NO PROTEIN LANGUAGE MODEL BEATS A LOOKUP TABLE AT FOLD-LEVEL RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Remote homology search finds proteins that share a fold after their sequences have diverged past the range where alignment scores work. The field solves it by embedding every protein with a pretrained protein language model. However, whether those embeddings are what carries fold-level ranking has not been tested across the released models, because published comparisons are small and leave the readout layer, label leakage and index size uncontrolled. Here we evaluate 115 released protein language models from 2.6M to 16.1B parameters on SCOPe40 retrieval under one protocol. We compare them with the Additive Normalised Gated K-mer index (ANGK), a static index that answers a query by table lookup with no forward pass. Fold-level AUROC falls as parameter count grows, with a Spearman correlation of -0.57 across the 115 models. The 39.9 MiB index reaches 0.9555 where no model reaches 0.9500. It also encodes eleven times faster than the fastest of them and holds first place at the topology level of a second taxonomy. A four-way position bin on 3-mers supplies that margin. The same index covers 2.41×10^8 AlphaFold structures in 7.4 hours. One pass over them finds 4,931 structures that connect 53 fold pairs SCOPe keeps apart. Fold-level retrieval now has a baseline with no forward pass.

1 INTRODUCTION

Most newly sequenced proteins have no experimentally characterised function. The standard way to assign one is to find a characterised protein similar enough for its annotation to transfer. When the two sequences are similar, BLAST finds that match in seconds. When the relationship is remote, sequence identity decays to the level of unrelated pairs and the similarity survives only in the structure. Drug target identification, microbial genome annotation and enzyme function prediction all rest on recovering that similarity. Structural classifications name three levels of relationship. A family holds domains with a clear evolutionary relationship that sequence identity establishes. A superfamily holds domains whose common ancestry is probable from structural and functional evidence. A fold holds domains whose major secondary structure elements share the same arrangement and topological connections, with no claim of common ancestry. Below twenty percent sequence identity, alignment scores stop separating related pairs from unrelated ones. Remote homology detection is therefore decided at the fold level.

Fold assignment in practice is a retrieval problem. Each database structure is embedded once and every query is ranked against them by cosine similarity. The representation therefore sets the quality of the ranking. The default representation is a mean-pooled protein language model embedding. This pipeline has not changed since the first protein language models appeared in 2019 (Alley et al., 2019; Heinzinger et al., 2019). Parameter counts have grown from a few million to sixteen billion (Sun et al., 2024). Structure tokens (Su et al., 2024), structure adapters (Tan et al., 2024) and diffusion objectives (Wang et al., 2024) have been added. Whether fold-level ranking improves as these models grow has not been measured across the released models. Published comparisons each report twelve models or fewer under different protocols.

Three decisions inside that measurement are rarely reported. Each one moves a score by more than the differences that papers claim. The first is which layer is read. Structural information peaks before the final layer that papers read by default (Beshkov & Malthe-Sørenssen, 2026). The second is what the checkpoint has already seen. Some released models were supervised with the fold labels

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.

the benchmark tests. Published comparisons place them beside zero-shot models without saying so. The last is how much storage each index costs. A comparison between a float embedding of more than a thousand dimensions and a compressed index of sixty-four dimensions rewards the larger one. Because these three decisions are left open, no published number separates the contribution of the representation from that of the readout, the training labels and the storage budget. No published study measures how much of the fold-level ranking remains when the forward pass is removed.

Here we suggest that the pretrained encoder can be replaced by a lookup table. We fix the three decisions above and apply one protocol to 115 released protein language models from 2.6M to 16.1B parameters. We then compare them with the Additive Normalised Gated K-mer index (ANGK). ANGK cuts each structure's Foldseek 3Di string into tokens under four views and reads one embedding row per token. It then sums, normalises, gates and concatenates the rows with no forward pass. ANGK scores above every model in the sweep and encodes about eleven times faster. A four-way position bin on 3-mers supplies that margin. The same index scans the whole AlphaFold database and finds 4,931 structures that connect 53 fold pairs, among them four beta-propeller folds that SCOPe separates by blade count.

2 Related work

Protein language models. Sequence-only models include ESM2 (Lin et al., 2023), ProtTrans (Elnaggar et al., 2022), Ankh (Elnaggar et al., 2023), ProGen2 (Nijkamp et al., 2023), CARP (Yang et al., 2024) and AIDO-Protein (Sun et al., 2024). Structure-token models add structure tokens to the vocabulary. SaProt (Su et al., 2024) uses Foldseek 3Di states and DPLM-2 (Wang et al., 2025) uses learned structure codes. Structure-supervised models were trained with SCOP fold labels, as in ESM-S (Zhang et al., 2024), or fine-tuned on CATH labels, as in TEDBench (Chen et al., 2026). Multimodal and retrieval-trained models include ProTrek (Su et al., 2026), TM-Vec (Hamamsy et al., 2024) and TM-Vec 2 (Keluskar et al., 2026).

Structural alphabets. Foldseek (van Kempen et al., 2024) turns tertiary contacts into the 20-state 3Di alphabet and prefilters with spaced 3Di k-mers. ProstT5 (Heinzinger et al., 2024) translates amino-acid sequence into 3Di.

Embedding-based homology search. ProtTucker (Heinzinger et al., 2022) and embedding-based annotation transfer (Schütze et al., 2022) find relatives by nearest neighbours in language-model embedding space. pLM-BLAST (Kaminski et al., 2023) aligns per-residue embeddings and NEAR (Olson et al., 2025) searches them by nearest neighbours. TM-Vec (Hamamsy et al., 2024) and TM-Vec 2 (Keluskar et al., 2026) regress structural similarity into one vector per protein. PROTOCOL (Cohn et al., 2026) keeps every residue vector for ColBERT-style (Khattab & Zaharia, 2020) late interaction at a higher storage cost. All of these keep a neural encoder in the query path. On sequence classification, bag-of-words histograms and hashed k-mer sketches are competitive with language-model embeddings (Ali et al., 2025; Papadopoulos et al., 2025). Neither study concerns structural retrieval.

Structure space. Whether protein structures form discrete folds or a continuum has been argued since the first maps of structure space (Holm & Sander, 1996). SCOPe and CATH encode the discrete reading. Alignment-based surveys found paths between folds (Kolodny et al., 2006; Skolnick et al., 2009). Later work described the space as discrete at high similarity and continuous at low similarity (Pascual-García et al., 2009). Predicted structures at the scale of the AlphaFold database reopened the question (Barrio-Hernandez et al., 2023).

3 Methodology

3.1 Dataset and metrics

We use the SCOPe40 retrieval benchmark (Chandonia et al., 2022), version 2.08, filtered at 40 percent identity. It holds 15,176 domains. Of these, 3,021 domains spanning 251 folds are queries and the rest form the database. Every model produces one vector per domain by mean pooling over residues. Domains are ranked by cosine similarity. We evaluate the family, superfamily and fold levels defined in Section 1. At the family level a true positive shares the family. At the superfamily

level it shares the superfamily and not the family. At the fold level it shares the fold and not the superfamily. Other pairs from the same fold are removed from the ranking. Negatives at every level are pairs from different folds. We report per-query AUROC and sensitivity to the first false positive. Appendix Table A1 adds AUPRC and top-1 accuracy. The fold level is the primary level, because its pairs share structure without a claim of common ancestry.

3.2 CONTROLS

An evaluation that picks the best layer on the test set reports a number that no held-out use reproduces. We probe a fixed ladder of layers for every model, which is layer 2, half depth, 0.6 depth, 0.8 depth and the last layer. We select the layer with the highest fold AUROC on 2,257 validation domains spanning 188 folds, none of which appears among the 251 test folds. We report that layer on the test folds. For SaProt-35M, the strongest zero-shot baseline, we evaluate all twelve layers and select among them, which can only help it.

ESM-S models were trained with SCOP fold supervision that covers 223 of the 251 test folds. TEDBench models were fine-tuned on CATH T-level labels with 965 classes. We do not drop these models. They give an upper bound on what fold supervision buys, and their numbers are not zero-shot.

A model with 1,280-dimension float32 vectors needs 5,120 bytes per structure. ANGK needs 64 bytes per structure at 64-dimension int8. We therefore repeat every comparison after projecting both sides to 64 dimensions with PCA and quantising to int8. On that equal-size index we test equivalence with two one-sided tests (Schuirmann, 1987) on paired per-query differences, with a margin of 0.01 and 90 percent confidence intervals. All other scores are uncompressed and at native dimension. In the figures a star marks ANGK and a dashed line marks its fold-level AUROC, unless a caption states otherwise. Error bars are standard deviations over three seeds.

3.3 MODEL

A backbone is converted to a 3Di string by Foldseek (van Kempen et al., 2024), whose pretrained encoder we keep frozen. We tokenise that string under four views (Figure 1a). The four views are contiguous 1-mers (40 types), contiguous 2-mers (649 types), gapped 3-mers at offsets (0,1,3) (8,462 types) and contiguous 3-mers tagged with a coarse position bin (31,724 types). The position bin is the only way global order enters the representation (Figure 1b). A token at relative position i in a 3Di string of n tokens gets the label $\lfloor 4i/n \rfloor$, so the same local motif near the N-terminus and near the C-terminus becomes two different tokens.

Each view owns an embedding table of size $|V|$ by 1,024. A structure's view vector is the normalised sum of its token rows weighted by token counts, computed as one embedding-bag reduction, and the view vectors are scaled by a learned positive gate, concatenated and normalised again. Writing c_t for the count of token t under view v,

$$u_v = \mathrm{normalise}\Big(\textstyle\sum_t c_t\, E_v[t]\Big), \qquad z = \mathrm{normalise}\big(\mathrm{concat}_v(g_v u_v)\big),$$
$$\cos(z, z') = \frac{\sum_v g_v^2 \cos(u_v, u'_v)}{\sum_v g_v^2}.$$

The second line makes the similarity a mixture over views with weights equal to the squared gates, which is why dropping a view costs something. Per-view normalisation is needed because views differ by an order of magnitude in tokens per structure, and without it the largest vocabulary dominates. Inference costs $4L$ table reads for a structure of length L, with no multiplication by a weight matrix at any point.

