LEARNING TO ALIGN THE KNOWN PROTEIN STRUCTURE UNIVERSE FROM SEQUENCE ALONE

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicted structure databases contain hundreds of millions to over a billion proteins, while sequence collections span tens of billions of nonredundant sequences. How small can a sequence model be while retaining structural alignment quality and compatibility with existing tools? We train ESM-2 encoders through a differentiable affine-gap alignment operator, learning residue scores and gap penalties from TM-align correspondences. On 68,431 SCOPe40-derived validation pairs, fully tuned 8M and 35M models without auxiliary masked language modeling (MLM) achieve full-reference Cα LDDT of 0.3893 and 0.4185 with maximum expected accuracy (MEA) decoding; a 150M model with MLM reaches 0.3902. All three exceed structure-based and ProstT5-based Foldseek in both 3Di-only and AA+3Di modes. Thresholding correspondence marginals at 0.5 gives F1 scores of 0.4801, 0.5639, and 0.4845 against TM-align, respectively. Frozen-backbone, adapter, LoRA, and initialization controls show that pretraining supplies useful features, but adapting them for alignment matters more than backbone size alone. Scaling beyond 35M offers no benefit under the tested full-tuning recipe with MLM. We then train a head to predict a distribution over Foldseek's 20 3Di states, using their expected score under its fixed substitution matrix. Alignment supervision trains the state assignments without requiring native 3Di labels. At inference, argmax produces pure 3Di states for unchanged Foldseek. In 3Di-only mode, the 8M model reaches full-reference Cα LDDT 0.3530 against structure-based Foldseek's 0.3506, while the 35M model exceeds both structure-based and ProstT5-based Foldseek on this metric and correspondence F1. These pairwise gains do not transfer uniformly to database retrieval. Compact sequence models can thus learn structural comparisons directly and express them through existing software suites without first predicting coordinates.

1 INTRODUCTION

Predicted structures have made structural comparison possible at a scale once reserved for sequence analysis. The new ESM Atlas includes over 1.1 billion ESMFold2 predictions across 6.8 billion sequences (Biohub, 2026), yet sequence collections are larger still: DIAMOND DeepClust retained 19.4 billion records after preprocessing and deduplication (Buchfink et al., 2026). These resources motivate a practical question: how small can a sequence model be while retaining structural alignment quality and compatibility with existing tools? The database counts and their different deduplication conventions are detailed in Appendix A.

Structural comparison from sequence has substantial precedent. Neural network aligners learn contextual residue scores (Llinares-López et al., 2023; Hamamsy et al., 2024), and ProstT5-based Foldseek predicts 3Di states from sequence (Heinzinger et al., 2024). Pretrained protein language models also contain structural information, including at the scale of ESM-2 8M (Lin et al., 2023; Simon & Zou, 2025). But information present in a representation need not be accessible through the residue similarities an aligner uses. A larger frozen encoder may contain richer features without supplying better correspondence scores. We therefore ask what alignment supervision must change: the scores above a fixed representation, or the representation itself.

We train the representation for the correspondence itself. Building on Orihime (d^2p) (Mogilevsky & Liang, 2026), we use its differentiable alignment operator to train 8M, 35M, and 150M ESM-2

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.

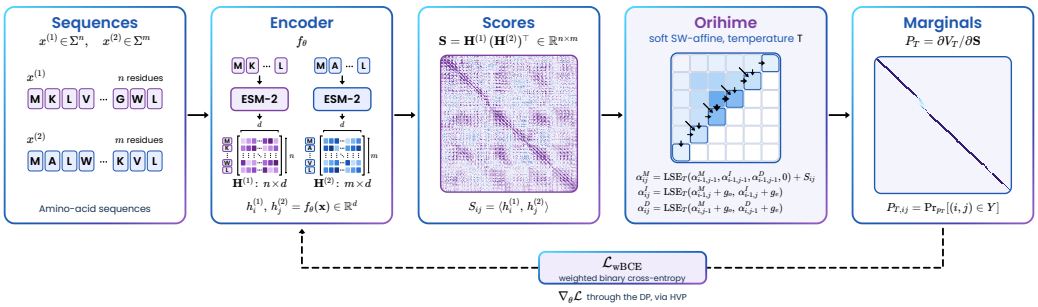


Figure 1: **Learning structural correspondences from sequence.** Each sequence passes through its own ESM-2 encoder instance, producing residue-row embeddings $H^{(1)} \in \mathbb{R}^{n\times d}$ and $H^{(2)} \in \mathbb{R}^{m\times d}$. Their matrix product $S = H^{(1)}(H^{(2)})^\top$ supplies pairwise scores to Orihime's differentiable affine-gap alignment operator. Weighted binary cross-entropy against TM-align correspondences trains the representations through the resulting marginals. Heatmaps are schematic; ellipses denote omitted residues and dimensions.

encoders against TM-align correspondences. Each protein is encoded independently; multiplying residue embedding matrices gives contextual substitution scores. Differentiable affine-gap alignment converts these scores into correspondence marginals, through which we learn the representation and gap penalties (Figure 1). Structures supply supervision; inference requires only sequences. Frozen-backbone, adapter, LoRA, and initialization controls test how much of the alignment ability comes from pretraining and how much requires adaptation.

The second question is whether this ability can survive a fixed discrete interface. We constrain the output to Foldseek's 20-state alphabet and keep its published substitution matrix unchanged. During training, each residue predicts a distribution over states, and residue pairs receive the expected substitution score between their distributions. The score remains differentiable, so alignment supervision can train the state assignments without native 3Di labels. At inference, argmax produces pure 3Di states for unchanged Foldseek. This tests whether a representation learned for correspondence can use an existing scoring system, even when its states do not reproduce the labels assigned from coordinates.

On 68,431 SCOPe40-derived validation pairs, the fully tuned 8M and 35M models exceed structure-based and ProstT5-based Foldseek in both 3Di-only and AA+3Di modes on full-reference Cα LDDT with MEA decoding. Even fully tuned 8M exceeds the strongest frozen 150M adapter on the reported LDDT and F1 endpoints. Scaling beyond 35M provides no benefit under the tested full-tuning recipe with MLM, which may reflect the limited size of our training dataset. The discrete models retain useful alignment information through native Foldseek, despite low agreement with native state labels. The experiments separate these findings from database retrieval: better correspondences within a protein pair do not uniformly improve the ranking of proteins across a database. Compactness, adaptation, and software compatibility are therefore tested together, while alignment quality and search quality are evaluated separately.

2 Related Work

Probabilistic alignment and differentiable inference. Probabilistic alignment and annealing predate neural network encoders (Kim et al., 1994; Bucher & Hofmann, 1996; Frith, 2020); LoL-align applies posterior alignment to structures (Reifenrath et al., 2025). Forward–backward and inside–outside inference can be expressed as backpropagation (Eisner, 2016). Soft-DTW, structured attention, and differentiable dynamic programming bring these constrained computations into learning systems (Cuturi & Blondel, 2017; Kim et al., 2017; Mensch & Blondel, 2018; Rush, 2020; Niculae et al., 2018). The d^2p framework differentiates through alignment marginals to learn structure encoders, gap parameters, and substitution matrices (Mogilevsky & Liang, 2026). We use its operator to learn from sequence and compare against its 64D and 128D ProteinMPNN aligners (Dauparas

et al., 2022). We do not evaluate its learned substitution matrices here; Appendix A explains how that comparison differs from the fixed-matrix experiment.

Learning representations and substitution scores. DEDAL jointly learns contextual embeddings, substitution scores, and alignment parameters (Llinares-López et al., 2023). DeepBLAST learns structural correspondences by differentiating through alignment of language-model representations; its companion TM-Vec addresses retrieval (Hamamsy et al., 2024). PEbA, pLM-BLAST, and PLMAlign align pretrained residue embeddings (Iovino & Ye, 2024; Kaminski et al., 2023; Liu et al., 2024). SMURF and learnMSA2 extend learned alignment to multiple sequences (Petti et al., 2023; Becker & Stanke, 2024). Our controls examine how the size and adaptation of a sequence encoder affect its usefulness for structural correspondence.

Structural alphabets. Foldseek encodes local structure in a 20-state alphabet for search and alignment (van Kempen et al., 2024). ProstT5-based Foldseek predicts native 3Di states from sequence (Heinzinger et al., 2024); SoftAlign learns a categorical alphabet from coordinate-based ProteinMPNN embeddings (Trinquier et al., 2025). Our discrete model learns sequence-to-state assignments from alignment correspondences while holding Foldseek's published matrix fixed. Recovering native structural labels is a separate supervision condition.

Structural information in pretrained sequence models. ESM-2 supports atomic structure prediction (Lin et al., 2023); ESMC underpins ESMFold2 and the new ESM Atlas (Candido et al., 2026). Sparse autoencoders recover structural features from ESMC and ESM-2, including the 8M model (Candido et al., 2026; Simon & Zou, 2025). Such features need not produce residue similarities suited to alignment. Frozen backbones, adapters, LoRA (Hu et al., 2021), and full tuning test this distinction; random initialization tests the contribution of pretraining itself.

3 Learning Structural Correspondences from Sequence

3.1 Contextual substitution operators

The same encoder processes each protein independently, so a residue's representation depends on its sequence context before the two proteins are compared. Let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_m)$ be amino acid sequences. A fixed substitution matrix B scores a residue pair as $e(x_i)^\top B e(y_j)$, where e is the one-hot encoding of amino acid identity. The same two amino acid identities always receive the same score. A contextual substitution operator (referred to from here on as the substitution operator) maps the sequences to a score matrix $S_\theta(x, y) \in \mathbb{R}^{n \times m}$. We encode the proteins independently as $H_x \in \mathbb{R}^{n \times d}$ and $H_y \in \mathbb{R}^{m \times d}$ and compute

$$S_\theta(x, y) = H_x H_y^\top, \qquad S_{ij} = \langle H_{x,i}, H_{y,j} \rangle. \tag{1}$$

The dot product compares the two residue contexts; no cross-protein attention precedes it. We use unnormalized dot products for the primary continuous models and evaluate cosine scores as an additional condition.

Training structures supply a binary TM-align correspondence matrix $Y \in \{0, 1\}^{n \times m}$ and validity mask M. Inference requires only sequences and padding masks. We evaluate both recovery of the reference correspondences and geometric agreement under the predicted mapping.

3.2 Representations and adaptation regimes

We use 8M, 35M, and 150M ESM-2 backbones. The *gaps-only* control learns two gap parameters over frozen MLM embeddings. *External adapters* map those embeddings to 64D or 128D; *LoRA* learns low-rank encoder updates; *full fine-tuning* updates the encoder itself. We test full tuning with and without auxiliary MLM and with random initialization (Appendix D).

3.3 Alignment supervision and decoding

Orihime applies differentiable affine-gap local alignment to the residue scores, producing correspondence marginals P_{ij}. We train the encoder and gap scores with class-weighted binary cross-entropy

against TM-align correspondences, differentiating through the marginals using the second derivatives supplied by d^{2}p (Mogilevsky & Liang, 2026). Runs with auxiliary MLM add its loss for each protein with unit weight. All training used a cluster of 24 NVIDIA H200 GPUs across three DGX pods. Appendix D.1 gives the operator, loss, and temperature schedule.

For correspondence F1, we threshold $P_{ij} > 0.5$. For structural quality, maximum expected accuracy (MEA) selects a monotone mapping maximizing summed marginals with free gaps. We also report mutual matching, which selects reciprocal row and column maxima without enforcing monotonicity. Appendix B.3 specifies all three decoders.

3.4 A FIXED 20-STATE INTERFACE

To use the learned representation in Foldseek, we replace continuous residue scores with scores expressed through its fixed alphabet. The discrete model starts from an alignment-fine-tuned sequence encoder and attaches a head $g_\phi : \mathbb{R}^d \to \mathbb{R}^{20}$. The head is a linear projection or an MLP. Let $\ell_i = g_\phi(H_{x,i})$ and $p_i = \mathrm{softmax}(\ell_i/\tau_c)$; define q_j analogously for the second protein. The mixture temperature $\tau_c = 1$ is distinct from the alignment temperature τ. With the fixed Foldseek 3Di substitution matrix $B \in \mathbb{R}^{20\times 20}$, the residue-pair score is

$$S_{ij}^{\mathrm{3Di}} = p_i^\top B q_j = \sum_{a=1}^{20}\sum_{b=1}^{20} p_{i,a} q_{j,b} B_{ab}. \tag{2}$$

This expected substitution score is differentiable in both state distributions. Alignment supervision can therefore teach useful mixtures while keeping the alphabet and score matrix fixed. Figure S1 shows this change of scorer within the shared alignment-training workflow.

At inference, argmax selects $a_i = \arg\max_a p_{i,a}$ and $b_j = \arg\max_b q_{j,b}$. The distributions become one-hot states, recovering the ordinary lookup score $B_{a_i b_j}$. These tokens can be passed directly to unchanged Foldseek. Quantization can change alignment paths, so we evaluate soft mixtures, hard states with Orihime, and hard states through Foldseek separately. Appendix E gives an exact embedding factorization of the fixed matrix and the full training and inference settings; Figure S2 shows the resulting mixture geometry.

The frozen-encoder experiment learns only the projection and gaps above an alignment-tuned backbone; unlike the continuous adapters, it starts from a representation already adapted to correspondence. Further conditions unfreeze the encoder, use straight-through selectors, or add masked structural-token supervision (Appendix E). Alignment-only training rewards a token's contribution to correspondence without requiring the native label; learned-to-native alignment is therefore a separate question. Native Foldseek evaluation here uses learned strings on both sides.

4 EXPERIMENTS

4.1 DATA, PROTOCOL, AND METRICS

We train on 297,601 SCOPe40-derived pairs and evaluate on 68,431 pairs after restricting lengths to 2,048 residues (Fox et al., 2014). Pairs come from an all-versus-all TM-align search and retain a maximum directional TM-score of at least 0.6; the split is made at the SCOPe fold level so that alignments, not proteins, are divided approximately 80/20 between training and validation. TM-align supplies the reference correspondences (Zhang & Skolnick, 2005). Retrieval and clustering use all 5,059 validation proteins without TM filtering (Appendix B).

We measure geometric agreement with four symmetric LDDT scores: Cα and all-atom, each with full-reference and aligned-only denominators (Mariani et al., 2013). Full-reference LDDT accounts for incomplete correspondence coverage; aligned-only LDDT measures agreement within mapped regions. Our all-atom variant uses chemically comparable atoms. We average each pair's directional scores, then weight eligible pairs equally. Correspondence precision, recall, and F1 instead pool edge counts against TM-align. Our principal endpoints use MEA maps for full-reference Cα LDDT and threshold maps for F1; they evaluate different decoded alignments. External aligners retain their native maps. Appendix B.3 specifies metric eligibility, including empty and unavailable maps.

Continuous models use seed 42 and are evaluated after 50 epochs. Token budgets are matched within backbone size but differ across sizes. Appendices D and E specify the checkpoint epochs.

4.2 SEQUENCE MODELS RECOVER STRUCTURAL CORRESPONDENCES

With MEA decoding, the fully tuned 8M and 35M models reach full-reference Cα LDDT of 0.3893 and 0.4185, exceeding both 3Di-only Foldseek baselines: 0.3506 for structure-based Foldseek and 0.3640 for ProstT5-based Foldseek. They also exceed the other sequence-input baselines in Table 1, with the same ordering on all-atom full-reference LDDT. Structure-input methods still lead: USalign fNS reaches 0.4803 and TM-align 0.4641 with their native maps; the original ProteinMPNN-64 and -128 models reach 0.4650 and 0.4655 with MEA. The sequence models narrow the gap to methods that receive coordinates at inference.

Full-reference and aligned-only scores reveal how that gain is obtained. The 8M model's aligned-only Cα LDDT is 0.5290, below structure-based Foldseek's 0.5616 in 3Di-only mode, despite its higher full-reference score. Its advantage therefore involves coverage as well as local geometry. Conversely, PAM30 reaches aligned-only LDDT of 0.7291 but full-reference LDDT of only 0.0178. A method can align a small region well while accounting for little of the full structure.

The correspondence decoder also changes this balance. For continuous 35M without MLM, threshold, mutual matching, and MEA yield full-reference Cα LDDT of 0.3334, 0.4135, and 0.4185, while F1 decreases from 0.5639 to 0.5352 and 0.5328. Mutual matching retains much of MEA's geometric quality; thresholding yields higher correspondence F1. These are different selections from the same marginals, so the score describes the representation together with its decoder (Appendix B.3).

Table 1: **Alignment quality across sequence and structure inputs.** All 32 external configurations plus principal ESM, ProteinMPNN, and SoftAlign models. Each block identifies the decoder above its metric columns. Native alignments come directly from the external tool. Full and aligned denote full-reference and aligned-only LDDT. Matrix labels include gap open/extend settings. Higher is better.

	Cα LDDT		All-atom LDDT		
Method	Full	Aligned	Full	Aligned	F1
Sequence: alignment-tuned ESM-2		*LDDT: MEA*			*F1: Threshold*
8M, no MLM	0.3893	0.5290	0.3508	0.5075	0.4801
35M, no MLM	0.4185	0.5677	0.3768	0.5451	0.5639
150M, with MLM	0.3902	0.5272	0.3523	0.5059	0.4845
Sequence: neural network baselines		*LDDT and F1: native alignment*			
ProstT5-based Foldseek, 3Di-only	0.3640	0.5647	0.3186	0.5445	0.4589
ProstT5-based Foldseek, AA+3Di	0.3329	0.5792	0.2852	0.5518	0.4711
PEbA	0.3321	0.5307	0.2924	0.5083	0.4222
DeepBLAST	0.3206	0.5022	0.2805	0.4845	0.4738
PLMAlign	0.3070	0.5157	0.2645	0.4954	0.4752
Global pLM-BLAST	0.2606	0.5334	0.2177	0.5117	0.4785
DEDAL (full length)	0.2080	0.6351	0.1761	0.5949	0.3734
Structure: classical aligners		*LDDT and F1: native alignment*			
USalign fNS	0.4803	0.6566	0.4235	0.6189	0.6276
USalign sNS	0.4780	0.6626	0.4242	0.6320	0.8423
USalign sequential	0.4641	0.6751	0.4066	0.6445	0.9834
TM-align	0.4641	0.6751	0.4066	0.6445	1.0000
DALI†	0.4572	0.6873	0.3923	0.6553	0.7191
SPalign	0.4530	0.6800	0.3935	0.6479	0.7511
SPalign-NS	0.4027	0.7237	0.3288	0.6710	0.5055
MUSTANG†	0.3951	0.6169	0.3444	0.5928	0.5395
MUSCLE3D	0.3657	0.5306	0.3251	0.5072	0.4394
GANGSTA+	0.3511	0.6365	0.2910	0.6025	0.4603
Foldseek 3Di-only	0.3506	0.5616	0.3045	0.5416	0.4339
Foldseek AA+3Di	0.3192	0.5814	0.2711	0.5537	0.4535
Reseek 3	0.3127	0.5677	0.2681	0.5333	0.4627
Reseek 2	0.2939	0.5386	0.2491	0.5092	0.4343
Structure: ProteinMPNN		*LDDT: MEA*			*F1: Threshold*
ProteinMPNN 64D	0.4650	0.6381	0.4166	0.6126	0.7460
ProteinMPNN 128D	0.4655	0.6403	0.4166	0.6144	0.7536
Structure: SoftAlign		*LDDT and F1: MEA*			
SoftAlign SW	0.4285	0.5858	0.3845	0.5634	0.6495
SoftAlign geometric mean	0.4165	0.6063	0.3669	0.5831	0.6639
Sequence: classical aligners		*LDDT and F1: native alignment*			
BLOSUM45 / 14/2	0.1203	0.5851	0.0972	0.5387	0.2338
BLOSUM50 / 13/2	0.1116	0.6062	0.0905	0.5560	0.2364
BLOSUM62 / 10/1	0.1119	0.5997	0.0910	0.5491	0.2363
BLOSUM62 / 11/1	0.1049	0.6222	0.0853	0.5688	0.2344
BLOSUM80 / 10/1	0.0774	0.6743	0.0631	0.6088	0.2062
BLOSUM90 / 10/1	0.0696	0.6803	0.0568	0.6108	0.1923
MAFFT	0.2322	0.3804	0.2013	0.3547	0.2185
MMseqs2	0.1021	0.6165	0.0830	0.5625	0.2313
PAM250 / 15/2	0.1212	0.5574	0.0969	0.5136	0.2147
PAM30 / 9/1	0.0178	0.7291	0.0144	0.6354	0.0709
PAM70 / 10/1	0.0462	0.6952	0.0372	0.6174	0.1422

† Missing maps out of 68,431 pairs: DALI 8,557; MUSTANG 77. Other rows cover the cohort; undefined aligned-only scores are excluded per metric. DEDAL includes the full-length extension. Coverage rules and matched-subset comparisons appear in Appendix C.

a Full Cα LDDT — MEA decoder

Architecture	64D 8M	64D 35M	64D 150M	128D 8M	128D 35M	128D 150M
Linear	0.267	0.275	0.290	0.266	0.275	0.286
MLP-2	0.278	0.306	0.317	0.280	0.309	0.318
MLP-3	0.285	0.310	0.312	0.289	0.310	0.314
MLP-4	0.291	0.305	0.306	0.295	0.306	0.307
CNN-3	0.311	0.316	0.315	0.313	0.316	0.317
CNN-5	0.320	0.319	0.320	0.321	0.320	0.322
CNN-7	0.324	0.324	0.325	0.325	0.324	0.326
CNN-9	0.325	0.324	0.326	0.328	0.328	0.327
BiGRU-1	0.323	0.340	0.341	0.324	0.340	0.342
BiGRU-2	0.336	0.344	0.346	0.338	0.344	0.346
BiLSTM-1	0.326	0.341	0.341	0.328	0.340	0.342
BiLSTM-2	0.336	0.347	0.349	0.342	0.345	0.346
Projectformer-2	0.305	0.314	0.320	0.316	0.318	0.325
Projectformer-4	0.321	0.322	0.333	0.325	0.335	0.335

b Pooled F1 — Threshold > 0.5 decoder

Architecture	64D 8M	64D 35M	64D 150M	128D 8M	128D 35M	128D 150M
Linear	0.145	0.113	0.183	0.154	0.143	0.166
MLP-2	0.118	0.241	0.282	0.140	0.255	0.289
MLP-3	0.170	0.263	0.277	0.192	0.270	0.283
MLP-4	0.201	0.256	0.258	0.217	0.263	0.262
CNN-3	0.255	0.280	0.277	0.266	0.283	0.282
CNN-5	0.284	0.287	0.288	0.289	0.291	0.293
CNN-7	0.295	0.297	0.299	0.299	0.298	0.303
CNN-9	0.298	0.297	0.305	0.305	0.306	0.304
BiGRU-1	0.256	0.340	0.345	0.260	0.340	0.350
BiGRU-2	0.312	0.354	0.365	0.321	0.356	0.364
BiLSTM-1	0.273	0.340	0.345	0.278	0.341	0.347
BiLSTM-2	0.315	0.359	0.369	0.338	0.359	0.357
Projectformer-2	0.245	0.281	0.294	0.278	0.286	0.303
Projectformer-4	0.284	0.292	0.319	0.299	0.326	0.324

Shared color scale: 0.00 (white) to 0.50 (purple); values rounded to 0.001

Figure 2: **All 84 primary adapters: 14 architectures, two widths, three backbones.** Left: full-reference Cα LDDT from MEA. Right: F1 from threshold decoding. Each model is evaluated after 50 epochs. Architecture and width can matter more than backbone size. All 84 adapter configurations, all three decoders and four LDDT endpoints continue in Appendix K.

4.3 Adaptation matters more than backbone size alone

The adaptation controls ask whether useful correspondences can be recovered from fixed pretrained features or require changes to the encoder. Frozen gaps-only models recover correspondences poorly. Adapters improve those features, but fully tuning 8M with MLM exceeds the selected frozen 150M adapters on both endpoints (Table 2). On threshold F1, LoRA exceeds these adapters at 8M and 35M, but not at 150M. Larger pretrained representations alone do not close the gap.

Table 2: **Pretraining and adaptation controls.** Full-reference Cα LDDT (MEA) and F1 (threshold) after 50 epochs. Both recurrent adapters are shown at 64D; Figure 2 reports all architectures and widths. Dashes indicate unavailable conditions.

	8M		35M		150M	
Adaptation	LDDT MEA	F1 Threshold	LDDT MEA	F1 Threshold	LDDT MEA	F1 Threshold
Frozen gaps only	0.2363	0.1421	0.2245	0.1340	0.2232	0.1321
Frozen BiLSTM-2, 64D	0.3355	0.3150	0.3466	0.3591	0.3492	0.3693
Frozen BiGRU-2, 64D	0.3357	0.3119	0.3436	0.3540	0.3464	0.3649
LoRA + MLM	0.3630	0.3877	0.3605	0.3966	0.3420	0.3590
Full + MLM	0.3860	0.4651	0.4154	0.5520	0.3902	0.4845
Full, no MLM	0.3893	0.4801	0.4185	0.5639	–	–
Random init + MLM	0.2960	0.1319	0.2857	0.1141	–	–

Architecture matters within a fixed backbone as well. At 8M, replacing a linear 64D projection with BiLSTM-2 raises threshold F1 from 0.1451 to 0.3150; retaining BiLSTM-2 and scaling the backbone to 150M reaches 0.3693. Neither recurrent family is uniformly best across LDDT and F1, and increasing output width gives no consistent gain. Figure 2 shows all 84 primary adapters; Appendix D.3 develops the architecture and width comparisons, with all earlier configurations retained in the complete figures.

Pretraining and continued MLM have different effects. Random initialization at 8M reduces threshold F1 from 0.4651 to 0.1319 and MEA LDDT from 0.3860 to 0.2960 under the same training

	Encoder / head / training selector	a Full Cα LDDT 3Di-only	a Full Cα LDDT AA+3Di	b Pooled F1 3Di-only	b Pooled F1 AA+3Di
ESM-2 8M	Frozen linear / softmax	0.328	0.300	0.291	0.351
	Frozen linear / ST argmax	0.269	0.247	0.310	0.362
	Frozen linear / ST Gumbel	0.328	0.306	0.328	0.380
	Frozen MLP-4 / softmax	0.334	0.319	0.299	0.352
	Frozen MLP-4 / ST argmax	0.294	0.271	0.319	0.362
	Frozen MLP-4 / ST Gumbel	0.330	0.304	0.356	0.397
	Trainable MLP-4 / softmax	0.353	0.343	0.357	0.399
	Trainable MLP-4 / ST argmax	0.299	0.284	0.383	0.415
	Trainable MLP-4 / ST Gumbel	0.339	0.323	0.397	0.431
	Trainable MLP-4 / softmax + masked	0.341	0.330	0.360	0.391
ESM-2 35M	Frozen linear / softmax	0.356	0.326	0.386	0.422
	Frozen linear / ST argmax	0.334	0.301	0.430	0.451
	Frozen linear / ST Gumbel	0.368	0.344	0.443	0.467
	Frozen MLP-4 / softmax	0.365	0.336	0.397	0.430
	Frozen MLP-4 / ST argmax	0.319	0.287	0.427	0.450
	Frozen MLP-4 / ST Gumbel	0.372	0.344	0.465	0.486
	Trainable MLP-4 / softmax	0.389	0.377	0.486	0.500
	Trainable MLP-4 / ST argmax	0.339	0.322	0.485	0.502
	Trainable MLP-4 / ST Gumbel	0.384	0.367	0.505	0.522
	Trainable MLP-4 / softmax + masked	0.379	0.371	0.487	0.498
	Native references				
	Structure-derived 3Di	0.351	0.319	0.434	0.453
	ProstT5-based Foldseek	0.364	0.333	0.459	0.471

Shared color scale: 0.00 (white) to 0.60 (purple)

Figure 3: **All 20 learned predictors through unchanged Foldseek.** Both metrics use native Foldseek alignments from hard tokens, in 3Di-only and AA+3Di modes. ST denotes straight-through training. All tokens are deterministic at inference. The 35M models can exceed both reference token sources; training selector and encoder adaptation affect the tradeoff. Soft/hard Orihime results appear in Appendix E.

duration. Removing auxiliary MLM from pretrained full tuning instead improves both metrics at 8M and 35M. With MLM, scaling from 35M to 150M reduces LDDT from 0.4154 to 0.3902. Pretraining supplies a useful starting point, but continuing its objective during alignment tuning is not necessary for these gains. Size-dependent budgets prevent attributing the scaling result solely to capacity.

4.4 Learned tokens retain quality through unchanged Foldseek

The discrete experiment asks whether correspondence learning can use Foldseek's fixed alphabet and scoring procedure. An unfrozen 35M encoder with an MLP-4 head trained on soft mixtures reaches full-reference Cα LDDT 0.3887 and F1 0.4861 through native 3Di-only Foldseek, exceeding both structure-based and ProstT5-based Foldseek. Both metrics use the same native alignment. The model exceeds both baselines in AA+3Di mode as well. Figure 3 compares all 20 predictors in both modes, using deterministic tokens at inference.

The two endpoints do not always improve together. In 3Di-only mode, the 8M softmax model reaches LDDT 0.3530 against structure-based Foldseek's 0.3506, with lower F1 (0.3568 versus 0.4339). At 35M, the unfrozen Gumbel model reaches higher F1 than the unfrozen softmax model (0.5053 versus 0.4861) with slightly lower LDDT (0.3843 versus 0.3887).

Replacing state distributions with argmax states is only part of the transfer. Holding Orihime MEA fixed, this replacement lowers LDDT from 0.4017 to 0.3768 for the 35M softmax model and from 0.3755 to 0.3400 at 8M. Five of the 20 predictors instead improve under hard states (Figure S3b). Switching those states to Foldseek then changes the scoring and alignment procedure as well, so the soft-to-hard and hard-to-native comparisons answer different questions (Appendix E).

Useful states also need not reproduce the coordinate tokenizer's labels. On 856,552 valid-descriptor residues, the 35M softmax predictor agrees with native 3Di only 5.19% of the time, compared with 61.30% for the ProstT5-based Foldseek predictor and 20.21% for the majority native state. Figure S3a shows that native-state identity and learned-string alignment quality rank these predictors differently. Yet the same fixed substitution matrix supports the alignments between learned strings reported above. Low state identity separates literal reconstruction from alignment quality; it does not establish a relabeling symmetry or compatibility with native-token databases (Appendix E.4).

The advantage also depends on structural similarity. The continuous 35M model exceeds 8M in all four reference-TM bins, but the native 3Di-only comparison reverses at high TM-score. The 35M softmax model reaches LDDT 0.3337 against structure-based Foldseek's 0.2857 in $[0.6, 0.7)$, which contains 69.8% of pairs. In $[0.9, 1.0]$, it falls below the same baseline, 0.6960 against 0.7433, with F1 reversing as well. The complete strata in Appendix G show where the aggregate advantage holds and where it does not.

4.5 Transfer from pairwise alignment to database retrieval

Good correspondences within a pair need not produce a good ranking across proteins. We test that transfer over all 5,059 validation proteins, ranking by Foldseek score and measuring sensitivity before the first different-fold false positive (van Kempen et al., 2024). In 3Di-only mode, the 35M softmax model reaches family sensitivity 0.6249, below structure-based Foldseek's 0.7446 and ProstT5-based Foldseek's 0.7993. Its superfamily and fold sensitivities are also lower. The unfrozen Gumbel model instead improves fold sensitivity over structure-based Foldseek (0.1004 versus 0.0830), while remaining below ProstT5-based Foldseek's 0.1088.

The pairwise gains therefore do not transfer uniformly to retrieval; the outcome depends on the predictor and the relationship being retrieved. Figure S4 compares all 20 predictors in both scoring modes, including cluster agreement at family, superfamily, and fold levels. Appendix F develops the full retrieval, clustering, and score-rank comparisons, including both Foldseek modes and all four TM-align normalizations. Appendix M includes all 24 TM-score graph clustering conditions.

5 Discussion

The distinction between pretraining and adaptation is central to these results. Without pretraining, our 8M and 35M controls fail to learn useful correspondences from alignment supervision within the tested training budget. Frozen pretrained features also leave much of the correspondence task unsolved. Full tuning connects the two: it starts from useful sequence features and changes their similarities to support structural correspondence. Within the tested budgets, this adaptation matters more than increasing the size of a frozen encoder. The lack of improvement beyond 35M under full tuning with MLM may reflect limited supervision, size-dependent optimization, or both; matched budgets and larger training sets would test these explanations.

The fixed alphabet exposes a second distinction. Alignment supervision constrains which states work together under the substitution matrix, without requiring them to recover the labels of the coordinate tokenizer. The learned strings can therefore support strong pairwise alignments despite low native-state identity. This is useful compatibility with an existing scoring procedure, but the evaluation uses learned tokens on both sides. Whether the same states can be compared directly with structure-derived database strings remains an empirical question.

Retrieval tests a further property: whether scores place related proteins ahead of unrelated candidates. Better geometry within selected pairs does not ensure that ordering. The high-TM reversal and the mixed retrieval results show where the pairwise advantage changes, while the different behavior of softmax and Gumbel predictors suggests that the training condition matters for transfer. Objectives that train rankings directly are a natural next test. Their benefit, and the calibration of native search statistics under learned tokens, require separate evaluation.

5.1 LIMITATIONS

Runs use one seed and size-dependent budgets. Generalization to multidomain proteins and public databases, pretraining overlap, and the effects of finite-precision marginal excursions remain uncharacterized. The inference runtime is included in the supplementary material as Hikoboshi, a standalone C++ implementation (`hikoboshi`), and Hikoboshi-Torch, its PyTorch version (`hikoboshi-torch`). Appendix H gives the reference-label and numerical audits.

6 CONCLUSION

Structural correspondence can be learned directly from sequence with compact encoders. Pretraining supplies useful features, but their presence does not by itself make residue similarities suitable for alignment: without pretraining, the 8M and 35M controls fail to learn useful correspondences within the tested training budget. Within the tested configurations, adapting a small encoder is more effective than retaining a much larger frozen one. Training through differentiable alignment learns those similarities and their gap penalties together. The same approach extends to a fixed 20-state alphabet, producing pure 3Di states that retain useful alignment information through unchanged Foldseek without requiring native 3Di labels.

