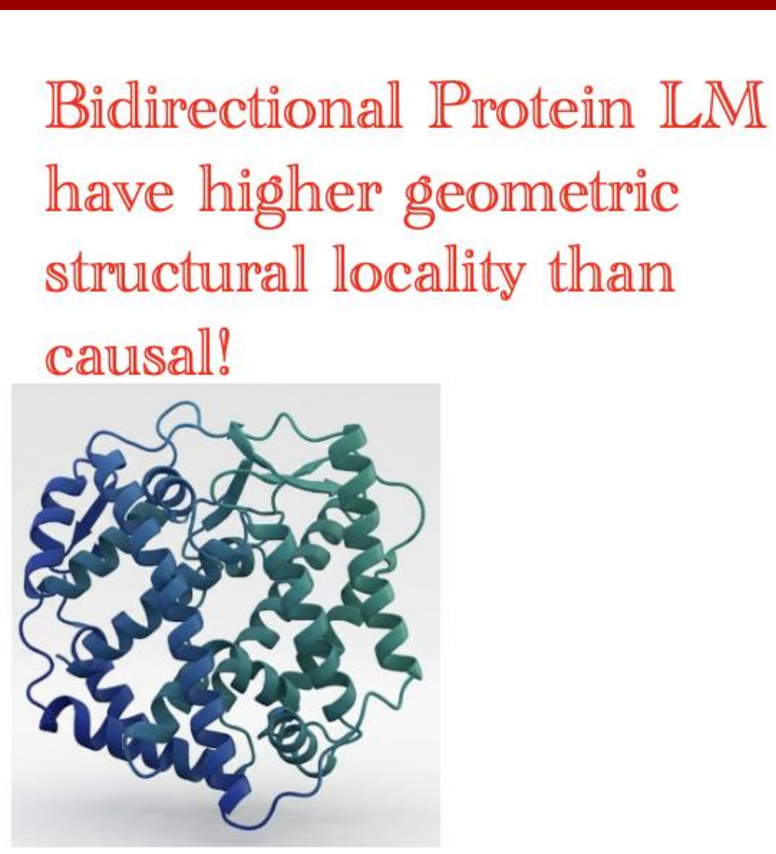




Teach people what you learned in 5 seconds (takeaway, not title)

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

Wei Gao, Alex Fedorec

1 TL;DR

1. Protein Language Models(PLM): ESM-2 (bidirectional) learns more spatially-local structural features than RITA (causal), at every depth.

2. The result holds under fold clustered bootstraps, null floors, varied seeds, and a plain probe.

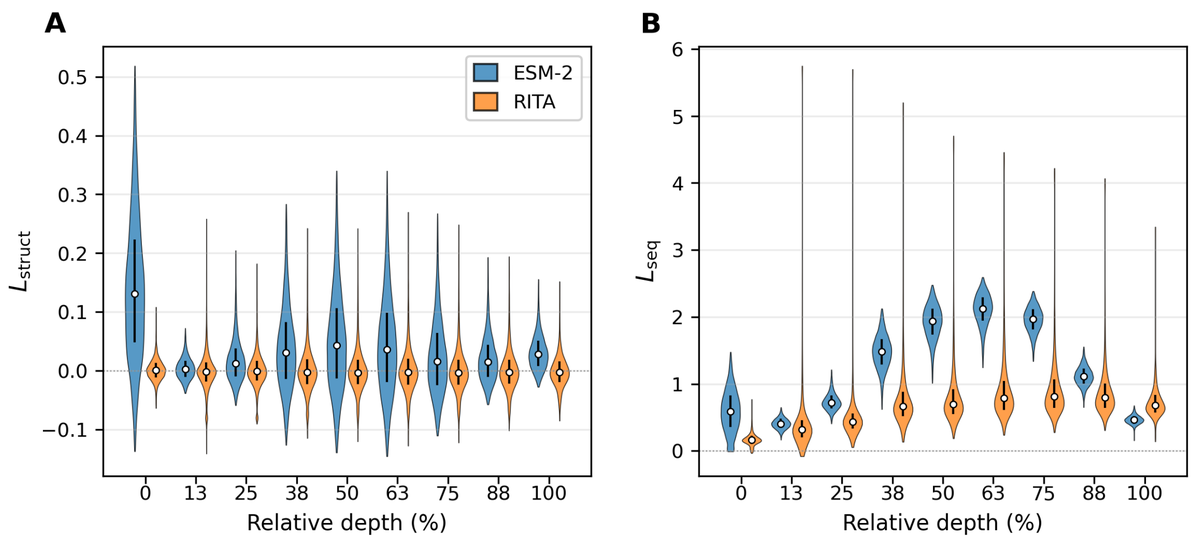
3. This demonstration is currently a family-level, geometric correlation, not objective only, and not causal proof.