3.4 TRAINING AND COST

The tables and four gates are the only trainable parameters. They are trained for 3,000 steps by distillation against a frozen TM-Vec 2 teacher on structural-similarity pairs. The teacher is absent at

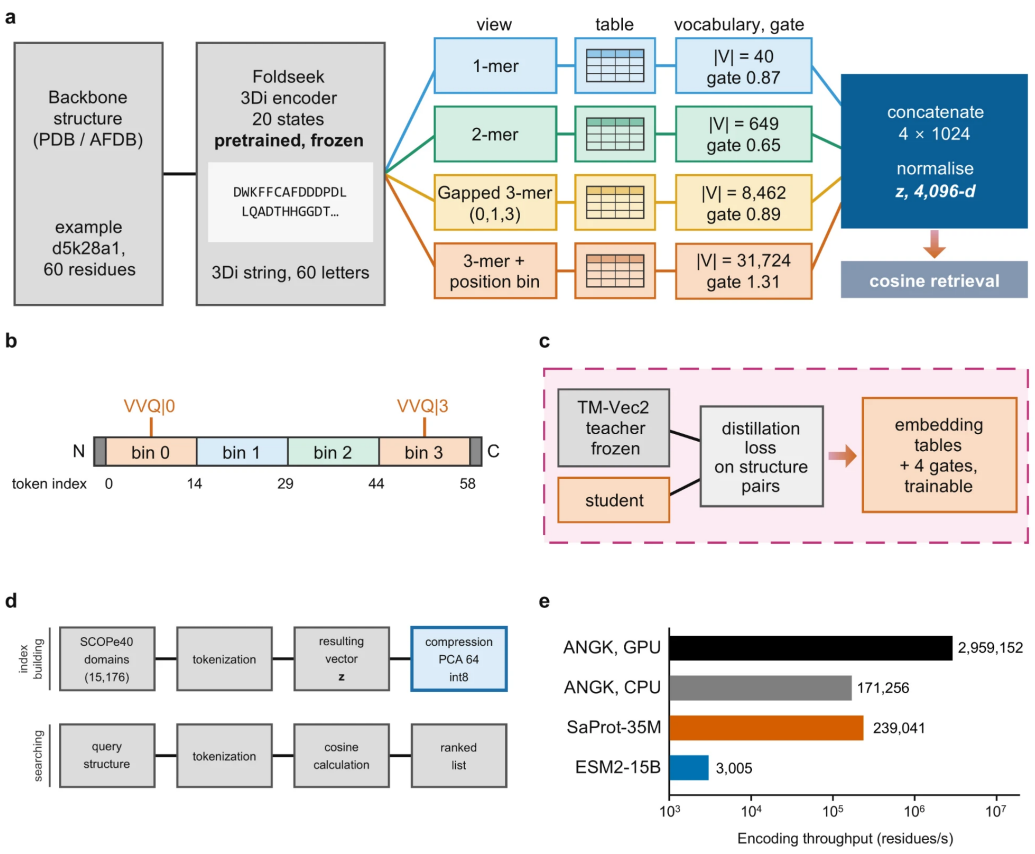

Figure 1: **Architecture of ANGK.** a, Inference. A 3Di string is tokenised under four views. Each token is looked up in the table of its view. The view vectors are normalised, gated and concatenated. b, The position bin, which tags each 3-mer with the quarter of the sequence it falls in. c, Training by distillation from a frozen teacher. d, Encoding of the database and of a query. e, Encoding throughput on one GPU.

inference. The index is a static int8 table of 40,875 by 1,024 entries and occupies 39.9 MiB. Adding a structure costs one tokenisation pass.

The sequence-only variant replaces the measured 3Di string with one that ProstT5 (Heinzinger et al., 2024) predicts from the amino-acid sequence. The tables are unchanged. The ablation trains three further variants with three seeds each. The first has a two-way position bin. The second has no position bin. The third is a capacity control that replaces bin labels with a deterministic hash, which enlarges the vocabulary without carrying position.

Encoding throughput was timed for every model on one RTX 5090 with 32 GB of VRAM, driven by an AMD Ryzen 9 9950X with 128 GB of system memory. Each run used the same 300 longest domains, the same batching rule and one warm-up batch. Load time was excluded. Models too large for 32 GB of VRAM were run with CPU offloading or a smaller token budget, so their throughput is a lower bound.

3.5 ECOD and the AlphaFold scan

ECOD classifies the same domains on evolutionary evidence. Its levels nest as X (possible homology), H (established homology), T (topology) and F (family). We map 14,893 of our 15,176 domains to ECOD and score every model again with the same vectors, queries and protocol.

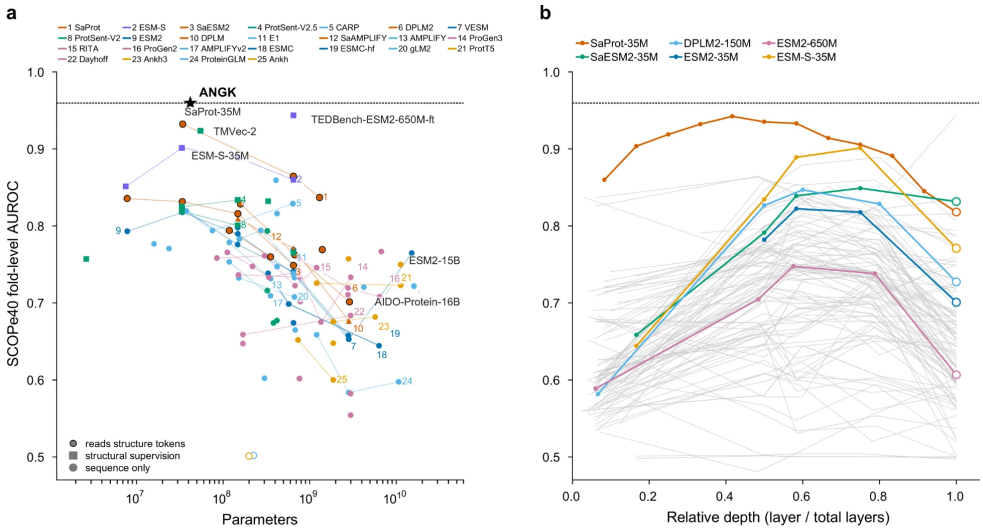

Figure 2: **Fold-level AUROC against model size and layer depth.** a, SCOPe40 fold-level AUROC against parameter count for 115 models. Colour marks the training objective. Lines join members of one family, and numbered families are named in the legend. Squares mark structure-supervised models and black edges mark structure-token models. b, Fold-level AUROC against relative layer depth for models probed at five or more layers, with six highlighted. Open circles mark final layers.

Table 1: AUROC at the family, superfamily and fold levels of SCOPe40 for ANGK, its sequence-only variant and four protein language models. Negatives at every level are pairs from different folds.

Level	ANGK	ANGK, sequence only	TEDBench	SaProt-35M	ProtXLNet	ESM2-15B
family	0.9876	0.9829	0.9895	0.9807	0.9709	0.9812
superfamily	0.9781	0.9723	0.9814	0.9630	0.9452	0.9528
fold	**0.9555**	0.9383	0.9435	0.9320	0.8591	0.7646

We build the same index over all 241,070,489 structures of the AlphaFold database as 256-dimension int8 vectors. One SCOPe representative per fold serves as an anchor and every structure serves as a probe. A probe bridges two folds when it attaches to both above a cosine of 0.70. The scan uses no SCOPe label beyond the choice of anchors. We anchor 916 of the 1,257 SCOPe40 folds after removing folds of short, low-complexity or membrane domains (Appendix A). The 341 excluded folds are untested and not isolated. Probes are filtered for length and low complexity in the same way. Fold pairs whose anchors are closer than 0.55 are excluded as already close without a probe. Membrane proteins are excluded. A probe within a cosine of 0.98 of an anchor counts as a near-duplicate.

As a null, we apply a random orthogonal transform to the query vectors. It preserves every inner product among the queries and destroys only their alignment with the corpus. Two robustness runs lower the length filter to 80 residues and raise the anchors to five per fold (Appendix Table A4). Bridging structures are aligned to their anchors in amino acids and in 3Di (Appendix Table A3).

4 Experiments

4.1 A lookup table with coarse position beats every protein language model at the fold level

Across the six entries of Table 1, family-level AUROC spans 0.019 and fold-level AUROC spans 0.191. ANGK scores 0.0120 fold AUROC above the strongest model, the CATH-supervised

Table 2: AUROC of ANGK and its sequence-only variant at the four ECOD levels, with the number of models that score higher, the largest gap and whether that gap falls inside an equivalence margin of 0.01.

ECOD level	ANGK	ANGK, sequence only	Models above	Largest gap	Equivalent at 0.01
X (remote)	0.9031	0.8954	1	SaProt-35M, 0.0039	yes
H (homology)	0.9188	0.9152	3	TM-Vec 2, 0.0171	no for TM-Vec 2 and TEDBench
T (topology)	**0.9684**	0.9595	**0**	none	-
F (family)	0.9660	0.9624	20	ProTrek-650M, 0.0062	yes

TEDBench-ESM2-650M-ft. It scores 0.0235 above the strongest zero-shot model, SaProt-35M, and 0.0964 above the strongest sequence-only model, ProtXLNet.