These results connect sequence representations to structural comparison without first predicting coordinates. They also distinguish useful correspondence from literal state reconstruction and from database ranking. Strong alignments between learned strings coexist with low native-state identity, and the pairwise gains do not transfer uniformly to retrieval. Learned-to-native token compatibility and objectives that address ranking are therefore concrete next tests. Compact sequence models can learn structural comparisons and express them through existing software; extending that capability to general database search requires evaluating how the learned representation behaves beyond the selected alignment population. Indeed, as it turns out, sequence really was all you needed for accurate alignments.

AI USE STATEMENT

Yes—to aid or polish writing. Generative AI tools assisted with editing for readability and formatting and with producing figures; all text, claims, results, and figures were reviewed and verified by the author. Details are described here.

Yes—for retrieval and discovery (e.g., finding related work). Generative AI tools assisted with sourcing, searching, and identifying relevant literature; all cited works were inspected against primary sources by the author.

LLM tools used: Claude Code, Codex, Muse Code, OpenCode, and Paperclip by GXL. The author takes responsibility for the final content, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies computational protein alignment using existing sequence and structure data; the reported experiments do not involve human participants. Predicted correspondences do not establish functional equivalence, clinical relevance, or biological safety. Biological applications require validation appropriate to their use and to the limitations of the evaluated domain population.

REPRODUCIBILITY STATEMENT

The appendix specifies architectures, optimization, objectives, decoders, checkpoint selection, and metric denominators. Tables are generated from saved validation records and per-pair scores, with source and checkpoint hashes retained for verification. External alignments use a common structural scorer. An independent evaluation of the 35M no-MLM checkpoint reproduced its four saved LDDT means (Appendix H).

REFERENCES

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*, 622:637–645, 2023. doi: 10.1038/s41586-023-06510-w. URL https://doi.org/10.1038/s41586-023-06510-w.

Felix Becker and Mario Stanke. learnMSA2: deep protein multiple alignments with large language and hidden Markov models. *Bioinformatics*, 2024. doi: 10.1093/bioinformatics/btae381. URL https://doi.org/10.1093/bioinformatics/btae381.

Biohub. Get started with ESM: ESMFold2, ESMC, and ESM atlas, 2026. URL https://biohub.ai/esm/protein/get-started. Official model and dataset documentation; accessed September 25, 2026.

Philipp Bucher and Kay Hofmann. A sequence similarity search algorithm based on a probabilistic interpretation of an alignment scoring system. In *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pp. 44–51, 1996.

Benjamin J. Buchfink, Émile Barbé, Haim Ashkenazy, Klaus Reuter, John A. Kennedy, and Hajk-Georg Drost. Clustering the protein universe of life using DIAMOND DeepClust. *Nature Methods*, 23:724–727, 2026. doi: 10.1038/s41592-026-03030-z. URL https://doi.org/10.1038/s41592-026-03030-z.

Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z. Wu, Jin Sub Lee, Elise S. Bruguera, Jehan A. Keval, Mykhailo Kopylov, John E. Pak, Wesley Wu, Neil Thomas, Samson Mataraso, Alvin Hsu, Ashton C. Trotman-Grant, Kilian Fatras, Allan dos Santos Costa, Rohil Badkundri, Halil Akın, Deniz Oktay, Jonathan Deaton, Elizabeth Montabana, Hrishita Sitwala, Yue Yu, Marius Wiggert, Dylan Alexander Carlin, Anthony W. Goering, Tomasz Blazejewski, McCullen Sandora, Michael Hla, Tina Z. Jia, Leon H. Kloker, Nicholas J. Sofroniew, Masatoshi Uehara, Jassi Pannu, Sharrol Bachas, Daniel S. Liu, Tom Sercu, and Alexander Rives. Language modeling materializes a world model of protein biology. *bioRxiv*, 2026. doi: 10.64898/2026.06.03.729735. URL https://doi.org/10.64898/2026.06.03.729735. Preprint.

Marco Cuturi and Mathieu Blondel. Soft-DTW: a differentiable loss function for time-series. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 894–903. PMLR, 2017. URL https://proceedings.mlr.press/v70/cuturi17a.html.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Alexis Courbet, Rob J. de Haas, Neville Bethel, Philip J. Y. Leung, Timothy F. Huddy, Samuel Pellock, Doug Tischer, Frederick Chan, Brian Koepnick, Hannah Nguyen, Alex Kang, Banumathi Sankaran, Asim K. Bera, Neil P. King, and David Baker. Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187. URL https://doi.org/10.1126/science.add2187.

Jason Eisner. Inside-outside and forward-backward algorithms are just backprop (tutorial paper). In *Proceedings of the Workshop on Structured Prediction for NLP*, pp. 1–17. Association for Computational Linguistics, 2016. doi: 10.18653/v1/W16-5901. URL https://doi.org/10.18653/v1/W16-5901.

Naomi K. Fox, Steven E. Brenner, and John-Marc Chandonia. SCOPe: Structural classification of proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research*, 42(D1):D304–D309, 2014. doi: 10.1093/nar/gkt1240. URL https://doi.org/10.1093/nar/gkt1240.

Martin C. Frith. How sequence alignment scores correspond to probability models. *Bioinformatics*, 36(2):408–415, 2020. doi: 10.1093/bioinformatics/btz576. URL https://doi.org/10.1093/bioinformatics/btz576.

Google DeepMind and EMBL-EBI. AlphaFold protein structure database, 2026. URL https://alphafold.ebi.ac.uk/. Resource homepage; accessed September 25, 2026.

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*, 42:975–985, 2024. doi: 10.1038/s41587-023-01917-2. URL https://doi.org/10.1038/s41587-023-01917-2.

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. URL https://doi.org/10.1093/nargab/lqae150.

Ian Holmes and Gerald M. Rubin. An expectation maximization algorithm for training hidden substitution models. *Journal of Molecular Biology*, 317(5):753–764, 2002. doi: 10.1006/jmbi.2002.5405. URL https://doi.org/10.1006/jmbi.2002.5405.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Benjamin Giovanni Iovino and Yuzhen Ye. Protein embedding based alignment. *BMC Bioinformatics*, 25:85, 2024. doi: 10.1186/s12859-024-05699-5. URL https://doi.org/10.1186/s12859-024-05699-5.

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. URL https://doi.org/10.1093/bioinformatics/btad579.

Jin Kim, Sakti Pramanik, and Moon Jung Chung. Multiple sequence alignment using simulated annealing. *Computer Applications in the Biosciences (CABIOS)*, 10(4):419–426, 1994. doi: 10.1093/bioinformatics/10.4.419. URL https://doi.org/10.1093/bioinformatics/10.4.419.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. Structured attention networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HkE0Nvqlg.

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. URL https://doi.org/10.1126/science.ade2574.

Wei Liu, Ziye Wang, Ronghui You, Chenghan Xie, Hong Wei, Yi Xiong, Jianyi Yang, and Shanfeng Zhu. PLMSearch: Protein language model powers accurate and fast sequence search for remote homology. *Nature Communications*, 15:2775, 2024. doi: 10.1038/s41467-024-46808-5. URL https://doi.org/10.1038/s41467-024-46808-5.

Felipe Llinares-López, Quentin Berthet, Mathieu Blondel, Olivier Teboul, and Jean-Philippe Vert. Deep embedding and alignment of protein sequences. *Nature Methods*, 20:104–111, 2023. doi: 10.1038/s41592-022-01700-2. URL https://doi.org/10.1038/s41592-022-01700-2.

Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. LDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013. doi: 10.1093/bioinformatics/btt473. URL https://doi.org/10.1093/bioinformatics/btt473.

Arthur Mensch and Mathieu Blondel. Differentiable dynamic programming for structured prediction and attention. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3462–3471, 2018. URL https://proceedings.mlr.press/v80/mensch18a.html.

Meta AI. ESM metagenomic atlas: About and march 2023 update, 2023. URL https://esmatlas.com/about. Release v2023_02; accessed September 25, 2026.

Casey Sumagaysay Mogilevsky and Kimberly Liang. d^2p: Structured soft attention is all you need, 2026. Manuscript.

Vlad Niculae, André F. T. Martins, Mathieu Blondel, and Claire Cardie. SparseMAP: Differentiable sparse structured inference. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3799–3808. PMLR, 2018. URL https://proceedings.mlr.press/v80/niculae18a.html.

Samantha Petti, Nicholas Bhattacharya, Roshan Rao, Justas Dauparas, Neil Thomas, Juannan Zhou, Alexander M. Rush, Peter K. Koo, and Sergey Ovchinnikov. End-to-end learning of multiple sequence alignments with differentiable Smith–Waterman. *Bioinformatics*, 39(1): btac724, 2023. doi: 10.1093/bioinformatics/btac724. URL https://doi.org/10.1093/bioinformatics/btac724.

Sebastian Prillo, Yun Deng, Pierre Boyeau, Xingyu Li, Po-Yen Chen, and Yun S. Song. CherryML: Scalable maximum likelihood estimation of phylogenetic models. *Nature Methods*, 20: 1232–1236, 2023. doi: 10.1038/s41592-023-01917-9. URL https://doi.org/10.1038/s41592-023-01917-9.

Lasse Reifenrath, Michel van Kempen, Gyuri Kim, Soo Hyun Kim, Mohammadreza Radnezhad, Milot Mirdita, Martin Steinegger, and Johannes Söding. LoL-align: Sensitive and fast probabilistic protein structure alignment. *bioRxiv*, 2025. doi: 10.1101/2025.11.24.690091. URL https://doi.org/10.1101/2025.11.24.690091.

Alexander Rush. Torch-struct: Deep structured prediction library. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 335–342. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-demos.38. URL https://aclanthology.org/2020.acl-demos.38/.

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. URL https://doi.org/10.3389/fbinf.2022.1033775.

Elana Simon and James Zou. InterPLM: discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, 22:2107–2117, 2025. doi: 10.1038/s41592-025-02836-7. URL https://doi.org/10.1038/s41592-025-02836-7.

Jeanne Trinquier, Samantha Petti, Sukhwan Park, Kithmini Herath, Michel van Kempen, Shihao Feng, Johannes Söding, Martin Steinegger, and Sergey Ovchinnikov. SoftAlign: End-to-end protein structures alignment. *bioRxiv*, 2025. doi: 10.1101/2025.05.09.653096. URL https://doi.org/10.1101/2025.05.09.653096. Preprint.

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*, 42:243–246, 2024. doi: 10.1038/s41587-023-01773-0. URL https://doi.org/10.1038/s41587-023-01773-0.

Jingi Yeo, Yewon Han, Nicola Bordin, Andy M. Lau, Shaun M. Kandathil, Hyunbin Kim, Eli Levy Karin, Milot Mirdita, David T. Jones, Christine Orengo, and Martin Steinegger. Metagenomic-scale analysis of the predicted protein structure universe. *bioRxiv*, 2025. doi: 10.1101/2025.04.23.650224. URL https://doi.org/10.1101/2025.04.23.650224. Preprint; revised March 2026.

Yang Zhang and Jeffrey Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309, 2005. doi: 10.1093/nar/gki524. URL https://doi.org/10.1093/nar/gki524.

Wei Zheng, Qiqige Wuyun, Yang Li, Chengxin Zhang, Lydia Freddolino, and Yang Zhang. Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data. *Nature Methods*, 21:279–289, 2024. doi: 10.1038/s41592-023-02130-4. URL https://doi.org/10.1038/s41592-023-02130-4.

Overview of the Appendix

The supplementary experiments ask how sequence representations recover residue correspondences, how those correspondences preserve local structure, and how the resulting scores recover relationships across a protein database. These questions depend on different parts of the model: the representation determines residue-pair scores, the decoder chooses a correspondence map, and the search procedure ranks whole proteins. We evaluate each stage with its own metric and population, specifying the checkpoint, inference inputs, and decoder so that changes in one stage can be distinguished from changes in another.

Appendix A expands the resource-scale motivation and the distinction between learning scores, states, and correspondences. Appendix B defines the evaluation population, correspondence decoders, and structural metrics. Appendix C gives the full external benchmark, and Appendix D describes training and adaptation controls. Appendix E specifies the fixed 3Di representation and discrete selectors. Appendix F defines the full-validation retrieval and clustering comparisons. Appendix G defines the TM-score strata and correspondence geometry, and Appendix H reports numerical and reproduction checks. Appendix I specifies our learned architectures and gives exact component and checkpoint parameter counts, including external neural network baselines. Appendix J compares overall and TM-score-stratified alignment quality. Appendix K shows complete aggregate comparisons and discrete alignment pathways; Appendix L gives decoder, retrieval and token diagnostics. Appendix M reports all TM-score graph thresholds and clustering diagnostics. The figures cover the 116 ESM-2 checkpoints in the main sweep, the two original ProteinMPNN checkpoints, and external baselines. ESM-2 training uses uncapped wBCE; the ProteinMPNN references use a positive-weight cap of 100. All applicable representations and decoders are retained.

Complete comparison	Figure series
Correspondence decoders	S5; S77–S81
Backbone size	S6–S8
Aggregate alignment quality	S66–S74
Discrete alignment pathways	S75–S76
TM-score strata	S22–S65
TM-score strata: external tools	S62–S65
Native Foldseek differences	S18–S21
Retrieval metrics	S10–S11
Clustering agreement	S12–S13
Native partition agreement	S14–S15
TM-score rank correlations	S16–S17
Precision–recall curves	S82–S86
Token identity and alignment	S9
Token distances	S87
TM-score threshold clustering	S88–S90
Clustering diagnostics	S91–S92

A Sequence Resources and Alignment Context

A.1 What the resource counts measure

The resource sizes in the introduction motivate sequence-only structural comparison, but their counts refer to different units and filtering procedures. AFDB provides over 260 million predicted structures (Google DeepMind and EMBL-EBI, 2026). The ESM Metagenomic Atlas grew from 617 million to 772 million (Lin et al., 2023; Meta AI, 2023), and the new ESM Atlas includes over 1.1 billion ESMFold2 predictions across 6.8 billion sequences (Biohub, 2026). Structure-based clustering organizes these predictions at a different level: Foldseek Cluster grouped 214 million AFDB structures into 2.30 million non-singleton clusters (Barrio-Hernandez et al., 2023), while AFESM extended structural clustering to 820 million predicted structures (Yeo et al., 2025). A cluster count is therefore not a count of individual sequences or structures.

Sequence collections are larger still. DeepMSA2 searches roughly 40 billion sequences, including 35.6 billion reported as nonredundant across TaraDB, MetaSourceDB, and JGIclust (Zheng et al., 2024). DIAMOND DeepClust retained 19.4 billion of 22.8 billion records after preprocessing and

deduplication (Buchfink et al., 2026). These counts exceed the ESMFold2 structure subset by over an order of magnitude, but they are not globally unique censuses. Exact deduplication removes identical sequences; similarity clustering groups distinct sequences. JGIclust clusters within projects at 90% identity and coverage, whereas the DeepClust count precedes similarity clustering. The counts establish the scale of available resources without establishing a globally deduplicated size across them.

Encoding cost is one reason model size matters for this setting. In one study, ProtT5 encoding took 7 hours 23 minutes while nearest-neighbor search took 77 seconds (Schütze et al., 2022). This example motivates measuring feature computation as part of a search pipeline. The inference runtime for our encoders is provided in the supplementary material as Hikoboshi, a standalone C++ implementation (`hikoboshi`), and Hikoboshi-Torch (`hikoboshi-torch`).

A.2 Learning scores, states, and correspondences

The d^2p framework learns structure encoders, gap parameters, and substitution matrices through alignment marginals (Mogilevsky & Liang, 2026). Our structure-input controls use its original 64D and 128D ProteinMPNN aligners. Its learned substitution matrices are not included in this comparison. In particular, its learned 3Di matrix still receives structure-derived states; it changes the scores assigned to state pairs. Our discrete sequence model instead holds Foldseek's published scores fixed and learns which state to assign to each residue from sequence. These choices change different parts of the alignment problem.

Learning substitution processes also has a distinct phylogenetic history (Holmes & Rubin, 2002; Prillo et al., 2023). Learning the parameters of a substitution process, recovering a coordinate tokenizer's labels, and learning residue correspondences provide different forms of supervision. Here TM-align supplies the target correspondences. The alignment-only discrete condition can change state assignments whenever that improves correspondence under the fixed matrix; the masked-3Di condition additionally asks the predictor to recover native labels. Appendix E specifies these objectives and their separate evaluation conditions.

B Data and Evaluation Definitions

B.1 Alignment and retrieval populations

We use the SCOPe40-derived corpus described in Section 4. All-versus-all TM-align pairs are restricted to 2,048 residues and to a maximum directional TM-score of at least 0.6, then split at the SCOPe fold level so that alignments, not proteins, are divided approximately 80/20 between training and validation; this leaves 297,601 training pairs and 68,431 validation pairs. Mapping the pair indices to SCOPe metadata gives 9,724 training domains from 949 folds and 4,624 validation domains from 600 folds. The complete validation split contains 5,059 proteins, including 435 that do not participate in the filtered alignment pairs. Retrieval and clustering use this complete population.

The 68,431 ordered validation pairs are structurally related by construction: the maximum of their two directional TM-align reference scores is at least 0.6. These reference scores precede model evaluation; the cohort manifest records protein identities, ordered pair indices, lengths, and reference directions. Of the selected pairs, 12,065 share a family, 17,686 share a superfamily but differ in family, 12,784 share a fold but differ in superfamily, and 25,896 differ in fold. The filter therefore spans several SCOPe relationships. Even its different-fold pairs have been selected for high TM-score, however, and cannot represent unrestricted database negatives. We evaluate retrieval separately over all 5,059 validation proteins (Appendix F).

B.2 Correspondence reference and aggregation

We score predicted correspondences against the dataset labels, excluding 29 out-of-bounds reference edges and recording those exclusions. The training collator could instead clip these edges to a batch-dependent limit, so F1 from the training logs and F1 from the per-pair scorer can differ slightly. All results use the shared per-pair scorer unless explicitly identified as training-log measurements. Structural evaluation uses each method's unchanged predicted map and the same original coordinates.

For pair q, let A_q be its predicted residue-edge set and R_q the saved reference-edge set. The primary correspondence statistics pool counts across available pairs:

$$\text{precision} = \frac{\sum_q |A_q \cap R_q|}{\sum_q |A_q|}, \qquad \text{recall} = \frac{\sum_q |A_q \cap R_q|}{\sum_q |R_q|}, \qquad F_1 = \frac{2\sum_q |A_q \cap R_q|}{\sum_q |A_q| + \sum_q |R_q|}. \quad (3)$$

These are micro averages: a protein pair contributing more edges can contribute more to the pooled score. Separately reported macro averages give equal weight to eligible pairs and need not preserve the same ordering. To reconstruct overall F1 from TM-score bins, we pool their edge counts before computing the ratio.

An empty prediction and an unavailable prediction have different meanings. An empty map remains in the full-reference population with score zero; an unavailable map is excluded and counted separately. An aligned-only score with no comparable reference contacts is undefined and excluded from that metric's mean. Noninjective predictions are retained for correspondence PRF and excluded from the geometric scorer, which requires an injective map. We evaluate these predictions as emitted. These rules apply separately to every method and metric, so the number of available maps is not a substitute for the valid-pair count of a structural score.

B.3 Decoders and structural metrics

The three decoders make different selections from the same residue-pair marginals. Threshold decoding retains every edge with $P_{ij} > 0.5$. Mutual matching retains reciprocal row and column maxima without a probability threshold and does not impose a monotone path. MEA uses dynamic programming to maximize total selected marginal mass along a monotone path, with zero gap cost at this hard-decoding stage. Its gap convention is distinct from the learned affine-gap distribution used to obtain the marginals. MEA resolves equal choices in diagonal, up, then left order. For discrete models, choosing a single 3Di state per residue by argmax occurs before the marginals are computed and can be combined with any of these correspondence decoders.

Because thresholding does not enforce injectivity, some predictions assign a residue more than once and cannot be evaluated geometrically. The SoftAlign geometric-mean model illustrates the consequence: threshold decoding gives 68,431 available maps but only 14,689 valid full-reference Cα scores, and 9,639 valid aligned-only Cα scores. Its corresponding means describe those smaller populations. Mutual matching and MEA decode the original marginals independently; they do not repair the thresholded map.

Local distance agreement. The structural scores follow the distance-agreement principle of LDDT (Mariani et al., 2013). For each reference direction, observed atom pairs within a 15 Å neighborhood contribute the fraction of distance-error thresholds $\{0.5, 1, 2, 4\}$ Å that they satisfy. The distance-error comparisons are inclusive. Cα reference contacts include the 15 Å boundary; the all-atom reference-contact condition is strictly below 15 Å. These comparisons require no rigid superposition. Cα scoring uses one atom per observed residue. The custom all-atom variant uses the atom37 representation and a chemical mapping between the two amino acid types. It excludes within-residue distances and adds no separate stereochemical penalties.

Full-reference scoring retains reference contacts involving unmapped residues; absent predicted contacts receive no agreement credit. Aligned-only scoring restricts the denominator to the mapped region. We compute both reference directions and average their scores within each protein pair, requiring a nonzero denominator in both directions. We then average equally across valid protein pairs. The four endpoints are therefore Cα full-reference, Cα aligned-only, all-atom full-reference, and all-atom aligned-only LDDT. They differ in both the compared atoms and the denominator. Reporting all four distinguishes local agreement within a selected region from agreement across the full reference structure.

The denominator matters even when the coordinates and mapping are unchanged. Consider a map covering only a small conserved region. Aligned-only LDDT asks how well the comparable distances within that region agree; full-reference LDDT also includes reference contacts for which the map supplies no counterpart. The first score can therefore be high while the second is low. Extending the map may increase reference coverage while introducing less accurate distances, so the

two scores need not improve together. The PAM30 example in the main text illustrates why a high aligned-only value cannot by itself establish good alignment of the whole structure.

Correspondence F1 asks a different question again. It compares selected residue edges with the TM-align labels, without assigning more credit to an edge because its distances agree better. Higher F1 can follow from a different precision–recall balance and does not alone imply that more reference edges were recovered. Comparisons of recovered edge counts require recall or true-positive counts. The decoder, geometric denominator, and correspondence endpoint must therefore remain explicit when interpreting a change in score.

Chemically comparable atoms. The all-atom mapping, `chemical_v1`, anchors backbone N, Cα, C, and O atoms, with terminal OXT where applicable. Side-chain matching uses a maximum connected induced common subgraph that preserves elements, aromaticity, coarse saturation classes, and bonds. Ties prefer unchanged atom names and then lexicographic order; reverse amino acid pairs use inverse mappings. Identical amino acid types use identity atom mappings before symmetry resolution. Symmetry candidates cover ASP, GLU, ARG, LEU, VAL, PHE, and TYR. Up to four candidate permutations are scored against nonambiguous anchors independently for each residue and reference direction, with canonical order breaking ties.

Under the `chemical_comparable_v1` reference policy, a matched residue excludes atoms without a chemically comparable atom in the other amino acid type. If a comparable counterpart is experimentally missing, its reference contact remains in the denominator and receives no credit. Unmatched residues retain their observed reference contacts. Thus a missing experimental atom, an incompatible chemical role, and an unaligned residue have different consequences. This is a custom extension for comparing different protein sequences; other all-atom LDDT implementations may use different conventions. All methods in a reported comparison use the same scorer and original coordinates. The per-pair results include source hashes for reproducibility.

Table S1: Full fine-tuning evaluated using three decoders: 8M and 35M each with and without auxiliary MLM, and 150M with MLM. All scores are 0–1 decimals. Full Cα and all-atom (AA) scores use the full-reference denominator. Aligned Cα restricts the denominator to mapped atoms.

Configuration	Decoder	F1	Full Cα	Full AA	Aligned Cα
8M without MLM	Threshold	0.4801	0.2737	0.2332	0.5786
8M without MLM	Mutual matching	0.4459	0.3818	0.3416	0.5337
8M without MLM	MEA	0.4424	0.3893	0.3508	0.5290
8M with MLM	Threshold	0.4651	0.2548	0.2158	0.5842
8M with MLM	Mutual matching	0.4302	0.3766	0.3366	0.5288
8M with MLM	MEA	0.4261	0.3860	0.3481	0.5230
35M without MLM	Threshold	0.5639	0.3334	0.2879	0.5995
35M without MLM	Mutual matching	0.5352	0.4135	0.3706	0.5711
35M without MLM	MEA	0.5328	0.4185	0.3768	0.5677
35M with MLM	Threshold	0.5520	0.3135	0.2691	0.5999
35M with MLM	Mutual matching	0.5205	0.4090	0.3663	0.5661
35M with MLM	MEA	0.5174	0.4154	0.3742	0.5619
150M with MLM	Threshold	0.4845	0.2746	0.2332	0.5767
150M with MLM	Mutual matching	0.4486	0.3825	0.3430	0.5313
150M with MLM	MEA	0.4452	0.3902	0.3523	0.5272

Figure S5 includes every full-fine-tuning configuration. Appendix L continues the same decoder and coverage comparisons for every other learned representation.

C Complete External Benchmark

C.1 Inputs, native outputs, and coverage

We compare external aligners by evaluating their native correspondences with the same four symmetric LDDT metrics, on a 0–1 scale. Missing maps and undefined metric denominators are excluded

under the rules in Appendix B. The principal continuous models use MEA; external methods retain their native correspondences. Continuous-model scores are measured after 50 training epochs. The shared scorer independently reproduces the 35M no-MLM row (Appendix H).

The MUSCLE3D configuration receives structure-derived Mega profiles generated from the original PDB inputs by Reseek 2.8. Its native profile-alignment implementation consumes those profiles; the saved source identifier `muscle5` does not denote an amino-acid-only input in this comparison.

Table S2: External aligners: native correspondences. All 32 configurations. Four symmetric LDDTs (Cα/all-atom; full-reference/aligned-only) and pooled correspondence F1 are fractions. Maps are used unchanged. BLOSUM/PAM suffixes identify the supplied gap-open/gap-extension configuration. MUSCLE3D has source identifier `muscle5`.

Configuration	Cα full	Cα aligned	All-atom full	All-atom aligned	Pooled F1
BLOSUM45 / 14/2	0.1203	0.5851	0.0972	0.5387	0.2338
BLOSUM50 / 13/2	0.1116	0.6062	0.0905	0.5560	0.2364
BLOSUM62 / 10/1	0.1119	0.5997	0.0910	0.5491	0.2363
BLOSUM62 / 11/1	0.1049	0.6222	0.0853	0.5688	0.2344
BLOSUM80 / 10/1	0.0774	0.6743	0.0631	0.6088	0.2062
BLOSUM90 / 10/1	0.0696	0.6803	0.0568	0.6108	0.1923
DEDAL (full length)	0.2080	0.6351	0.1761	0.5949	0.3734
DeepBLAST	0.3206	0.5022	0.2805	0.4845	0.4738
ProstT5-based Foldseek, 3Di-only	0.3640	0.5647	0.3186	0.5445	0.4589
ProstT5-based Foldseek, AA+3Di	0.3329	0.5792	0.2852	0.5518	0.4711
Structure / Foldseek 3Di	0.3506	0.5616	0.3045	0.5416	0.4339
Structure / Foldseek AA+3Di	0.3192	0.5814	0.2711	0.5537	0.4535
GANGSTA+	0.3511	0.6365	0.2910	0.6025	0.4603
DALI (available subset)	0.4572	0.6873	0.3923	0.6553	0.7191
MUSTANG (available subset)	0.3951	0.6169	0.3444	0.5928	0.5395
TM-align	0.4641	0.6751	0.4066	0.6445	1.0000
USalign sequential	0.4641	0.6751	0.4066	0.6445	0.9834
USalign fNS	0.4803	0.6566	0.4235	0.6189	0.6276
USalign sNS	0.4780	0.6626	0.4242	0.6320	0.8423
MAFFT	0.2322	0.3804	0.2013	0.3547	0.2185
MMseqs2	0.1021	0.6165	0.0830	0.5625	0.2313
MUSCLE3D	0.3657	0.5306	0.3251	0.5072	0.4394
PAM250 / 15/2	0.1212	0.5574	0.0969	0.5136	0.2147
PAM30 / 9/1	0.0178	0.7291	0.0144	0.6354	0.0709
PAM70 / 10/1	0.0462	0.6952	0.0372	0.6174	0.1422
PEbA	0.3321	0.5307	0.2924	0.5083	0.4222
PLMAlign	0.3070	0.5157	0.2645	0.4954	0.4752
Global pLM-BLAST	0.2606	0.5334	0.2177	0.5117	0.4785
Reseek 2	0.2939	0.5386	0.2491	0.5092	0.4343
Reseek 3	0.3127	0.5677	0.2681	0.5333	0.4627
SPalign	0.4530	0.6800	0.3935	0.6479	0.7511
SPalign-NS	0.4027	0.7237	0.3288	0.6710	0.5055

Missing maps out of 68,431 pairs: DALI 8,557; MUSTANG 77. Other methods cover the cohort. Undefined scores are excluded per metric.

The table footer identifies missing native maps. Undefined aligned-only scores, including those from GANGSTA+, Reseek 3, and DEDAL, are excluded; empty GANGSTA+ maps receive zero full-reference scores. Per-metric denominators are retained in the machine-readable results. Structure-derived Foldseek comparisons use the current full-cohort runs in 3Di-only and AA+3Di modes.

The saved common-map comparison uses 59,860 pairs. Restricting comparisons to this subset removes differences in map availability, although noninjective predictions and empty aligned-only denominators still require metric-specific exclusions. Learned-token native Foldseek results have not been reaggregated on this intersection and are reported separately on the full cohort. The DEDAL

result includes an extension beyond its natively supported lengths; performance within its supported length range is not reported separately.

C.2 Structure-input neural network controls

The two original ProteinMPNN-based alignment checkpoints provide a structure-input comparison (Dauparas et al., 2022; Mogilevsky & Liang, 2026). Their node, edge, and hidden dimensions are 64 or 128, with three encoder layers, 64 neighbors, dropout 0.1, and no coordinate augmentation in the recorded configuration. They use dot-product scoring and the end of epoch 50, with a nominal 32,768-token budget and a wBCE positive-weight cap of 100. The evaluated model restores the checkpoint gap scores at temperature one.

With MEA, the 64D and 128D checkpoints reach Cα full-reference LDDT 0.4650 and 0.4655 and correspondence F1 0.7248 and 0.7334. Threshold decoding instead gives F1 0.7460 and 0.7536, with full-reference LDDT 0.3926 and 0.4034. As with the sequence models, the decoder with stronger reference-edge recovery is not the decoder with the highest full-reference geometry. These models receive structure at inference, so their results do not extend the sequence-only claim.

The native SoftAlign checkpoints (Trinquier et al., 2025) include distinct Smith–Waterman and geometric-mean configurations. Under MEA, they reach Cα full-reference LDDT 0.4285 and 0.4165, with F1 0.6495 and 0.6639, respectively. With threshold decoding, the geometric-mean model has the restricted structural eligibility described in Appendix B.3; its LDDT mean therefore cannot support a full-cohort ranking. We report both configurations with all three decoders to show how correspondence selection changes each model's performance.

Table S3: Structure-input neural network aligners. Four symmetric LDDTs and pooled F1 for all three correspondence decoders. Noninjective maps are excluded from geometric scoring without repair.

Model / decoder	Cα full	Cα aligned	All-atom full	All-atom aligned	Pooled F1
ProteinMPNN 64 (cap 100) / Threshold	0.3926	0.6941	0.3330	0.6646	0.7460
ProteinMPNN 64 (cap 100) / MM	0.4586	0.6443	0.4080	0.6184	0.7294
ProteinMPNN 64 (cap 100) / MEA	0.4650	0.6381	0.4166	0.6126	0.7248
ProteinMPNN 128 (cap 100) / Threshold	0.4034	0.6921	0.3437	0.6625	0.7536
ProteinMPNN 128 (cap 100) / MM	0.4597	0.6459	0.4088	0.6197	0.7376
ProteinMPNN 128 (cap 100) / MEA	0.4655	0.6403	0.4166	0.6144	0.7334
SoftAlign SW / Threshold	0.3148	0.6923	0.2616	0.6623	0.6786
SoftAlign SW / MM	0.4069	0.5943	0.3598	0.5712	0.6605
SoftAlign SW / MEA	0.4285	0.5858	0.3845	0.5634	0.6495
SoftAlign geometric mean / Threshold	0.2365	0.6676	0.2077	0.6367	0.6534
SoftAlign geometric mean / MM	0.3573	0.5851	0.3070	0.5619	0.6447
SoftAlign geometric mean / MEA	0.4165	0.6063	0.3669	0.5831	0.6639

Geometric-mean SoftAlign with threshold decoding has valid full-reference LDDT for 14,689 of 68,431 pairs; the remaining maps are noninjective.

D Adaptation, Pretraining, and Optimization

We evaluate continuous models at the end of epoch 50 (zero-based index 49), using the complete validation dataset and alignment temperature 1. A common training duration makes the comparison independent of intermediate checkpoint selection. Repeated records of an experiment agree in their metrics and are counted once. Frozen and tuned models use the same nominal token budget within each backbone size; at 35M this is 65,536 rather than the 131,072 used in an earlier cohort. The primary continuous models in this comparison use dot-product residue scores.

Let P_{train} count all trainable parameters, including adapters or the MLM head when present and the two gap parameters. Let T be the nominal one-side padded residue budget per global optimizer update. The peak learning rate is resolved once at initialization as

$$\eta_{\max} = \min\left\{10^{-3},\ 10^{-3}\sqrt{\frac{256432}{P_{\text{train}}}}\sqrt{\frac{T}{32768}}\right\}. \tag{4}$$

The token convention is pair count times bucket length times accumulation and worker count; it does not multiply by two for the two sequences. Length buckets are 128, 256, 512, 1,024, and 2,048. Global nominal pair counts per bucket are $(1024, 512, 256, 128, 64)$ for 8M, $(512, 256, 128, 64, 32)$ for 35M, and $(384, 192, 96, 48, 24)$ for 150M. Terminal batches can be smaller. The 150M full-tuning run distributes the logical batch across two workers. Microbatch splitting preserves logical-batch loss counts.