2 Background / Motivating Question

- **Protein LMs encode 3D structure, but how, in their features?**
- **Do features capture long-range 3D contacts, not just sequence neighbours?**
- **Does bidirectional (ESM-2) vs causal (RITA) family change this?**

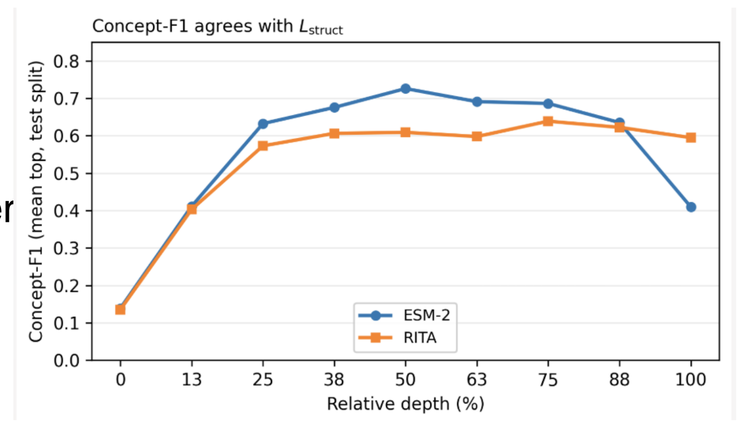
3 Method & Main Result

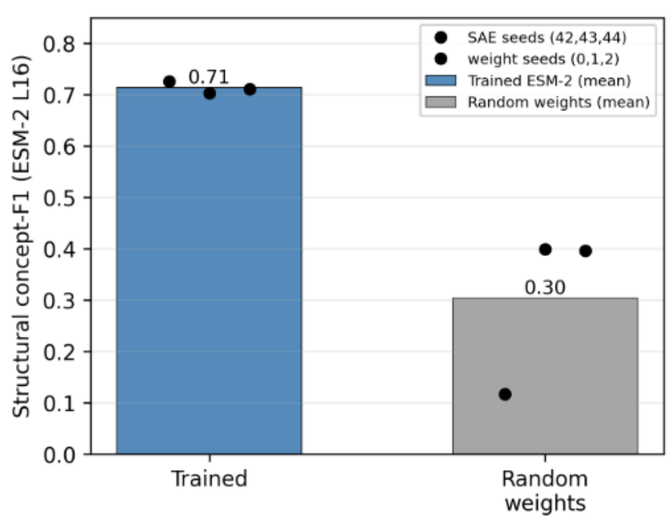
- **Structural Locality:** TopK SAEs on residue activations; per feature, test if co-active residues are long-range 3D contacts (Cα < 8 Å, seq gap ≥ 12), shuffle-corrected.
- **H1: ESM-2 > RITA** at all 9 matched depths; survives fold-cluster bootstrap and sits ESM2 ≈150–1350×; RITA ≈10–22× over the shuffled-graph floor.



4 Convergent evidence

- **Multiple lenses agree:** ESM-2 SAE concept-F1 0.73 vs RITA 0.64; ESM-2-raw > RITA-raw under a plain linear probe too — lenses near-independent (Spearman ρ ≤ ~0.11). Other analysis agree, see paper.





- **Controls hold:** Randomizing ESM-2 SAE weights collapses structural concept-F1 **0.71→0.30** (matched protocol); joint k×expansion sweep vindicates **k=256 / 8×** expansion.

5 Limitations and Discussion

- **Family-level, not objective-only:** ESM-2 and RITA differ in architecture, data, and objective — we scope the claim as a family contrast, not proof that bidirectionality causes structural locality.
- **Correlational:** ablating top-L_struct features hurts contacts (p=0.022) but weakly, and steering is directionally right yet not significant (p≈0.10) — geometric locality, not established causal control.
- **Next:** a second residue-tokenised causal PLM and ProtT5 cross-attention experiments to disentangle the confound and strengthen the mechanistic account.

Original PAPER

Extension progress for camera-ready see code below

Github & progress track

Mechanistic Interpretability Workshop, ICML 2026

zcbtwga@ucl.ac.uk