Fold-level AUROC falls with parameter count, with a Spearman correlation of -0.57 across the 115 models (Figure 2a). Inside most families the largest member is not the best. The three largest models in the sweep score below many smaller ones. No training objective in the sweep produces a model above ANGK. Most layer curves peak before the final layer (Figure 2b). The exception is TEDBench, which was fine-tuned at its final layer. For SaProt-35M, reading the final layer loses 0.124 AUROC against the best layer. Validation-fold selection costs a baseline 0.0079 AUROC on average against test-set selection. It picks the test-optimal layer for 63 of 115 models.

Compressing ANGK to 64-dimension int8 raises its fold AUROC by 0.0039. At equal index size every fold-level difference is positive, and all but one 90 percent interval excludes zero (Appendix Figure A1). The one difference inside the margin is against SaProt-35M read at its test-optimal layer. ANGK scores 0.0253 above its teacher TM-Vec 2 and 0.1670 above the distilled student TM-Vec 2s, and both intervals exclude zero. At the superfamily level TEDBench-ESM2-650M-ft is equivalent to ANGK. On sensitivity to the first false positive, 18 models rank above ANGK. Six rank above it on AUPRC and three on top-1 accuracy (Appendix Table A1).

The sequence-only variant scores 0.017 below ANGK. Its predicted strings agree with the Foldseek strings at a median of 0.693 per residue. It still scores above every sequence-only and structure-token model.

ANGK encodes about eleven times faster than the fastest protein language model and twelve times faster than SaProt-35M (Figure 3). Encoding the AlphaFold database takes 7.4 hours with ANGK on one GPU, 12 times less than with SaProt-35M and 990 times less than with ESM2-15B. Its index of that database needs 4 to 40 times less storage than float16 embeddings of 480 to 5,120 dimensions. On the CPU alone ANGK is faster than every 150M-parameter transformer on the GPU.

Adding a four-way position bin to the 3-mer view raises fold AUROC by 0.012 (Figure 4a, Appendix Table A5). A two-way bin gives 70 percent of that gain. The capacity control scores 0.0025 below the unbinned index. The learned gates give the position-binned view twice the weight of plain 2-mers (Figure 4b).

4.2 An independent taxonomy agrees, and 4,931 of 241 million structures cross fold boundaries

At the ECOD topology level ANGK is first of 116 (Table 2). At the family level twenty models edge past it, and every difference tested there falls inside the equivalence margin. SaProt-35M is higher at X and H within the margin. TM-Vec 2 and the CATH-supervised TEDBench model are higher at H outside the margin.

No sequence-only model passes ANGK at SCOPe fold or at ECOD X, H or T. Fourteen pass it at ECOD F. At the topology level the sequence-only variant of ANGK scores 0.0054 above the structure-token SaProt-35M.

The scan over the AlphaFold database takes 13 minutes on one GPU. It finds 4,931 structures that bridge 53 fold pairs, about one structure in 48,900 (Appendix Table A2).

Of the 916 anchored folds, 867 (94.7 percent) attract no bridging structure (Figure 5a). The 49 bridged folds fall into eight connected components (Figure 5b, Appendix Table A6). Of the 53

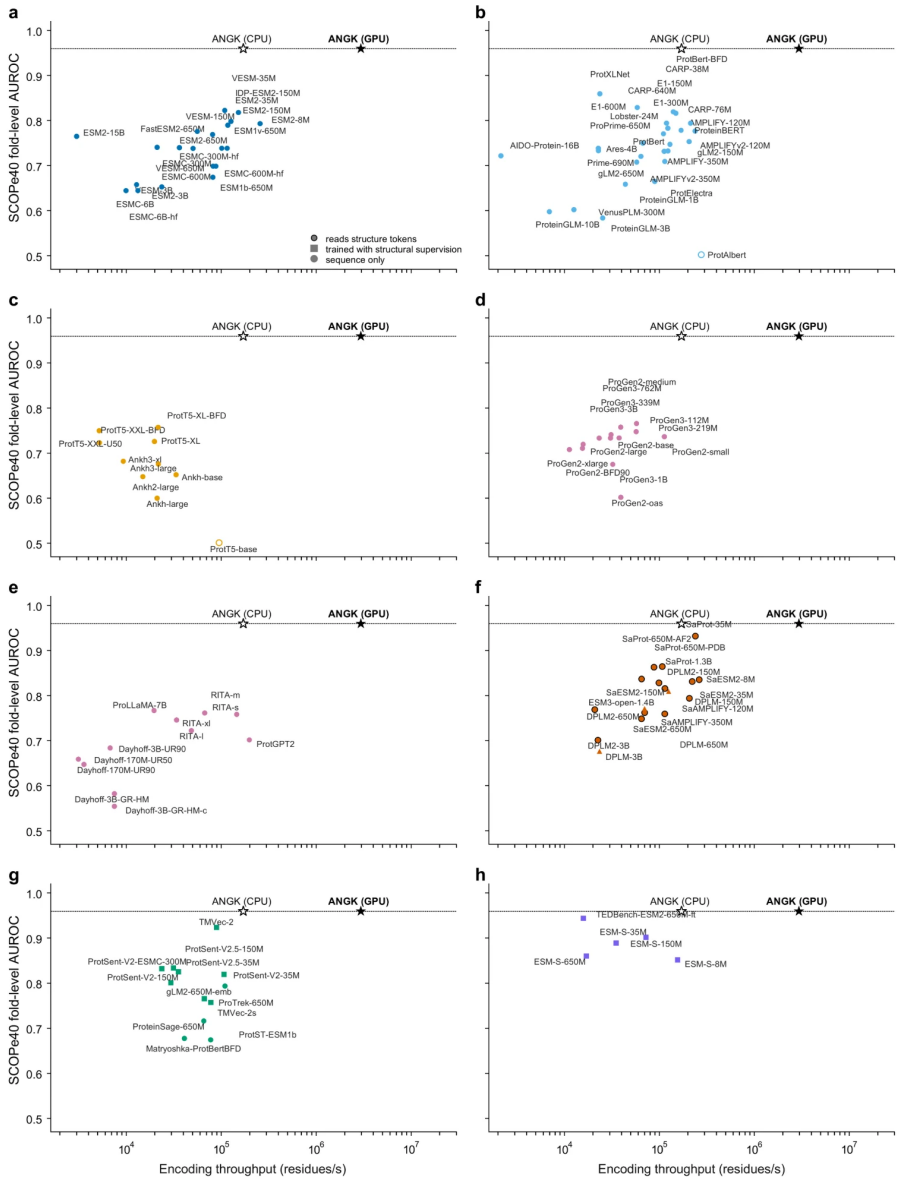

Figure 3: **Fold-level AUROC against encoding throughput.** Each panel shows one group of models. a, ESM masked language models. b, Other masked encoders. c, Encoder-decoder models. d, ProGen autoregressive models. e, Other autoregressive models. f, Discrete diffusion models (triangles) and structure-token models (circles). g, Retrieval-trained models. h, Label-supervised models. Open symbols mark two models with near-constant embeddings. The open star is ANGK on the CPU.

bridged pairs, 52 lie above the 99th percentile of cosine between anchor pairs (Figure 5c). The exclusion at 0.55 sets the upper end of that range.

The largest entry, six-bladed against seven-bladed beta-propellers, has 2,127 bridging structures. The four propeller folds that SCOPe separates by blade count fall into one component (Figure 5d). A second pair joins right-handed and left-handed single-stranded beta-helices, which differ only in handedness. Lowering the length filter to 80 residues keeps all 53 bridged pairs. With five anchors per fold the propeller folds stay in one component (Appendix Table A4). The median amino-acid identity of a bridging structure to its anchors is 0.056, against 0.048 for random structures (Appendix Table A3).

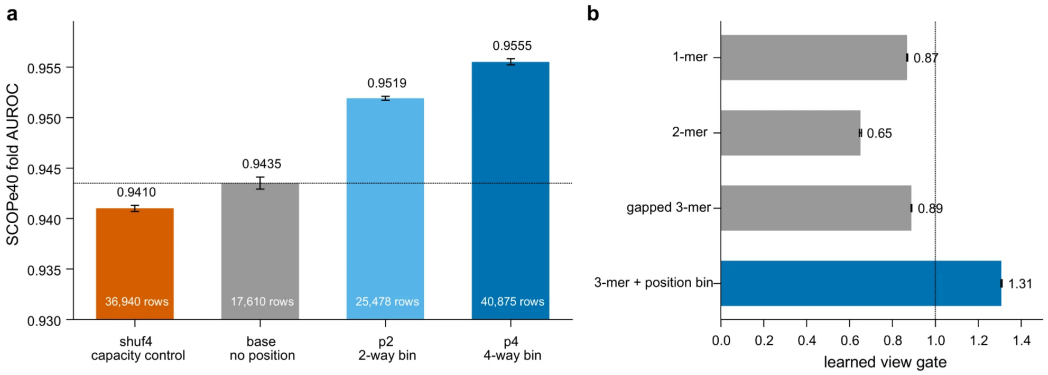

Figure 4: **Position-bin ablation and learned view gates.** a, SCOPe40 fold-level AUROC of four ANGK variants. The number inside each bar is the vocabulary size. The dashed line marks the variant without position information. b, Learned view gates of the final model.