AdamW uses $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 0.01 on regular parameters, and zero decay on gap parameters. Gradient clipping uses norm 1. Learning-rate warmup occupies the first 5% of planned optimizer updates, rounded up, followed by cosine decay to 1% of the peak. The alignment temperature follows a separate cosine schedule from 5 to 1. Training uses bfloat16 autocast; the alignment wrapper requests float32 outputs (Appendix H). Checkpoint tensor shapes, storage types, and hashes accompany the exact parameter inventory in Appendix I. Seed 42 is used throughout. Table S4 gives representative trainable counts and resolved peak rates.

Table S4: Trainable parameters and resolved peak learning rates. Counts include the two learned gap parameters. Adapter widths and initialization settings vary between configurations; these rows are not parameter-matched comparisons.

Backbone	Regime	Trainable parameters	Peak learning rate
8M	BiLSTM-2 (64D)	633,858	1.00e-03
8M	LoRA	276,482	1.00e-03
8M	Full + MLM	7,511,071	3.70e-04
35M	BiLSTM-2 (64D)	1,273,858	6.35e-04
35M	LoRA	829,442	7.86e-04
35M	Full + MLM	33,499,231	1.24e-04
150M	BiLSTM-2 (64D)	2,319,874	4.07e-04
150M	LoRA	2,764,802	3.73e-04
150M	Full + MLM	148,136,991	5.10e-05

D.1 Differentiable affine-gap alignment and correspondence loss

Let $\mathcal{A}$ denote the local alignment paths admitted by the affine-gap operator. For a path a, let a_{ij} indicate a matched residue pair, and let $o(a)$ and $e(a)$ count gap-opening and gap-extension transitions under the operator's convention. With learned gap scores g_o, g_e, the path score and soft alignment value are

$$q(a; S, g_o, g_e) = \sum_{ij} a_{ij} S_{ij} + g_o o(a) + g_e e(a), \tag{5}$$

$$F_\tau(S, g_o, g_e) = \tau \log \sum_{a \in \mathcal{A}} \exp(q(a; S, g_o, g_e)/\tau)\,. \tag{6}$$

Differentiating this soft value with respect to a residue-pair score gives that pair's marginal probability under the alignment distribution:

$$P_{ij} = \frac{\partial F_\tau}{\partial S_{ij}} = \mathbb{E}_{a \sim p_\tau}[a_{ij}]. \tag{7}$$

The marginals are themselves derivatives. Training them requires differentiating again, both through residue scores and through gaps; Orihime supplies these second derivatives (Mogilevsky & Liang, 2026). Continuous models initialize gap scores at -3 and -1 and anneal alignment temperature from 5 to 1; evaluation uses temperature 1.

Alignment temperature controls the distribution over complete paths: at a given score scale, lower temperature concentrates probability more strongly on higher-scoring paths. It does not choose a single 3Di state for a residue. The categorical temperature in the discrete model acts earlier, on its state logits, before residue-pair scores enter the alignment operator. Consequently, hard-state inference and low-temperature alignment are different operations. The checkpoint-specific discrete alignment temperatures are given in Appendix E.

Correspondence supervision. Only a small fraction of residue pairs correspond. Let $N = \sum M_{ij}$ count valid cells and $K = \sum M_{ij} Y_{ij}$ count reference matches across the logical training batch. With detached positive weight $w_+ = \max\{1, (N-K)/(K+1)\}$, the correspondence loss is

$$\mathcal{L}_{\text{align}} = -\frac{1}{N} \sum_{ij} M_{ij} \left[w_+ Y_{ij} \log \widetilde{P}_{ij} + (1 - Y_{ij}) \log(1 - \widetilde{P}_{ij}) \right], \tag{8}$$

where $\widetilde{P}$ is clamped to $[10^{-6}, 1 - 10^{-6}]$ for the logarithms. Positive weighting is uncapped. The marginals describe the learned alignment distribution; this weighted objective does not calibrate them as probabilities of biological correctness.

Positive weighting addresses the imbalance between matched and unmatched residue pairs. It changes the penalty for errors on reference matches while leaving the path constraints inside the alignment operator. The loss therefore supervises a globally constrained marginal matrix, even though its binary terms are summed over cells. Detaching the weight prevents gradients through the batch-dependent class count; the encoder and gap scores are trained through the marginals themselves.

Auxiliary MLM adds $\mathcal{L}_{\text{MLM}}(x) + \mathcal{L}_{\text{MLM}}(y)$ with unit coefficients. Appendix D specifies masking, corruption, and optimization.

D.2 Encoder adaptation and training objectives

We use the 8M, 35M, and 150M ESM-2 backbones, with native residue dimensions 320, 480, and 640. The same encoder processes each protein independently. The proteins first interact through the pairwise score matrix, with no cross-protein attention before that stage. The four adaptation regimes change which parameters can learn from correspondence supervision.

Frozen backbone. The gaps-only control uses native pretrained representations and learns the gap-open and gap-extension parameters. External adapters instead learn a mapping above the frozen encoder, with 64- or 128-dimensional outputs. The fourteen adapter architectures comprise a linear projection; sequential SwiGLU MLPs of depths two, three, and four; two-layer sequence CNNs with kernel widths three, five, seven, and nine; one- and two-layer bidirectional GRUs and LSTMs; and two- and four-block Projectformer projection/refinement adapters. Depth labels count the corresponding projection or recurrent stages as defined in Appendix I.2. The frozen encoder is evaluated without training dropout. Comparing these adapters tests how much correspondence information can be recovered by transforming fixed pretrained features, either residue by residue or with additional processing along the sequence.

Low-rank adaptation. LoRA updates attention query, key, value, and output projections and both feed-forward projections in every transformer layer. Each adapted matrix has the form $W = W_0 + (\alpha/r)BA$, where W_0 is frozen, $r = 8$, and $\alpha = 16$. The low-rank branch uses dropout 0.1. The MLM head remains frozen, while gradients through it can reach the low-rank parameters. Including the two gap parameters, the 8M, 35M, and 150M runs train 276,482, 829,442, and 2,764,802 parameters, respectively. Outputs retain the backbone's native dimension.

Full fine-tuning and initialization controls. Full fine-tuning updates the sequence encoder, gap parameters, and the MLM head when present. Removing MLM at 8M and 35M tests the effect of retaining the language-modeling objective during alignment training. A separate 8M control starts from random weights with the same architecture and training duration, measuring the benefit of pretraining under a fixed training budget.

Auxiliary objective and corruption. For runs with MLM, the loss is $\mathcal{L}_{\text{align}} + \mathcal{L}_{\text{MLM}}(x) + \mathcal{L}_{\text{MLM}}(y)$, with unit coefficients. Training selects 15% of positions for MLM, using 80% mask replacement, 10% random replacement, and 10% unchanged tokens. A disjoint 5% alignment-noise selection uses 80% mask and 20% random replacement. Validation alignment inputs are clean; MLM evaluation retains its masking procedure. Removing MLM changes both the auxiliary supervision and the trainable head, which slightly changes the parameter-dependent learning-rate calculation.

D.3 Adapter architectures and full frozen-backbone results

The adapters span residuewise linear and gated MLP projections, gated sequence CNNs, bidirectional GRUs and LSTMs, and Projectformer attention blocks. Changing output dimension can also change intermediate widths. Their complete layer dimensions, activations, normalization, masking, and initialization are specified in Appendix I.2, with exact parameter counts in Table S25. Output normalization follows each architecture's implementation and is not varied as an independent control. Adapter dropout is 0.1; frozen backbone dropout is disabled. Table S5 contains all 84 primary adapter conditions: 14 architectures, two output widths, and three backbone sizes. The gaps-only control retains native dimension and appears separately in Table S6. The primary 35M experiments share a 65,536-token budget; the earlier 131,072-token 35M cohort is excluded. The reported sweep contains 96 continuous ESM-2 checkpoints and 20 discrete predictors; additional historical recipes are excluded.

Table S5: Complete frozen-backbone sweep: threshold correspondence F1 (0–1) under the shared scorer. Each width contains all 14 adapter architectures at all three backbone sizes. Training budgets are matched within each size; gap-only controls retain native widths and are tabulated separately.

	64D adapters			128D adapters		
Architecture	8M	35M	150M	8M	35M	150M
Linear	0.1451	0.1131	0.1829	0.1538	0.1434	0.1658
MLP-2	0.1178	0.2413	0.2822	0.1405	0.2549	0.2890
MLP-3	0.1696	0.2635	0.2773	0.1920	0.2697	0.2832
MLP-4	0.2012	0.2564	0.2580	0.2166	0.2625	0.2619
CNN-3	0.2549	0.2803	0.2772	0.2661	0.2829	0.2821
CNN-5	0.2835	0.2873	0.2877	0.2894	0.2910	0.2935
CNN-7	0.2953	0.2968	0.2986	0.2994	0.2981	0.3028
CNN-9	0.2983	0.2973	0.3049	0.3054	0.3058	0.3041
BiGRU-1	0.2555	0.3402	0.3448	0.2598	0.3402	0.3497
BiGRU-2	0.3119	0.3540	0.3649	0.3210	0.3558	0.3644
BiLSTM-1	0.2729	0.3403	0.3450	0.2781	0.3407	0.3469
BiLSTM-2	0.3150	0.3591	0.3693	0.3377	0.3594	0.3572
Projectformer-2	0.2451	0.2808	0.2945	0.2780	0.2861	0.3031
Projectformer-4	0.2836	0.2923	0.3192	0.2988	0.3265	0.3242

D.4 Correspondence endpoints and adaptation details

Threshold F1 measures recovery of reference correspondences and complements the MEA structural comparison in the main text.

D.5 Adapting the backbone improves correspondence recovery

How much correspondence information can be recovered without changing the pretrained encoder? Learning only the gaps yields threshold F1 between 0.1321 and 0.1421, with no improvement from a larger backbone (Table S6). The best 64D adapters raise F1 to 0.3150, 0.3591, and 0.3693 at 8M, 35M, and 150M. Updating the encoder as well, with auxiliary MLM, raises it to 0.4651, 0.5520, and 0.4845: absolute F1 gains of 0.1500, 0.1929, and 0.1152 over those adapters. The fully tuned 8M model exceeds even the best 150M 64D frozen adapter by 0.0958 F1. Within this training budget, adapting the representation contributes more than increasing the capacity of a frozen encoder.

Table S6: Threshold correspondence F1 (0–1). BiLSTM-2 is the best 64D frozen adapter at each size under this metric. Each checkpoint is evaluated with the shared per-pair scorer. LoRA and full tuning use MLM unless indicated.

Adaptation regime	8M	35M	150M
Frozen gaps only	0.1421	0.1340	0.1321
Frozen BiLSTM-2, 64D	0.3150	0.3591	0.3693
Frozen BiGRU-2, 64D	0.3119	0.3540	0.3649
LoRA + MLM	0.3877	0.3966	0.3590
Full + MLM	0.4651	0.5520	0.4845
Full, no MLM	0.4801	0.5639	–
Random init + MLM	0.1319	–	–

LoRA restricts encoder adaptation to low-rank updates and reaches 0.3877, 0.3966, and 0.3590. It exceeds every 64D frozen adapter at 8M and 35M but falls 0.0103 F1 below the best such adapter at 150M. Full tuning also peaks at 35M. These results do not isolate capacity: learning rate, batch budget, and updates per epoch vary with backbone size. They establish the ordering of the tested configurations, without establishing an intrinsic optimum at 35M.

The controls distinguish two uses of extra parameters. An external adapter learns how to read fixed residue features; LoRA and full tuning can change the features being read. The initialization control asks whether the same architecture can learn from the available alignment supervision without pretrained weights. These comparisons answer related but different questions. Within a backbone size the token budget is matched, whereas comparisons across sizes also change the optimization conditions. The results support the value of adaptation under the tested recipe, rather than a general ranking of backbone capacities.

The backbone-size figures in Appendix O include every continuous ESM-2 checkpoint, grouped by adaptation family. All 14 adapter architectures and both output widths are retained, together with the MLM, LoRA, gaps-only and random-initialization controls. The checkpoint IDs distinguish conditions with the same architecture.

D.6 Sequence-processing adapters outperform simple projections

The backbone-size figures and Table S5 show substantial variation among adapters with the same output dimension. At 8M, a linear projection reaches 0.1451 F1, a width-nine CNN reaches 0.2983, and a two-layer bidirectional LSTM reaches 0.3150. The corresponding 150M values are 0.1829, 0.3049, and 0.3693. For the width-nine CNN, increasing the frozen backbone from 8M to 150M improves F1 by only 0.0066 F1, whereas changing the 8M adapter from linear to CNN improves it by 0.1532 F1. The two-layer bidirectional recurrent adapters are the strongest 64D architectures at each size.

These gains show that correspondence information in frozen features becomes more accessible after additional processing along the sequence. The architectures also differ in parameter count, initialization, and normalization, so the experiment does not isolate sequence mixing alone. Increasing output width to 128D has a smaller and less consistent effect: 8M BiLSTM-2 improves from 0.3150 to 0.3377, while 150M BiGRU-2 changes from 0.3649 to 0.3644. Both widths are evaluated for every architecture and backbone size.

The preferred architecture also depends on the metric. At 8M/64D, BiGRU-2 has slightly higher MEA full-reference Cα LDDT (0.3357 versus 0.3355), although BiLSTM-2 has higher threshold F1 (0.3150 versus 0.3119). At 150M/128D, BiGRU-2 has higher threshold F1 (0.3644 versus 0.3572), while BiLSTM-2 has slightly higher LDDT (0.3465 versus 0.3459). The complete graphical comparisons retain every adapter rather than representing the family by its strongest architecture.

Appendix K combines all four LDDT endpoints and F1 for every adapter and training recipe in shared matrices, with all three correspondence decoders. This includes the complete 84-condition adapter sweep and the 12 continuous controls.

D.7 PRETRAINING AND AUXILIARY MASKED LANGUAGE MODELING

Starting the 8M architecture from random weights reduces threshold F1 from 0.4651 to 0.1319 with auxiliary MLM. Its MEA full-reference Cα LDDT is 0.2960, compared with 0.3860 from pretrained initialization. The 35M architecture shows the same effect: random initialization leaves threshold F1 at 0.1141 and MEA full-reference Cα LDDT at 0.2857, far below the pretrained 35M model. Pretraining thus makes the same alignment supervision more effective within the fixed training budget, with a larger relative benefit for threshold F1 than for MEA geometry.

Removing MLM improves threshold F1 by 0.0150 at 8M and 0.0120 at 35M (Table S7). MEA F1 and the two full-reference structural scores also increase modestly in both cases. Under the tested weighting and corruption schedule, alignment supervision alone is therefore sufficient to match and modestly improve these endpoints.

Table S7: MLM ablation under full fine-tuning. All metrics are 0–1 decimals; structural scores use the full-reference denominator and MEA mappings. The learning-rate rule changes slightly when the MLM head is removed.

Size	Objective	Threshold F1	MEA F1	Cα LDDT	All-atom LDDT
8M	With MLM	0.4651	0.4261	0.3860	0.3481
8M	Without MLM	0.4801	0.4424	0.3893	0.3508
35M	With MLM	0.5520	0.5174	0.4154	0.3742
35M	Without MLM	0.5639	0.5328	0.4185	0.3768

D.8 THE DECODER CHANGES THE MEASURED OUTCOME

Changing the decoder can reverse the ordering of correspondence and structural scores, even for a fixed checkpoint (Table S1). For the 35M fully tuned model without MLM, threshold decoding yields 0.5639 F1, mutual matching 0.5352, and MEA 0.5328. Full-reference Cα LDDT follows the reverse ordering: 0.3334, 0.4135, and 0.4185. Threshold decoding has the highest aligned-only Cα score, 0.5995, after excluding undefined predictions. Mutual matching and MEA reach 0.5711 and 0.5677, respectively. A favorable aligned-only average can therefore coexist with reduced coverage and a smaller evaluation population.

Mutual matching preserves one-to-one residue assignments but does not require sequence order. For this checkpoint, 2,750 pairs contain nonmonotone matches, compared with none under MEA. Residue coverage is 77.19% with mutual matching, close to MEA's 78.13% and above thresholding's 61.41%. These are different correspondence selections from the same learned scores, so the comparison concerns both coverage and alignment constraints.

Thresholding selects high-confidence edges from the marginal matrix. MEA can select a monotone path with broader coverage and improve full-reference geometry even when pooled correspondence F1 decreases. The reported score therefore describes the representation together with its decoder. We specify both the decoder and denominator for each metric and apply the same structural scorer, coordinate filters, and empty-alignment rules to all external methods.

D.9 COMPLETE ADAPTER COMPARISONS

Table S27 compares every architecture and width under all three correspondence decoders, with all four structural metrics and correspondence F1. The shared aggregate matrices in Appendix K place the decoders side by side.

D.10 CONTINUOUS CONTROLS AND ALTERNATIVE TRAINING CONFIGURATIONS

The continuous controls comprise three gaps-only, five full-tuning, three LoRA, and one random-initialization checkpoint. Tables S8 and S9 report all twelve under all three correspondence decoders. All use uncapped wBCE and the affine SW operator. The 35M conditions use a 65,536-token budget.

Table S8: Continuous control condition key. All 12 gaps-only, full-tuning, LoRA and random-initialization checkpoints. Rows specify the score, operator, alignment-noise fraction, and positive-weight cap; None means uncapped weighting. SW and SV identify the recorded operator families. The cosine-scoring full-tuning condition is explicit. Token budget is the nominal non-padding residue budget; a dash indicates unavailable provenance, not zero.

Condition	Short ID	Score	Operator	Noise	Weight cap	Tokens
8M Frozen gaps only	71222270	dot	sw	0.05	None	131072
8M Full, no MLM	5675f423	dot	sw	0.05	None	131072
8M Full + MLM	903cf320	dot	sw	0.05	None	131072
8M LoRA + MLM	11273f46	dot	sw	0.05	None	131072
8M Random init + MLM	435d65ae	dot	sw	0.05	None	131072
35M Frozen gaps only	3041e38b	dot	sw	0.05	None	65536
35M Full + MLM	f7fbcb88	dot	sw	0.05	None	65536
35M Full, no MLM	9278fe7b	dot	sw	0.05	None	65536
35M LoRA + MLM	5995efb6	dot	sw	0.05	None	65536
150M Frozen gaps only	86ff5a2d	dot	sw	0.05	None	49152
150M Full + MLM	00728588	dot	sw	0.05	None	49152
150M LoRA + MLM	9976e4d4	dot	sw	0.05	None	49152

Table S9: Continuous controls: every checkpoint and decoder. All 12 controls with threshold, mutual matching and MEA decoding (36 routes). Four symmetric LDDTs and pooled F1 in each row use the same map. Condition IDs resolve to Table S8; distinct recipes are not averaged.

Condition [ID] / decoder	Cα full	Cα aligned	All-atom full	All-atom aligned	Pooled F1
8M Frozen gaps only [71222270] / Threshold	0.2168	0.3440	0.1923	0.3232	0.1421
8M Frozen gaps only [71222270] / MM	0.2343	0.3306	0.2122	0.3117	0.1395
8M Frozen gaps only [71222270] / MEA	0.2363	0.3309	0.2144	0.3122	0.1393
8M Full, no MLM [5675f423] / Threshold	0.2737	0.5786	0.2332	0.5536	0.4801
8M Full, no MLM [5675f423] / MM	0.3818	0.5337	0.3416	0.5120	0.4459
8M Full, no MLM [5675f423] / MEA	0.3893	0.5290	0.3508	0.5075	0.4424
8M Full + MLM [903cf320] / Threshold	0.2548	0.5842	0.2158	0.5587	0.4651
8M Full + MLM [903cf320] / MM	0.3766	0.5288	0.3366	0.5072	0.4302
8M Full + MLM [903cf320] / MEA	0.3860	0.5230	0.3481	0.5017	0.4261
8M LoRA + MLM [11273f46] / Threshold	0.2108	0.5861	0.1774	0.5599	0.3877
8M LoRA + MLM [11273f46] / MM	0.3490	0.5023	0.3115	0.4815	0.3631
8M LoRA + MLM [11273f46] / MEA	0.3630	0.4945	0.3283	0.4742	0.3584
8M Random init + MLM [435d65ae] / Threshold	0.1016	0.5019	0.0833	0.4772	0.1319
8M Random init + MLM [435d65ae] / MM	0.2902	0.3981	0.2618	0.3806	0.1453
8M Random init + MLM [435d65ae] / MEA	0.2960	0.3957	0.2689	0.3784	0.1441
35M Frozen gaps only [3041e38b] / Threshold	0.2092	0.3240	0.1868	0.3041	0.1340
35M Frozen gaps only [3041e38b] / MM	0.2227	0.3136	0.2020	0.2953	0.1319
35M Frozen gaps only [3041e38b] / MEA	0.2245	0.3138	0.2039	0.2956	0.1317
35M Full + MLM [f7fbcb88] / Threshold	0.3135	0.5999	0.2691	0.5748	0.5520
35M Full + MLM [f7fbcb88] / MM	0.4090	0.5661	0.3663	0.5436	0.5205
35M Full + MLM [f7fbcb88] / MEA	0.4154	0.5619	0.3742	0.5395	0.5174
35M Full, no MLM [9278fe7b] / Threshold	0.3334	0.5995	0.2879	0.5746	0.5639
35M Full, no MLM [9278fe7b] / MM	0.4135	0.5711	0.3706	0.5483	0.5352
35M Full, no MLM [9278fe7b] / MEA	0.4185	0.5677	0.3768	0.5451	0.5328
35M LoRA + MLM [5995efb6] / Threshold	0.1981	0.6052	0.1648	0.5771	0.3966
35M LoRA + MLM [5995efb6] / MM	0.3464	0.5012	0.3088	0.4804	0.3706
35M LoRA + MLM [5995efb6] / MEA	0.3605	0.4934	0.3256	0.4732	0.3650
150M Frozen gaps only [86ff5a2d] / Threshold	0.2088	0.3208	0.1865	0.3007	0.1321
150M Frozen gaps only [86ff5a2d] / MM	0.2215	0.3112	0.2007	0.2928	0.1301
150M Frozen gaps only [86ff5a2d] / MEA	0.2232	0.3114	0.2026	0.2931	0.1299
150M Full + MLM [00728588] / Threshold	0.2746	0.5767	0.2332	0.5517	0.4845
150M Full + MLM [00728588] / MM	0.3825	0.5313	0.3430	0.5097	0.4486
150M Full + MLM [00728588] / MEA	0.3902	0.5272	0.3523	0.5059	0.4452
150M LoRA + MLM [9976e4d4] / Threshold	0.2179	0.5520	0.1842	0.5264	0.3590
150M LoRA + MLM [9976e4d4] / MM	0.3302	0.4776	0.2949	0.4578	0.3327

Continued on next page

Continuous controls: every checkpoint and decoder. (continued)

Condition [ID] / decoder	Cα full	Cα aligned	All-atom full	All-atom aligned	Pooled F1
150M LoRA + MLM [9976e4d4] / MEA	0.3420	0.4723	0.3088	0.4528	0.3287

E Learning and Evaluating a Fixed 3Di Representation

E.1 A fixed score matrix and learned residue states

We fix the 20-state substitution matrix to Foldseek revision `463739e`, with bit factor 2.1 and score bias zero. The state order is `ACDEFGHIKLMNPQRSTVWY`; these letters identify structural states, not amino acids. A checksum identifies the exact matrix, and its full signed factorization reconstructs it within absolute tolerance 10^{-5}. Training uses this structural score without an amino acid term or composition correction. Native Foldseek can include both, so transferring the predicted states to Foldseek also changes the scoring procedure.

For categorical distributions p_i and q_j, the soft residue-pair score is $p_i^\top B q_j$. For deterministic states a_i and b_j, it reduces exactly to $B_{a_i b_j}$. The encoder learns which states support alignment while every state-pair score remains fixed. Correspondence supervision therefore constrains the alignment induced by the states without requiring them to reconstruct the labels assigned by Foldseek's coordinate tokenizer.

An exact embedding factorization. The continuous model compares residue embeddings by a dot product. The fixed-matrix model can also be expressed through embeddings, but preserving the published score requires retaining both positive and negative eigenmodes. For symmetric B, write $B = Q\Lambda Q^\top$, let $E = Q|\Lambda|^{1/2}$, and set $J = \operatorname{diag}(\operatorname{sign} \lambda_k)$. Then

$$z_i = E^\top p_i, \qquad w_j = E^\top q_j, \qquad z_i^\top J w_j = p_i^\top B q_j. \tag{9}$$

Each row of E is the embedding of a pure state, and $E^\top p_i$ is their probability-weighted mixture. Since $EJE^\top = B$, the signed bilinear product between two mixtures reproduces their expected substitution score exactly. When each distribution becomes one-hot, the same identity recovers the ordinary lookup score. Replacing this factorization by an ordinary positive-semidefinite Gram factorization would change the matrix and therefore the alignment problem. The fixed signed metric is part of the score definition, not a learned correction.

Corollary 1 (Exact relaxation of fixed-table substitution)**.** *Let $B = EJE^\top$ be the full signed factorization of a symmetric substitution matrix. For logits ℓ, m and $\tau_c > 0$, set $p = \operatorname{softmax}(\ell/\tau_c)$ and $q = \operatorname{softmax}(m/\tau_c)$. Then $z(p) = E^\top p$ satisfies*

$$z(p)^\top J z(q) = p^\top B q = \mathbb{E}_{a\sim p,\, b\sim q}[B_{ab}], \tag{10}$$

where the two state draws are independent. Replacing p, q by one-hot states e_a, e_b recovers B_{ab} exactly. Thus the fixed embedding map preserves both the soft substitution score and its derivatives with respect to the logits.

Proof. Associativity gives $(E^\top p)^\top J (E^\top q) = p^\top (EJE^\top) q = p^\top B q$. Expanding the last bilinear form gives the expectation; smoothness of softmax gives equality of derivatives. □

The probabilities, not the raw logits, enter the fixed embedding map. Negative eigenmodes must retain their signs. This is a substitution-level identity: argmax need not preserve an alignment, and straight-through selectors use surrogate gradients rather than derivatives of argmax.

How hardening can change a substitution. Let $a^\star = \arg\max_a p_a$, $b^\star = \arg\max_b q_b$, and $\Delta_B = \max_{a,b} B_{ab} - \min_{a,b} B_{ab}$. Then

$$\left| p^\top B q - B_{a^\star b^\star} \right| \le \Delta_B \left(1 - p_{a^\star} q_{b^\star}\right). \tag{11}$$

To see this, separate the probability mass $p_{a^\star} q_{b^\star}$ at the selected pair. Every other table entry differs from $B_{a^\star b^\star}$ by at most Δ_B. Thus concentrated mixtures approach their hard substitution score, but this bound alone does not preserve the best alignment or retrieval ranking.

Gradients and the role of the signs. For $s = p^\top Bq$ with $p = \mathrm{softmax}(\ell/\tau_c)$,

$$\nabla_\ell s = \frac{1}{\tau_c}\, p \odot (Bq - s\mathbf{1})\,. \tag{12}$$

The signed implementation has the same gradient and higher derivatives. The pinned runtime matrix has 17 positive and three negative eigenvalues; retaining all 20 modes gives a change of coordinates, not a low-rank approximation. Using $EE^\top$ instead of $EJE^\top$, or discarding the negative modes, changes the matrix and hence the model.

Scope of the expectation. The score $\sum_{a,b} p_a q_b B_{ab}$ averages substitution scores. It is generally different from $\log \sum_{a,b} p_a q_b \exp(B_{ab})$. Applying the alignment operator to expected substitutions does not marginalize a joint distribution over all state assignments and alignment paths. The corollary also does not identify the entire native Foldseek scoring procedure with the training operator: amino acid terms, composition corrections, learned versus fixed gaps, and correspondence decoding remain distinct.

The factors in Equation (9) are fixed buffers. We compute softmax probabilities and the signed bilinear product in float32, even when the prediction head uses bfloat16 autocast, to avoid rounding cancellation between signed modes to bfloat16. The linear head predicts 20 logits directly. The four-stage MLP uses three SwiGLU hidden projections followed by a linear 20-logit output; dropout acts on logits before categorical selection.

State selection during training and inference. There are two distinct selections in the discrete model. The state selector chooses a 3Di representation for each residue; the correspondence decoder later chooses which residues match. Softmax training propagates the expected score between state distributions. Straight-through training instead uses hard states in its forward computation while retaining a surrogate gradient through the probabilities. Neither choice specifies the correspondence decoder. We evaluate threshold, mutual matching, and MEA on both soft and hard scores to keep the two decisions separate.

For deterministic straight-through training, let p be the soft probabilities and h the one-hot argmax. The implemented selector is $h + (p - \mathrm{stopgrad}(p))$, which has a hard forward value and a soft surrogate gradient. The Gumbel variant adds sampled Gumbel noise before forming the training selector. Evaluation uses deterministic argmax without noise. Diagnostic soft mixtures use temperature one and the same logits; categorical surrogate temperature and alignment temperature remain separately recorded quantities.

The categorical logit projection is initialized orthogonally with unit scale and zero bias. In the masked-3Di control, the same MLP head serves both alignment and token supervision. The alignment pass receives clean amino acid inputs. A second pass selects 15% of amino acid positions and uses 80% mask, 10% random, and 10% unchanged corruption; cross-entropy targets are valid native 3Di labels in the range 0–19 at those selected positions. The masked loss and alignment loss have unit coefficients. This control disables the separate 5% alignment-noise corruption. Its differences from alignment-only training therefore include direct structural-token supervision and the input-corruption schedule.

The frozen-backbone discrete experiments start from an encoder already fine-tuned for alignment. This gives their projection heads a different starting representation from the external adapters trained above the original pretrained encoder. The twelve frozen-backbone models train for ten epochs, initialize gap-open/extension scores at $(-10, -1)$, and start alignment temperature at one. The eight unfrozen runs initialize gaps at $(-3, -1)$ and start alignment temperature at five. Comparisons between these groups therefore change initialization and temperature as well as backbone freezing. Evaluation restores the learned gaps from each specified checkpoint.

Figures S1 and S2 show how categorical states replace continuous residue vectors while retaining the alignment objective, and how the fixed signed codebook scores soft mixtures and hard states.

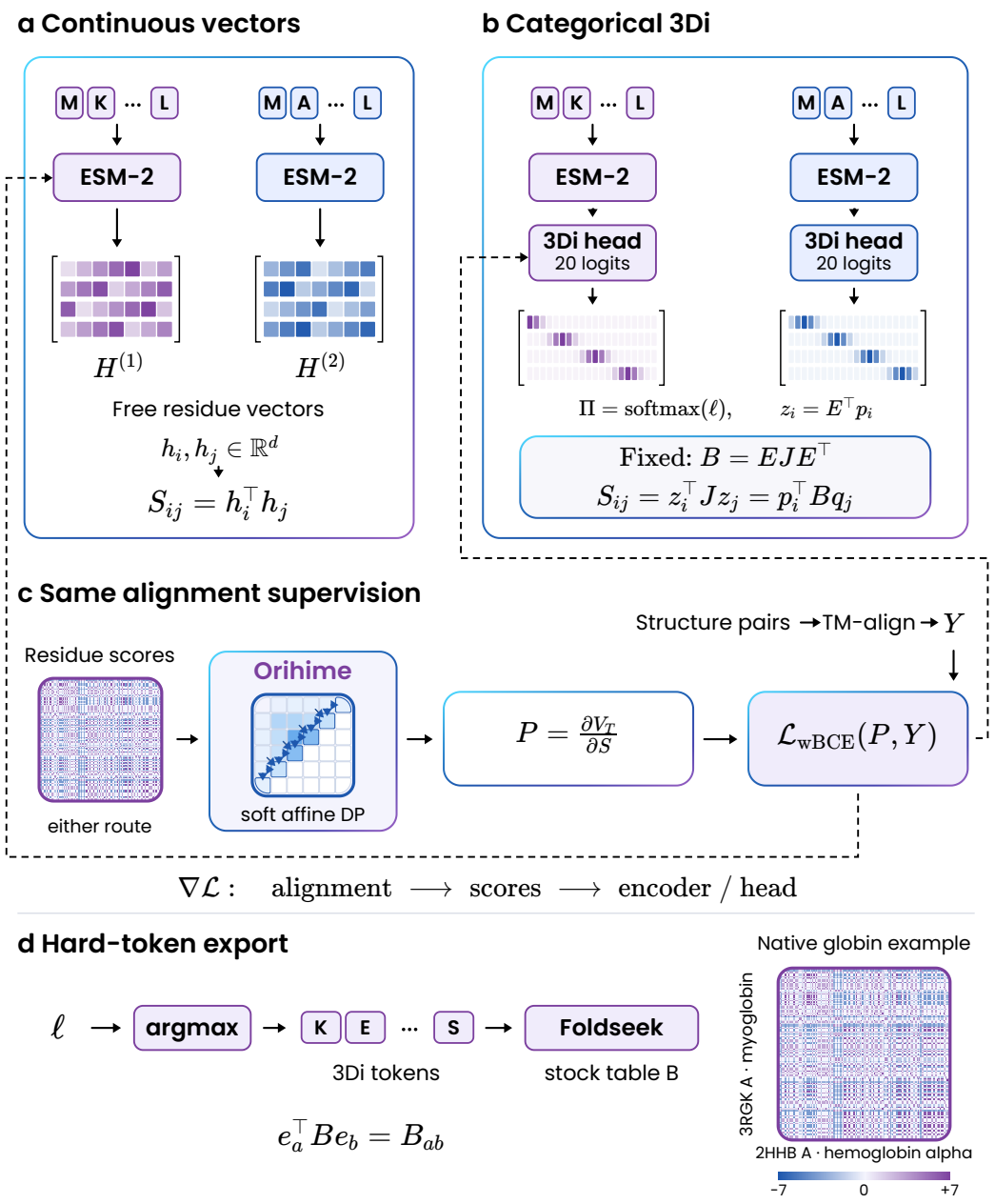


Figure S1: **Changing the representation, retaining the alignment objective.** (a) The continuous route learns residue vectors and their dot-product scores. (b) The 3Di route instead learns 20-state assignments. A fixed codebook maps each assignment to $z = E^\top p$; the signed product recovers the expected Foldseek table score because $B = EJE^\top$. Paired branches share parameters. (c) Either score construction feeds the same differentiable alignment and TM-align correspondence loss. The head, and an unfrozen encoder, receive gradients; E and J do not change. This is alignment supervision, not native-state classification. The soft path is shown; straight-through conditions use hard forward states and surrogate gradients. Auxiliary objectives are omitted. (d) Argmax exports native-format tokens for stock Foldseek. Native globin scores illustrate the input in (c) and inset in (d), not learned predictions: 3RGK A versus 2HHB A, 149×141 exact lookups, without AA or composition corrections. Blue/white/purple denote $-7/0/+7$. Arrays and DP artwork are schematic.

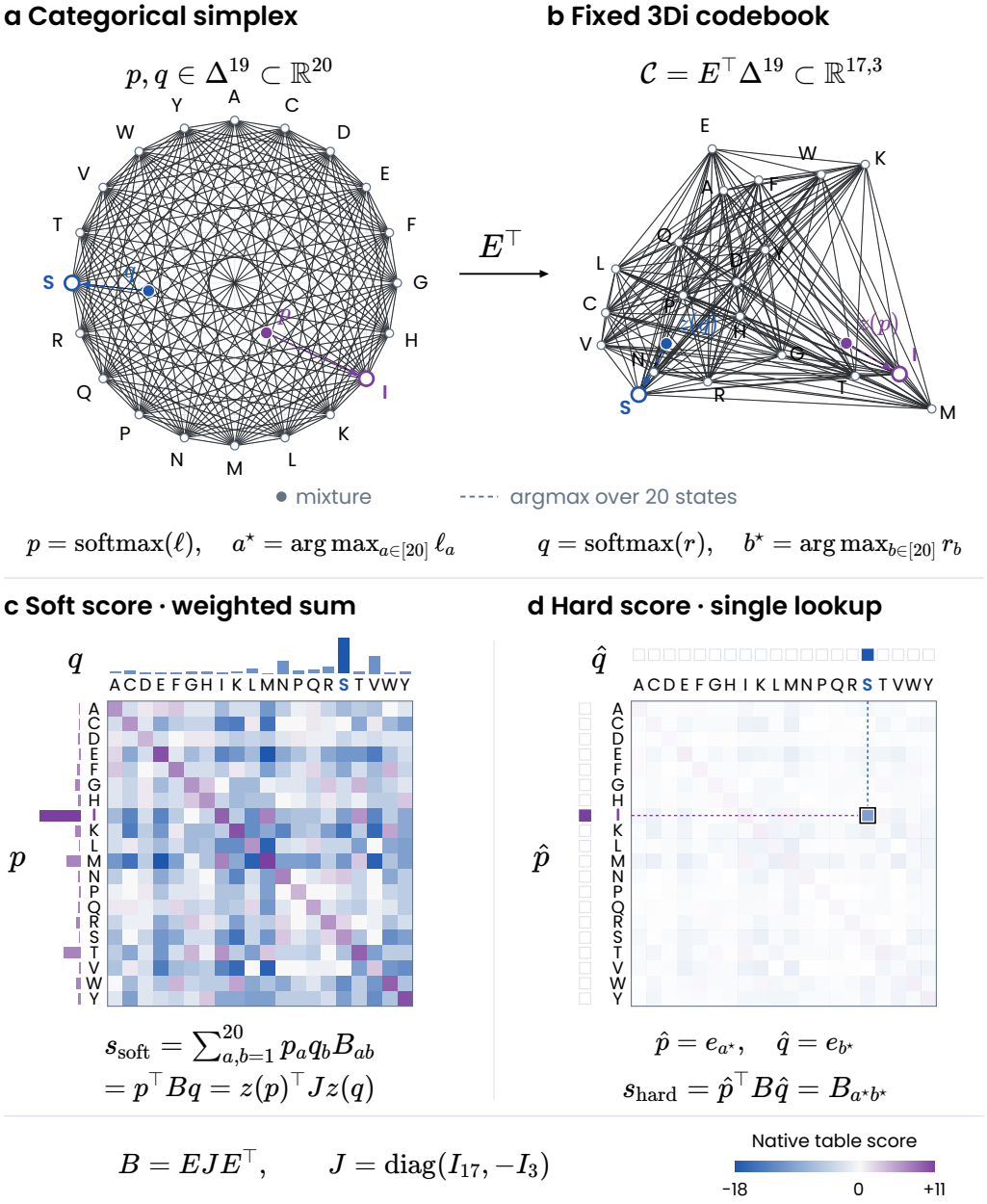


Figure S2: **Full-alphabet mixture geometry and native scoring.** (a) Two 20-dimensional logit vectors ℓ, r produce assignments $p, q \in \Delta^{19}$. All 20 vertices and 190 edges of the simplex are shown. (b) The fixed map $E^\top$ sends vertices to native state embeddings and assignments to their barycenters. Filled markers are mixtures; outlined colored vertices are the full-alphabet argmax choices. Dashed arrows indicate hard selection, not nearest-neighbor projection. Both views are two-dimensional linear display projections: the left uses an orthonormal Fourier pair; the right shows the first two codebook coordinates. Projected barycenters are exact, but distances and scores are not preserved. (c) Soft scoring weights all 400 table entries by $p_a q_b$; bars show the complete marginals. (d) One-hot selectors $\hat{p}, \hat{q}$ identify a single lookup in the same table. Unused entries are faded, not changed; table colors are shared across panels. Signed scoring uses all 20 coordinates, including the three negative modes. Letters identify 3Di states, not amino acids. The illustrative logits are not model predictions; no states are truncated.

E.2 CHECKPOINTS, TEMPERATURES, AND NATIVE ALIGNMENT

The discrete sweep contains ten 8M and ten 35M predictors. Each size includes frozen linear and four-layer MLP heads trained with softmax, deterministic straight-through, or Gumbel straight-through selection; an unfrozen four-layer MLP under each selector; and an unfrozen softmax model with masked 3Di supervision. All native Foldseek evaluations use deterministic tokens.

Each predictor is evaluated in eight ways: soft and hard state scores through each of threshold, mutual matching, and MEA correspondence decoding, plus hard strings through native Foldseek in 3Di-only and AA+3Di modes. This gives 160 pairwise configurations across the 20 predictors. Hard-state evaluation can use any of the three correspondence decoders. All six Orihime evaluations use the same checkpoint and its learned gaps; the two native evaluations use Foldseek's fixed settings and the correspondences it emits.

The softmax-trained examples in Section 4.4 are checkpoint `16132e1d` at zero-based epoch 38 (8M) and `f7815ea1` at epoch 28 (35M). The Gumbel-trained 35M example is `f32e1543` at epoch 21. These short identifiers match Table S10. Each uses an unfrozen encoder and a four-layer head. For each model, the soft, hard/Orihime, and native Foldseek scores all come from this same checkpoint.

All predictions are evaluated at alignment temperature one. Categorical softmax and surrogate temperature remain one. Native Foldseek uses none of these learned-operator temperatures. Within-checkpoint soft/hard comparisons hold alignment temperature fixed; comparisons between predictors do not isolate the selector from training epoch and training temperature.

Native Foldseek uses revision `463739e0014a1549a527de589102cde98f802f37`, hard token strings, gap penalties 10/1, 3Di scale 2.1, and amino acid scale 1.4 in AA+3Di mode or zero in 3Di-only mode. Composition correction is 0.5 and reverse-query correction is enabled. Learned gaps, Orihime marginals, and coordinate-based reranking are absent from native evaluation. All 68,431 requested pairs are explicit candidates, without prefiltering, and all return a native correspondence map. Prediction-derived inputs contain amino acid and 3Di strings, with no coordinate database. These experiments compare learned strings with learned strings.

Table S10 identifies all twenty checkpoints. The same short identifiers link their alignment, fidelity, retrieval, and clustering results.

Table S10: Discrete predictor checkpoint key. All 20 predictors. Epochs are zero-based. The short ID is the last eight hexadecimal characters of the stable run ID. DP T is the actual evaluated Orihime alignment temperature, rounded to four decimals; exact values and checkpoint hashes are in the companion CSV. ST denotes the training straight-through selector. Native Foldseek uses deterministic hard tokens; its aligner does not use this DP temperature. Frozen/unfrozen describes the backbone; +mask denotes the masked recipe.

Backbone / head	Short ID	Epoch	DP T	Training selector
8M frozen MLP4	87d2c509	9	1.0000	ST Gumbel
8M frozen MLP4	d58f9e14	9	1.0000	ST argmax
8M frozen MLP4	7b541c60	9	1.0000	Softmax
8M frozen linear	31bba998	9	1.0000	ST Gumbel
8M frozen linear	fbc46ff2	9	1.0000	ST argmax
8M frozen linear	1cf7945c	9	1.0000	Softmax
8M unfrozen MLP4	d50ef818	15	1.0000	ST Gumbel
8M unfrozen MLP4	650078f8	16	1.0000	ST argmax
8M unfrozen MLP4	16132e1d	38	1.0000	Softmax
8M unfrozen MLP4 + masked	9d792c09	27	1.0000	Softmax
35M frozen MLP4	72459d15	9	1.0000	ST Gumbel
35M frozen MLP4	9fceaf26	9	1.0000	ST argmax
35M frozen MLP4	0eb8395a	9	1.0000	Softmax
35M frozen linear	4b1a9c9d	9	1.0000	ST Gumbel
35M frozen linear	f8f5cad4	9	1.0000	ST argmax
35M frozen linear	c336a672	9	1.0000	Softmax
35M unfrozen MLP4	f32e1543	21	1.0000	ST Gumbel
35M unfrozen MLP4	39efc845	21	1.0000	ST argmax
35M unfrozen MLP4	f7815ea1	28	1.0000	Softmax

Continued on next page

Discrete predictor checkpoint key. (continued)

Backbone / head	Short ID	Epoch	DP T	Training selector
35M unfrozen MLP4 + masked	929752d4	17	1.0000	Softmax

E.3 What changes under discretization and native decoding

Transferring the learned representation to native Foldseek involves two steps. First, replacing each state distribution by argmax changes the residue-pair score while keeping Orihime's learned gaps, temperature, and correspondence decoder fixed. Second, passing those same hard strings to native Foldseek changes the alignment procedure and scoring corrections. Comparing a soft Orihime result directly with a native Foldseek result combines both changes. The intermediate hard/Orihime evaluation makes their effects separately visible within each checkpoint.

The effect of discretization depends on how the states were trained. With MEA and full-reference Cα LDDT, hard states outperform soft mixtures in five of the twenty predictors: the frozen linear and MLP-4 models trained with straight-through argmax at 8M and 35M, and the unfrozen 8M straight-through argmax model. These models already use hard states in the training forward pass; soft evaluation is a diagnostic mixture. Their behavior differs from that of the softmax-trained models and rules out a universal loss from discretization.

For the unfrozen softmax-trained examples, replacing soft mixtures by argmax decreases MEA Cα full-reference LDDT from 0.3755 to 0.3400 at 8M and from 0.4017 to 0.3768 at 35M. Keeping those hard strings and switching to native 3Di-only Foldseek gives 0.3530 and 0.3887. The switch to Foldseek also replaces learned gaps, marginal computation, and correspondence decoding, and enables native scoring corrections. The resulting change measures how the fixed strings behave under the complete native procedure; the preceding soft-to-hard comparison isolates the change in state representation within Orihime.

The same distinction applies to the auxiliary amino acid score. For the 35M softmax model, native 3Di-only and AA+3Di give F1 0.4861 and 0.5002, but full-reference Cα LDDT 0.3887 and 0.3774. Adding the amino acid term improves correspondence recovery while reducing this geometric endpoint. The 35M Gumbel model reaches higher native F1, 0.5053 and 0.5218 in the two modes, while its LDDT values are 0.3843 and 0.3667. Between these two checkpoints, stronger correspondence recovery therefore accompanies lower full-reference Cα LDDT.

The native tests use learned strings on both sides of each alignment. They establish that these strings can support useful alignments under the fixed native matrix and procedure. They do not test mixed comparisons between learned strings and structure-derived database strings, nor the calibration of native E-values under the learned token distribution.

Tables S30 and S11 give the complete within-checkpoint comparisons and native outputs, with separate soft, hard, and native results under each predictor's recorded evaluation settings.

Table S11: Discrete predictors: unchanged native Foldseek. All 20 predictors in both native modes. The hard tokens are inputs to Foldseek; its output maps are retained unchanged and are not decoded by Orihime. Four symmetric LDDTs and pooled F1 are fractions. Checkpoint key: Table S10; structure-derived and ProstT5-based Foldseek references: Table S2.

Predictor [ID] / mode	Cα full	Cα aligned	All-atom full	All-atom aligned	Pooled F1
8M frozen MLP4 ST-Gumbel [87d2c509] / 3Di-only	0.3295	0.5022	0.2880	0.4822	0.3558
8M frozen MLP4 ST-Gumbel [87d2c509] / AA+3Di	0.3043	0.5311	0.2599	0.5043	0.3973
8M frozen MLP4 ST-argmax [d58f9e14] / 3Di-only	0.2936	0.4906	0.2515	0.4720	0.3193
8M frozen MLP4 ST-argmax [d58f9e14] / AA+3Di	0.2715	0.5169	0.2287	0.4917	0.3619
8M frozen MLP4 softmax [7b541c60] / 3Di-only	0.3345	0.4604	0.3009	0.4417	0.2991
8M frozen MLP4 softmax [7b541c60] / AA+3Di	0.3195	0.4847	0.2800	0.4600	0.3518
8M frozen linear ST-Gumbel [31bba998] / 3Di-only	0.3278	0.4908	0.2876	0.4711	0.3280
8M frozen linear ST-Gumbel [31bba998] / AA+3Di	0.3057	0.5183	0.2617	0.4924	0.3797
8M frozen linear ST-argmax [fbc46ff2] / 3Di-only	0.2691	0.4953	0.2275	0.4747	0.3096
8M frozen linear ST-argmax [fbc46ff2] / AA+3Di	0.2470	0.5404	0.2057	0.5114	0.3619

Continued on next page

Discrete predictors: unchanged native Foldseek. (continued)

Predictor [ID] / mode	Cα full	Cα aligned	All-atom full	All-atom aligned	Pooled F1
8M frozen linear softmax [1cf7945c] / 3Di-only	0.3278	0.4605	0.2933	0.4422	0.2915
8M frozen linear softmax [1cf7945c] / AA+3Di	0.3001	0.4949	0.2603	0.4693	0.3514
8M unfrozen MLP4 ST-Gumbel [d50ef818] / 3Di-only	0.3388	0.5238	0.2959	0.5032	0.3970
8M unfrozen MLP4 ST-Gumbel [d50ef818] / AA+3Di	0.3228	0.5442	0.2786	0.5185	0.4315
8M unfrozen MLP4 ST-argmax [650078f8] / 3Di-only	0.2992	0.5234	0.2560	0.5028	0.3830
8M unfrozen MLP4 ST-argmax [650078f8] / AA+3Di	0.2841	0.5493	0.2417	0.5228	0.4155
8M unfrozen MLP4 softmax [16132e1d] / 3Di-only	0.3530	0.4873	0.3169	0.4677	0.3568
8M unfrozen MLP4 softmax [16132e1d] / AA+3Di	0.3427	0.5066	0.3016	0.4819	0.3991
8M unfrozen MLP4 softmax+mask [9d792c09] / 3Di-only	0.3405	0.4965	0.3018	0.4772	0.3599
8M unfrozen MLP4 softmax+mask [9d792c09] / AA+3Di	0.3301	0.5104	0.2882	0.4866	0.3907
35M frozen MLP4 ST-Gumbel [72459d15] / 3Di-only	0.3721	0.5494	0.3283	0.5286	0.4652
35M frozen MLP4 ST-Gumbel [72459d15] / AA+3Di	0.3437	0.5676	0.2962	0.5406	0.4860
35M frozen MLP4 ST-argmax [9fceaf26] / 3Di-only	0.3192	0.5530	0.2741	0.5313	0.4271
35M frozen MLP4 ST-argmax [9fceaf26] / AA+3Di	0.2867	0.5814	0.2420	0.5513	0.4501
35M frozen MLP4 softmax [0eb8395a] / 3Di-only	0.3651	0.5131	0.3262	0.4929	0.3969
35M frozen MLP4 softmax [0eb8395a] / AA+3Di	0.3359	0.5349	0.2918	0.5079	0.4302
35M frozen linear ST-Gumbel [4b1a9c9d] / 3Di-only	0.3683	0.5462	0.3245	0.5252	0.4430
35M frozen linear ST-Gumbel [4b1a9c9d] / AA+3Di	0.3445	0.5641	0.2968	0.5369	0.4667
35M frozen linear ST-argmax [f8f5cad4] / 3Di-only	0.3344	0.5550	0.2875	0.5336	0.4296
35M frozen linear ST-argmax [f8f5cad4] / AA+3Di	0.3006	0.5819	0.2535	0.5527	0.4505
35M frozen linear softmax [c336a672] / 3Di-only	0.3558	0.4986	0.3178	0.4791	0.3860
35M frozen linear softmax [c336a672] / AA+3Di	0.3262	0.5257	0.2824	0.4992	0.4223
35M unfrozen MLP4 ST-Gumbel [f32e1543] / 3Di-only	0.3843	0.5724	0.3390	0.5506	0.5053
35M unfrozen MLP4 ST-Gumbel [f32e1543] / AA+3Di	0.3667	0.5841	0.3195	0.5574	0.5218
35M unfrozen MLP4 ST-argmax [39efc845] / 3Di-only	0.3394	0.5650	0.2945	0.5431	0.4853
35M unfrozen MLP4 ST-argmax [39efc845] / AA+3Di	0.3222	0.5800	0.2775	0.5525	0.5022
35M unfrozen MLP4 softmax [f7815ea1] / 3Di-only	0.3887	0.5514	0.3465	0.5303	0.4861
35M unfrozen MLP4 softmax [f7815ea1] / AA+3Di	0.3774	0.5569	0.3323	0.5316	0.5002
35M unfrozen MLP4 softmax+mask [929752d4] / 3Di-only	0.3785	0.5535	0.3346	0.5328	0.4866
35M unfrozen MLP4 softmax+mask [929752d4] / AA+3Di	0.3711	0.5573	0.3251	0.5334	0.4981

The shared matrices in Appendix K report all four LDDT endpoints and F1 for every soft, hard and native route, preserving all three Orihime decoders.

E.4 Literal state fidelity

Do states learned from alignment recover Foldseek's coordinate-derived labels? To test this directly, we count each of the 4,624 proteins once, independent of its frequency in the alignment cohort. The primary analysis contains 856,552 residues with valid native conformational descriptors; 68 missing-Cα positions and 9,307 fallback descriptors are excluded. An alternative mask analysis includes the fallback tokens. Native descriptor reconstruction reproduces the saved structural tokens exactly, and the training-side labels agree at 856,549 of the 856,552 primary positions. Three differences caused by coordinate serialization cannot explain the low learned-token identity.

Across all twenty predictors, literal accuracy on the primary mask ranges from 5.19% to 15.47%, below the 20.21% majority-state predictor. The ProstT5-based Foldseek predictor reaches 61.30%. The best literal accuracy within the learned sweep belongs to the frozen 8M MLP-4 Gumbel predictor, not the 35M predictor highlighted for native structural agreement. Low identity therefore occurs across heads and selectors, and literal reconstruction ranks the predictors differently from native structural agreement. Uniform random identity is 5%; the majority-state baseline also accounts for the imbalance in the reference alphabet.

We report residue-micro and per-protein accuracy, per-state confusion, state frequencies, and balanced accuracy. A positional substitution-weighted dissimilarity uses $d(a, b) = (B_{aa}+B_{bb})/2-B_{ab}$ under the native matrix. Its off-diagonal costs are positive, but it violates the triangle inequality; neither it nor its normalized complement is a metric distance or a token-identity score. Edit distance is an additional diagnostic that can absorb errors through shifts. Literal identity and downstream alignment remain separate evaluations.

The independent-marginals chance agreement is $\sum_a f_a g_a$, where f and g are native and predicted state frequencies. Predicted entropy is $H = -\sum_a g_a \log_2 g_a$, with effective state count 2^H. These diagnostics describe state use without assuming uniformly distributed tokens. Balanced accuracy averages native-state recalls rather than weighting common states more heavily. For edit distance, both same-protein strings receive the same reference mask; valid-descriptor positions are concatenated before global Levenshtein distance is computed. Its unit insertion, deletion, and substitution costs differ from the positional matrix dissimilarity. The latter has maximum 28.5 under the fixed matrix and 358 ordered triangle-inequality violations. Dividing by 28.5 yields a normalized dissimilarity, not a percentage of correctly predicted states.

Allowing shifts or weighting substitutions by the native matrix does not remove the difference from literal reconstruction. For the 35M softmax model, normalized edit distance is 0.9124 and normalized positional matrix dissimilarity is 0.3229, compared with 0.3763 and 0.0694 for the ProstT5-based Foldseek predictor. The 35M Gumbel model uses 13.64 effective states, close to the ProstT5-based Foldseek predictor's 13.86, despite literal accuracy of only 7.35%. Its low state identity therefore does not mean that its predictions occupy a single state. The Cα-present mask changes the softmax model's accuracy from 5.19% to 5.34%, which does not resolve the discrepancy with native labels.

Low identity permits several explanations, including changed state frequencies and alternative sequence-dependent assignments. It does not by itself demonstrate a permutation symmetry of the fixed score matrix. The per-state confusion and occupancy diagnostics are descriptive; a mechanism claim would require testing the proposed transformation and its score preservation explicitly.

This distinction follows from what the alignment objective observes. A state is rewarded through the correspondences its scores support with states in the other learned string. Literal fidelity instead compares that state with the native label at the same residue, without asking which correspondences the string induces. The two measurements can therefore disagree without either being redundant. Mixed learned/native alignments would test a third property: whether states assigned under the two conventions can be compared using the same fixed matrix.

Figure S3 compares native-state identity with native alignment quality and isolates soft-to-hard changes under the same MEA correspondence decoder across all 20 models.

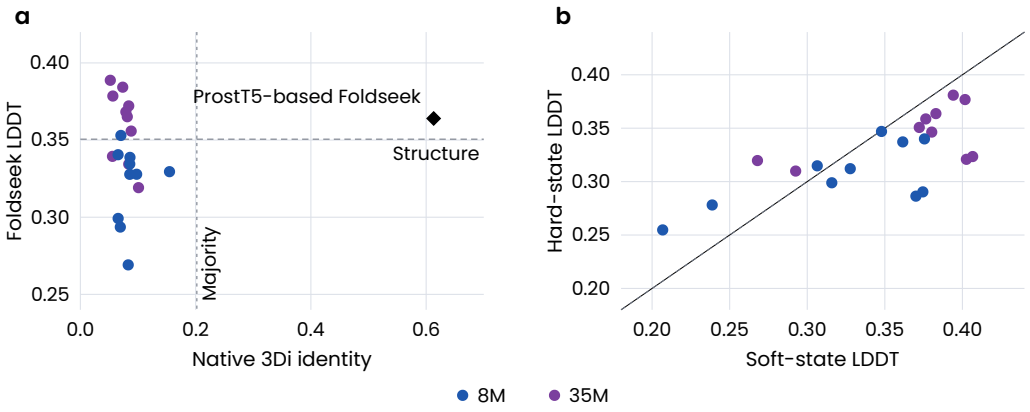


Figure S3: **Native-state identity and hardening.** All 20 learned checkpoints are shown. (a) Native-state identity uses 856,552 valid descriptors from 4,624 proteins; 3Di-only Foldseek alignment quality is symmetric full-reference Cα LDDT on 68,431 related development pairs. Dashed lines mark majority-state identity and structure-derived alignment quality. (b) Soft and hard states share the same checkpoint, gaps, temperature and Orihime MEA decoder. The diagonal denotes equality. Five checkpoints improve under hardening. Coincident points are retained without jitter.

Table S12 reports the valid-descriptor and Cα-present reference masks for every predictor.

Table S12: Native-state fidelity under both residue masks. All 21 predictors (20 learned plus the ProstT5-based Foldseek predictor), each evaluated once per unique protein: 4,624 proteins. Main valid-descriptor view (V): 856,552 residues; alternative native Cα-present view (C): 865,859. Q20 is pooled literal state accuracy, BA balanced accuracy, and MF1 macro F1. Both majority baselines use the source-reported majority state V. Dashes denote unreported baseline metrics, not zero. State identity alone does not establish the mechanism of alignment quality.

Predictor [ID]	V Q20	V BA	V MF1	C Q20	C BA	C MF1
ProstT5-based Foldseek	0.6130	0.5728	0.5829	0.6150	0.5730	0.5834
8M frozen MLP4 ST-Gumbel [87d2c509]	0.1547	0.0671	0.0549	0.1549	0.0666	0.0547
8M frozen MLP4 ST-argmax [d58f9e14]	0.0693	0.0469	0.0356	0.0692	0.0470	0.0356
8M frozen MLP4 softmax [7b541c60]	0.0857	0.0548	0.0289	0.0925	0.0557	0.0299
8M frozen linear ST-Gumbel [31bba998]	0.0976	0.0504	0.0389	0.0977	0.0501	0.0388
8M frozen linear ST-argmax [fbc46ff2]	0.0830	0.0508	0.0425	0.0848	0.0510	0.0428
8M frozen linear softmax [1cf7945c]	0.0857	0.0594	0.0349	0.0910	0.0604	0.0357
8M unfrozen MLP4 ST-Gumbel [d50ef818]	0.0860	0.0515	0.0439	0.0883	0.0517	0.0443
8M unfrozen MLP4 ST-argmax [650078f8]	0.0655	0.0434	0.0398	0.0660	0.0436	0.0399
8M unfrozen MLP4 softmax [16132e1d]	0.0703	0.0594	0.0336	0.0730	0.0594	0.0340
8M unfrozen MLP4 softmax+mask [9d792c09]	0.0654	0.0669	0.0474	0.0652	0.0664	0.0470
35M frozen MLP4 ST-Gumbel [72459d15]	0.0839	0.0397	0.0321	0.0835	0.0393	0.0319
35M frozen MLP4 ST-argmax [9fceaf26]	0.1009	0.0514	0.0451	0.1000	0.0511	0.0448
35M frozen MLP4 softmax [0eb8395a]	0.0814	0.0601	0.0381	0.0850	0.0606	0.0388
35M frozen linear ST-Gumbel [4b1a9c9d]	0.0790	0.0419	0.0330	0.0803	0.0423	0.0333
35M frozen linear ST-argmax [f8f5cad4]	0.0839	0.0581	0.0542	0.0843	0.0578	0.0540
35M frozen linear softmax [c336a672]	0.0882	0.0551	0.0348	0.0931	0.0554	0.0355
35M unfrozen MLP4 ST-Gumbel [f32e1543]	0.0735	0.0522	0.0456	0.0759	0.0527	0.0462
35M unfrozen MLP4 ST-argmax [39efc845]	0.0557	0.0402	0.0374	0.0572	0.0405	0.0377
35M unfrozen MLP4 softmax [f7815ea1]	0.0519	0.0619	0.0358	0.0534	0.0621	0.0360
35M unfrozen MLP4 softmax+mask [929752d4]	0.0563	0.0791	0.0521	0.0561	0.0790	0.0518
Majority-state accuracy	0.2021	–	–	0.2000	–	–

Table S13: Token state-usage diagnostics. Valid-descriptor mask; all 21 predictors. Each uses 4,624 proteins and 856,552 residues, with 9,375 excluded residues and 0 missing proteins. Protein mean averages per-protein token accuracy. Chance agreement is the dot product of reference and predicted state frequencies. Entropy H is in bits; effective states equals 2^H. These describe token usage, not alignment quality.

Predictor	Protein mean accuracy	Chance agreement	Entropy bits	Effective states
ProstT5-based Foldseek	0.6201	0.0926	3.7932	13.8637
8M frozen MLP4 ST-Gumbel [87d2c509]	0.1666	0.1032	2.9039	7.4842
8M frozen MLP4 ST-argmax [d58f9e14]	0.0641	0.0735	3.4178	10.6869
8M frozen MLP4 softmax [7b541c60]	0.0905	0.0758	2.3641	5.1484
8M frozen linear ST-Gumbel [31bba998]	0.1015	0.1027	2.8248	7.0849
8M frozen linear ST-argmax [fbc46ff2]	0.0894	0.0782	3.4978	11.2961
8M frozen linear softmax [1cf7945c]	0.0928	0.0768	2.5316	5.7822
8M unfrozen MLP4 ST-Gumbel [d50ef818]	0.0872	0.0786	3.5199	11.4707
8M unfrozen MLP4 ST-argmax [650078f8]	0.0643	0.0719	3.9145	15.0798
8M unfrozen MLP4 softmax [16132e1d]	0.0745	0.0582	2.7320	6.6435
8M unfrozen MLP4 softmax+mask [9d792c09]	0.0689	0.0481	3.4268	10.7540
35M frozen MLP4 ST-Gumbel [72459d15]	0.0896	0.1067	2.9411	7.6801
35M frozen MLP4 ST-argmax [9fceaf26]	0.1038	0.0922	3.5183	11.4584
35M frozen MLP4 softmax [0eb8395a]	0.0822	0.0650	2.9197	7.5668
35M frozen linear ST-Gumbel [4b1a9c9d]	0.0809	0.0997	3.1264	8.7323
35M frozen linear ST-argmax [f8f5cad4]	0.0898	0.0682	3.8805	14.7280
35M frozen linear softmax [c336a672]	0.0996	0.0776	2.7530	6.7413
35M unfrozen MLP4 ST-Gumbel [f32e1543]	0.0738	0.0715	3.7697	13.6393
35M unfrozen MLP4 ST-argmax [39efc845]	0.0580	0.0609	4.0289	16.3232
35M unfrozen MLP4 softmax [f7815ea1]	0.0539	0.0437	3.4324	10.7956
35M unfrozen MLP4 softmax+mask [929752d4]	0.0545	0.0357	3.8772	14.6946

Table S14: Token state-usage diagnostics. C-alpha-present mask; all 21 predictors. Each uses 4,624 proteins and 865,859 residues, with 68 excluded residues and 0 missing proteins. Protein mean averages per-protein token accuracy. Chance agreement is the dot product of reference and predicted state frequencies. Entropy H is in bits; effective states equals 2^H. These describe token usage, not alignment quality.

Predictor	Protein mean accuracy	Chance agreement	Entropy bits	Effective states
ProstT5-based Foldseek	0.6226	0.0938	3.7810	13.7462
8M frozen MLP4 ST-Gumbel [87d2c509]	0.1668	0.1051	2.8975	7.4516
8M frozen MLP4 ST-argmax [d58f9e14]	0.0642	0.0732	3.4184	10.6917
8M frozen MLP4 softmax [7b541c60]	0.0995	0.0798	2.3574	5.1243
8M frozen linear ST-Gumbel [31bba998]	0.1015	0.1041	2.8204	7.0635
8M frozen linear ST-argmax [fbc46ff2]	0.0918	0.0797	3.4939	11.2661
8M frozen linear softmax [1cf7945c]	0.1004	0.0800	2.5251	5.7562
8M unfrozen MLP4 ST-Gumbel [d50ef818]	0.0902	0.0797	3.5189	11.4628
8M unfrozen MLP4 ST-argmax [650078f8]	0.0652	0.0718	3.9139	15.0730
8M unfrozen MLP4 softmax [16132e1d]	0.0778	0.0604	2.7276	6.6237
8M unfrozen MLP4 softmax+mask [9d792c09]	0.0688	0.0489	3.4211	10.7116
35M frozen MLP4 ST-Gumbel [72459d15]	0.0890	0.1083	2.9356	7.6508
35M frozen MLP4 ST-argmax [9fceaf26]	0.1026	0.0927	3.5132	11.4175
35M frozen MLP4 softmax [0eb8395a]	0.0870	0.0668	2.9139	7.5366
35M frozen linear ST-Gumbel [4b1a9c9d]	0.0827	0.1007	3.1228	8.7108
35M frozen linear ST-argmax [f8f5cad4]	0.0906	0.0689	3.8756	14.6779
35M frozen linear softmax [c336a672]	0.1069	0.0809	2.7468	6.7121
35M unfrozen MLP4 ST-Gumbel [f32e1543]	0.0769	0.0722	3.7696	13.6380
35M unfrozen MLP4 ST-argmax [39efc845]	0.0601	0.0617	4.0279	16.3123
35M unfrozen MLP4 softmax [f7815ea1]	0.0556	0.0447	3.4271	10.7562
35M unfrozen MLP4 softmax+mask [929752d4]	0.0542	0.0357	3.8728	14.6499

Table S15: Unweighted token differences. Valid-descriptor mask; all 21 predictors, 4,624 proteins and 856,552 scored residues per predictor. Identity and mismatch compare corresponding residue positions. Edit is unit-cost global Levenshtein distance divided by scored length. Micro pools residues; macro averages proteins. The same reference mask is applied to both strings before scoring.