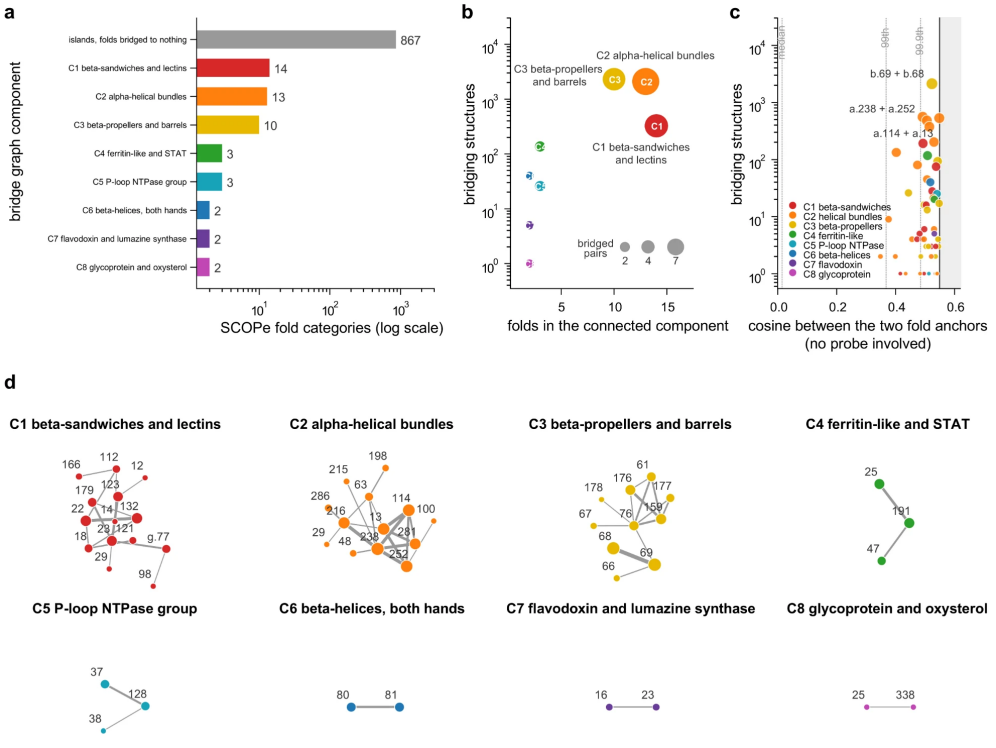

Figure 5: **Bridges between SCOPe folds in the AlphaFold database.** a, The 916 anchored SCOPe folds divided into isolated folds and the eight connected components C1 to C8. The same colours mark the components in every panel and in Figure 6. b, Folds per component against bridging structures. Point area scales with the number of bridged pairs. c, Anchor cosine of each bridged pair against its bridging structures. Point area scales with the number of distinct taxa. Dashed lines mark the median, the 99th and the 99.9th percentile of all anchor pairs. The shaded band marks pairs excluded by design. d, The eight components. Node spacing follows one minus the anchor cosine on one shared scale. Node area and line width scale with bridging structures.

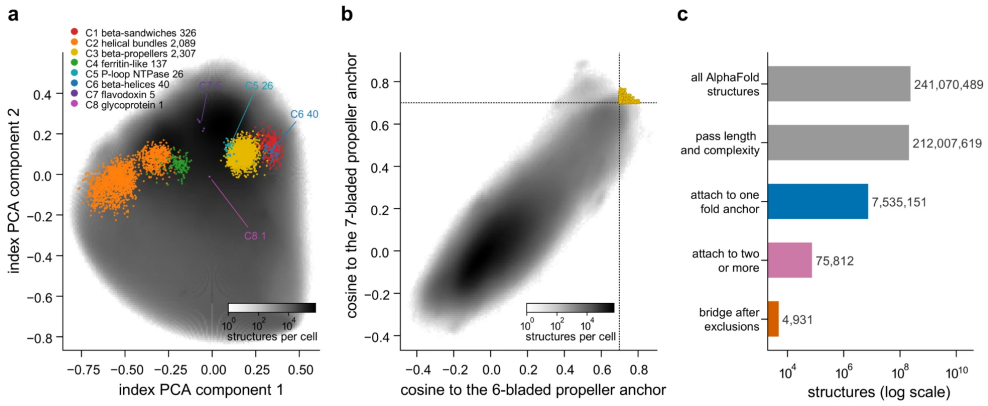

Figure 6: **The AlphaFold database in the space of the index.** a, Density of all 241,070,489 AlphaFold structures on the first two principal components of the index, with bridging structures coloured by component. b, The same structures against the six-bladed and seven-bladed propeller anchors. Dashed lines mark the attachment threshold of 0.70. c, Structures remaining after each step of the scan.

The corpus narrows to the 4,931 bridging structures in four steps (Figure 6c). On the first two principal components of the index the bridging structures occupy a few compact regions away from the densest part of the map (Figure 6a). Against the six-bladed and seven-bladed propeller anchors, every structure that passes 0.70 on both axes belongs to the propeller component (Figure 6b). Under the orthogonal null the full pass finds no bridging structure.

5 Discussion

Masked-token pretraining scores the identity of a residue given its neighbours. Mean pooling then averages the residue vectors into one vector per protein and removes their order along the sequence. A pooled embedding therefore keeps the local context that separates close relatives and loses the global arrangement that defines a fold. The narrow family-level spread and the wide fold-level spread in Table 1 show this pattern. The final layer sits closest to the masked-token objective and in most models carries less fold-level signal than the middle layers (Figure 2b). Larger models reach a lower masked-token loss (Lin et al., 2023). This suggests that added parameters go to local context. This is consistent with the fall of fold-level AUROC with parameter count (Figure 2a) and with earlier doubts about scaling (Fournier et al., 2024).

The index keeps token composition and adds global order at a resolution of one quarter of the sequence. Its margin comes from the position bin and not from table capacity (Figure 4a). Fold-level rank therefore depends on composition together with coarse order. Both quantities are additive in sequence length and need no contextualisation. ANGK has no access to the sequence homology and functional annotation that ECOD uses at its X and H levels. It is first at ECOD T and at SCOPe fold, where both schemes classify structural arrangement (Table 2).

Contextual models rank above the index on sensitivity, on top-1 accuracy and at the ECOD family level (Appendix Table A1, Table 2). Each of these measures rewards placing one close relative first. The fold-level order of the whole database does not depend on contextualisation.

A bridge marks two folds that the representation already places near each other (Figure 5c). Because ANGK does not count repeats, propellers that differ only in blade count look alike to it. The propeller component can come from this insensitivity as well as from structural intermediates. Structural alignment of the 4,931 bridging structures is needed to decide between them.

AI use statement

A large language model was used only to edit the English of this paper.

Ethics statement

This work evaluates released models on public structural data. It introduces no new biological capability and it lowers the compute cost of homology search.

Reproducibility statement

The code and the source data for every figure and table are provided as anonymised supplementary material.

References

Sarwan Ali, Taslim Murad, Imdadullah Khan, and Safi Faizullah. When do cheap embeddings beat protein language models? A theoretically-grounded hashing sketch for biological sequence classification. *arXiv preprint arXiv:2512.10147*, 2025.

Ethan C. Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M. Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, 16(12):1315–1322, 2019. doi: 10.1038/s41592-019-0598-1.

Inigo Barrio-Hernandez, Jingi Yeo, Jürgen Jänes, Milot Mirdita, Cameron L. M. Gilchrist, Tanita Wein, Mihaly Varadi, Sameer Velankar, Pedro Beltrao, and Martin Steinegger. Clustering predicted structures at the scale of the known protein universe. *Nature*, 622(7983):637–645, 2023. doi: 10.1038/s41586-023-06510-w.

Kosio Beshkov and Anders Malthe-Sørenssen. Towards understanding the shape of representations in protein language models. In *International Conference on Learning Representations*, 2026.

John-Marc Chandonia, Lindsey Guan, Shiangyi Lin, Changhua Yu, Naomi K Fox, and Steven E Brenner. SCOPe: improvements to the structural classification of proteins – extended database to facilitate variant interpretation and machine learning. *Nucleic Acids Research*, 50(D1):D553–D559, 2022. doi: 10.1093/nar/gkab1054.

Dexiong Chen, Andrei Manolache, Mathias Niepert, and Karsten Borgwardt. Protein fold classification at scale: Benchmarking and pretraining. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Gabrielle Cohn, Rohan Gumaste, Minh Hoang, and Vihan Lakshman. PROTOCOL: Late interaction retrieval for protein homolog search. *arXiv preprint arXiv:2605.29158*, 2026.

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

Ahmed Elnaggar, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, and Burkhard Rost. Ankh: Optimized protein language model unlocks general-purpose modelling. *bioRxiv*, 2023. doi: 10.1101/2023.01.16.524265.

Quentin Fournier, Robert M. Vernon, Almer van der Sloot, Benjamin Schulz, Sarath Chandar, and Christopher James Langmead. Protein language models: Is scaling necessary? *bioRxiv*, 2024. doi: 10.1101/2024.09.23.614603.

Tymor Hamamsy, James T. Morton, Robert Blackwell, Daniel Berenberg, Nicholas Carriero, Vladimir Gligorijevic, Charlie E. M. Strauss, Julia Koehler Leman, Kyunghyun Cho, and Richard Bonneau. Protein remote homology detection and structural alignment using deep learning. *Nature Biotechnology*, 42(6):975–985, 2024. doi: 10.1038/s41587-023-01917-2.

Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, and Burkhard Rost. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, 20(1):723, 2019. doi: 10.1186/s12859-019-3220-8.

Michael Heinzinger, Maria Littmann, Ian Sillitoe, Nicola Bordin, Christine Orengo, and Burkhard Rost. Contrastive learning on protein embeddings enlightens midnight zone. *NAR Genomics and Bioinformatics*, 4(2):lqac043, 2022. doi: 10.1093/nargab/lqac043.

Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, 6(4):lqae150, 2024. doi: 10.1093/nargab/lqae150.

Liisa Holm and Chris Sander. Mapping the protein universe. *Science*, 273(5275):595–602, 1996. doi: 10.1126/science.273.5275.595.

Kamil Kaminski, Jan Ludwiczak, Kamil Pawlicki, Vikram Alva, and Stanislaw Dunin-Horkawicz. pLM-BLAST: distant homology detection based on direct comparison of sequence representations from protein language models. *Bioinformatics*, 39(10):btad579, 2023. doi: 10.1093/bioinformatics/btad579.