Predictor	Identity	Mismatch	Edit micro	Edit macro
ProstT5-based Foldseek	0.6130	0.3870	0.3763	0.3694
8M frozen MLP4 ST-Gumbel [87d2c509]	0.1547	0.8453	0.7913	0.7807
8M frozen MLP4 ST-argmax [d58f9e14]	0.0693	0.9307	0.8787	0.8903
8M frozen MLP4 softmax [7b541c60]	0.0857	0.9143	0.8739	0.8696
8M frozen linear ST-Gumbel [31bba998]	0.0976	0.9024	0.8390	0.8381
8M frozen linear ST-argmax [fbc46ff2]	0.0830	0.9170	0.8573	0.8525
8M frozen linear softmax [1cf7945c]	0.0857	0.9143	0.8697	0.8638
8M unfrozen MLP4 ST-Gumbel [d50ef818]	0.0860	0.9140	0.8541	0.8530
8M unfrozen MLP4 ST-argmax [650078f8]	0.0655	0.9345	0.8852	0.8910
8M unfrozen MLP4 softmax [16132e1d]	0.0703	0.9297	0.8925	0.8890
8M unfrozen MLP4 softmax+mask [9d792c09]	0.0654	0.9346	0.8939	0.8922
35M frozen MLP4 ST-Gumbel [72459d15]	0.0839	0.9161	0.8366	0.8344
35M frozen MLP4 ST-argmax [9fceaf26]	0.1009	0.8991	0.8430	0.8405
35M frozen MLP4 softmax [0eb8395a]	0.0814	0.9186	0.8740	0.8761
35M frozen linear ST-Gumbel [4b1a9c9d]	0.0790	0.9210	0.8537	0.8580
35M frozen linear ST-argmax [f8f5cad4]	0.0839	0.9161	0.8639	0.8606
35M frozen linear softmax [c336a672]	0.0882	0.9118	0.8641	0.8554
35M unfrozen MLP4 ST-Gumbel [f32e1543]	0.0735	0.9265	0.8725	0.8738
35M unfrozen MLP4 ST-argmax [39efc845]	0.0557	0.9443	0.8917	0.8936
35M unfrozen MLP4 softmax [f7815ea1]	0.0519	0.9481	0.9124	0.9120
35M unfrozen MLP4 softmax+mask [929752d4]	0.0563	0.9437	0.9092	0.9127

Table S16: Weighted token differences. Valid-descriptor mask; all 21 predictors, 4,624 proteins and 856,552 scored residues per predictor. Position-wise difference uses $d(a, b) = (B_{aa} + B_{bb})/2 - B_{ab}$; micro pools residues and macro averages proteins. Normalized difference divides the micro mean by 28.5, the maximum state cost; it is not a mismatch fraction. Raw score is the mean native matrix entry. No realignment is used for these weighted quantities.

Predictor	Difference micro	Difference macro	Normalized difference	Mean raw score
ProstT5-based Foldseek	1.9782	1.9131	0.0694	3.5146
8M frozen MLP4 ST-Gumbel [87d2c509]	6.0515	5.8257	0.2123	-1.0139
8M frozen MLP4 ST-argmax [d58f9e14]	7.7762	7.5939	0.2729	-2.2190
8M frozen MLP4 softmax [7b541c60]	6.7272	6.6507	0.2360	-1.4495
8M frozen linear ST-Gumbel [31bba998]	6.4650	6.3742	0.2268	-1.4825
8M frozen linear ST-argmax [fbc46ff2]	7.0374	6.8597	0.2469	-1.6848
8M frozen linear softmax [1cf7945c]	6.3671	6.2857	0.2234	-1.1737
8M unfrozen MLP4 ST-Gumbel [d50ef818]	7.1547	7.1188	0.2510	-1.7772
8M unfrozen MLP4 ST-argmax [650078f8]	8.6548	8.4268	0.3037	-2.9603
8M unfrozen MLP4 softmax [16132e1d]	7.7045	7.6176	0.2703	-2.0689
8M unfrozen MLP4 softmax+mask [9d792c09]	7.6053	7.4795	0.2669	-1.7651
35M frozen MLP4 ST-Gumbel [72459d15]	6.9876	6.8189	0.2452	-1.9762
35M frozen MLP4 ST-argmax [9fceaf26]	7.8168	7.4369	0.2743	-2.3424
35M frozen MLP4 softmax [0eb8395a]	6.5166	6.4986	0.2287	-1.1464
35M frozen linear ST-Gumbel [4b1a9c9d]	7.1982	7.0242	0.2526	-2.0931
35M frozen linear ST-argmax [f8f5cad4]	7.6777	7.2655	0.2694	-2.0419
35M frozen linear softmax [c336a672]	6.2156	6.0761	0.2181	-1.0005
35M unfrozen MLP4 ST-Gumbel [f32e1543]	8.1444	8.0725	0.2858	-2.6026
35M unfrozen MLP4 ST-argmax [39efc845]	9.5206	9.1725	0.3341	-3.5774
35M unfrozen MLP4 softmax [f7815ea1]	9.2015	9.1312	0.3229	-3.0315
35M unfrozen MLP4 softmax+mask [929752d4]	9.6073	9.5201	0.3371	-3.1176

Table S17: Unweighted token differences. C-alpha-present mask; all 21 predictors, 4,624 proteins and 865,859 scored residues per predictor. Identity and mismatch compare corresponding residue positions. Edit is unit-cost global Levenshtein distance divided by scored length. Micro pools residues; macro averages proteins. The same reference mask is applied to both strings before scoring.

Predictor	Identity	Mismatch	Edit micro	Edit macro
ProstT5-based Foldseek	0.6150	0.3850	0.3744	0.3670
8M frozen MLP4 ST-Gumbel [87d2c509]	0.1549	0.8451	0.7906	0.7798
8M frozen MLP4 ST-argmax [d58f9e14]	0.0692	0.9308	0.8786	0.8899
8M frozen MLP4 softmax [7b541c60]	0.0925	0.9075	0.8669	0.8603
8M frozen linear ST-Gumbel [31bba998]	0.0977	0.9023	0.8383	0.8373
8M frozen linear ST-argmax [fbc46ff2]	0.0848	0.9152	0.8551	0.8496
8M frozen linear softmax [1cf7945c]	0.0910	0.9090	0.8640	0.8557
8M unfrozen MLP4 ST-Gumbel [d50ef818]	0.0883	0.9117	0.8517	0.8499
8M unfrozen MLP4 ST-argmax [650078f8]	0.0660	0.9340	0.8845	0.8899
8M unfrozen MLP4 softmax [16132e1d]	0.0730	0.9270	0.8895	0.8852
8M unfrozen MLP4 softmax+mask [9d792c09]	0.0652	0.9348	0.8936	0.8915
35M frozen MLP4 ST-Gumbel [72459d15]	0.0835	0.9165	0.8359	0.8335
35M frozen MLP4 ST-argmax [9fceaf26]	0.1000	0.9000	0.8433	0.8408
35M frozen MLP4 softmax [0eb8395a]	0.0850	0.9150	0.8703	0.8711
35M frozen linear ST-Gumbel [4b1a9c9d]	0.0803	0.9197	0.8518	0.8554
35M frozen linear ST-argmax [f8f5cad4]	0.0843	0.9157	0.8634	0.8597
35M frozen linear softmax [c336a672]	0.0931	0.9069	0.8587	0.8477
35M unfrozen MLP4 ST-Gumbel [f32e1543]	0.0759	0.9241	0.8698	0.8702
35M unfrozen MLP4 ST-argmax [39efc845]	0.0572	0.9428	0.8896	0.8907
35M unfrozen MLP4 softmax [f7815ea1]	0.0534	0.9466	0.9108	0.9100
35M unfrozen MLP4 softmax+mask [929752d4]	0.0561	0.9439	0.9095	0.9130

Table S18: Weighted token differences. C-alpha-present mask; all 21 predictors, 4,624 proteins and 865,859 scored residues per predictor. Position-wise difference uses $d(a, b) = (B_{aa} + B_{bb})/2 - B_{ab}$; micro pools residues and macro averages proteins. Normalized difference divides the micro mean by 28.5, the maximum state cost; it is not a mismatch fraction. Raw score is the mean native matrix entry. No realignment is used for these weighted quantities.

Predictor	Difference micro	Difference macro	Normalized difference	Mean raw score
ProstT5-based Foldseek	1.9709	1.9050	0.0692	3.5056
8M frozen MLP4 ST-Gumbel [87d2c509]	6.0299	5.7998	0.2116	-1.0052
8M frozen MLP4 ST-argmax [d58f9e14]	7.7510	7.5631	0.2720	-2.2033
8M frozen MLP4 softmax [7b541c60]	6.6681	6.5735	0.2340	-1.4015
8M frozen linear ST-Gumbel [31bba998]	6.4346	6.3347	0.2258	-1.4621
8M frozen linear ST-argmax [fbc46ff2]	7.0064	6.8174	0.2458	-1.6640
8M frozen linear softmax [1cf7945c]	6.3142	6.2142	0.2215	-1.1324
8M unfrozen MLP4 ST-Gumbel [d50ef818]	7.1151	7.0660	0.2497	-1.7474
8M unfrozen MLP4 ST-argmax [650078f8]	8.6093	8.3692	0.3021	-2.9272
8M unfrozen MLP4 softmax [16132e1d]	7.6548	7.5533	0.2686	-2.0293
8M unfrozen MLP4 softmax+mask [9d792c09]	7.5694	7.4323	0.2656	-1.7377
35M frozen MLP4 ST-Gumbel [72459d15]	6.9674	6.7938	0.2445	-1.9686
35M frozen MLP4 ST-argmax [9fceaf26]	7.8007	7.4232	0.2737	-2.3381
35M frozen MLP4 softmax [0eb8395a]	6.4695	6.4359	0.2270	-1.1119
35M frozen linear ST-Gumbel [4b1a9c9d]	7.1577	6.9730	0.2511	-2.0631
35M frozen linear ST-argmax [f8f5cad4]	7.6430	7.2263	0.2682	-2.0202
35M frozen linear softmax [c336a672]	6.1667	6.0099	0.2164	-0.9621
35M unfrozen MLP4 ST-Gumbel [f32e1543]	8.0936	8.0065	0.2840	-2.5620
35M unfrozen MLP4 ST-argmax [39efc845]	9.4638	9.1027	0.3321	-3.5340
35M unfrozen MLP4 softmax [f7815ea1]	9.1467	9.0605	0.3209	-2.9885
35M unfrozen MLP4 softmax+mask [929752d4]	9.5718	9.4746	0.3359	-3.0910

F Retrieval and Clustering over the Full Validation Population

Pairwise correspondence quality does not determine whether a score ranks related proteins first. We test that transfer over the full validation population: 5,059 domains, 2,414 families, 1,241 superfamilies, and 790 folds. Excluding self-comparisons leaves 25,588,422 directed candidate pairs. We evaluate all 20 learned predictors, structure-derived Foldseek, and ProstT5-based Foldseek in both scoring modes, plus amino-acid-only MMseqs2: 45 native configurations. The original TM-align corpus supplies all 12,794,211 unordered pairs; we separately rank their minimum, maximum, geometric-mean, and arithmetic-mean directional normalizations as four structural reference rankings. Exhaustive candidate scoring removes prefilter sensitivity from this comparison.

Native rankings use the emitted bit-score field. The explicit candidate database contains every ordered pair, including self-pairs during scoring, and self-hits are removed before analysis. Foldseek uses alignment mode 3, gap-open/extension 10/1, no structure-bit reranking, zero minimum identity, length, and coverage cutoffs, no alternative alignments, and composition correction scale 0.5. The permissive E-value threshold is 10^{100}, with maximum accepted and rejected candidates both set to $2^{31} - 1$; these settings preserve the requested exhaustive score population. All 25,593,481 entries are checked for presence before excluding 5,059 self-hits. Exhaustive scoring lets us compare ranking quality without variation in which candidates pass a prefilter; it does not measure the throughput of ordinary prefiltered search.

The population change is essential to this test. The alignment benchmark starts from pairs selected for structural similarity and asks which residues correspond. Retrieval places each candidate among the other validation proteins, including those outside that selected pair population. Improving the correspondence within a related pair need not increase its score relative to all competing candidates. Exhaustive scoring holds candidate availability fixed so that these comparisons concern the ranking produced by the scores.

F.1 Published retrieval sensitivity and nearest-hit classification

For retrieval, we use the Foldseek authors' benchmark scripts with their definitions and tie rules (revision `654100b11242e581f9e6d43798b07a778903862e`). Family positives share a family.

Superfamily positives share a superfamily but differ in family; fold positives share a fold but differ in superfamily. Different-fold hits are false positives. Other same-fold relationships are ignored where they do not count as positives. Sensitivity is measured before the first false positive and averaged over the 1,328 queries eligible at all three levels. We preserve the scripts' superfamily/fold denominator convention, including its added one. Score ties are ordered by target domain identifier. Class-size-weighted precision–recall curves use the published weighting and eligibility rules.

More explicitly, let n_f, n_s, n_d count the query's family, superfamily, and fold members, including the query itself. Eligibility requires $n_f > 1$, $n_s > n_f$, and $n_d > n_s$. If h_f, h_s, h_d count the three disjoint positive relationships before the first false positive, the published per-query sensitivities are

$$\frac{h_f}{n_f - 1}, \qquad \frac{h_s}{n_s - n_f + 1}, \qquad \frac{h_d}{n_d - n_s + 1}. \tag{13}$$

The added one in the latter two denominators follows the published script. These denominators therefore exceed the number of available positives by one.

The weighted precision–recall calculation follows the corresponding published definition. It globally orders nonself pairs by score, then query and target identifiers for ties. Eligible family, superfamily, and fold positives receive weights $1/(n_f - 1)$, $1/(n_s - n_f)$, and $1/(n_d - n_s)$, respectively. A different-fold false positive receives weight $1/\max\{n_d - n_s, n_f - 1, 1\}$, including for queries outside the common positive-eligibility set. Weighted recall divides accumulated positive mass by 1,328; precision divides positive mass by positive plus false-positive mass. These are published class-weighting conventions, not ordinary unweighted PR curves. Their global score ordering also differs from the within-query correlation analysis below.

These definitions progressively exclude closer relationships. A same-family hit counts for family retrieval; superfamily retrieval asks for hits outside that family, and fold retrieval asks for hits outside that superfamily. Fold sensitivity can therefore be low even when many top hits share the query's fold. Stopping at the first different-fold hit also makes this statistic sensitive to an early ranking error. Top-hit classification asks a complementary question about only the highest-scoring candidate or tie group. Both evaluate protein ranking rather than residue correspondence F1.

Top-hit classification asks whether the highest-ranked candidate shares the query's family, superfamily, or fold. It uses 3,459, 4,415, and 4,696 eligible queries, respectively, after excluding singletons at each level: 1,600, 644, and 363 queries are excluded, respectively. These denominators differ from the retrieval-sensitivity denominator. Deterministic top-hit accuracy uses the specified domain-ID tie break. Tie-averaged accuracy averages over the highest-scoring targets; class-macro accuracy gives equal weight to eligible classes. Together, these summaries show how the result depends on ties and on the unequal sizes of SCOPe classes.

The selected 35M softmax model improves family retrieval sensitivity from 0.6249 in 3Di-only mode to 0.7114 in AA+3Di, but both remain below the corresponding structure-derived values of 0.7446 and 0.8176, and ProstT5-based Foldseek values of 0.7993 and 0.8352. The 35M Gumbel model instead improves fold-level sensitivity over structure-derived tokens in both modes: 0.1004 versus 0.0830 in 3Di-only and 0.0981 versus 0.0962 in AA+3Di. It remains below ProstT5-based Foldseek's 0.1088 and 0.1208. Transfer therefore depends on the predictor and the SCOPe relationship being retrieved.

Nearest-hit predictions offer a complementary view. The 35M softmax model's AA+3Di tie-averaged family, superfamily, and fold accuracies are 0.5366, 0.7955, and 0.7844. Class-macro accuracies using the deterministic tie break are 0.4874, 0.6469, and 0.6386. This changes the weighting of queries and classes as well as the treatment of ties, over the larger classification populations rather than the 1,328-query retrieval subset. All forty-five native configurations and four TM-align rankings use these same population and eligibility rules.

Table S32 reports every score configuration at all three SCOPe levels.

The retrieval figures in Appendix O include every learned predictor in both native scoring modes, both modes of structure-based and ProstT5-based Foldseek, MMseqs2, and all four TM-align score normalizations. Each retrieval metric is shown at all three SCOPe levels.

F.2 NATIVE CLUSTERING AND ITS TWO REFERENCE PARTITIONS

Clustering tests whether the token scores also support coherent groups of proteins. We use the native clustering procedure, including its candidate selection, independently of exhaustive retrieval scoring. For Foldseek we invoke the same revision's `cluster` command, changing only the alignment type (0 for 3Di-only, 2 for AA+3Di), disabling structure-bit reranking, and setting the thread count to 24. Other clustering thresholds remain at that binary's defaults. We check that every domain occurs exactly once in the cluster memberships. Adjusted Rand index (ARI), adjusted mutual information (AMI), and normalized mutual information (NMI) compare the resulting partitions with SCOPe labels at family, superfamily, and fold levels. We also compare each native partition with each of the two structure-derived Foldseek partitions. The latter analysis measures preservation of a native clustering result. Comparing against both references distinguishes agreement with SCOPe classification from agreement with the grouping produced by structure-derived Foldseek tokens.

We additionally cluster graphs from all four TM-score normalizations at six fixed thresholds. The following comparisons include the prespecified 0.5 references; Appendix M reports every threshold, the graph protocol, and the additional partition diagnostics for all native configurations.

The configuration ordering also depends on the clustering metric. In AA+3Di mode, the 35M frozen linear and frozen four-layer heads trained with straight-through argmax reach family adjusted Rand indices of 0.4057 and 0.4021, respectively, compared with 0.4002 from structure-derived tokens and 0.3559 from ProstT5-based Foldseek. Their adjusted mutual information is lower than the structure-derived baseline (0.5336 and 0.5309 versus 0.6004). The checkpoints highlighted for pairwise alignment therefore do not summarize every aspect of the discrete-model sweep.

The highlighted 35M softmax model forms 1,226 clusters in AA+3Di mode, compared with 2,290 for structure-derived tokens and 1,925 for ProstT5-based Foldseek. Its family ARI is 0.1068, below 0.4002 and 0.3559, respectively. Against the structure-derived AA+3Di partition itself, its ARI is 0.1708; ProstT5-based Foldseek reaches 0.5365. The structure-derived 3Di-only and AA+3Di partitions also differ from one another, with ARI 0.6484. Cluster counts help explain the granularity of these partitions, but ARI and AMI remain valid comparisons when those counts differ. We report the counts together with the agreement scores rather than impose a common cluster count after evaluation.

Figure S4 compares retrieval sensitivity and cluster agreement for the complete 20-model 3Di sweep, using the same model order and reference methods at all three SCOPe hierarchy levels.

a 3Di-only

	Encoder / head / training selector	ROC1 Fam.	ROC1 SF	ROC1 Fold	ARI Fam.	ARI SF	ARI Fold	AMI Fam.	AMI SF	AMI Fold
ESM-2 8M	Frozen linear / softmax	0.273	0.183	0.017	0.195	0.169	0.099	0.231	0.238	0.211
	Frozen linear / ST argmax	0.228	0.154	0.015	0.302	0.237	0.131	0.333	0.317	0.270
	Frozen linear / ST Gumbel	0.270	0.185	0.021	0.285	0.259	0.151	0.339	0.333	0.288
	Frozen MLP-4 / softmax	0.297	0.199	0.018	0.172	0.167	0.106	0.223	0.230	0.207
	Frozen MLP-4 / ST argmax	0.241	0.171	0.020	0.263	0.226	0.134	0.315	0.321	0.286
	Frozen MLP-4 / ST Gumbel	0.305	0.211	0.026	0.321	0.257	0.154	0.374	0.359	0.313
	Trainable MLP-4 / softmax	0.370	0.253	0.036	0.211	0.241	0.188	0.323	0.336	0.310
	Trainable MLP-4 / ST argmax	0.335	0.251	0.049	0.143	0.208	0.184	0.339	0.383	0.368
	Trainable MLP-4 / ST Gumbel	0.412	0.284	0.059	0.103	0.173	0.215	0.284	0.346	0.345
	Trainable MLP-4 / softmax + masked	0.343	0.240	0.021	0.231	0.261	0.200	0.355	0.394	0.366
ESM-2 35M	Frozen linear / softmax	0.480	0.298	0.015	0.289	0.290	0.181	0.395	0.423	0.374
	Frozen linear / ST argmax	0.498	0.332	0.038	0.375	0.354	0.222	0.498	0.534	0.476
	Frozen linear / ST Gumbel	0.517	0.341	0.036	0.333	0.324	0.217	0.489	0.527	0.465
	Frozen MLP-4 / softmax	0.519	0.338	0.034	0.210	0.276	0.207	0.364	0.422	0.395
	Frozen MLP-4 / ST argmax	0.457	0.315	0.045	0.381	0.349	0.221	0.491	0.497	0.437
	Frozen MLP-4 / ST Gumbel	0.527	0.349	0.061	0.317	0.333	0.230	0.490	0.527	0.472
	Trainable MLP-4 / softmax	0.625	0.425	0.071	0.093	0.175	0.169	0.355	0.456	0.441
	Trainable MLP-4 / ST argmax	0.589	0.416	0.086	0.228	0.310	0.248	0.465	0.556	0.521
	Trainable MLP-4 / ST Gumbel	0.640	0.453	0.100	0.066	0.135	0.190	0.309	0.409	0.418
	Trainable MLP-4 / softmax + masked	0.624	0.424	0.083	0.046	0.097	0.137	0.257	0.350	0.362
	Native references									
	Structure-derived 3Di	0.745	0.466	0.083	0.404	0.415	0.279	0.578	0.653	0.601
	ProstT5-based Foldseek	0.799	0.532	0.109	0.345	0.401	0.298	0.558	0.653	0.617

b AA+3Di

	Encoder / head / training selector	ROC1 Fam.	ROC1 SF	ROC1 Fold	ARI Fam.	ARI SF	ARI Fold	AMI Fam.	AMI SF	AMI Fold
ESM-2 8M	Frozen linear / softmax	0.498	0.278	0.016	0.232	0.227	0.136	0.357	0.380	0.332
	Frozen linear / ST argmax	0.418	0.250	0.017	0.343	0.290	0.162	0.428	0.424	0.361
	Frozen linear / ST Gumbel	0.467	0.278	0.020	0.336	0.296	0.170	0.442	0.444	0.381
	Frozen MLP-4 / softmax	0.516	0.288	0.019	0.225	0.221	0.134	0.336	0.353	0.310
	Frozen MLP-4 / ST argmax	0.421	0.252	0.022	0.316	0.301	0.191	0.406	0.432	0.387
	Frozen MLP-4 / ST Gumbel	0.493	0.295	0.028	0.393	0.346	0.201	0.512	0.506	0.434
	Trainable MLP-4 / softmax	0.551	0.331	0.037	0.272	0.295	0.189	0.394	0.416	0.368
	Trainable MLP-4 / ST argmax	0.451	0.306	0.051	0.180	0.231	0.181	0.375	0.427	0.398
	Trainable MLP-4 / ST Gumbel	0.526	0.340	0.062	0.102	0.167	0.182	0.276	0.344	0.346
	Trainable MLP-4 / softmax + masked	0.491	0.296	0.024	0.255	0.289	0.201	0.418	0.473	0.432
ESM-2 35M	Frozen linear / softmax	0.642	0.382	0.022	0.338	0.372	0.237	0.497	0.556	0.490
	Frozen linear / ST argmax	0.636	0.398	0.040	0.406	0.386	0.232	0.534	0.583	0.520
	Frozen linear / ST Gumbel	0.666	0.411	0.044	0.396	0.394	0.241	0.553	0.599	0.527
	Frozen MLP-4 / softmax	0.656	0.407	0.038	0.230	0.300	0.232	0.427	0.509	0.474
	Frozen MLP-4 / ST argmax	0.612	0.385	0.050	0.402	0.368	0.228	0.531	0.551	0.481
	Frozen MLP-4 / ST Gumbel	0.679	0.420	0.066	0.319	0.357	0.263	0.535	0.590	0.525
	Trainable MLP-4 / softmax	0.711	0.476	0.073	0.107	0.192	0.191	0.371	0.469	0.454
	Trainable MLP-4 / ST argmax	0.682	0.466	0.087	0.251	0.323	0.246	0.503	0.584	0.539
	Trainable MLP-4 / ST Gumbel	0.726	0.498	0.098	0.084	0.158	0.160	0.321	0.423	0.419
	Trainable MLP-4 / softmax + masked	0.689	0.467	0.081	0.040	0.080	0.097	0.257	0.352	0.358
	Native references									
	Structure-derived 3Di	0.818	0.535	0.096	0.400	0.419	0.308	0.600	0.671	0.614
	ProstT5-based Foldseek	0.835	0.578	0.121	0.356	0.412	0.335	0.567	0.663	0.632

Shared color scale: 0.00 (white) to 1.00 (purple) · Fam. = family; SF = superfamily

Figure S4: **Native Foldseek retrieval and clustering.** All 20 learned 3Di models and two references are evaluated in (a) 3Di-only and (b) AA+3Di modes, in the same order in both panels. Each metric reports family (Fam.), superfamily (SF), and fold performance. Cells share a fixed zero-to-one color scale; numbers are rounded to three decimal places. ROC1 uses exhaustive scoring of 1,328 eligible queries from 5,059 validation domains. ARI and AMI compare native clustering with SCOPe family, superfamily and fold labels for all 5,059 domains; each partition is reused across levels. The configurations and ordering match the checkpoint inventory; exact identities and zero-based epochs are retained in the companion data. No models are selected per metric.

The complete SCOPe comparisons appear in Tables S34, S34, and S34. Tables S36 and S36 instead use the two native reference partitions.

F.3 Similarity rankings across all candidate pairs

To compare the full rankings, we compute a Spearman correlation between each query's native scores and each of the four TM-align reference similarities, using average ranks for ties. Queries with constant scores are excluded and counted separately. If a and b are the two directional TM-align normalizations, these four reference similarities are $\min(a, b)$, $\max(a, b)$, $\sqrt{ab}$, and $(a + b)/2$. Reversing the query and target swaps a and b; it does not require a second alignment solve. Correlations are computed within each query, then summarized by their mean and median across valid queries. This does not assume that native scores are calibrated across unrelated queries. Structure-derived 3Di-only excludes one constant query and retains 5,058; the exclusion count is part of the result.

The relative performance depends on the reference normalization. In AA+3Di mode, the 35M softmax predictor has mean Spearman correlation 0.2707 with geometric-mean TM similarity, compared with 0.2357 for structure-derived tokens and 0.2667 for ProstT5-based Foldseek. Against maximum directional TM similarity, its correlation is 0.1417, below 0.1957 and 0.2155. The 35M Gumbel predictor similarly reaches 0.2953 for the geometric mean and 0.1676 for the maximum. These comparisons concern ranks over the entire validation candidate population, not just the TM-filtered alignment pairs.

These rank and clustering comparisons require neither a positive-semidefinite kernel nor a metric distance. They measure the quality of the scores on this validation population. Public-database throughput and native Foldseek E-value calibration under the learned token distribution require separate evaluation.

Table S38 reports all four structural reference normalizations for every native score configuration, together with the valid-query counts.

G Reference TM-score Strata and Correspondence Geometry

To test how alignment quality changes with structural similarity, we group pairs by the maximum of their two directional reference TM-align scores. The bins are left-closed and right-open, except that the last includes 1.0: 47,756 pairs lie in $[0.6, 0.7)$, 13,934 in $[0.7, 0.8)$, 5,775 in $[0.8, 0.9)$, and 966 in $[0.9, 1.0]$. These strata are shared across methods; missing maps and undefined metrics reduce a method's available and valid counts within each bin. Appendix J plots the overall aggregate and all four bins for eight alignment metrics, with separate comparisons for sequence-input neural networks, classical sequence methods, structure-based methods, learned 3Di states, and correspondence decoders.

The lowest bin contains 69.8% of the alignment cohort, while the highest contains 1.4%. Overall means therefore emphasize the lower end of the selected TM-score range. The 35M continuous no-MLM model improves MEA full-reference Cα LDDT over the 8M model in every bin: 0.3637 versus 0.3348, 0.4906 versus 0.4570, 0.6424 versus 0.6180, and 0.7475 versus 0.7348. Their residue coverage is similar in each bin, so these gains do not follow simply from a uniformly longer predicted alignment.

The native-token comparison has a different pattern. The 35M softmax predictor exceeds structure-derived 3Di-only Foldseek in the first two bins, with Cα full-reference LDDT 0.3337 versus 0.2857 and 0.4652 versus 0.4353. It falls below in the last two: 0.6085 versus 0.6173 and 0.6960 versus 0.7433. Correspondence F1 shows the same reversal. The overall advantage is therefore concentrated in the larger, less structurally similar bins. Learned tokens have slightly higher residue coverage even in the last two bins, so broader coverage alone does not explain the reversal in geometric agreement.

Recombining the external-method bins with the appropriate denominators reproduces their overall means and pooled correspondence counts. We also assess each unchanged correspondence map after Kabsch superposition, reporting Cα RMSD and the mean of two chain-normalized TM-like scores. The length-dependent distance scale is $d_0(L) = \max\{1.24\sqrt[3]{L - 15} - 1.8, 0.5\}$ for $L > 15$ and

0.5 otherwise, with L the observed-Cα chain length. These scores use the Kabsch fit and do not optimize a new TM-align path or TM-score superposition. RMSD is undefined below three matched residues. Empty alignments receive zero TM-like score and residue coverage.

Residue coverage reports the mean matched-residue fraction across the two proteins. Contact coverage instead reports aligned-reference contacts divided by full-reference contacts, averaged across reference directions. All LDDT results here use symmetric scoring. The evaluation tables report 492 evaluation configurations, each overall and in four bins, for 2,460 rows with metric-specific valid counts. These include 32 external-tool configurations, original ProteinMPNN checkpoints, SoftAlign configurations, continuous sequence models, and discrete models. The strict common-map historical comparison has 59,860 pairs; that subset has not been reaggregated for learned-token native Foldseek, whose results use the full cohort.

H Numerical Checks and Reproducibility

H.1 Precision and consistency checks

The alignment wrapper casts pair scores to float32, bounds scores and gaps to $\pm 79\tau$, and requests float32 operator outputs, where τ is the alignment temperature. The fixed 3Di softmax, codebook mixture, and signed bilinear score also use float32. These choices distinguish the structured computation from the sequence encoder's bfloat16 autocast. The inspected evaluation records specify bfloat16 autocast with TF32 disabled. Cα geometric calculations use float32, while aggregate sums use float64. The all-atom scorer preserves double-precision reference distances for its usual-size path and uses float32 model distances. Its reference-distance path switches to float32 above 20,000 observed atoms. Precision therefore differs between sequence encoding, structured alignment, geometric scoring, and aggregation.

For the 150M with-MLM endpoint, 294,468 of 2,621,539,017 valid marginal entries are logged outside the unit interval before clamping, approximately 0.0112%. This count alone does not distinguish small floating-point excursions from larger errors. An exact reference comparison and the distribution of excursion magnitudes are needed before attributing them to a particular numerical cause. The loss clamps values using $\epsilon = \max(10^{-6}, \epsilon_{\mathrm{dtype}})$; this stabilizes logarithms but does not by itself verify the operator's accuracy.

An independent evaluation of the 35M no-MLM checkpoint after 50 epochs recovered all 68,431 MEA maps. With the shared external-benchmark scorer, Cα full/aligned and all-atom full/aligned means exactly matched the saved endpoint: 0.4184912323, 0.5677381728, 0.3767816578, and 0.5451201865, respectively at the displayed precision. Per-pair hashes and bin recombination audits accompany the aggregate tables. The independent evaluation therefore reproduces the saved endpoint for this checkpoint at the displayed precision.

Across all 492 configurations, summing the four bins reproduces available/missing-map counts and pooled true-positive, prediction, and reference counts exactly. Valid-count-weighted means reproduce all four LDDTs and Kabsch geometry to floating-point precision. Pooled F1 recomputed from these counts agrees with the saved aggregates. Checkpoint and source checksums identify the inputs to each evaluation. These checks establish numerical consistency of the aggregation. Replication across training seeds and evaluation on an external population remain separate requirements for assessing generalization.

H.2 Experimental coverage and remaining limitations

The reported evaluation covers 120 checkpoint records: 84 frozen-adapter conditions, three gaps-only conditions, five full-tuning conditions, three LoRA conditions, one random-initialization condition, twenty discrete predictors, two ProteinMPNN checkpoints, and two SoftAlign checkpoints. Each continuous checkpoint is evaluated with three correspondence decoders; each discrete predictor is evaluated in six Orihime and two native Foldseek configurations. Together with 32 external native configurations, these give 492 combinations of training configuration, checkpoint, representation, and decoder. Models sharing an architecture name can differ in these other settings and are not necessarily replicates.

The 84 frozen-adapter conditions span 14 architectures, two output widths and three backbone sizes. Each evaluation records the exact checkpoint hash, epoch, training settings, and per-metric counts. Retrieval, fidelity, and clustering use their separately defined populations. All tables and plots use these fixed evaluations, without selecting a different best epoch for each metric.

The sweep uses a single seed. Figure labels give nominal backbone sizes; Appendix I provides exact encoder, head, trainable, and inference counts for every model. Replication and evaluation on an independent population would test how these findings generalize beyond the validation cohort.

H.3 Additional diagnostics

Per-metric and per-bin eligibility counts are retained in the machine-readable results. Missing maps and undefined scores follow the exclusion rules in Appendix B.

The companion atlas displays all 49 sampled precision–recall configurations at three levels and all 42 state confusion matrices. Appendix L includes all precision–recall curves, clustering and score-rank comparisons, and token-distance diagnostics. Source data also retain macro correspondence and directional coverage metrics, per-query outcomes, and cluster assignments.

I Complete Model Architecture Specifications

This reference specifies our 118 learned checkpoints: 116 ESM-2 models and two original ProteinMPNN models. External architectures are described in their respective papers. The ESM-2 set contains 84 frozen adapters, three gaps-only controls, five full-tuning conditions, three LoRA models, one random-initialization control, and 20 3Di predictors. A single encoder can support several correspondence decoders without introducing another encoder or another set of weights. Tables S24–S26 give integer parameter counts and identify the evaluated checkpoints. The familiar 8M, 35M, and 150M names below identify backbone configurations; they are not parameter counts.

I.1 ESM-2 backbone, tokenization, and language-model head

The sequence encoder uses the ESM-2 configurations in Table S19. The same encoder weights process the two proteins independently, producing one vector per residue. There is no attention between the proteins. The backbone contains a token embedding, L transformer blocks, and a final LayerNorm. Each block has pre-attention LayerNorm; separate, biased $D \times D$ query, key, value, and output projections; an attention residual connection; pre-feed-forward LayerNorm; and a biased $D \to 4D \to D$ feed-forward network with GELU between the two linear maps and a second residual connection. LayerNorm has learned scale and offset, with $\epsilon = 10^{-5}$. Attention is bidirectional and masks padded keys. No contact-prediction head is included.

Table S19: ESM-2 layer specifications used by the evaluated sequence models. The final column is the exact number of backbone parameters, excluding any external adapter, MLM head, or gap parameters.

Configuration	Blocks L	Width D	Heads	FFN width	Backbone parameters
`t6_8M_UR50D`	6	320	20	1,280	7,407,680
`t12_35M_UR50D`	12	480	20	1,920	33,267,360
`t30_150M_UR50D`	30	640	20	2,560	147,725,440

The implementation remaps pretrained ESM-2 weights to a 29-token vocabulary. Indices 0–19 encode `ARNDCQEGHILKMFPSTWYV`; indices 20–24 encode `BUZOX`; and indices 25–28 are padding, beginning-of-sequence, end-of-sequence, and mask tokens. A beginning token precedes each sequence and an end token follows its valid residues. They participate in backbone attention but are removed from the returned residue embeddings. Padded residue outputs are zero. The embedding table is therefore $29 \times D$, rather than the released model's 33-row table.

Rotary position embeddings act on the entire query/key head dimension, respectively 16, 24, and 32 channels. Their base is 10,000, with frequencies $10{,}000^{-2j/d_h}$ for head width d_h; they use

token positions including the beginning token. They introduce fixed frequency buffers, not learned position embeddings. Mask-token embeddings are zeroed, and embedding magnitudes are rescaled by $0.88/(1-f_{\text{mask}})$, where f_{mask} is the observed mask-token fraction among valid tokens including the two special tokens. This embedding calibration also applies to clean inputs.

Fully tuned backbones use dropout 0.1 on attention weights, on the attention output before its residual addition, and after the feed-forward GELU. Frozen backbones run in evaluation mode with dropout disabled. LoRA backbones disable these ordinary dropout operations and apply dropout only inside their low-rank branches. All dropout is disabled during evaluation. These operations are separate from the training input-corruption rules in Appendix D.

The optional amino-acid MLM head is a biased $D \to D$ linear map, GELU, LayerNorm, and a projection to 29 logits using the encoder's embedding table transposed, plus a learned 29-vector bias. The embedding table is shared, so it is counted once. The head adds exactly $D^2 + 3D + 29$ parameters: 103,389, 231,869, and 411,549 for the three widths. It is needed for the auxiliary MLM objective, but not for residue embedding or alignment inference. The backbone alone has

$$N_{\text{ESM}} = L(12D^2 + 13D) + 31D, \tag{14}$$

which includes all linear biases and LayerNorm parameters.

I.2 Continuous frozen-backbone adapters

Every continuous adapter preserves sequence length. Its output width O is 64 or 128, and the resulting residue vectors are compared by their unscaled dot product. Table S20 gives every intermediate channel width. Here A, B, C are taper widths, R is the concatenated width of a one-layer bidirectional recurrent adapter, and M is the intermediate Projectformer width. The fourteen architectures apply to every backbone/output-width combination in this table.

Table S20: Complete channel-width specification for the 84 primary frozen adapters. For 64D outputs, $R = B$; the 8M/128D one-layer recurrent adapter uses $R = 192$ rather than $B = 224$.

Backbone	D	O	A	B	C	R	M
8M	320	64	256	192	128	192	144
35M	480	64	384	256	128	256	176
150M	640	64	512	320	192	320	200
8M	320	128	256	224	192	192	200
35M	480	128	384	256	192	256	248
150M	640	128	512	320	192	320	288

Linear and MLP adapters. Linear is a single bias-free $D \to O$ projection. MLP-2, MLP-3, and MLP-4 use the respective width sequences $D \to C \to O$, $D \to B \to C \to O$, and $D \to A \to B \to C \to O$. Every hidden stage $a \to b$ first applies a biased linear map to $2b$ channels and then splits its result into gate and value vectors:

$$[u, v] = Wx + b_0, \qquad \text{SwiGLU}_{a\to b}(x) = c\ \text{SiLU}(u) \odot v, \qquad c = (0.3557755198)^{-1/2}. \tag{15}$$

The last stage is a bias-free linear projection to O. There are no skip connections, intermediate normalizations, or hidden-stage dropout in these adapters. Thus the depth label counts the hidden gated stages plus the output projection. The hidden weights initialize from $\mathcal{N}(0, 1/a)$ and their biases initialize to zero. The final projection, including Linear, initializes with orthogonal rows scaled by $\sqrt{D/O}$. A hidden stage has $2b(a+1)$ parameters, and a final $a \to O$ map has aO.

Sequence CNN adapters. CNN-k, for $k \in \{3, 5, 7, 9\}$, uses two gated one-dimensional convolutions followed by a pointwise output convolution, with channel sequence $D \to B \to C \to O$. Each hidden convolution emits twice its stated output width and applies the gate/value product in Equation (15). Both hidden kernels have width k, stride one, dilation one, and zero padding $(k-1)/2$; the final kernel has width one and no bias. The receptive fields added above the contextual backbone are therefore 5, 9, 13, and 17 residues. Padding is masked before the first convolution and after every

stage. There is no residual path or hidden normalization. Hidden convolution weights initialize with variance $1/(ak)$ and zero bias; the final projection follows the MLP output initialization. A gated convolution $a \to b$ contributes $2b(ak+1)$ parameters.

Bidirectional recurrent adapters. BiGRU-1 and BiLSTM-1 contain one recurrent layer with $R/2$ hidden channels in each direction, concatenate the directions to width R, and apply a bias-free $R \to O$ output projection. BiGRU-2 and BiLSTM-2 stack two such layers with per-direction widths $A/2$ and $C/2$, yielding $D \to A \to C \to O$. GRUs use the standard reset, update, and candidate gates; LSTMs use input, forget, candidate, and output gates. Reset, update, input, forget, and output gates use sigmoid; candidate states use $\tanh$, and the LSTM cell-to-hidden readout also uses $\tanh$. Each direction owns separate input and recurrent weight matrices and both bias vectors. Valid residues are packed, excluding padding from both recurrent directions; internally masked positions are compacted and restored after processing. Recurrent states start at zero independently for each protein. There are no skip connections, inter-layer normalizations, or recurrent dropout. The final output map has the same initialization as the MLP output. A bidirectional layer with input width a and per-direction hidden width h has $2g(ah + h^2 + 2h)$ parameters, where $g = 3$ for GRU and $g = 4$ for LSTM.

Dropout and normalization. All primary adapters apply dropout 0.1 once to their final residue vectors. The Linear, MLP, CNN, and recurrent outputs have no additional normalization; in particular, the shared configuration field naming a final RMS norm does not add one to these four families. The hidden-RMS option is disabled in all evaluated primary configurations. Projectformer has the internal and final normalization described next. Consequently, adapter comparisons vary processing, capacity, and normalization together.

I.3 Projectformer projection and refinement

Projectformer-2 contains a width-changing projection block $D \to O$ followed by a width-preserving transformer refinement at width O. Projectformer-4 contains projection $D \to M$, refinement at M, projection $M \to O$, and refinement at O. The middle width is $M = 8\,\mathrm{round}(\sqrt{DO}/8)$, as enumerated in Table S20. Every block preserves residue positions and attends globally within one protein. There is no sequence pooling, strided downsampling, or learned target-token bank in these evaluated configurations.

A projection block maps query, key, value, and residual streams separately from width a to width b. Each of these four channel maps is a dense skip projection plus a gated SwiGLU correction, with correction hidden width $h = 2\min(a, b)$:

$$\mathcal{C}_{a\to b}(x) = W_s x + \tanh(\gamma) \odot W_2\big[\mathrm{SiLU}(W_g x) \odot W_v x\big], \tag{16}$$

where γ is a learned b-vector initialized to 0.25. These channel maps have no linear biases and no variance multiplier c from Equation (15). Query input, source input, and residual input have separate learned RMS norms before their channel maps. Queries and keys then receive learned RMS normalization within each attention head, followed by partial RoPE. The projected residual supplies the skip path across the width change.

A refinement block uses one shared input RMS norm, ordinary bias-free linear query/key/value projections, and an identity residual path. In both block types, attention context is multiplied elementwise by a learned sigmoid gate obtained from a bias-free projection of the normalized query input. A bias-free output projection and RMS norm precede the attention residual addition. A second RMS-normalized branch applies a bias-free SwiGLU feed-forward network, an output RMS norm, and a residual addition. Thus the blocks use normalization both before and after each residual branch. There are no extra scalar gates on the attention or feed-forward residual branches.

All Projectformer RMS norms have $\epsilon = 10^{-6}$ and a learned scale. RoPE uses residue indices, base 10,000, and the largest even prefix no greater than one quarter of each head's channels. Table S21 specifies every realized attention and feed-forward dimension; the feed-forward hidden width is $64\lceil(8b/3)/64\rceil$. A projection block and the refinement that follows it share these output-width specifications. Attention and internal residual dropout are zero. One final RMS norm at width O precedes the common adapter-output dropout. Linear weights initialize with standard deviation 0.02,

except attention-output and feed-forward-output matrices, whose standard deviation is $0.02/\sqrt{2n}$ for a stack of n blocks.

Table S21: Every attention width used in the Projectformer adapters. The rotary dimension is per head. Intermediate widths occur in four-block adapters; 64 and 128 are the final widths.

Block width	Heads	Channels/head	Rotary channels/head	FFN hidden width
64	2	32	8	192
128	4	32	8	384
144	4	36	8	384
176	4	44	10	512
200	5	40	10	576
248	4	62	14	704
288	9	32	8	768

I.4 GAPS-ONLY CONTROLS AND BACKBONE ADAPTATION

Table S22: Recipe coverage of all 69 earlier 8M/64D adapter checkpoints. Each uses raw-dot scoring and the affine SW operator. These are checkpoint counts, not replicate counts.

Output dropout	Alignment token noise	Positive-weight cap	Checkpoints
0	0	100	12
0	0.05	100	14
0	0.05	None	20
0.1	0	None	6
0.1	0.05	None	17

The seven gaps-only checkpoints apply no adapter and retain widths 320, 480, or 640. They freeze the encoder and learn only the two scalar affine-gap scores. Three are the primary size controls; four are earlier 8M controls with the recorded corruption/loss settings. The affine gaps are used directly as learned scores, without an additional neural network for gap prediction. Alignment temperature is scheduled or fixed and is not a trainable parameter in these evaluated conditions.

Full tuning updates the backbone, both affine gaps when the chosen operator uses them, and the MLM head when present. Five primary full-tuning configurations comprise all three backbone sizes with MLM and 8M/35M without MLM. Three further 8M conditions change raw-dot scoring to cosine scoring, replace affine SW by row softmax, or use the recorded affine SV operator. These changes do not add an encoder projection. The random-initialization control uses the same 8M backbone and MLM-head architecture initialized from random weights. Its architecture and parameter count therefore match the pretrained 8M full-tuning condition with MLM.

LoRA replaces all six projection types in every backbone block: query, key, value, attention output, and the two feed-forward linear maps. For a base matrix $W_0 \in \mathbb{R}^{b\times a}$, the branch computes

$$y = W_0 x + b_0 + 2BA\,\mathrm{Dropout}_{0.1}(x), \qquad A \in \mathbb{R}^{8\times a}, \quad B \in \mathbb{R}^{b\times 8}. \tag{17}$$

The base matrix and bias stay frozen. A uses Kaiming-uniform initialization and B starts at zero. Rank is eight and the multiplier is $\alpha/r = 16/8 = 2$. All backbone normalization and embedding parameters, and the MLM head, remain frozen. Gradients through them still reach the low-rank matrices. There is no external output adapter. The low-rank matrices contribute exactly $18rDL$ parameters; the affine gaps contribute two more. For inference, the low-rank update can be merged into the base matrix after disabling dropout, removing the separate branch tensors from a merged inference representation without changing its function.

I.5 DISCRETE 3DI PREDICTORS AND FIXED SCORING BUFFERS

The 20 discrete models use either a biased $D \to 20$ linear head or an MLP-4 head with three gated hidden stages and a biased 20-logit output. The exact MLP ladders are $320 \to 256 \to 128 \to 64 \to$

20 at 8M and $480 \to 384 \to 192 \to 64 \to 20$ at 35M. Hidden stages use Equation (15); there are no residual connections or hidden/output norms. Unlike the continuous adapter output, the final logit weights initialize orthogonally at unit scale, with zero bias. Dropout 0.1 acts on logits before state selection. The linear heads have 6,420 and 9,620 parameters; the MLP-4 heads have 247,956 and 543,252 parameters, respectively.

At each backbone size, six predictors freeze an alignment-tuned backbone and cross the two head architectures with softmax, deterministic straight-through argmax, and Gumbel straight-through selection. Three more update the backbone and MLP-4 jointly under the same three selectors. The tenth updates the backbone and MLP-4 with softmax selection and auxiliary masked 3Di supervision. Frozen heads use dropout-free backbones; unfrozen heads use backbone dropout 0.1. In the masked-3Di condition the alignment and auxiliary passes share the same 20-logit head, so no separate MLM decoder is added. Appendix E gives the selector equations, masking, and gaps.

Let the fixed Foldseek matrix have full signed eigendecomposition $B = U\Lambda U^\top$. The head stores $E = U|\Lambda|^{1/2}$ as a 20×20 buffer. A selected distribution or one-hot vector p becomes $z = p^\top E$. The scorer stores the 20 entries of $s = \operatorname{sign}(\Lambda)$ and returns $z_i^\top \operatorname{diag}(s) z_j$, exactly $p_i^\top B p_j$ within the stored factorization tolerance. All 20 modes are retained. These 420 fixed scalars are buffers, not learned codebook or metric parameters. State selection and codebook multiplication use float32; straight-through evaluation uses deterministic argmax, and the diagnostic soft mixture uses softmax temperature one without Gumbel noise. Hard token output needs only the state index; applying native Foldseek to those indices introduces no additional learned prediction head.

I.6 Original ProteinMPNN structural encoders

The original MPNN-64 and MPNN-128 checkpoints use a ProteinMPNN-style structural encoder with three message-passing layers, 64 nearest Cα neighbors (or sequence length when shorter), and node/edge/hidden widths $H = 64$ or 128. They consume N, Cα, C, and O coordinates, residue indices, and chain labels. A virtual Cβ is computed from the backbone:

$$b = C_\alpha - N, \quad c = C - C_\alpha, \qquad C_\beta = C_\alpha - 0.58273431(b \times c) + 0.56802827b - 0.54067466c. \tag{18}$$

For every graph edge, all 25 ordered distance types among these five atoms receive 16 Gaussian radial basis functions, with centers evenly spaced from 2 to 22 Å and width $20/16$ Å. Relative residue offsets clipped to $[-32, 32]$, together with a separate inter-chain category, form a 66-class one-hot vector mapped by a bias-free linear layer to 16 channels. Concatenating the 400 distance channels and 16 relative-position channels gives a 416-vector, projected without bias to H, followed by LayerNorm and a biased $H \to H$ edge projection. Node states start at zero; there is no amino-acid embedding or sequence-prediction decoder in these alignment encoders.

Each message-passing layer applies a biased $3H \to H \to H \to H$ MLP with GELU after its first two maps to concatenated source-node, edge, and neighbor-node states. Masked neighbor messages are summed and divided by 30, then added through dropout to the node state and LayerNorm. A biased $H \to 4H \to H$ GELU feed-forward branch supplies a second dropout/residual/LayerNorm update. The updated nodes feed another independently parameterized $3H \to H \to H \to H$ message MLP, whose output updates each edge through dropout, a residual addition, and LayerNorm. LayerNorm has scale and offset. The original MPNN configurations record dropout 0.1; evaluation disables dropout and coordinate augmentation. The final node states, without a separate output projection, are the alignment vectors. The encoder contributes

$$N_{\mathrm{MPNN}} = 55H^2 + 470H + 1{,}056, \tag{19}$$

or 256,416 and 962,336 parameters. Their two restored gap scalars give total alignment-model counts 256,418 and 962,338. The recovered evaluation configurations freeze the restored gaps.

In both original ProteinMPNN encoders, the last layer constructs an edge update after its final node update. Alignment returns only the node states, so that last edge update cannot affect the output. Its three linear maps and LayerNorm remain registered in the checkpoints, contributing 20,800 parameters at width 64 and 82,560 at width 128. They are included in stored-model counts and can be omitted from an encoder used only to produce the returned alignment embeddings.

Parameter accounting. Counts distinguish the encoder, adapter or categorical head, auxiliary MLM head, low-rank updates, and alignment scalars. A tensor shared between both

proteins, or tied between input embeddings and MLM readout, contributes once. Fixed frequency/codebook/signature buffers do not contribute to parameter totals. Trainable counts follow the training configuration; inference counts exclude auxiliary-only heads and distinguish stored tensors from a merged LoRA representation. All integer counts and checkpoint identities are given in Tables S24–S26.

I.7 External neural network parameter counts

Table S23 counts the unique learned scalar weights used to produce residue correspondences. A shared encoder is counted once per model, regardless of the number of proteins encoded. Tied tensor aliases are counted once; fixed positional tables and weights unused by the evaluated route are listed separately below. The counts come from the evaluated PyTorch checkpoint tensor shapes, the Foldseek GGUF tensor directory, and the DEDAL TensorFlow checkpoint index, rather than nominal model-size labels. All external weights remain frozen in our benchmark.

Table S23: Exact parameter counts for external sequence-input neural network baselines. Each route uses its native correspondence decoder. The ProstT5-based Foldseek count applies to both native Foldseek 3Di-only and AA+3Di scoring. Its task head predicts 3Di states; the DeepBLAST and DEDAL heads parameterize alignment scores. DEDAL's fixed positional table and three-scalar homology head are specified in the text.

Evaluated model	Encoder	Task head	Active learned total
PEbA	1,208,141,824	0	1,208,141,824
PLMAlign global	1,208,141,824	0	1,208,141,824
pLM-BLAST global	1,208,141,824	0	1,208,141,824
DeepBLAST v3	1,208,141,824	35,657,728	1,243,799,552
ProstT5 $\rightarrow$ Foldseek	1,208,164,352	233,908	1,208,398,260
DEDAL v3	42,627,072	1,769,475	44,396,547

Architectural descriptions are given in the original publications (Iovino & Ye, 2024; Liu et al., 2024; Kaminski et al., 2023; Hamamsy et al., 2024; Heinzinger et al., 2024; Llinares-López et al., 2023). The two released structure-input SoftAlign models follow Trinquier et al. (2025); their checkpoint counts appear in Table S26.

For parameter accounting, the evaluated ProstT5-based Foldseek route uses its encoder and CNN, without a Transformer decoder. Its GGUF additionally stores 153,600 output weights unused by this route; these are excluded from the active total. DEDAL's three-scalar homology head does not affect residue correspondences, and its fixed positional table is not learned. Extending that table from 786,432 to 1,572,864 entries changes storage but adds no learned parameters. Evaluated checkpoint identifiers, tensor shapes, and source revisions accompany the parameter inventory.

I.8 EXACT COMPONENT AND CHECKPOINT PARAMETER COUNTS

Table S24: Exact parameter counts for the implemented ESM-2 backbones and optional components. Family names identify architectures; every count is an integer. The MLM column excludes its output embedding matrix, which is tied to the backbone embedding. LoRA lists the added rank-8 matrices across all layers. Two affine-gap scalars are additional when the operator uses gaps.

Backbone	Layers	Width	Backbone parameters	MLM head	LoRA
ESM-2 8M	6	320	7,407,680	103,389	276,480
ESM-2 35M	12	480	33,267,360	231,869	829,440
ESM-2 150M	30	640	147,725,440	411,549	2,764,800

Table S25: Exact learned adapter parameters for all 84 primary architecture settings. Column pairs give output width 64 and 128 for each backbone. These counts exclude the frozen backbone and two learned gap scalars. Thus the trainable count is the tabulated value plus 2, and the total and alignment-inference counts additionally include the backbone in Table S24. Each selected checkpoint is also counted separately in Table S26.

Adapter	ESM-2 8M		ESM-2 35M		ESM-2 150M	
	64	128	64	128	64	128
Linear	20,480	40,960	30,720	61,440	40,960	81,920
MLP-2	90,368	147,840	131,328	209,280	258,432	270,720
MLP-3	180,864	254,784	320,256	369,536	545,792	558,080
MLP-4	320,640	390,464	640,512	689,792	1,120,256	1,132,544
CNN-3	524,928	713,536	942,848	1,057,664	1,610,752	1,623,040
CNN-5	868,992	1,172,288	1,565,440	1,745,792	2,675,712	2,688,000
CNN-7	1,213,056	1,631,040	2,188,032	2,433,920	3,740,672	3,752,960
CNN-9	1,557,120	2,089,792	2,810,624	3,122,048	4,805,632	4,817,920
BiGRU-1	253,056	265,344	484,864	501,248	790,400	810,880
BiGRU-2	477,440	574,080	957,440	1,078,656	1,742,976	1,755,264
BiLSTM-1	333,312	345,600	641,024	657,408	1,047,040	1,067,520
BiLSTM-2	633,856	757,248	1,273,856	1,430,016	2,319,872	2,332,160
Projectformer-2	563,008	1,387,072	778,528	1,817,632	994,048	2,248,192
Projectformer-4	1,918,384	3,662,672	3,121,008	5,606,184	4,344,848	7,665,216

Table S26: Exact counts for the 116 ESM-2 checkpoints, two ProteinMPNN references and two SoftAlign baselines. IDs are the unique 16-character suffixes of the checkpoint IDs used in the results tables; epochs are zero-based. Head excludes LoRA; MLM is the auxiliary amino-acid head. Trainable counts registered trainable scalars in the training model, and Total counts all registered parameters, including frozen weights and unused branches. Inference counts parameters required for alignment after removing MLM and unused final edge updates, and merging LoRA matrices into the backbone. Buffers are excluded throughout. SoftAlign epochs are not supplied. Its geometric-mean route also omits the two unused stored gap scalars from Inference. The component and tensor inventories retain full IDs and checkpoint SHA-256 values.

Checkpoint ID	Epoch	Head	MLM	Trainable	Total	Inference
Frozen backbone, learned gaps / ESM-2 8M						
`5534128071222270`	49	0	0	2	7,407,682	7,407,682
Frozen backbone, learned gaps / ESM-2 35M						
`0f1579403041e38b`	49	0	0	2	33,267,362	33,267,362
Frozen backbone, learned gaps / ESM-2 150M						
`4dcb4aa786ff5a2d`	49	0	0	2	147,725,442	147,725,442

Continued on next page

Table S26 (continued)

Checkpoint ID	Epoch	Head	MLM	Trainable	Total	Inference
Full tuning / ESM-2 8M						
1df1da235675f423	49	0	0	7,407,682	7,407,682	7,407,682
b50fb493903cf320	49	0	103,389	7,511,071	7,511,071	7,407,682
Full tuning / ESM-2 35M						
227888e0f7fbcb88	49	0	231,869	33,499,231	33,499,231	33,267,362
60a7d7149278fe7b	49	0	0	33,267,362	33,267,362	33,267,362
Full tuning / ESM-2 150M						
6d8f48f100728588	49	0	411,549	148,136,991	148,136,991	147,725,442
LoRA / ESM-2 8M						
ce54842011273f46	49	0	103,389	276,482	7,787,551	7,407,682
LoRA / ESM-2 35M						
fe715bb95995efb6	49	0	231,869	829,442	34,328,671	33,267,362
LoRA / ESM-2 150M						
8286d8e69976e4d4	49	0	411,549	2,764,802	150,901,791	147,725,442
Random initialization / ESM-2 8M						
3f0d804b435d65ae	49	0	103,389	7,511,071	7,511,071	7,407,682
Frozen adapters / ESM-2 8M						
009c0c0dc203b430	49	20,480	0	20,482	7,428,162	7,428,162
106f56ad4e16faca	49	2,089,792	0	2,089,794	9,497,474	9,497,474
23f604efabdee92f	49	3,662,672	0	3,662,674	11,070,354	11,070,354
2cc01917762e1c9a	49	1,387,072	0	1,387,074	8,794,754	8,794,754
2dfaaf70dd52dd69	49	868,992	0	868,994	8,276,674	8,276,674
2f54a965824ccff8	49	1,213,056	0	1,213,058	8,620,738	8,620,738
30e9ea2c5c762f69	49	320,640	0	320,642	7,728,322	7,728,322
3246609d9e6edc2a	49	524,928	0	524,930	7,932,610	7,932,610
39880cee70dc0520	49	1,918,384	0	1,918,386	9,326,066	9,326,066
53e3a4d6bd73acd5	49	147,840	0	147,842	7,555,522	7,555,522
55e8bc490ca12fc7	49	477,440	0	477,442	7,885,122	7,885,122
775f0ab7ea28b452	49	180,864	0	180,866	7,588,546	7,588,546
973d696756a38fec	49	253,056	0	253,058	7,660,738	7,660,738
97c97051b5ae1da1	49	40,960	0	40,962	7,448,642	7,448,642
a432124b9d50e6e2	49	1,172,288	0	1,172,290	8,579,970	8,579,970
a68449555ce20326	49	345,600	0	345,602	7,753,282	7,753,282
b66739ba293fb2db	49	254,784	0	254,786	7,662,466	7,662,466
c1d649115ac00304	49	1,557,120	0	1,557,122	8,964,802	8,964,802
c5c9ab1dfd8eabfd	49	633,856	0	633,858	8,041,538	8,041,538
c9e2b0f1a882d927	49	713,536	0	713,538	8,121,218	8,121,218
d022f3d85ac3e73e	49	757,248	0	757,250	8,164,930	8,164,930
e166c9c0977f5e2e	49	390,464	0	390,466	7,798,146	7,798,146
e4e79360cfb98e58	49	563,008	0	563,010	7,970,690	7,970,690
e7f761341fd522dd	49	574,080	0	574,082	7,981,762	7,981,762
ebb61e72ddac38b6	49	1,631,040	0	1,631,042	9,038,722	9,038,722
efc5e46565456b5e	49	265,344	0	265,346	7,673,026	7,673,026
f55ab5ffbee2c633	49	90,368	0	90,370	7,498,050	7,498,050
fce0a74517569f0a	49	333,312	0	333,314	7,740,994	7,740,994
Frozen adapters / ESM-2 35M						
106520c70743300e	49	131,328	0	131,330	33,398,690	33,398,690
2581201e4ee8a15f	49	641,024	0	641,026	33,908,386	33,908,386
2894ea8cecb8903e	49	501,248	0	501,250	33,768,610	33,768,610
424b3045867bd048	49	3,121,008	0	3,121,010	36,388,370	36,388,370
46953354c0b23985	49	1,745,792	0	1,745,794	35,013,154	35,013,154
47e719989afb0060	49	1,273,856	0	1,273,858	34,541,218	34,541,218
4ad2c14a6f383993	49	1,817,632	0	1,817,634	35,084,994	35,084,994
4ad7a3df3eac231a	49	778,528	0	778,530	34,045,890	34,045,890
5f1f15d21cd75735	49	369,536	0	369,538	33,636,898	33,636,898

Continued on next page

Table S26 (continued)

Checkpoint ID	Epoch	Head	MLM	Trainable	Total	Inference
5fb9ea6e3906f03e	49	1,430,016	0	1,430,018	34,697,378	34,697,378
65078243a4b511ac	49	957,440	0	957,442	34,224,802	34,224,802
671c640bd9f0e5f5	49	1,057,664	0	1,057,666	34,325,026	34,325,026
686205a63ad9fb74	49	484,864	0	484,866	33,752,226	33,752,226
71c2433548c6cdd9	49	5,606,184	0	5,606,186	38,873,546	38,873,546
738f1be4b136b6d2	49	320,256	0	320,258	33,587,618	33,587,618
79f6463f0469b803	49	1,078,656	0	1,078,658	34,346,018	34,346,018
7c8cccd6cee62e7a	49	1,565,440	0	1,565,442	34,832,802	34,832,802
7e49001e02433d0d	49	2,810,624	0	2,810,626	36,077,986	36,077,986
9870ded032e58231	49	3,122,048	0	3,122,050	36,389,410	36,389,410
9a4f933e6626171f	49	61,440	0	61,442	33,328,802	33,328,802
a7ea17ab08980ed6	49	2,433,920	0	2,433,922	35,701,282	35,701,282
ba8e1b4b9208eb3e	49	209,280	0	209,282	33,476,642	33,476,642
bb548b308438f4c1	49	640,512	0	640,514	33,907,874	33,907,874
d8219a8bfcb59fc4	49	657,408	0	657,410	33,924,770	33,924,770
ee28a52d6eb6a42a	49	30,720	0	30,722	33,298,082	33,298,082
f0d0c784160bc752	49	942,848	0	942,850	34,210,210	34,210,210
ffb9320456e70f6b	49	2,188,032	0	2,188,034	35,455,394	35,455,394
ffce05aae38bebdc	49	689,792	0	689,794	33,957,154	33,957,154
Frozen adapters / ESM-2 150M						
00fd5ae7fbe4f32b	49	1,120,256	0	1,120,258	148,845,698	148,845,698
11c41dc4dd577721	49	1,755,264	0	1,755,266	149,480,706	149,480,706
1be0ebf0391f6bf3	49	3,740,672	0	3,740,674	151,466,114	151,466,114
2064580783edcc2e	49	994,048	0	994,050	148,719,490	148,719,490
209189ebaa56efbb	49	4,805,632	0	4,805,634	152,531,074	152,531,074
33ec2d0614450876	49	4,344,848	0	4,344,850	152,070,290	152,070,290
3a4e6e6b3331453e	49	2,675,712	0	2,675,714	150,401,154	150,401,154
43c9b9d78b0c8276	49	558,080	0	558,082	148,283,522	148,283,522
5cbd7c6539c8183a	49	7,665,216	0	7,665,218	155,390,658	155,390,658
615c7a8f2f83cfe4	49	81,920	0	81,922	147,807,362	147,807,362
67340e5e5a61434d	49	1,742,976	0	1,742,978	149,468,418	149,468,418
728cb738cb67a288	49	4,817,920	0	4,817,922	152,543,362	152,543,362
731e638a0635dcf3	49	2,332,160	0	2,332,162	150,057,602	150,057,602
74d42584b8ac53ce	49	270,720	0	270,722	147,996,162	147,996,162
77976811fcf3f7a3	49	790,400	0	790,402	148,515,842	148,515,842
81500da22d6a17e6	49	810,880	0	810,882	148,536,322	148,536,322
861338c93d693811	49	1,047,040	0	1,047,042	148,772,482	148,772,482
8fb7a527008ef6b8	49	1,610,752	0	1,610,754	149,336,194	149,336,194
99a539bfe4491234	49	40,960	0	40,962	147,766,402	147,766,402
9fc7c1c013248f80	49	1,132,544	0	1,132,546	148,857,986	148,857,986
a29628f4ef9f6864	49	1,067,520	0	1,067,522	148,792,962	148,792,962
c13ee7fa9c44568f	49	1,623,040	0	1,623,042	149,348,482	149,348,482
c7dda068b5202b86	49	2,248,192	0	2,248,194	149,973,634	149,973,634
d532ad6a421ceedb	49	545,792	0	545,794	148,271,234	148,271,234
ed45e08e18ceefba	49	258,432	0	258,434	147,983,874	147,983,874
f249037287a0cb31	49	2,688,000	0	2,688,002	150,413,442	150,413,442
f4579f3e9162d15d	49	2,319,872	0	2,319,874	150,045,314	150,045,314
fa0f7bab1d246182	49	3,752,960	0	3,752,962	151,478,402	151,478,402
Discrete structural-token predictors / ESM-2 8M						
04eb067e7b541c60	9	247,956	0	247,958	7,655,638	7,655,638
1908ddfc16132e1d	38	247,956	0	7,655,638	7,655,638	7,655,638
3756578fd58f9e14	9	247,956	0	247,958	7,655,638	7,655,638
394e75eb650078f8	16	247,956	0	7,655,638	7,655,638	7,655,638
44e38631d50ef818	15	247,956	0	7,655,638	7,655,638	7,655,638
466e23429d792c09	27	247,956	0	7,655,638	7,655,638	7,655,638

Continued on next page

Table S26 (continued)

Checkpoint ID	Epoch	Head	MLM	Trainable	Total	Inference
483214ea31bba998	9	6,420	0	6,422	7,414,102	7,414,102
61977e2987d2c509	9	247,956	0	247,958	7,655,638	7,655,638
842549c71cf7945c	9	6,420	0	6,422	7,414,102	7,414,102
cd987876fbc46ff2	9	6,420	0	6,422	7,414,102	7,414,102
Discrete structural-token predictors / ESM-2 35M						
04a03c3672459d15	9	543,252	0	543,254	33,810,614	33,810,614
04a05661929752d4	17	543,252	0	33,810,614	33,810,614	33,810,614
0f471cc1f8f5cad4	9	9,620	0	9,622	33,276,982	33,276,982
147f87299fceaf26	9	543,252	0	543,254	33,810,614	33,810,614
43d6384c39efc845	21	543,252	0	33,810,614	33,810,614	33,810,614
805760c0c336a672	9	9,620	0	9,622	33,276,982	33,276,982
bee5597af7815ea1	28	543,252	0	33,810,614	33,810,614	33,810,614
c49115d7f32e1543	21	543,252	0	33,810,614	33,810,614	33,810,614
d97278c30eb8395a	9	543,252	0	543,254	33,810,614	33,810,614
f5520efd4b1a9c9d	9	9,620	0	9,622	33,276,982	33,276,982
Original ProteinMPNN (wBCE cap 100)						
83f865d90731829b	49	0	0	962,338	962,338	879,778
e9c9454d16ce6ed6	49	0	0	256,418	256,418	235,618
SoftAlign						
b8b6263b9e794c93	–	0	0	256,434	256,434	235,632
1e5793985ad6a335	–	0	0	256,434	256,434	235,634

J OVERALL AND TM-SCORE-STRATIFIED ALIGNMENT QUALITY

The figure series in Appendix O covers every learned checkpoint, representation, correspondence decoder, and external-tool configuration. Each route appears overall and in all four reference TM-score intervals. Colors and line styles identify model configurations within each comparison. Each page uses a fixed named correspondence decoder, with full-size overlaid curves on eight metric axes and all five points per metric. Native reference tools retain their native correspondence maps. These include all frozen adapters, gaps-only controls, full fine-tuning with and without MLM, LoRA, random initialization, both original ProteinMPNN widths, both SoftAlign checkpoints, and every soft, hard and native Foldseek route of the learned 3Di models.