Aryan Keluskar, Paarth Batra, Valentyn Bezshapkin, James T. Morton, and Qiyun Zhu. TM-Vec 2s: Accelerated protein remote homology detection. *bioRxiv*, 2026. doi: 10.64898/2026.02.05.704073.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pp. 39–48, 2020. doi: 10.1145/3397271.3401075.

Rachel Kolodny, Donald Petrey, and Barry Honig. Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. *Current Opinion in Structural Biology*, 16(3):393–398, 2006. doi: 10.1016/j.sbi.2006.04.007.

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

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

Daniel Olson, Thomas Colligan, Daphne Demekas, Jack W Roddy, Ken Youens-Clark, and Travis J Wheeler. NEAR: neural embeddings for amino acid relationships. *Bioinformatics*, 41(Supplement 1):i449–i457, 2025. doi: 10.1093/bioinformatics/btaf198.

Frixos Papadopoulos, Tilman Sanchez-Elsner, Mahesan Niranjan, and Ashley I. Heinson. Bag-of-words is competitive with sum-of-embeddings language-inspired representations on protein inference. *PLOS One*, 20(8):e0325531, 2025. doi: 10.1371/journal.pone.0325531.

Alberto Pascual-García, David Abia, Ángel R. Ortiz, and Ugo Bastolla. Cross-over between discrete and continuous protein structure space: Insights into automatic classification and networks of protein structures. *PLoS Computational Biology*, 5(3):e1000331, 2009. doi: 10.1371/journal.pcbi.1000331.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987. doi: 10.1007/BF01068419.

Konstantin Schütze, Michael Heinzinger, Martin Steinegger, and Burkhard Rost. Nearest neighbor search on embeddings rapidly identifies distant protein relations. *Frontiers in Bioinformatics*, 2:1033775, 2022. doi: 10.3389/fbinf.2022.1033775.

Jeffrey Skolnick, Adrian K. Arakaki, Seung Yup Lee, and Michal Brylinski. The continuity of protein structure space is an intrinsic property of proteins. *Proceedings of the National Academy of Sciences*, 106(37):15690–15695, 2009. doi: 10.1073/pnas.0907683106.

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

Jin Su, Yan He, Shiyang You, Shiyu Jiang, Xibin Zhou, Xuting Zhang, Yuxuan Wang, Xining Su, Igor Tolstoy, Xing Chang, Hongyuan Lu, and Fajie Yuan. A trimodal protein language model enables advanced protein searches. *Nature Biotechnology*, 44(8):1310–1316, 2026. doi: 10.1038/s41587-025-02836-0.

Ning Sun, Shuxian Zou, Tianhua Tao, Sazan Mahbub, Dian Li, Yonghao Zhuang, Hongyi Wang, Xingyi Cheng, Le Song, and Eric P. Xing. Mixture of experts enable efficient and effective protein understanding and design. *bioRxiv*, 2024. doi: 10.1101/2024.11.29.625425.

Yang Tan, Mingchen Li, Bingxin Zhou, Bozitao Zhong, Lirong Zheng, Pan Tan, Ziyi Zhou, Huiqun Yu, Guisheng Fan, and Liang Hong. Simple, efficient, and scalable structure-aware adapter boosts protein language models. *Journal of Chemical Information and Modeling*, 64(16):6338–6349, 2024. doi: 10.1021/acs.jcim.4c00689.

Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, 42(2):243–246, 2024. doi: 10.1038/s41587-023-01773-0.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52309–52333, 2024.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. DPLM-2: A multimodal diffusion protein language model. In *International Conference on Learning Representations*, 2025.

Kevin K. Yang, Nicolo Fusi, and Alex X. Lu. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Systems*, 15(3):286–294.e2, 2024. doi: 10.1016/j.cels.2024.01.008.

Zuobai Zhang, Jiarui Lu, Vijil Chenthamarakshan, Aurélie Lozano, Payel Das, and Jian Tang. Structure-informed protein language model. *arXiv preprint arXiv:2402.05856*, 2024.

A Additional analyses at 2.41×10^8 structures

Excluded folds. Folds whose domains are all shorter than 100 residues (271 folds), all low-complexity in 3Di (1) or in SCOPe class f (69) are not used as anchors. Short domains, low-complexity strings and transmembrane helices attach to many unrelated anchors at once. These filters can only remove bridges.

Taxonomic distribution. Every bridging structure carries a UniProt accession and an NCBI taxonomy identifier. We recovered both for all 4,931 of them. The 2,127 structures that join six-bladed and seven-bladed propellers come from 1,416 distinct taxa. The most frequent taxon contributes 39 of them, which is 1.8 percent. Among the 19 pairs with at least 20 bridging structures, the most frequent taxon contributes between 1.5 and 23.1 percent. We compare against a random one percent sample of the corpus, 2,409,109 structures, because AlphaFold covers some proteomes far more deeply than others and a single over-represented organism could produce a bridge on its own. The most frequent taxon for the propeller pair, an uncultured Bacteroidota bacterium, is enriched 7.7 fold over that background on 39 observations, and the other leading taxa rest on single-digit counts. The propeller bridges are therefore spread across the tree of life. They do not come from one deeply sequenced proteome. We release the accession list so the individual structures can be examined.

Helical bundles. Long helical bundles fill much of the bridge list. Regressing embedding similarity on 3Di composition alone explains 27.4 percent of the variance. Removing that component leaves the ordering unchanged. Helical bundles resemble each other in arrangement and not only in content. This is a property of the architecture and not evidence of descent. We report the list with that caveat attached.

B Equivalence testing at equal index size

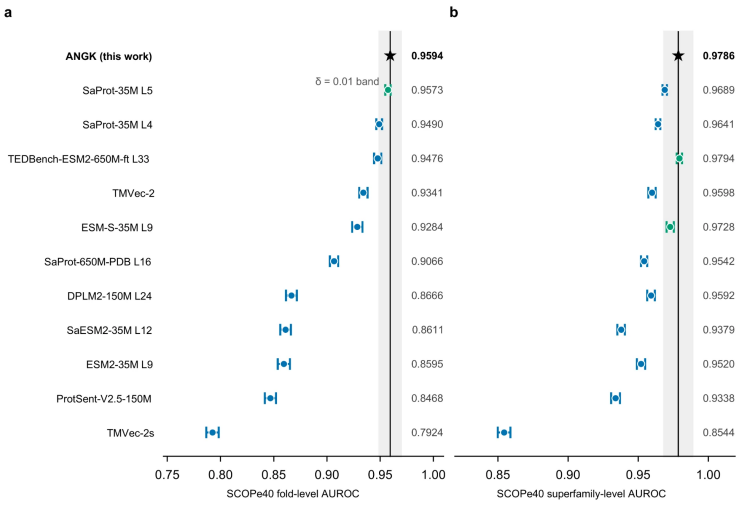

Figure A1: **Equivalence tests at equal index size.** Mean per-query AUROC of ANGK and ten models at equal index size, with 90 percent confidence intervals of the paired difference. The shaded band is the equivalence margin of 0.01, and green marks models inside it. a, Fold level. b, Superfamily level. SaProt-35M appears at layers 4 and 5.

C ALL 116 ENTRIES UNDER ONE PROTOCOL

Table A1: All 116 entries ordered by SCOPe40 fold-level AUROC. Layer is selected on validation folds. Oracle layer is the best layer on the test folds. The last column lists the metrics on which a model is ahead of ANGK. ESMC checkpoints appear twice because the same weights are distributed through two channels.

#	Model	Group	Params	Dim	Layer	SENS	AUROC	AUPRC	Top-1	Oracle layer	Oracle AUROC	Residues s-1	Ahead of ANGK on
0	ANGK (this work)	This work	42M	4096	-	0.0440	0.9555	0.3482	0.5585	-		2,959,152	
1	TEDBench-ESM2-650M-ft	Fold-supervised (label leakage)	652M	1280	L33	0.1387	0.9435	0.5013	0.6474	L33	0.9435	15,875	SENS, AUPRC, Top-1
2	SaProt-35M	Diffusion / structure-token	34M	480	L4	0.0677	0.9320	0.3637	0.5743	L5	0.9422	239,041	SENS, AUPRC, Top-1
3	TMVec-2	Retrieval-trained	55M	512	OUT	0.0828	0.9233	0.3911	0.4901	OUT	0.9233	89,524	SENS, AUPRC
4	ESM-S-35M	Fold-supervised (label leakage)	34M	480	L9	0.1765	0.9012	0.4420	0.5520	L9	0.9012	72,521	SENS, AUPRC
5	ESM-S-150M	Fold-supervised (label leakage)	148M	640	L24	0.1176	0.8888	0.4376	0.5830	L24	0.8888	35,134	SENS, AUPRC, Top-1
6	SaProt-650M-PDB	Diffusion / structure-token	652M	1280	L16	0.0533	0.8645	0.2632	0.5257	L16	0.8645	107,250	SENS
7	SaProt-650M-AF2	Diffusion / structure-token	652M	1280	L16	0.0405	0.8631	0.2498	0.5047	L16	0.8631	87,722	
8	ESM-S-650M	Fold-supervised (label leakage)	651M	1280	L26	0.0985	0.8598	0.3946	0.5550	L26	0.8598	17,070	SENS, AUPRC
9	ProtXLNet	Masked LM, other	409M	1024	L15	0.0362	0.8591	0.2228	0.3415	L15	0.8591	23,684	
10	ESM-S-8M	Fold-supervised (label leakage)	7.5M	320	L4	0.1017	0.8511	0.2877	0.4292	L4	0.8511	156,098	SENS
11	SaProt-1.3B	Diffusion / structure-token	1.3B	1280	L2	0.0041	0.8368	0.1235	0.2515	L2	0.8368	64,914	
12	SaESM2-8M	Diffusion / structure-token	7.8M	320	L6	0.0306	0.8354	0.1790	0.3333	L6	0.8354	262,251	
13	ProtSent-V2.5-150M	Retrieval-trained	148M	640	OUT	0.0446	0.8334	0.2291	0.3971	OUT	0.8334	31,475	SENS
14	ProtSent-V2-ESMC-300M	Retrieval-trained	333M	960	OUT	0.0549	0.8319	0.2359	0.3848	OUT	0.8319	23,672	SENS
15	SaESM2-35M	Diffusion / structure-token	34M	480	L12	0.0285	0.8315	0.1694	0.3678	L9	0.8488	221,660	
16	CARP-640M	Masked LM, other	643M	1280	L33	0.0745	0.8287	0.2236	0.4386	L33	0.8287	58,678	SENS
17	DPLM2-150M	Diffusion / structure-token	159M	640	L24	0.0497	0.8284	0.2640	0.4450	L18	0.8468	98,981	SENS
18	ProtSent-V2.5-35M	Retrieval-trained	34M	480	OUT	0.0289	0.8250	0.2172	0.4263	OUT	0.8250	35,507	
19	VESM-35M	Masked LM, ESM lineage	34M	480	L9	0.0538	0.8223	0.2095	0.4082	L7	0.8321	109,097	SENS
20	CARP-38M	Masked LM, other	38M	1024	L9	0.0271	0.8191	0.1942	0.3398	L12	0.8202	140,971	
21	ProtSent-V2-35M	Retrieval-trained	34M	480	OUT	0.0285	0.8190	0.2169	0.4123	OUT	0.8190	107,351	
22	ESM2-35M	Masked LM, ESM lineage	34M	480	L9	0.0202	0.8176	0.2403	0.4152	L7	0.8222	152,340	
23	ProtBert-BFD	Masked LM, other	421M	1024	L18	0.0408	0.8159	0.2130	0.3392	L18	0.8159	149,548	
24	SaESM2-150M	Diffusion / structure-token	149M	640	L30	0.0239	0.8157	0.1540	0.3509	L18	0.8357	114,259	
25	DPLM-150M	Diffusion / structure-token	149M	640	L24	0.0429	0.8091	0.2438	0.3304	L18	0.8420	125,518	
26	ProtSent-V2-150M	Retrieval-trained	148M	640	OUT	0.0467	0.8005	0.2118	0.3737	OUT	0.8005	29,512	SENS
27	IDP-ESM2-150M	Masked LM, ESM lineage	148M	640	L24	0.0247	0.7978	0.2236	0.3895	L18	0.8380	127,003	
28	E1-300M	Masked LM, other	274M	1024	L12	0.0384	0.7940	0.1612	0.4649	L12	0.7940	119,456	
29	SaAMPLIFY-120M	Diffusion / structure-token	118M	640	L24	0.0167	0.7939	0.1328	0.2988	L19	0.8012	207,024	
30	CARP-76M	Masked LM, other	76M	1024	L19	0.0761	0.7938	0.2122	0.4018	L19	0.7938	213,758	SENS
31	gLM2-650M-emb	Retrieval-trained	326M	1280	L8	0.0028	0.7933	0.0911	0.1813	L9	0.8018	109,670	
32	ESM2-8M	Masked LM, ESM lineage	7.8M	320	L4	0.0144	0.7929	0.1997	0.3041	L4	0.7929	257,378	
33	ESM2-150M	Masked LM, ESM lineage	149M	640	L24	0.0449	0.7895	0.2306	0.3596	L18	0.8128	117,424	SENS
34	E1-150M	Masked LM, other	154M	768	L16	0.0361	0.7828	0.1921	0.3971	L16	0.7828	123,279	
35	AMPLIFY-120M	Masked LM, other	118M	640	L19	0.0199	0.7782	0.1557	0.2579	L19	0.7782	170,011	
36	ProteinBERT	Masked LM, other	16M	128	L4	0.0091	0.7766	0.1051	0.2532	L3	0.7964	238,998	
37	VESM-150M	Masked LM, ESM lineage	148M	640	L24	0.0424	0.7756	0.1949	0.4444	L18	0.8059	56,047	
38	DPLM-650M	Diffusion / structure-token	652M	1280	L26	0.0333	0.7709	0.1719	0.4368	L19	0.7923	70,109	
39	Lobster-24M	Masked LM, other	24M	408	L6	0.0183	0.7703	0.1445	0.2608	L8	0.7992	110,713	
40	ESM3-open-1.4B	Diffusion / structure-token	1.4B		L24	0.0273	0.7691	0.1289	0.2754	L24	0.7691	20,825	
41	ESM1v-650M	Masked LM, ESM lineage	652M	1280	L19	0.0160	0.7688	0.1706	0.4216	L16	0.7716	81,274	
42	ProLLaMA-7B	Autoregressive, other	6.7B	4096	L19	0.0108	0.7666	0.1015	0.2257	L16	0.7809	19,638	
43	ProGen3-112M	Autoregressive, ProGen	112M	384	L8	0.0121	0.7654	0.1273	0.2731	L8	0.7654	57,682	
44	ProTrek-650M	Retrieval-trained	651M	1280	L26	0.0396	0.7652	0.1527	0.4877	L19	0.7757	66,486	
45	ESM2-15B	Masked LM, ESM lineage	15.1B	5120	L38	0.0251	0.7646	0.1285	0.4041	L28	0.8023	3,005	