Figures S62–S65 show the external aligners under the same TM-score stratification. These include TM-align, USalign, DALI, MUSTANG, SPalign, SPalign-NS, GANGSTA+, MUSCLE3D, Reseek, and structure-based Foldseek; the sequence-input neural network aligners, including ProstT5-based Foldseek; and classical sequence aligners, including MMseqs2 and affine Smith–Waterman substitution matrices. Both 3Di-only and AA+3Di modes are shown for each Foldseek input. The 35M ESM-2 MEA result provides a common learned reference on these pages.

Each curve uses the same correspondence map for all eight endpoints: full-reference and aligned-only Cα and all-atom LDDT, pooled F1, residue coverage, Cα RMSD, and symmetric Kabsch TM-score. The last two metrics evaluate the predicted map after Kabsch fitting; the horizontal axis instead uses the reference maximum directional TM-score. The overall point is read directly from the saved aggregate and separated from the four interval points. The first three intervals exclude their upper endpoints; the last includes 1.0. Undefined geometric values retain their metric-specific exclusion policy. Scores use a 0–1 scale except RMSD, in Å.

K COMPLETE AGGREGATE QUALITY AND DISCRETE ALIGNMENT PATHWAYS

The aggregate matrices report every alignment route on the full validation cohort. Each learned representation has three columns per metric: threshold, mutual matching (MM), and MEA. All four LDDT endpoints, F1 and residue coverage therefore appear together for the same checkpoint. Native tools retain their own alignments and both Foldseek scoring modes. The model inventory is the 116-run ESM-2 sweep and the two original ProteinMPNN references. These matrices also include the complete adapter and discrete-route sweeps.

The pathway figures follow all 20 discrete predictors through four routes: soft structural-state mixtures and hard state argmax, each followed by Orihime MEA, then native Foldseek in 3Di-only and AA+3Di modes. Each row follows the same checkpoint; all four LDDT definitions and F1 appear side by side. Hard state argmax is a representation operation, not a correspondence decoder. Moving from Orihime to Foldseek changes both scoring and decoding. Reference segments retain both modes of structure-based and ProstT5-based Foldseek. Checkpoint IDs and epochs are given in Table S10.

L COMPLETE DECODER AND REPRESENTATION DIAGNOSTICS

The decoder comparisons begun in Figure S5 continue here for every learned continuous representation and each soft or hard 3Di representation. Every checkpoint is shown under threshold, mutual matching and MEA in the coverage–geometry and correspondence–geometry panels. The aggregate matrices report the four LDDT endpoints, F1 and residue coverage for those same maps.

L.1 CLASS-WEIGHTED RETRIEVAL CURVES

These curves report every evaluated ranking, including all learned predictors in both Foldseek modes, both modes of the structure-based and ProstT5-based references, MMseqs2, and all four TM-align normalizations. Every supplied sampled point is retained.

L.2 TOKEN FIDELITY UNDER BOTH REFERENCE MASKS

The token-distance figures include all predictors under both valid-descriptor and native-Cα masks. These compare predicted and structure-derived states of the same protein. Their evaluation does not use a Foldseek alignment mode. The earlier identity–alignment figures separately show both native scoring modes for each predictor.

M TM-SCORE GRAPH CLUSTERING AND PARTITION DIAGNOSTICS

We also cluster the complete validation population using graphs obtained by thresholding each of the four symmetric TM-score combinations. Thresholds 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9 were fixed before inspecting the results, with 0.5 designated as the reference. The unchanged Foldseek set-cover core (`clust -cluster-mode 0 -similarity-type 1`) operates on each graph. Edges are selected using the original floating-point scores. Native tie and reassignment scores use the nearest integer to 30,000 times the similarity. Every member has a threshold-passing edge to its representative, although members need not all be pairwise connected. The procedure uses one thread and does not apply the stock Foldseek cascade, E-value cutoff or coverage filter.

All 24 graph conditions are reported at family, superfamily and fold levels. ARI and AMI adjust partition agreement for chance. Homogeneity measures label mixing; completeness measures the fragmentation of reference classes. Protein-pair precision, recall and F1 count unordered pairs of distinct proteins assigned to the same cluster, and are separate from residue-correspondence F1. The additional diagnostic figures include all 45 native configurations and all four TM-align references at the fixed 0.5 threshold. Full threshold sweeps are shown separately.

N COMBINED RESULT TABLES

The following tables place related metrics and decoders side by side for each configuration. The preceding sections define their populations, scoring conventions and checkpoint identities.

Table S27: All 84 frozen adapters, with every structural endpoint and correspondence F1. T = threshold; MM = mutual matching. Each metric group gives all three decoders for the same checkpoint.

Configuration	Cα full LDDT			Cα aligned LDDT			All-atom full LDDT			All-atom aligned LDDT			Correspondence F1		
	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA
8M 64D BiGRU-1	0.1265	0.3097	0.3229	0.5582	0.4491	0.4407	0.1036	0.2756	0.2918	0.5338	0.4303	0.4223	0.2555	0.2488	0.2452
8M 64D BiGRU-2	0.1669	0.3242	0.3357	0.5542	0.4663	0.4595	0.1366	0.2890	0.3032	0.5294	0.4471	0.4406	0.3119	0.2900	0.2864
8M 64D BiLSTM-1	0.1413	0.3140	0.3264	0.5521	0.4530	0.4454	0.1161	0.2797	0.2950	0.5276	0.4340	0.4268	0.2729	0.2586	0.2551
8M 64D BiLSTM-2	0.1717	0.3239	0.3355	0.5486	0.4669	0.4603	0.1405	0.2886	0.3028	0.5237	0.4476	0.4413	0.3150	0.2919	0.2885
8M 64D CNN-3	0.1628	0.3005	0.3113	0.4979	0.4309	0.4253	0.1354	0.2684	0.2816	0.4759	0.4131	0.4078	0.2549	0.2316	0.2287
8M 64D CNN-5	0.1804	0.3101	0.3202	0.5057	0.4431	0.4380	0.1503	0.2774	0.2897	0.4835	0.4250	0.4202	0.2835	0.2573	0.2543
8M 64D CNN-7	0.1934	0.3148	0.3239	0.5022	0.4482	0.4438	0.1614	0.2818	0.2928	0.4804	0.4300	0.4258	0.2953	0.2692	0.2664
8M 64D CNN-9	0.2009	0.3167	0.3249	0.4965	0.4498	0.4457	0.1683	0.2838	0.2939	0.4751	0.4317	0.4278	0.2983	0.2732	0.2706
8M 64D Linear	0.0848	0.2582	0.2668	0.5223	0.3721	0.3679	0.0715	0.2305	0.2410	0.4945	0.3553	0.3514	0.1451	0.1375	0.1357
8M 64D MLP-2	0.0495	0.2601	0.2777	0.5397	0.3874	0.3774	0.0435	0.2302	0.2518	0.5229	0.3706	0.3612	0.1178	0.1428	0.1397
8M 64D MLP-3	0.0804	0.2698	0.2847	0.5054	0.3963	0.3881	0.0680	0.2397	0.2579	0.4861	0.3794	0.3716	0.1696	0.1668	0.1636
8M 64D MLP-4	0.1161	0.2783	0.2913	0.4843	0.4051	0.3979	0.0967	0.2476	0.2636	0.4635	0.3879	0.3810	0.2012	0.1851	0.1822
8M 64D Projectformer-2	0.1340	0.2923	0.3054	0.5152	0.4224	0.4150	0.1122	0.2606	0.2766	0.4924	0.4045	0.3976	0.2451	0.2190	0.2156
8M 64D Projectformer-4	0.1818	0.3108	0.3205	0.4986	0.4426	0.4375	0.1522	0.2780	0.2898	0.4761	0.4240	0.4192	0.2836	0.2564	0.2536
8M 128D BiGRU-1	0.1263	0.3104	0.3236	0.5597	0.4493	0.4412	0.1036	0.2763	0.2925	0.5352	0.4306	0.4228	0.2598	0.2516	0.2480
8M 128D BiGRU-2	0.1732	0.3268	0.3384	0.5557	0.4698	0.4631	0.1418	0.2914	0.3057	0.5305	0.4506	0.4441	0.3210	0.2978	0.2942
8M 128D BiLSTM-1	0.1439	0.3156	0.3278	0.5545	0.4546	0.4472	0.1184	0.2812	0.2962	0.5297	0.4357	0.4286	0.2781	0.2623	0.2588
8M 128D BiLSTM-2	0.1832	0.3309	0.3421	0.5543	0.4750	0.4684	0.1505	0.2950	0.3087	0.5293	0.4554	0.4491	0.3377	0.3119	0.3081
8M 128D CNN-3	0.1756	0.3037	0.3134	0.4936	0.4335	0.4287	0.1464	0.2716	0.2835	0.4717	0.4156	0.4111	0.2661	0.2413	0.2385
8M 128D CNN-5	0.1941	0.3121	0.3210	0.4946	0.4439	0.4396	0.1623	0.2796	0.2903	0.4731	0.4258	0.4217	0.2894	0.2639	0.2612
8M 128D CNN-7	0.2063	0.3172	0.3249	0.4926	0.4481	0.4447	0.1733	0.2846	0.2941	0.4715	0.4300	0.4268	0.2994	0.2746	0.2722
8M 128D CNN-9	0.2122	0.3199	0.3277	0.4917	0.4521	0.4484	0.1786	0.2870	0.2965	0.4705	0.4339	0.4304	0.3054	0.2814	0.2790
8M 128D Linear	0.1006	0.2573	0.2662	0.5429	0.3737	0.3694	0.0839	0.2295	0.2402	0.5096	0.3567	0.3528	0.1538	0.1421	0.1401
8M 128D MLP-2	0.0606	0.2643	0.2805	0.5307	0.3911	0.3817	0.0526	0.2344	0.2543	0.5130	0.3743	0.3655	0.1405	0.1530	0.1498
8M 128D MLP-3	0.0990	0.2750	0.2889	0.4993	0.4018	0.3938	0.0830	0.2446	0.2617	0.4791	0.3847	0.3772	0.1920	0.1805	0.1774
8M 128D MLP-4	0.1349	0.2839	0.2955	0.4831	0.4093	0.4032	0.1124	0.2532	0.2674	0.4615	0.3919	0.3861	0.2166	0.1962	0.1934
8M 128D Projectformer-2	0.1675	0.3052	0.3159	0.5096	0.4376	0.4322	0.1396	0.2724	0.2854	0.4861	0.4191	0.4141	0.2780	0.2494	0.2464
8M 128D Projectformer-4	0.1944	0.3161	0.3252	0.5015	0.4498	0.4454	0.1630	0.2827	0.2937	0.4789	0.4311	0.4269	0.2988	0.2703	0.2677

Table S28: All 84 frozen adapters, with every structural endpoint and correspondence F1. T = threshold; MM = mutual matching. Each metric group gives all three decoders for the same checkpoint. Continuation 2.

Configuration	Cα full LDDT			Cα aligned LDDT			All-atom full LDDT			All-atom aligned LDDT			Correspondence F1		
	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA
35M 64D BiGRU-1	0.1461	0.3237	0.3402	0.6149	0.4750	0.4664	0.1173	0.2875	0.3074	0.5857	0.4553	0.4473	0.3402	0.3216	0.3161
35M 64D BiGRU-2	0.1780	0.3301	0.3436	0.5745	0.4793	0.4722	0.1447	0.2938	0.3101	0.5481	0.4598	0.4531	0.3540	0.3266	0.3222
35M 64D BiLSTM-1	0.1551	0.3252	0.3405	0.5999	0.4750	0.4669	0.1255	0.2890	0.3076	0.5716	0.4555	0.4479	0.3403	0.3174	0.3122
35M 64D BiLSTM-2	0.1940	0.3344	0.3466	0.5586	0.4822	0.4759	0.1593	0.2979	0.3126	0.5332	0.4626	0.4566	0.3591	0.3309	0.3270
35M 64D CNN-3	0.1590	0.3037	0.3157	0.5293	0.4386	0.4331	0.1305	0.2711	0.2856	0.5053	0.4207	0.4155	0.2803	0.2536	0.2498
35M 64D CNN-5	0.1861	0.3101	0.3193	0.5001	0.4426	0.4383	0.1550	0.2776	0.2888	0.4781	0.4246	0.4206	0.2873	0.2614	0.2587
35M 64D CNN-7	0.2038	0.3162	0.3242	0.4919	0.4482	0.4446	0.1710	0.2834	0.2932	0.4708	0.4301	0.4266	0.2968	0.2720	0.2695
35M 64D CNN-9	0.2078	0.3170	0.3244	0.4901	0.4479	0.4445	0.1746	0.2844	0.2934	0.4692	0.4298	0.4266	0.2973	0.2736	0.2712
35M 64D Linear	0.0430	0.2572	0.2749	0.5785	0.3843	0.3752	0.0374	0.2278	0.2492	0.5603	0.3671	0.3585	0.1131	0.1462	0.1428
35M 64D MLP-2	0.0973	0.2887	0.3059	0.6050	0.4273	0.4183	0.0799	0.2564	0.2772	0.5779	0.4092	0.4007	0.2413	0.2322	0.2266
35M 64D MLP-3	0.1142	0.2919	0.3095	0.5913	0.4320	0.4238	0.0927	0.2591	0.2802	0.5628	0.4138	0.4061	0.2635	0.2438	0.2385
35M 64D MLP-4	0.1277	0.2895	0.3049	0.5523	0.4258	0.4188	0.1035	0.2572	0.2757	0.5253	0.4078	0.4013	0.2564	0.2325	0.2282
35M 64D Projectformer-2	0.1332	0.2978	0.3139	0.5642	0.4375	0.4294	0.1092	0.2648	0.2841	0.5371	0.4192	0.4116	0.2808	0.2534	0.2484
35M 64D Projectformer-4	0.1960	0.3135	0.3220	0.4920	0.4456	0.4416	0.1646	0.2805	0.2907	0.4701	0.4271	0.4233	0.2923	0.2656	0.2633
35M 128D BiGRU-1	0.1459	0.3230	0.3397	0.6146	0.4745	0.4661	0.1171	0.2868	0.3069	0.5853	0.4548	0.4469	0.3402	0.3212	0.3157
35M 128D BiGRU-2	0.1803	0.3305	0.3435	0.5745	0.4790	0.4723	0.1465	0.2943	0.3101	0.5482	0.4595	0.4531	0.3558	0.3271	0.3227
35M 128D BiLSTM-1	0.1551	0.3245	0.3400	0.5993	0.4741	0.4659	0.1252	0.2884	0.3070	0.5710	0.4545	0.4469	0.3407	0.3179	0.3127
35M 128D BiLSTM-2	0.1972	0.3340	0.3453	0.5538	0.4808	0.4747	0.1622	0.2976	0.3114	0.5287	0.4611	0.4553	0.3594	0.3303	0.3266
35M 128D CNN-3	0.1697	0.3047	0.3158	0.5154	0.4384	0.4332	0.1400	0.2723	0.2857	0.4920	0.4205	0.4156	0.2829	0.2558	0.2524
35M 128D CNN-5	0.1967	0.3122	0.3203	0.4939	0.4425	0.4390	0.1646	0.2800	0.2897	0.4725	0.4246	0.4212	0.2910	0.2658	0.2631
35M 128D CNN-7	0.2094	0.3169	0.3242	0.4887	0.4477	0.4445	0.1762	0.2844	0.2933	0.4678	0.4296	0.4266	0.2981	0.2741	0.2718
35M 128D CNN-9	0.2179	0.3207	0.3277	0.4879	0.4516	0.4484	0.1841	0.2881	0.2965	0.4672	0.4334	0.4304	0.3058	0.2826	0.2804
35M 128D Linear	0.0560	0.2585	0.2752	0.5802	0.3847	0.3773	0.0480	0.2294	0.2494	0.5596	0.3674	0.3606	0.1434	0.1617	0.1582
35M 128D MLP-2	0.1055	0.2919	0.3088	0.6021	0.4306	0.4218	0.0865	0.2594	0.2798	0.5744	0.4125	0.4042	0.2549	0.2398	0.2343
35M 128D MLP-3	0.1183	0.2926	0.3100	0.5882	0.4331	0.4250	0.0960	0.2598	0.2807	0.5596	0.4150	0.4074	0.2697	0.2476	0.2422
35M 128D MLP-4	0.1354	0.2914	0.3063	0.5436	0.4272	0.4204	0.1104	0.2593	0.2771	0.5172	0.4092	0.4029	0.2625	0.2363	0.2321
35M 128D Projectformer-2	0.1773	0.3076	0.3177	0.5030	0.4394	0.4342	0.1478	0.2750	0.2873	0.4802	0.4212	0.4162	0.2861	0.2565	0.2535
35M 128D Projectformer-4	0.2188	0.3280	0.3355	0.5043	0.4634	0.4600	0.1845	0.2939	0.3029	0.4822	0.4444	0.4411	0.3265	0.2999	0.2975

Table S29: All 84 frozen adapters, with every structural endpoint and correspondence F1. T = threshold; MM = mutual matching. Each metric group gives all three decoders for the same checkpoint. Continuation 3.

Configuration	Cα full LDDT			Cα aligned LDDT			All-atom full LDDT			All-atom aligned LDDT			Correspondence F1		
	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA
150M 64D BiGRU-1	0.1543	0.3248	0.3407	0.6054	0.4760	0.4679	0.1244	0.2885	0.3076	0.5767	0.4564	0.4487	0.3448	0.3225	0.3172
150M 64D BiGRU-2	0.1968	0.3346	0.3464	0.5613	0.4823	0.4761	0.1619	0.2983	0.3125	0.5360	0.4627	0.4568	0.3649	0.3350	0.3310
150M 64D BiLSTM-1	0.1672	0.3270	0.3409	0.5830	0.4746	0.4673	0.1360	0.2909	0.3077	0.5561	0.4551	0.4482	0.3450	0.3186	0.3140
150M 64D BiLSTM-2	0.2147	0.3393	0.3492	0.5424	0.4855	0.4800	0.1782	0.3026	0.3148	0.5184	0.4657	0.4604	0.3693	0.3400	0.3367
150M 64D CNN-3	0.1787	0.3053	0.3153	0.5021	0.4374	0.4330	0.1479	0.2731	0.2852	0.4796	0.4195	0.4153	0.2772	0.2523	0.2493
150M 64D CNN-5	0.2031	0.3123	0.3203	0.4880	0.4438	0.4402	0.1702	0.2799	0.2895	0.4670	0.4258	0.4224	0.2877	0.2649	0.2625
150M 64D CNN-7	0.2163	0.3176	0.3246	0.4858	0.4484	0.4455	0.1823	0.2850	0.2934	0.4651	0.4304	0.4276	0.2986	0.2765	0.2743
150M 64D CNN-9	0.2213	0.3196	0.3263	0.4869	0.4509	0.4481	0.1867	0.2870	0.2950	0.4663	0.4329	0.4302	0.3049	0.2828	0.2808
150M 64D Linear	0.0627	0.2723	0.2903	0.6332	0.4062	0.3973	0.0524	0.2415	0.2630	0.6081	0.3880	0.3797	0.1829	0.2006	0.1952
150M 64D MLP-2	0.1057	0.2967	0.3167	0.6392	0.4428	0.4335	0.0855	0.2630	0.2868	0.6078	0.4241	0.4155	0.2822	0.2704	0.2639
150M 64D MLP-3	0.1205	0.2938	0.3117	0.5933	0.4361	0.4283	0.0970	0.2607	0.2819	0.5637	0.4178	0.4105	0.2773	0.2555	0.2502
150M 64D MLP-4	0.1502	0.2931	0.3059	0.5173	0.4259	0.4204	0.1228	0.2612	0.2766	0.4927	0.4080	0.4028	0.2580	0.2325	0.2291
150M 64D Projectformer-2	0.1438	0.3051	0.3204	0.5646	0.4451	0.4377	0.1178	0.2714	0.2898	0.5376	0.4266	0.4196	0.2945	0.2661	0.2614
150M 64D Projectformer-4	0.2131	0.3255	0.3334	0.5018	0.4608	0.4570	0.1793	0.2915	0.3011	0.4799	0.4420	0.4384	0.3192	0.2933	0.2907
150M 128D BiGRU-1	0.1583	0.3272	0.3422	0.6038	0.4773	0.4695	0.1277	0.2910	0.3091	0.5754	0.4577	0.4504	0.3497	0.3266	0.3214
150M 128D BiGRU-2	0.1972	0.3346	0.3459	0.5612	0.4817	0.4758	0.1621	0.2984	0.3121	0.5359	0.4621	0.4565	0.3644	0.3339	0.3300
150M 128D BiLSTM-1	0.1693	0.3280	0.3417	0.5838	0.4758	0.4685	0.1379	0.2920	0.3086	0.5567	0.4563	0.4494	0.3469	0.3200	0.3152
150M 128D BiLSTM-2	0.2077	0.3360	0.3465	0.5435	0.4818	0.4764	0.1719	0.2996	0.3124	0.5191	0.4621	0.4569	0.3572	0.3289	0.3257
150M 128D CNN-3	0.1820	0.3069	0.3166	0.5023	0.4387	0.4345	0.1509	0.2746	0.2864	0.4799	0.4208	0.4168	0.2821	0.2564	0.2536
150M 128D CNN-5	0.2059	0.3142	0.3220	0.4883	0.4451	0.4418	0.1728	0.2818	0.2912	0.4673	0.4271	0.4239	0.2935	0.2699	0.2675
150M 128D CNN-7	0.2186	0.3191	0.3261	0.4871	0.4500	0.4470	0.1845	0.2865	0.2949	0.4664	0.4318	0.4290	0.3028	0.2800	0.2778
150M 128D CNN-9	0.2220	0.3200	0.3266	0.4862	0.4502	0.4475	0.1876	0.2874	0.2954	0.4656	0.4322	0.4296	0.3041	0.2820	0.2799
150M 128D Linear	0.0594	0.2686	0.2861	0.6140	0.4002	0.3913	0.0501	0.2383	0.2593	0.5909	0.3822	0.3739	0.1658	0.1852	0.1805
150M 128D MLP-2	0.1103	0.2982	0.3177	0.6363	0.4440	0.4348	0.0892	0.2646	0.2877	0.6051	0.4253	0.4168	0.2890	0.2736	0.2671
150M 128D MLP-3	0.1242	0.2962	0.3136	0.5936	0.4379	0.4303	0.1004	0.2630	0.2838	0.5641	0.4196	0.4125	0.2832	0.2593	0.2538
150M 128D MLP-4	0.1517	0.2946	0.3073	0.5205	0.4275	0.4219	0.1242	0.2626	0.2778	0.4957	0.4096	0.4043	0.2619	0.2356	0.2320
150M 128D Projectformer-2	0.1840	0.3150	0.3250	0.5161	0.4504	0.4455	0.1533	0.2814	0.2936	0.4926	0.4318	0.4270	0.3031	0.2735	0.2706
150M 128D Projectformer-4	0.2190	0.3275	0.3350	0.5032	0.4631	0.4597	0.1845	0.2934	0.3024	0.4811	0.4441	0.4409	0.3242	0.2983	0.2959

Table S30: All 20 3Di predictors, with separate soft and hard state rows. T = threshold; MM = mutual matching; MEA = maximum expected accuracy. State argmax is separate from correspondence decoding. IDs follow Table S10.

Configuration	Cα full LDDT			Cα aligned LDDT			All-atom full LDDT			All-atom aligned LDDT			Correspondence F1		
	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA
35M frozen MLP4 ST-Gumbel [72459d15] / soft	0.2885	0.3523	0.3764	0.6307	0.5722	0.5492	0.2394	0.3050	0.3331	0.6048	0.5502	0.5286	0.5332	0.5164	0.5003
35M frozen MLP4 ST-Gumbel [72459d15] / hard	0.3086	0.3520	0.3587	0.5574	0.5180	0.5166	0.2615	0.3114	0.3191	0.5363	0.4999	0.4985	0.4708	0.4540	0.4530
35M frozen MLP4 ST-argmax [9fceaf26] / soft	0.1070	0.2005	0.2925	0.7689	0.6410	0.4916	0.0804	0.1600	0.2547	0.7305	0.6122	0.4740	0.3555	0.4479	0.3979
35M frozen MLP4 ST-argmax [9fceaf26] / hard	0.2574	0.3017	0.3099	0.5270	0.4850	0.4860	0.2133	0.2616	0.2701	0.5080	0.4691	0.4700	0.4097	0.3958	0.3962
35M frozen MLP4 softmax [0eb8395a] / soft	0.2685	0.3919	0.4066	0.6451	0.5642	0.5575	0.2224	0.3480	0.3657	0.6175	0.5425	0.5362	0.5668	0.5334	0.5272
35M frozen MLP4 softmax [0eb8395a] / hard	0.2646	0.3151	0.3235	0.5235	0.4751	0.4746	0.2222	0.2784	0.2878	0.5024	0.4578	0.4573	0.4083	0.3913	0.3912
35M frozen linear ST-Gumbel [4b1a9c9d] / soft	0.2489	0.3301	0.3722	0.6474	0.6041	0.5527	0.2028	0.2816	0.3293	0.6201	0.5807	0.5320	0.5198	0.5136	0.4888
35M frozen linear ST-Gumbel [4b1a9c9d] / hard	0.2905	0.3424	0.3507	0.5580	0.5145	0.5125	0.2433	0.3016	0.3112	0.5363	0.4962	0.4942	0.4439	0.4274	0.4266
35M frozen linear ST-argmax [f8f5cad4] / soft	0.0249	0.0897	0.2679	0.8486	0.6763	0.4448	0.0163	0.0660	0.2346	0.8030	0.6420	0.4304	0.1527	0.3215	0.3304
35M frozen linear ST-argmax [f8f5cad4] / hard	0.2557	0.3082	0.3197	0.5414	0.4977	0.4956	0.2108	0.2666	0.2794	0.5212	0.4810	0.4789	0.4089	0.3950	0.3953
35M frozen linear softmax [c336a672] / soft	0.2335	0.3829	0.4026	0.6673	0.5621	0.5527	0.1902	0.3385	0.3620	0.6384	0.5409	0.5319	0.5474	0.5241	0.5160
35M frozen linear softmax [c336a672] / hard	0.2583	0.3124	0.3209	0.5201	0.4669	0.4661	0.2155	0.2761	0.2858	0.4989	0.4499	0.4491	0.4027	0.3844	0.3840
35M unfrozen MLP4 ST-Gumbel [f32e1543] / soft	0.2405	0.3631	0.3941	0.6573	0.5610	0.5502	0.1962	0.3172	0.3528	0.6294	0.5405	0.5301	0.5446	0.5240	0.5116
35M unfrozen MLP4 ST-Gumbel [f32e1543] / hard	0.2496	0.3577	0.3809	0.6473	0.5411	0.5370	0.2041	0.3142	0.3405	0.6204	0.5220	0.5179	0.5330	0.5034	0.4963
35M unfrozen MLP4 ST-argmax [39efc845] / soft	0.1731	0.3014	0.3802	0.7163	0.5807	0.5427	0.1359	0.2566	0.3385	0.6829	0.5589	0.5227	0.4900	0.5245	0.5011
35M unfrozen MLP4 ST-argmax [39efc845] / hard	0.2152	0.3199	0.3464	0.6322	0.5071	0.5066	0.1735	0.2780	0.3066	0.6048	0.4902	0.4895	0.4953	0.4694	0.4650
35M unfrozen MLP4 softmax [f7815ea1] / soft	0.2653	0.3842	0.4017	0.6450	0.5638	0.5555	0.2197	0.3396	0.3606	0.6175	0.5426	0.5347	0.5634	0.5313	0.5240
35M unfrozen MLP4 softmax [f7815ea1] / hard	0.2913	0.3653	0.3768	0.5976	0.5326	0.5299	0.2446	0.3238	0.3373	0.5733	0.5131	0.5105	0.5155	0.4870	0.4839
35M unfrozen MLP4 softmax+mask [929752d4] / soft	0.2219	0.3545	0.3829	0.6592	0.5466	0.5387	0.1792	0.3101	0.3428	0.6308	0.5271	0.5194	0.5251	0.5037	0.4954
35M unfrozen MLP4 softmax+mask [929752d4] / hard	0.2343	0.3406	0.3636	0.6358	0.5214	0.5179	0.1901	0.2985	0.3247	0.6094	0.5034	0.4999	0.5048	0.4771	0.4729
8M frozen MLP4 ST-Gumbel [87d2c509] / soft	0.2265	0.2960	0.3277	0.5937	0.5388	0.5033	0.1839	0.2518	0.2872	0.5683	0.5178	0.4842	0.4217	0.4086	0.3889
8M frozen MLP4 ST-Gumbel [87d2c509] / hard	0.2492	0.3033	0.3121	0.5119	0.4687	0.4670	0.2065	0.2653	0.2754	0.4914	0.4515	0.4500	0.3596	0.3441	0.3434
8M frozen MLP4 ST-argmax [d58f9e14] / soft	0.0568	0.1269	0.2387	0.8011	0.6447	0.4405	0.0401	0.0966	0.2061	0.7554	0.6121	0.4239	0.2180	0.3167	0.2696
8M frozen MLP4 ST-argmax [d58f9e14] / hard	0.2321	0.2711	0.2781	0.4749	0.4419	0.4431	0.1911	0.2343	0.2418	0.4578	0.4272	0.4282	0.3029	0.2936	0.2944
8M frozen MLP4 softmax [7b541c60] / soft	0.2025	0.3553	0.3745	0.6459	0.5234	0.5151	0.1641	0.3146	0.3375	0.6166	0.5032	0.4954	0.4716	0.4413	0.4344
8M frozen MLP4 softmax [7b541c60] / hard	0.2225	0.2814	0.2904	0.4809	0.4280	0.4265	0.1847	0.2487	0.2592	0.4606	0.4118	0.4102	0.3107	0.2951	0.2947
8M frozen linear ST-Gumbel [31bba998] / soft	0.1737	0.2625	0.3158	0.6122	0.5465	0.4975	0.1371	0.2194	0.2758	0.5843	0.5244	0.4784	0.3933	0.3896	0.3700
8M frozen linear ST-Gumbel [31bba998] / hard	0.2306	0.2891	0.2989	0.5015	0.4538	0.4525	0.1887	0.2522	0.2633	0.4812	0.4373	0.4360	0.3254	0.3107	0.3105

Table S31: All 20 3Di predictors, with separate soft and hard state rows. T = threshold; MM = mutual matching; MEA = maximum expected accuracy. State argmax is separate from correspondence decoding. IDs follow Table S10. Continuation 2.

Configuration	Cα full LDDT			Cα aligned LDDT			All-atom full LDDT			All-atom aligned LDDT			Correspondence F1		
	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA
8M frozen linear ST-argmax [fbc46ff2] / soft	0.0129	0.0387	0.2068	0.8548	0.6881	0.3991	0.0085	0.0269	0.1759	0.8057	0.6474	0.3847	0.0888	0.1744	0.2136
8M frozen linear ST-argmax [fbc46ff2] / hard	0.1939	0.2445	0.2548	0.4798	0.4295	0.4303	0.1552	0.2077	0.2182	0.4614	0.4150	0.4156	0.2989	0.2880	0.2887
8M frozen linear softmax [1cf7945c] / soft	0.1744	0.3442	0.3700	0.6689	0.5212	0.5096	0.1400	0.3027	0.3332	0.6385	0.5011	0.4901	0.4468	0.4303	0.4216
8M frozen linear softmax [1cf7945c] / hard	0.2195	0.2771	0.2864	0.4833	0.4291	0.4290	0.1806	0.2440	0.2544	0.4633	0.4135	0.4134	0.3051	0.2911	0.2908
8M unfrozen MLP4 ST-Gumbel [d50ef818] / soft	0.2944	0.3308	0.3479	0.5677	0.5345	0.5143	0.2505	0.2880	0.3079	0.5444	0.5132	0.4941	0.4249	0.4122	0.3985
8M unfrozen MLP4 ST-Gumbel [d50ef818] / hard	0.3187	0.3421	0.3470	0.5267	0.5054	0.5024	0.2745	0.3018	0.3076	0.5059	0.4860	0.4831	0.4018	0.3920	0.3896
8M unfrozen MLP4 ST-argmax [650078f8] / soft	0.2093	0.2475	0.3064	0.6328	0.5772	0.4893	0.1723	0.2074	0.2671	0.6030	0.5508	0.4701	0.4160	0.4120	0.3720
8M unfrozen MLP4 ST-argmax [650078f8] / hard	0.2831	0.3086	0.3148	0.5162	0.4850	0.4837	0.2399	0.2682	0.2749	0.4961	0.4669	0.4657	0.3834	0.3726	0.3704
8M unfrozen MLP4 softmax [16132e1d] / soft	0.2937	0.3696	0.3755	0.5504	0.5242	0.5205	0.2520	0.3298	0.3369	0.5276	0.5034	0.4999	0.4465	0.4217	0.4190
8M unfrozen MLP4 softmax [16132e1d] / hard	0.3062	0.3370	0.3400	0.5002	0.4770	0.4766	0.2657	0.3016	0.3051	0.4796	0.4582	0.4578	0.3718	0.3615	0.3610
8M unfrozen MLP4 softmax+mask [9d792c09] / soft	0.3208	0.3551	0.3616	0.5487	0.5245	0.5182	0.2778	0.3149	0.3224	0.5266	0.5039	0.4980	0.4192	0.4060	0.4012
8M unfrozen MLP4 softmax+mask [9d792c09] / hard	0.3160	0.3348	0.3372	0.4993	0.4860	0.4862	0.2756	0.2977	0.3003	0.4797	0.4674	0.4676	0.3687	0.3627	0.3627

Table S32: All 49 retrieval configurations on 5,059 proteins. SF = superfamily. ROC1 uses the published-script convention and 1,328 eligible queries. Top-1 uses 3,459 family, 4,415 superfamily and 4,696 fold queries; deterministic ties use target ID.