#	Model	Group	Params	Dim	Layer	SENS	AUROC	AUPRC	Top-1	Oracle layer	Oracle AUROC	Residues s-1	Ahead of ANGK on
46	DPLM2-650M	Diffusion / structure-token	673M	1280	L26	0.0406	0.7623	0.1816	0.4357	L19	0.7764	69,960	
47	RITA-m	Autoregressive, other	302M	1024	L14	0.0100	0.7612	0.0954	0.2789	L12	0.7682	67,149	
48	SaAMPLIFY-350M	Diffusion / structure-token	354M	960	L32	0.0134	0.7596	0.1012	0.3193	L32	0.7596	114,067	
49	RITA-s	Autoregressive, other	85M	768	L7	0.0134	0.7582	0.1096	0.2585	L6	0.7717	145,845	
50	ProGen2-medium	Autoregressive, ProGen	765M	1536	L16	0.0172	0.7574	0.1131	0.3357	L13	0.7698	39,314	
51	ProtT5-XL-BFD	Encoder-decoder	2.8B	1024	L24	0.0264	0.7571	0.1410	0.3392	L12	0.7758	21,673	
52	TMVec-2s	Retrieval-trained	2.6M	512	OUT	0.0019	0.7569	0.0879	0.1766	OUT	0.7569	77,898	
53	AMPLIFYv2-120M	Masked LM, other	118M	640	L19	0.0348	0.7531	0.1366	0.3895	L19	0.7531	206,745	
54	ProtT5-XXL-BFD	Encoder-decoder	11.3B	1024	L24	0.0304	0.7496	0.1529	0.3456	L24	0.7496	5,187	
55	E1-600M	Masked LM, other	641M	1280	L24	0.0187	0.7495	0.1321	0.3719	L24	0.7495	67,316	
56	SaESM2-650M	Diffusion / structure-token	652M	1280	L33	0.0145	0.7486	0.1109	0.3363	L26	0.7632	64,880	
57	ProGen3-219M	Autoregressive, ProGen	220M	512	L8	0.0164	0.7473	0.1140	0.2649	L8	0.7473	57,138	
58	ProtBert	Masked LM, other	421M	1024	L24	0.0094	0.7469	0.1015	0.3012	L24	0.7469	129,702	
59	RITA-xl	Autoregressive, other	1.2B	2048	L14	0.0204	0.7456	0.0928	0.3035	L12	0.7519	33,948	
60	ProGen3-762M	Autoregressive, ProGen	763M	768	L13	0.0190	0.7407	0.1117	0.2877	L8	0.7490	31,079	
61	VESM-650M	Masked LM, ESM lineage	652M	1280	L26	0.0329	0.7403	0.1535	0.4439	L19	0.7504	21,146	
62	FastESM2-650M	Masked LM, ESM lineage	651M	1280	L26	0.0318	0.7396	0.1552	0.4561	L19	0.7503	36,328	
63	ESMC-300M	Masked LM, ESM lineage	333M	960	L24	0.0443	0.7383	0.1634	0.4129	L24	0.7383	101,407	SENS
64	ESMC-300M-hf	Masked LM, ESM lineage	333M	960	L24	0.0443	0.7383	0.1634	0.4129	L24	0.7383	115,634	SENS
65	ProPrime-650M	Masked LM, other	651M	1280	L26	0.0309	0.7380	0.1501	0.4292	L19	0.7475	22,852	
66	ESM2-650M	Masked LM, ESM lineage	652M	1280	L26	0.0309	0.7379	0.1500	0.4287	L19	0.7475	50,542	
67	ProGen2-small	Autoregressive, ProGen	151M	1024	L9	0.0084	0.7362	0.0817	0.2591	L6	0.7677	113,332	
68	ProGen2-base	Autoregressive, ProGen	765M	1536	L13	0.0064	0.7335	0.0623	0.2240	L13	0.7335	37,827	
69	ProGen3-3B	Autoregressive, ProGen	3.0B	1280	L19	0.0145	0.7332	0.0844	0.3058	L12	0.7378	23,281	
70	ProGen3-339M	Autoregressive, ProGen	339M	512	L13	0.0139	0.7331	0.1293	0.2982	L8	0.7477	30,664	
71	Prime-690M	Masked LM, other	669M	1280	L26	0.0288	0.7327	0.1439	0.4409	L19	0.7445	22,894	
72	gLM2-150M	Masked LM, other	152M	640	L24	0.0157	0.7322	0.1034	0.2626	L18	0.7404	123,581	
73	AMPLIFY-350M	Masked LM, other	354M	960	L25	0.0145	0.7315	0.1266	0.2608	L25	0.7315	112,927	
74	ProtT5-XL	Encoder-decoder	1.2B	1024	L24	0.0180	0.7255	0.1193	0.3053	L12	0.7805	19,856	
75	ProtT5-XXL-U50	Encoder-decoder	11.3B	1024	L24	0.0241	0.7228	0.1393	0.3234	L24	0.7228	5,234	
76	RITA-l	Autoregressive, other	680M	1536	L19	0.0111	0.7221	0.0643	0.2743	L12	0.7542	48,580	
77	AIDO-Protein-16B	Masked LM, other	16.1B		L28	0.0171	0.7215	0.0769	0.3257	L28	0.7215	2,146	
78	Ares-4B	Masked LM, other	4.2B	1024	L10	0.0272	0.7203	0.1501	0.3538	L10	0.7203	64,114	
79	ProGen2-large	Autoregressive, ProGen	2.8B	2560	L16	0.0190	0.7193	0.0850	0.2749	L16	0.7193	15,794	
80	ProteinSage-650M	Retrieval-trained	326M	1280	L26	0.0217	0.7158	0.1159	0.3713	L26	0.7158	65,578	
81	ProGen2-BFD90	Autoregressive, ProGen	2.8B	2560	L16	0.0093	0.7104	0.0592	0.2795	L16	0.7104	15,527	
82	AMPLIFYv2-350M	Masked LM, other	354M	960	L25	0.0240	0.7090	0.1001	0.3251	L25	0.7090	114,380	
83	ProGen2-xlarge	Autoregressive, ProGen	6.4B	4096	L16	0.0084	0.7077	0.0590	0.2883	L16	0.7077	11,314	
84	gLM2-650M	Masked LM, other	671M	1280	L19	0.0138	0.7075	0.1189	0.2655	L19	0.7075	57,912	
85	ProtGPT2	Autoregressive, other	784M	1280	L18	0.0019	0.7015	0.0531	0.1345	L18	0.7015	198,089	
86	DPLM2-3B	Diffusion / structure-token	2.9B	2560	L28	0.0297	0.7013	0.1119	0.4175	L18	0.7025	22,507	
87	ESMC-600M-hf	Masked LM, ESM lineage	575M	1152	L28	0.0435	0.6984	0.1466	0.4158	L28	0.6984	88,117	
88	ESMC-600M	Masked LM, ESM lineage	575M	1152	L28	0.0435	0.6984	0.1466	0.4158	L28	0.6984	82,217	
89	Dayhoff-3B-UR90	Autoregressive, other	3.0B	1280	L25	0.0032	0.6835	0.0437	0.1550	L25	0.6835	6,776	
90	Ankh3-xl	Encoder-decoder	5.7B	2560	L2	0.0009	0.6816	0.0608	0.1187	L2	0.6816	9,315	
91	Matryoshka-ProtBertBFD	Retrieval-trained	420M	1024	OUT	0.0063	0.6769	0.0874	0.2263	OUT	0.6769	40,986	
92	DPLM-3B	Diffusion / structure-token	2.8B	2560	L28	0.0247	0.6764	0.0955	0.3860	L18	0.6859	23,504	
93	Ankh3-large	Encoder-decoder	1.9B	1536	L24	0.0267	0.6754	0.1025	0.2655	L28	0.6828	21,805	
94	ProGen3-1B	Autoregressive, ProGen	1.4B	1024	L13	0.0095	0.6748	0.0878	0.2737	L8	0.7239	32,364	
95	ESM1b-650M	Masked LM, ESM lineage	652M	1280	L16	0.0262	0.6739	0.1038	0.4146	L2	0.6749	81,930	
96	ProtST-ESM1b	Retrieval-trained	383M	1280	L16	0.0277	0.6739	0.1074	0.4064	L2	0.6785	77,543	
97	ProtElectra	Masked LM, other	683M	1280	L16	0.0000	0.6647	0.0318	0.0000	L19	0.6843	89,582	
98	Dayhoff-170M-UR50	Autoregressive, other	170M	256	L12	0.0009	0.6587	0.0460	0.1257	L12	0.6587	3,136	
99	ProteinGLM-1B	Masked LM, other	1.2B	2048	L19	0.0257	0.6585	0.1038	0.3930	L19	0.6585	44,018	

#	Model	Group	Params	Dim	Layer	SENS	AUROC	AUPRC	Top-1	Oracle layer	Oracle AUROC	Residues s-1	Ahead of ANGK on
100	VESM-3B	Masked LM, ESM lineage	2.8B	2560	L28	0.0231	0.6575	0.0852	0.4123	L28	0.6575	12,868	
101	ESM2-3B	Masked LM, ESM lineage	2.8B	2560	L28	0.0205	0.6525	0.0801	0.3930	L28	0.6525	23,697	
102	Ankh-base	Encoder-decoder	737M	768	L2	0.0004	0.6517	0.0366	0.0702	L2	0.6517	33,521	
103	Ankh2-large	Encoder-decoder	1.9B		L48	0.0200	0.6475	0.0926	0.3292	L24	0.6609	14,961	
104	Dayhoff-170M-UR90	Autoregressive, other	170M	256	L12	0.0013	0.6470	0.0451	0.1292	L12	0.6470	3,584	
105	ESMC-6B-hf	Masked LM, ESM lineage	6.4B	2560	L48	0.0220	0.6443	0.0885	0.3111	L48	0.6443	13,281	
106	ESMC-6B	Masked LM, ESM lineage	6.4B	2560	L48	0.0221	0.6443	0.0885	0.3111	L48	0.6443	9,948	
107	VenusPLM-300M	Masked LM, other	302M	1024	L2	0.0004	0.6019	0.0309	0.0825	L2	0.6019	12,612	
108	ProGen2-oas	Autoregressive, ProGen	765M	1536	L27	0.0003	0.6016	0.0225	0.0556	L27	0.6016	39,330	
109	Ankh-large	Encoder-decoder	1.9B	1536	L2	0.0006	0.5998	0.0342	0.0936	L2	0.5998	21,165	
110	ProteinGLM-10B	Masked LM, other	10.7B	4352	L24	0.0311	0.5975	0.1006	0.3503	L28	0.6117	6,957	
111	ProteinGLM-3B	Masked LM, other	2.8B	2560	L28	0.0203	0.5835	0.0608	0.3626	L18	0.6253	25,390	
112	Dayhoff-3B-GR-HM	Autoregressive, other	3.0B	1280	L32	0.0006	0.5821	0.0229	0.0942	L19	0.6243	7,504	
113	Dayhoff-3B-GR-HM-c	Autoregressive, other	3.0B	1280	L32	0.0005	0.5541	0.0218	0.0749	L16	0.6083	7,504	
114	ProtAlbert	Masked LM, other	224M	4096	L6	0.0000	0.5020	0.0131	0.0000	L7	0.5037	276,963	
115	ProtT5-base	Encoder-decoder	199M	768	L6	0.0000	0.5010	0.0129	0.0000	L6	0.5010	95,214	

D ALL 53 BRIDGED FOLD PAIRS

Table A2: Every bridged fold pair at cosine 0.70, ordered by the number of bridging structures. Anchor cosine is the similarity between the two SCOPe representatives with no probe involved. Distinct taxa counts the NCBI taxonomy identifiers among the bridging structures of that pair.

Pair	Folds	Names	Class	Component	Bridging structures	Anchor cosine	Distinct taxa
1	b.69 + b.68	7-bladed beta-propeller / 6-bladed beta-propeller	b	C3	2,127	0.523	1,416
2	a.238 + a.252	BAR/IMD domain-like / Mediator hinge subcomplex-like	a	C2	555	0.492	491
3	a.114 + a.13	Interferon-induced guanylate-b / RAP domain-like	a	C2	531	0.548	446
4	a.238 + a.216	BAR/IMD domain-like / I/LWEQ domain	a	C2	480	0.506	408
5	a.238 + a.114	BAR/IMD domain-like / Interferon-induced guanylate-b	a	C2	374	0.515	311
6	a.238 + a.281	BAR/IMD domain-like / YheA-like	a	C2	202	0.530	190
7	b.22 + b.132	TNF-like / Supernatant protein factor (SP	b	C1	191	0.493	158
8	a.13 + a.252	RAP domain-like / Mediator hinge subcomplex-like	a	C2	133	0.402	126
9	a.25 + a.191	Ferritin-like / Methenyltetrahydrofolate cyclo	a	C4	117	0.508	107
10	b.176 + b.159	AttH-like / AOC barrel-like	b	C3	93	0.542	76
11	a.13 + a.281	RAP domain-like / YheA-like	a	C2	80	0.474	76
12	b.23 + b.123	CUB-like / Hypothetical protein TM1070	b	C1	75	0.537	68
13	a.114 + a.252	Interferon-induced guanylate-b / Mediator hinge subcomplex-like	a	C2	45	0.507	44
14	b.81 + b.80	Single-stranded left-handed be / Single-stranded right-handed b	b	C6	40	0.518	36
15	b.179 + b.23	Anthrax protective antigen N-t / CUB-like	b	C1	28	0.524	26
16	b.76 + b.159	open-sided beta-meander / AOC barrel-like	b	C3	26	0.444	20
17	c.37 + c.128	P-loop containing nucleoside t / DNA polymerase III chi subunit	c	C5	25	0.541	25
18	a.114 + a.281	Interferon-induced guanylate-b / YheA-like	a	C2	21	0.528	21
19	a.47 + a.191	STAT-like / Methenyltetrahydrofolate cyclo	a	C4	20	0.532	20
20	b.61 + b.159	Streptavidin-like / AOC barrel-like	b	C3	17	0.548	17
21	b.61 + b.76	Streptavidin-like / open-sided beta-meander	b	C3	16	0.497	16
22	b.23 + g.77	CUB-like / Resistin	b	C1	16	0.504	14
23	b.177 + b.159	YmcC-like / AOC barrel-like	b	C3	13	0.507	11
24	a.63 + a.238	Apolipophorin-III / BAR/IMD domain-like	a	C2	9	0.376	9
25	b.121 + b.18	Nucleoplasmin-like/VP (viral c / Galactose-binding domain-like	b	C1	6	0.498	6
26	a.286 + a.216	Sama2622-like / I/LWEQ domain	a	C2	6	0.530	6
27	b.22 + b.112	TNF-like / C-terminal domain of mollusc h	b	C1	5	0.481	5
28	c.23 + c.16	Flavodoxin-like / Lumazine synthase	c	C7	5	0.531	4
29	a.238 + a.48	BAR/IMD domain-like / N-cbl like	a	C2	4	0.456	4
30	b.76 + b.177	open-sided beta-meander / YmcC-like	b	C3	4	0.544	4
31	b.18 + b.22	Galactose-binding domain-like / TNF-like	b	C1	4	0.473	4
32	b.18 + b.132	Galactose-binding domain-like / Supernatant protein factor (SP	b	C1	4	0.491	3
33	a.63 + a.198	Apolipophorin-III / YcfC-like	a	C2	4	0.496	4
34	b.179 + b.132	Anthrax protective antigen N-t / Supernatant protein factor (SP	b	C1	3	0.522	3
35	b.67 + b.76	5-bladed beta-propeller / open-sided beta-meander	b	C3	3	0.503	3
36	b.69 + b.66	7-bladed beta-propeller / 4-bladed beta-propeller	b	C3	3	0.512	3
37	b.76 + b.176	open-sided beta-meander / AttH-like	b	C3	3	0.546	3
38	b.112 + b.166	C-terminal domain of mollusc h / MAL13P1.257-like	b	C1	3	0.536	3
39	b.69 + b.76	7-bladed beta-propeller / open-sided beta-meander	b	C3	2	0.485	2
40	a.13 + a.216	RAP domain-like / I/LWEQ domain	a	C2	2	0.400	2
41	b.61 + b.176	Streptavidin-like / AttH-like	b	C3	2	0.535	2
42	a.63 + a.216	Apolipophorin-III / I/LWEQ domain	a	C2	2	0.521	2
43	a.238 + a.215	BAR/IMD domain-like / A middle domain of Talin 1	a	C2	2	0.349	2
44	b.98 + g.77	Zn aminopeptidase N-terminal d / Resistin	b	C1	1	0.497	1
45	b.29 + b.23	Concanavalin A-like lectins/gl / CUB-like	b	C1	1	0.534	1
46	b.14 + b.112	Calpain large subunit, middle / C-terminal domain of mollusc h	b	C1	1	0.416	1
47	c.38 + c.128	PTS IIb component / DNA polymerase III chi subunit	c	C5	1	0.513	1
48	a.100 + a.252	6-phosphogluconate dehydrogena / Mediator hinge subcomplex-like	a	C2	1	0.483	1
49	b.12 + b.123	Lipase/lipooxygenase domain (P / Hypothetical protein TM1070	b	C1	1	0.536	1
50	d.25 + d.338	Mitochondrial glycoprotein MAM / Oxysterol-binding protein-like	d	C8	1	0.534	1
51	b.178 + b.76	Spiral beta-roll / open-sided beta-meander	b	C3	1	0.537	1
52	a.29 + a.216	Bromodomain-like / I/LWEQ domain	a	C2	1	0.541	1
53	a.63 + a.13	Apolipophorin-III / RAP domain-like	a	C2	1	0.434	1

E SEQUENCE AND STRUCTURAL-ALPHABET IDENTITY OF THE BRIDGING STRUCTURES

One alternative is that a bridge is a sequence homolog of both folds it joins. Aligning each bridging structure to its two anchors gives a median amino-acid identity of 0.056, close to the 0.048 of random AlphaFold structures aligned to an anchor, and only 0.7 percent of those alignments reach 30 percent identity. At 25 percent the bridges do exceed the background, 1.6 percent against 0.03 percent, so a small tail shares sequence with an anchor. The same alignments in 3Di reach 0.635 at the 90th percentile, which is expected because the bridges were selected by structural-alphabet similarity. The bridges are mostly not sequence homologs of the folds they join, and they are not explained by memorised sequence similarity (Appendix Table A3).

Each bridging structure is aligned against each of its two anchors by Smith-Waterman local alignment, with BLOSUM62 for amino acids and a match-mismatch matrix of 2 and -1 for 3Di. Identity is the number of identical aligned positions over the length of the shorter sequence. The background draws random AlphaFold structures of at least 100 residues and aligns each to a random anchor.

Table A3: Amino-acid and 3Di identity between each bridging structure and the two anchors it joins, and between random AlphaFold structures and random anchors.

Comparison	Alignments	Median	Mean	90th percentile	Reaching 30 percent
Bridging structure against its two anchors, amino acid	10,236	0.056	0.072	0.137	0.7
Bridging structure against its two anchors, 3Di	10,236	0.114	0.296	0.635	46.9
Random AlphaFold structure against an anchor, amino acid	3,000	0.048	0.056	0.097	0.0
Random AlphaFold structure against an anchor, 3Di	3,000	0.064	0.076	0.127	0.8

F ROBUSTNESS OF THE BRIDGE ANALYSIS

Two controls test whether the anchors or the filters produce the result (Appendix Table A4). Lowering the length filter from 100 to 80 residues keeps all 53 bridged pairs, adds two, and leaves the six-bladed against seven-bladed propeller pair first with exactly 2,127 structures. Raising the anchors to five per fold makes attachment easier, since a probe now needs to match any one of five representatives, so the counts rise by two orders and are not comparable. What is comparable is which pairs come out strongest, and there the propeller relationship grows. The eight-bladed propeller b.70 joins, b.69 with b.70 becomes the heaviest pair, and the four propeller folds stay in one component. The right-handed against left-handed beta-helix pair also survives. The finding does not rest on the choice of representative.

With five representatives per fold a probe attaches to a fold when it matches any one of them. The structure and pair counts of that row are therefore not on the same scale as the other two rows.

Table A4: The full pass over 2.41×10^8 structures under three settings. Anchors counts the SCOPe representatives and Folds counts the fold categories they cover.

Run	Setting	Anchors	Folds	Islands	Bridging structures	Bridged pairs	Components
baseline	one representative per fold, domains of 100 residues or more	916	916	867	4,931	53	8
length 80	same, with the length filter lowered to 80	916	916	866	5,002	55	8
five anchors	five representatives per fold, exclusion list held at the 87 pairs of the baseline	2,618	916	768	465,418	262	17

G POSITION-BIN CONTROLS

Table A5: SCOPe40 fold-level AUROC over three seeds for four ANGK variants that differ in position information and vocabulary size.

Variant	Vocabulary	Position information	fold AUROC (3 seeds)
p4 (final)	40,875	4-way bin	0.9555 ± 0.0003
p2	25,478	2-way bin	0.9519 ± 0.0002
base	17,610	none	0.9435 ± 0.0006
shuf4 (capacity control)	36,940	none, labels hashed	0.9410 ± 0.0003

H CONNECTED COMPONENTS OF THE BRIDGE GRAPH

Table A6: The eight connected components of the bridge graph. Fold ids are SCOPe fold identifiers. A structure that bridges two pairs inside one component is counted once.

Component	Name	Folds	Bridged pairs	Bridging structures	Fold ids
C1	beta-sandwiches and lectins	14	14	326	b.112 b.12 b.121 b.123 b.132 b.14 b.166 b.179 b.18 b.22 b.23 b.29 b.98 g.77
C2	alpha-helical bundles	13	19	2,089	a.100 a.114 a.13 a.198 a.215 a.216 a.238 a.252 a.281 a.286 a.29 a.48 a.63
C3	beta-propellers and barrels	10	13	2,307	b.159 b.176 b.177 b.178 b.61 b.66 b.67 b.68 b.69 b.76
C4	ferritin-like and STAT	3	2	137	a.191 a.25 a.47
C5	P-loop NTPase group	3	2	26	c.128 c.37 c.38
C6	beta-helices, both hands	2	1	40	b.80 b.81
C7	flavodoxin and lumazine synthase	2	1	5	c.16 c.23
C8	glycoprotein and oxysterol	2	1	1	d.25 d.338

I THRESHOLD SENSITIVITY

The attachment threshold was fixed before looking at AlphaFold. Inside SCOPe40 we lower a cosine cut and track the connected components, and the partition that best matches expert fold labels appears at 0.70 with an adjusted Rand index of 0.374. One step lower the benchmark itself collapses, with 66.9 percent of all domains in a single component and the index falling to 0.013, so values below that cannot be read. We transfer 0.70 to AlphaFold without refitting it. Counts then depend strongly on that choice and the ordering does not (Appendix Table A7).

Table A7: Bridging structures, bridged fold pairs and the rank of the six-bladed against seven-bladed propeller pair at five cosine thresholds.

Cosine threshold	Bridging structures	Fold pairs	Rank of the propeller pair
0.60	2,091,271	2,326	1
0.65	244,307	428	1
0.70	**4,931**	**53**	**1**
0.75	2	1	absent
0.80	0	0	absent

Between 0.60 and 0.70 the number of bridging structures changes by a factor of 424 while the six-bladed against seven-bladed propeller pair stays first at every one of those thresholds. Above 0.75 the phenomenon disappears, so it appears only in a narrow band of resolution and we report that band.