Configuration	ROC1			Tie-averaged top-1			Deterministic top-1			Macro-class top-1		
	Family	SF	Fold	Family	SF	Fold	Family	SF	Fold	Family	SF	Fold
MMseqs2 / AA	0.3762	0.1534	0.0006	0.5241	0.7089	0.6725	0.5244	0.7089	0.6729	0.4943	0.6219	0.6049
ProstT5-based Foldseek / 3Di-only	0.7993	0.5316	0.1088	0.6106	0.9008	0.8930	0.6094	0.9010	0.8933	0.6011	0.8229	0.8126
ProstT5-based Foldseek / AA+3Di	0.8352	0.5781	0.1208	0.6288	0.9229	0.9168	0.6285	0.9230	0.9167	0.6265	0.8634	0.8452
Structure / 3Di-only	0.7446	0.4659	0.0830	0.6016	0.8973	0.8901	0.6019	0.8976	0.8903	0.5920	0.8173	0.8076
Structure / AA+3Di	0.8176	0.5352	0.0962	0.6281	0.9281	0.9213	0.6273	0.9280	0.9214	0.6274	0.8712	0.8513
TM-align arithmetic mean / TM score	0.8773	0.6330	0.2122	0.6300	0.9411	0.9495	0.6300	0.9411	0.9495	0.6291	0.8927	0.8879
TM-align geometric mean / TM score	0.8759	0.6360	0.2200	0.6308	0.9425	0.9500	0.6308	0.9425	0.9500	0.6307	0.8946	0.8903
TM-align max / TM score	0.7242	0.4746	0.0962	0.5863	0.8686	0.8680	0.5863	0.8686	0.8680	0.5761	0.7851	0.7661
TM-align min / TM score	0.8423	0.5925	0.2012	0.6311	0.9311	0.9427	0.6311	0.9311	0.9427	0.6345	0.8744	0.8774
8M frozen MLP4 ST-Gumbel [87d2c509] / 3Di-only	0.3049	0.2106	0.0264	0.2894	0.4307	0.4325	0.2905	0.4315	0.4331	0.1727	0.2087	0.2102
8M frozen MLP4 ST-Gumbel [87d2c509] / AA+3Di	0.4925	0.2950	0.0281	0.4942	0.7086	0.6884	0.4947	0.7094	0.6889	0.4408	0.5733	0.5588
8M frozen MLP4 ST-argmax [d58f9e14] / 3Di-only	0.2412	0.1713	0.0196	0.2350	0.3700	0.3776	0.2345	0.3692	0.3765	0.1281	0.1523	0.1519
8M frozen MLP4 ST-argmax [d58f9e14] / AA+3Di	0.4213	0.2523	0.0216	0.4433	0.6360	0.6215	0.4443	0.6358	0.6214	0.3745	0.4555	0.4494
8M frozen MLP4 softmax [7b541c60] / 3Di-only	0.2972	0.1988	0.0180	0.3265	0.4745	0.4724	0.3278	0.4750	0.4730	0.2160	0.2701	0.2699
8M frozen MLP4 softmax [7b541c60] / AA+3Di	0.5162	0.2877	0.0190	0.5166	0.7300	0.7072	0.5172	0.7302	0.7072	0.4717	0.6066	0.5912
8M frozen linear ST-Gumbel [31bba998] / 3Di-only	0.2701	0.1851	0.0214	0.2720	0.4085	0.4140	0.2712	0.4079	0.4135	0.1659	0.1962	0.1948
8M frozen linear ST-Gumbel [31bba998] / AA+3Di	0.4671	0.2777	0.0203	0.4870	0.6932	0.6763	0.4868	0.6940	0.6770	0.4298	0.5484	0.5404
8M frozen linear ST-argmax [fbc46ff2] / 3Di-only	0.2279	0.1539	0.0146	0.2595	0.3927	0.4005	0.2596	0.3932	0.4012	0.1509	0.1779	0.1807
8M frozen linear ST-argmax [fbc46ff2] / AA+3Di	0.4185	0.2503	0.0165	0.4602	0.6598	0.6470	0.4605	0.6596	0.6465	0.3878	0.5012	0.4935
8M frozen linear softmax [1cf7945c] / 3Di-only	0.2732	0.1831	0.0167	0.3010	0.4476	0.4479	0.3007	0.4473	0.4476	0.1919	0.2308	0.2421
8M frozen linear softmax [1cf7945c] / AA+3Di	0.4984	0.2780	0.0163	0.5100	0.7173	0.6955	0.5082	0.7176	0.6953	0.4609	0.5912	0.5738
8M unfrozen MLP4 ST-Gumbel [d50ef818] / 3Di-only	0.4118	0.2844	0.0589	0.3293	0.5070	0.5206	0.3287	0.5074	0.5217	0.2073	0.2651	0.2795
8M unfrozen MLP4 ST-Gumbel [d50ef818] / AA+3Di	0.5264	0.3404	0.0621	0.4738	0.6827	0.6755	0.4727	0.6824	0.6753	0.4038	0.5125	0.5103
8M unfrozen MLP4 ST-argmax [650078f8] / 3Di-only	0.3352	0.2507	0.0494	0.2684	0.4306	0.4500	0.2680	0.4313	0.4502	0.1400	0.1793	0.1863
8M unfrozen MLP4 ST-argmax [650078f8] / AA+3Di	0.4515	0.3060	0.0515	0.4110	0.5996	0.5960	0.4108	0.5993	0.5954	0.3086	0.3886	0.3834
8M unfrozen MLP4 softmax [16132e1d] / 3Di-only	0.3697	0.2528	0.0361	0.3440	0.4957	0.5010	0.3432	0.4956	0.5011	0.2246	0.2671	0.2751
8M unfrozen MLP4 softmax [16132e1d] / AA+3Di	0.5511	0.3305	0.0374	0.5146	0.7311	0.7126	0.5149	0.7314	0.7127	0.4673	0.5979	0.5840
8M unfrozen MLP4 softmax+mask [9d792c09] / 3Di-only	0.3433	0.2397	0.0211	0.3216	0.4960	0.5034	0.3209	0.4960	0.5036	0.1943	0.2706	0.2878

Table S33: All 49 retrieval configurations on 5,059 proteins. SF = superfamily. ROC1 uses the published-script convention and 1,328 eligible queries. Top-1 uses 3,459 family, 4,415 superfamily and 4,696 fold queries; deterministic ties use target ID. Continuation 2.

Configuration	ROC1			Tie-averaged top-1			Deterministic top-1			Macro-class top-1		
	Family	SF	Fold	Family	SF	Fold	Family	SF	Fold	Family	SF	Fold
8M unfrozen MLP4 softmax+mask [9d792c09] / AA+3Di	0.4911	0.2963	0.0236	0.4701	0.6817	0.6664	0.4695	0.6809	0.6655	0.3935	0.5153	0.5123
35M frozen MLP4 ST-Gumbel [72459d15] / 3Di-only	0.5273	0.3494	0.0606	0.4306	0.6476	0.6449	0.4299	0.6473	0.6446	0.3206	0.4062	0.4058
35M frozen MLP4 ST-Gumbel [72459d15] / AA+3Di	0.6786	0.4198	0.0660	0.5468	0.8005	0.7821	0.5473	0.8005	0.7824	0.5020	0.6558	0.6361
35M frozen MLP4 ST-argmax [9fceaf26] / 3Di-only	0.4569	0.3152	0.0452	0.3713	0.5852	0.5843	0.3709	0.5851	0.5841	0.2452	0.3241	0.3205
35M frozen MLP4 ST-argmax [9fceaf26] / AA+3Di	0.6124	0.3849	0.0500	0.4907	0.7335	0.7179	0.4903	0.7339	0.7181	0.4189	0.5321	0.5125
35M frozen MLP4 softmax [0eb8395a] / 3Di-only	0.5189	0.3384	0.0344	0.4362	0.6648	0.6619	0.4365	0.6646	0.6614	0.3323	0.4282	0.4318
35M frozen MLP4 softmax [0eb8395a] / AA+3Di	0.6563	0.4071	0.0381	0.5471	0.8023	0.7855	0.5473	0.8023	0.7853	0.4999	0.6592	0.6453
35M frozen linear ST-Gumbel [4b1a9c9d] / 3Di-only	0.5166	0.3413	0.0363	0.4297	0.6521	0.6499	0.4284	0.6519	0.6499	0.3173	0.4069	0.4098
35M frozen linear ST-Gumbel [4b1a9c9d] / AA+3Di	0.6658	0.4106	0.0439	0.5405	0.7958	0.7782	0.5409	0.7952	0.7781	0.4875	0.6439	0.6261
35M frozen linear ST-argmax [f8f5cad4] / 3Di-only	0.4981	0.3318	0.0380	0.4265	0.6416	0.6419	0.4270	0.6419	0.6422	0.3135	0.4002	0.3989
35M frozen linear ST-argmax [f8f5cad4] / AA+3Di	0.6363	0.3980	0.0402	0.5311	0.7808	0.7645	0.5319	0.7807	0.7645	0.4790	0.6213	0.6046
35M frozen linear softmax [c336a672] / 3Di-only	0.4797	0.2980	0.0147	0.4357	0.6507	0.6430	0.4377	0.6512	0.6437	0.3371	0.4275	0.4199
35M frozen linear softmax [c336a672] / AA+3Di	0.6419	0.3821	0.0218	0.5663	0.8060	0.7857	0.5663	0.8059	0.7858	0.5263	0.6754	0.6575
35M unfrozen MLP4 ST-Gumbel [f32e1543] / 3Di-only	0.6402	0.4531	0.1004	0.4626	0.7102	0.7191	0.4631	0.7103	0.7191	0.3685	0.4739	0.4909
35M unfrozen MLP4 ST-Gumbel [f32e1543] / AA+3Di	0.7265	0.4977	0.0981	0.5380	0.8050	0.7950	0.5383	0.8050	0.7951	0.4906	0.6499	0.6396
35M unfrozen MLP4 ST-argmax [39efc845] / 3Di-only	0.5893	0.4156	0.0862	0.4143	0.6598	0.6649	0.4140	0.6600	0.6655	0.2952	0.3975	0.4087
35M unfrozen MLP4 ST-argmax [39efc845] / AA+3Di	0.6824	0.4664	0.0872	0.5008	0.7529	0.7449	0.5007	0.7529	0.7449	0.4232	0.5554	0.5482
35M unfrozen MLP4 softmax [f7815ea1] / 3Di-only	0.6249	0.4255	0.0714	0.4654	0.7130	0.7120	0.4657	0.7135	0.7123	0.3709	0.4874	0.4966
35M unfrozen MLP4 softmax [f7815ea1] / AA+3Di	0.7114	0.4761	0.0730	0.5366	0.7955	0.7844	0.5366	0.7955	0.7843	0.4874	0.6469	0.6386
35M unfrozen MLP4 softmax+mask [929752d4] / 3Di-only	0.6243	0.4242	0.0827	0.4480	0.6942	0.6974	0.4484	0.6940	0.6974	0.3520	0.4575	0.4634
35M unfrozen MLP4 softmax+mask [929752d4] / AA+3Di	0.6891	0.4675	0.0809	0.5256	0.7788	0.7674	0.5250	0.7785	0.7670	0.4690	0.6149	0.6057

Table S34: All 45 native clusterings compared with SCOPe at three levels. Each configuration has one cluster count and three partition scores per level. All partitions use the complete 5,059-protein validation population.

Configuration	Clusters	Family			Superfamily			Fold		
	Count	ARI	AMI	NMI	ARI	AMI	NMI	ARI	AMI	NMI
MMseqs2 / AA	3962	0.1972	0.3386	0.9291	0.1051	0.3003	0.8696	0.0511	0.2380	0.8139
ProstT5-based Foldseek / 3Di-only	1878	0.3448	0.5577	0.9066	0.4007	0.6526	0.8946	0.2979	0.6166	0.8585
ProstT5-based Foldseek / AA+3Di	1925	0.3559	0.5668	0.9105	0.4118	0.6628	0.8994	0.3351	0.6324	0.8663
Structure / 3Di-only	2305	0.4039	0.5779	0.9216	0.4153	0.6526	0.9046	0.2791	0.6012	0.8649
Structure / AA+3Di	2290	0.4002	0.6004	0.9261	0.4188	0.6710	0.9099	0.3075	0.6137	0.8696
8M frozen MLP4 ST-Gumbel [87d2c509] / 3Di-only	3557	0.3208	0.3739	0.9159	0.2568	0.3590	0.8612	0.1535	0.3126	0.8102
8M frozen MLP4 ST-Gumbel [87d2c509] / AA+3Di	3135	0.3930	0.5124	0.9271	0.3461	0.5063	0.8846	0.2010	0.4338	0.8332
8M frozen MLP4 ST-argmax [d58f9e14] / 3Di-only	2967	0.2629	0.3147	0.8933	0.2259	0.3214	0.8366	0.1342	0.2863	0.7842
8M frozen MLP4 ST-argmax [d58f9e14] / AA+3Di	2573	0.3163	0.4065	0.8985	0.3013	0.4318	0.8533	0.1908	0.3874	0.8033
8M frozen MLP4 softmax [7b541c60] / 3Di-only	2681	0.1723	0.2233	0.8686	0.1668	0.2302	0.8031	0.1061	0.2074	0.7481
8M frozen MLP4 softmax [7b541c60] / AA+3Di	2448	0.2251	0.3356	0.8811	0.2212	0.3531	0.8273	0.1342	0.3105	0.7723
8M frozen linear ST-Gumbel [31bba998] / 3Di-only	3524	0.2847	0.3393	0.9096	0.2592	0.3328	0.8537	0.1511	0.2876	0.8013
8M frozen linear ST-Gumbel [31bba998] / AA+3Di	3192	0.3361	0.4418	0.9166	0.2956	0.4436	0.8701	0.1700	0.3814	0.8180
8M frozen linear ST-argmax [fbc46ff2] / 3Di-only	3544	0.3025	0.3329	0.9081	0.2371	0.3174	0.8496	0.1307	0.2703	0.7956
8M frozen linear ST-argmax [fbc46ff2] / AA+3Di	3138	0.3434	0.4285	0.9140	0.2897	0.4240	0.8647	0.1617	0.3613	0.8110
8M frozen linear softmax [1cf7945c] / 3Di-only	2633	0.1950	0.2309	0.8713	0.1686	0.2383	0.8066	0.0990	0.2108	0.7504
8M frozen linear softmax [1cf7945c] / AA+3Di	2540	0.2325	0.3574	0.8870	0.2268	0.3798	0.8367	0.1361	0.3315	0.7820
8M unfrozen MLP4 ST-Gumbel [d50ef818] / 3Di-only	1391	0.1032	0.2838	0.7935	0.1730	0.3457	0.7476	0.2154	0.3446	0.7037
8M unfrozen MLP4 ST-Gumbel [d50ef818] / AA+3Di	1172	0.1022	0.2758	0.7884	0.1671	0.3441	0.7431	0.1822	0.3460	0.6993
8M unfrozen MLP4 ST-argmax [650078f8] / 3Di-only	2478	0.1432	0.3387	0.8628	0.2082	0.3829	0.8169	0.1838	0.3682	0.7731
8M unfrozen MLP4 ST-argmax [650078f8] / AA+3Di	2445	0.1803	0.3754	0.8761	0.2314	0.4268	0.8352	0.1809	0.3977	0.7888
8M unfrozen MLP4 softmax [16132e1d] / 3Di-only	2579	0.2106	0.3228	0.8786	0.2408	0.3364	0.8231	0.1881	0.3095	0.7729
8M unfrozen MLP4 softmax [16132e1d] / AA+3Di	2441	0.2725	0.3940	0.8889	0.2951	0.4163	0.8415	0.1894	0.3685	0.7886
8M unfrozen MLP4 softmax+mask [9d792c09] / 3Di-only	2497	0.2314	0.3549	0.8788	0.2611	0.3940	0.8325	0.2004	0.3663	0.7850
8M unfrozen MLP4 softmax+mask [9d792c09] / AA+3Di	2318	0.2549	0.4181	0.8870	0.2892	0.4729	0.8505	0.2013	0.4315	0.8025
35M frozen MLP4 ST-Gumbel [72459d15] / 3Di-only	2937	0.3167	0.4897	0.9154	0.3334	0.5274	0.8811	0.2296	0.4719	0.8346
35M frozen MLP4 ST-Gumbel [72459d15] / AA+3Di	2668	0.3194	0.5354	0.9181	0.3573	0.5899	0.8919	0.2629	0.5249	0.8451
35M frozen MLP4 ST-argmax [9fceaf26] / 3Di-only	3311	0.3813	0.4909	0.9243	0.3491	0.4968	0.8830	0.2210	0.4365	0.8349

Table S35: All 45 native clusterings compared with SCOPe at three levels. Each configuration has one cluster count and three partition scores per level. All partitions use the complete 5,059-protein validation population. Continuation 2.

Configuration	Clusters	Family			Superfamily			Fold		
	Count	ARI	AMI	NMI	ARI	AMI	NMI	ARI	AMI	NMI
35M frozen MLP4 ST-argmax [9fceaf26] / AA+3Di	3086	0.4021	0.5309	0.9276	0.3683	0.5508	0.8926	0.2279	0.4806	0.8440
35M frozen MLP4 softmax [0eb8395a] / 3Di-only	1810	0.2098	0.3637	0.8556	0.2760	0.4225	0.8151	0.2068	0.3954	0.7666
35M frozen MLP4 softmax [0eb8395a] / AA+3Di	1718	0.2305	0.4266	0.8679	0.2996	0.5089	0.8407	0.2323	0.4743	0.7946
35M frozen linear ST-Gumbel [4b1a9c9d] / 3Di-only	2975	0.3330	0.4888	0.9185	0.3239	0.5273	0.8842	0.2169	0.4653	0.8361
35M frozen linear ST-Gumbel [4b1a9c9d] / AA+3Di	2671	0.3960	0.5533	0.9239	0.3938	0.5986	0.8967	0.2412	0.5274	0.8488
35M frozen linear ST-argmax [f8f5cad4] / 3Di-only	2815	0.3754	0.4976	0.9165	0.3541	0.5344	0.8824	0.2221	0.4763	0.8351
35M frozen linear ST-argmax [f8f5cad4] / AA+3Di	2585	0.4057	0.5336	0.9201	0.3859	0.5834	0.8923	0.2320	0.5202	0.8457
35M frozen linear softmax [c336a672] / 3Di-only	2356	0.2889	0.3953	0.8889	0.2899	0.4228	0.8428	0.1806	0.3741	0.7896
35M frozen linear softmax [c336a672] / AA+3Di	2341	0.3385	0.4969	0.9052	0.3720	0.5562	0.8768	0.2374	0.4898	0.8258
35M unfrozen MLP4 ST-Gumbel [f32e1543] / 3Di-only	876	0.0657	0.3086	0.7567	0.1353	0.4089	0.7320	0.1898	0.4176	0.6969
35M unfrozen MLP4 ST-Gumbel [f32e1543] / AA+3Di	865	0.0843	0.3208	0.7737	0.1582	0.4226	0.7484	0.1603	0.4192	0.7072
35M unfrozen MLP4 ST-argmax [39efc845] / 3Di-only	2319	0.2282	0.4648	0.8901	0.3099	0.5562	0.8688	0.2476	0.5206	0.8277
35M unfrozen MLP4 ST-argmax [39efc845] / AA+3Di	2300	0.2511	0.5027	0.8999	0.3232	0.5838	0.8788	0.2463	0.5388	0.8363
35M unfrozen MLP4 softmax [f7815ea1] / 3Di-only	1399	0.0925	0.3552	0.8147	0.1748	0.4564	0.7913	0.1687	0.4407	0.7487
35M unfrozen MLP4 softmax [f7815ea1] / AA+3Di	1226	0.1068	0.3706	0.8119	0.1921	0.4688	0.7891	0.1908	0.4537	0.7468
35M unfrozen MLP4 softmax+mask [929752d4] / 3Di-only	731	0.0459	0.2573	0.7011	0.0965	0.3497	0.6707	0.1367	0.3620	0.6348
35M unfrozen MLP4 softmax+mask [929752d4] / AA+3Di	708	0.0403	0.2574	0.7032	0.0804	0.3517	0.6746	0.0968	0.3579	0.6357

Table S36: All 45 native clusterings compared with both structure-based Foldseek reference partitions. These scores compare partitions, rather than measuring agreement with SCOPe labels.

Configuration	Structure 3Di-only reference			Structure AA+3Di reference		
	ARI	AMI	NMI	ARI	AMI	NMI
MMseqs2 / AA	0.2202	0.4114	0.9289	0.2151	0.4228	0.9306
ProstT5-based Foldseek / 3Di-only	0.5318	0.7347	0.9397	0.5306	0.7391	0.9409
ProstT5-based Foldseek / AA+3Di	0.5149	0.7239	0.9386	0.5365	0.7491	0.9444
Structure / 3Di-only	1.0000	1.0000	1.0000	0.6484	0.8256	0.9651
Structure / AA+3Di	0.6484	0.8256	0.9651	1.0000	1.0000	1.0000
8M frozen MLP4 ST-Gumbel [87d2c509] / 3Di-only	0.3886	0.4406	0.9169	0.4229	0.4601	0.9201
8M frozen MLP4 ST-Gumbel [87d2c509] / AA+3Di	0.4851	0.6040	0.9350	0.4980	0.6199	0.9379
8M frozen MLP4 ST-argmax [d58f9e14] / 3Di-only	0.3365	0.3795	0.8943	0.3340	0.3905	0.8966
8M frozen MLP4 ST-argmax [d58f9e14] / AA+3Di	0.4032	0.4894	0.9050	0.4182	0.5123	0.9096
8M frozen MLP4 softmax [7b541c60] / 3Di-only	0.2240	0.2695	0.8654	0.2301	0.2749	0.8670
8M frozen MLP4 softmax [7b541c60] / AA+3Di	0.2863	0.4007	0.8836	0.2923	0.4166	0.8871
8M frozen linear ST-Gumbel [31bba998] / 3Di-only	0.3424	0.4083	0.9106	0.3445	0.4152	0.9120
8M frozen linear ST-Gumbel [31bba998] / AA+3Di	0.4257	0.5364	0.9240	0.4225	0.5491	0.9264
8M frozen linear ST-argmax [fbc46ff2] / 3Di-only	0.3571	0.3923	0.9076	0.3593	0.3974	0.9088
8M frozen linear ST-argmax [fbc46ff2] / AA+3Di	0.4397	0.5164	0.9201	0.4314	0.5231	0.9216
8M frozen linear softmax [1cf7945c] / 3Di-only	0.2180	0.2753	0.8679	0.2262	0.2788	0.8690
8M frozen linear softmax [1cf7945c] / AA+3Di	0.3028	0.4378	0.8926	0.2956	0.4440	0.8942
8M unfrozen MLP4 ST-Gumbel [d50ef818] / 3Di-only	0.1453	0.3536	0.8027	0.1660	0.3612	0.8058
8M unfrozen MLP4 ST-Gumbel [d50ef818] / AA+3Di	0.1437	0.3396	0.7956	0.1641	0.3541	0.8009
8M unfrozen MLP4 ST-argmax [650078f8] / 3Di-only	0.2039	0.4169	0.8700	0.2282	0.4317	0.8738
8M unfrozen MLP4 ST-argmax [650078f8] / AA+3Di	0.2460	0.4600	0.8846	0.2806	0.4852	0.8904
8M unfrozen MLP4 softmax [16132e1d] / 3Di-only	0.2860	0.3850	0.8805	0.3196	0.4011	0.8841
8M unfrozen MLP4 softmax [16132e1d] / AA+3Di	0.3468	0.4653	0.8938	0.3754	0.4847	0.8980
8M unfrozen MLP4 softmax+mask [9d792c09] / 3Di-only	0.3070	0.4374	0.8857	0.3396	0.4552	0.8897
8M unfrozen MLP4 softmax+mask [9d792c09] / AA+3Di	0.3332	0.5096	0.8972	0.3573	0.5322	0.9023
35M frozen MLP4 ST-Gumbel [72459d15] / 3Di-only	0.4480	0.6235	0.9320	0.4644	0.6349	0.9343
35M frozen MLP4 ST-Gumbel [72459d15] / AA+3Di	0.4571	0.6727	0.9373	0.4796	0.6921	0.9413
35M frozen MLP4 ST-argmax [9fceaf26] / 3Di-only	0.5424	0.6172	0.9375	0.5067	0.6159	0.9376

Table S37: All 45 native clusterings compared with both structure-based Foldseek reference partitions. These scores compare partitions, rather than measuring agreement with SCOPe labels. Continuation 2.

Configuration	Structure 3Di-only reference			Structure AA+3Di reference		
	ARI	AMI	NMI	ARI	AMI	NMI
35M frozen MLP4 ST-argmax [9fceaf26] / AA+3Di	0.5244	0.6495	0.9407	0.5295	0.6653	0.9437
35M frozen MLP4 softmax [0eb8395a] / 3Di-only	0.3016	0.4537	0.8672	0.3003	0.4606	0.8694
35M frozen MLP4 softmax [0eb8395a] / AA+3Di	0.3107	0.5253	0.8829	0.3363	0.5401	0.8870
35M frozen linear ST-Gumbel [4b1a9c9d] / 3Di-only	0.4764	0.6257	0.9348	0.4799	0.6314	0.9360
35M frozen linear ST-Gumbel [4b1a9c9d] / AA+3Di	0.5403	0.6858	0.9418	0.5426	0.6992	0.9445
35M frozen linear ST-argmax [f8f5cad4] / 3Di-only	0.5065	0.6248	0.9320	0.4921	0.6305	0.9333
35M frozen linear ST-argmax [f8f5cad4] / AA+3Di	0.5485	0.6666	0.9378	0.5435	0.6763	0.9399
35M frozen linear softmax [c336a672] / 3Di-only	0.3887	0.4771	0.8959	0.3940	0.4892	0.8987
35M frozen linear softmax [c336a672] / AA+3Di	0.4576	0.6092	0.9203	0.4695	0.6248	0.9238
35M unfrozen MLP4 ST-Gumbel [f32e1543] / 3Di-only	0.0991	0.3924	0.7756	0.1161	0.4026	0.7802
35M unfrozen MLP4 ST-Gumbel [f32e1543] / AA+3Di	0.1213	0.3941	0.7874	0.1300	0.4045	0.7918
35M unfrozen MLP4 ST-argmax [39efc845] / 3Di-only	0.3411	0.5830	0.9079	0.3542	0.5989	0.9117
35M unfrozen MLP4 ST-argmax [39efc845] / AA+3Di	0.3728	0.6162	0.9167	0.3737	0.6341	0.9209
35M unfrozen MLP4 softmax [f7815ea1] / 3Di-only	0.1392	0.4492	0.8325	0.1514	0.4629	0.8373
35M unfrozen MLP4 softmax [f7815ea1] / AA+3Di	0.1518	0.4525	0.8271	0.1708	0.4705	0.8334
35M unfrozen MLP4 softmax+mask [929752d4] / 3Di-only	0.0693	0.3295	0.7182	0.0797	0.3383	0.7230
35M unfrozen MLP4 softmax+mask [929752d4] / AA+3Di	0.0590	0.3307	0.7209	0.0667	0.3365	0.7243

Table S38: Within-query Spearman correlations for all 45 native score configurations against all four TM-score normalizations: minimum, maximum, geometric and arithmetic means. Constant-score queries are excluded; all other query counts are 5,059. Structure / 3Di-only: 1 excluded query.

Configuration	Mean				Median			
	Min	Max	Geom.	Arith.	Min	Max	Geom.	Arith.
MMseqs2 / AA	0.0432	0.0317	0.0407	0.0328	-0.0047	0.0643	0.0050	0.0108
ProstT5-based Foldseek / 3Di-only	0.2570	0.2130	0.2972	0.2957	0.2102	0.2019	0.2617	0.2730
ProstT5-based Foldseek / AA+3Di	0.2213	0.2155	0.2667	0.2715	0.1704	0.1972	0.2203	0.2363
Structure / 3Di-only	0.2274	0.1989	0.2658	0.2664	0.1883	0.1837	0.2366	0.2476
Structure / AA+3Di	0.1947	0.1957	0.2357	0.2411	0.1508	0.1740	0.1941	0.2076
8M frozen MLP4 ST-Gumbel [87d2c509] / 3Di-only	0.1510	0.0740	0.1523	0.1437	0.1183	0.0542	0.1196	0.1139
8M frozen MLP4 ST-Gumbel [87d2c509] / AA+3Di	0.1614	0.0784	0.1628	0.1532	0.1213	0.0555	0.1234	0.1178
8M frozen MLP4 ST-argmax [d58f9e14] / 3Di-only	0.1038	0.0547	0.1044	0.0991	0.0708	0.0312	0.0686	0.0641
8M frozen MLP4 ST-argmax [d58f9e14] / AA+3Di	0.1139	0.0610	0.1153	0.1094	0.0757	0.0333	0.0731	0.0682
8M frozen MLP4 softmax [7b541c60] / 3Di-only	0.1377	0.0547	0.1321	0.1204	0.0994	0.0296	0.0861	0.0758
8M frozen MLP4 softmax [7b541c60] / AA+3Di	0.1508	0.0596	0.1449	0.1321	0.1019	0.0301	0.0905	0.0795
8M frozen linear ST-Gumbel [31bba998] / 3Di-only	0.1438	0.0595	0.1402	0.1307	0.1060	0.0460	0.1051	0.0999
8M frozen linear ST-Gumbel [31bba998] / AA+3Di	0.1574	0.0714	0.1569	0.1470	0.1051	0.0513	0.1113	0.1086
8M frozen linear ST-argmax [fbc46ff2] / 3Di-only	0.1294	0.0499	0.1238	0.1138	0.1070	0.0360	0.0985	0.0929
8M frozen linear ST-argmax [fbc46ff2] / AA+3Di	0.1387	0.0637	0.1375	0.1279	0.1032	0.0444	0.1018	0.0958
8M frozen linear softmax [1cf7945c] / 3Di-only	0.1124	0.0365	0.1030	0.0945	0.0952	0.0200	0.0792	0.0698
8M frozen linear softmax [1cf7945c] / AA+3Di	0.1316	0.0451	0.1227	0.1122	0.1028	0.0234	0.0884	0.0776
8M unfrozen MLP4 ST-Gumbel [d50ef818] / 3Di-only	0.2917	0.1473	0.3024	0.2841	0.2811	0.1389	0.2760	0.2588
8M unfrozen MLP4 ST-Gumbel [d50ef818] / AA+3Di	0.2964	0.1449	0.3051	0.2856	0.2874	0.1332	0.2734	0.2579
8M unfrozen MLP4 ST-argmax [650078f8] / 3Di-only	0.2018	0.0954	0.2015	0.1853	0.1589	0.0766	0.1550	0.1423
8M unfrozen MLP4 ST-argmax [650078f8] / AA+3Di	0.2205	0.1043	0.2210	0.2033	0.1869	0.0855	0.1775	0.1636
8M unfrozen MLP4 softmax [16132e1d] / 3Di-only	0.1509	0.0943	0.1611	0.1529	0.0954	0.0668	0.1088	0.1016
8M unfrozen MLP4 softmax [16132e1d] / AA+3Di	0.1732	0.0997	0.1818	0.1715	0.1057	0.0700	0.1205	0.1145
8M unfrozen MLP4 softmax+mask [9d792c09] / 3Di-only	0.1578	0.0766	0.1605	0.1511	0.1322	0.0603	0.1379	0.1282
8M unfrozen MLP4 softmax+mask [9d792c09] / AA+3Di	0.1669	0.0799	0.1695	0.1591	0.1341	0.0601	0.1431	0.1321
35M frozen MLP4 ST-Gumbel [72459d15] / 3Di-only	0.1775	0.1009	0.1839	0.1735	0.1360	0.0675	0.1324	0.1191
35M frozen MLP4 ST-Gumbel [72459d15] / AA+3Di	0.1927	0.1080	0.1998	0.1887	0.1445	0.0699	0.1427	0.1330
35M frozen MLP4 ST-argmax [9fceaf26] / 3Di-only	0.1001	0.0755	0.1089	0.1066	0.0522	0.0385	0.0532	0.0520

Table S39: Within-query Spearman correlations for all 45 native score configurations against all four TM-score normalizations: minimum, maximum, geometric and arithmetic means. Constant-score queries are excluded; all other query counts are 5,059. Structure / 3Di-only: 1 excluded query. Continuation 2.

Configuration	Mean				Median			
	Min	Max	Geom.	Arith.	Min	Max	Geom.	Arith.
35M frozen MLP4 ST-argmax [9fceaf26] / AA+3Di	0.1043	0.0810	0.1150	0.1130	0.0547	0.0397	0.0580	0.0560
35M frozen MLP4 softmax [0eb8395a] / 3Di-only	0.1324	0.0673	0.1329	0.1249	0.1044	0.0470	0.0998	0.0924
35M frozen MLP4 softmax [0eb8395a] / AA+3Di	0.1447	0.0740	0.1460	0.1369	0.1034	0.0517	0.1017	0.0951
35M frozen linear ST-Gumbel [4b1a9c9d] / 3Di-only	0.1497	0.0821	0.1527	0.1451	0.1099	0.0635	0.1103	0.1057
35M frozen linear ST-Gumbel [4b1a9c9d] / AA+3Di	0.1654	0.0880	0.1686	0.1594	0.1220	0.0667	0.1250	0.1175
35M frozen linear ST-argmax [f8f5cad4] / 3Di-only	0.1619	0.0860	0.1651	0.1554	0.1325	0.0612	0.1273	0.1185
35M frozen linear ST-argmax [f8f5cad4] / AA+3Di	0.1560	0.0939	0.1631	0.1554	0.1199	0.0661	0.1202	0.1135
35M frozen linear softmax [c336a672] / 3Di-only	0.1133	0.0560	0.1122	0.1049	0.0761	0.0326	0.0692	0.0633
35M frozen linear softmax [c336a672] / AA+3Di	0.1300	0.0626	0.1292	0.1201	0.0805	0.0403	0.0782	0.0732
35M unfrozen MLP4 ST-Gumbel [f32e1543] / 3Di-only	0.2747	0.1636	0.2959	0.2818	0.2269	0.1596	0.2670	0.2545
35M unfrozen MLP4 ST-Gumbel [f32e1543] / AA+3Di	0.2725	0.1676	0.2953	0.2827	0.2256	0.1649	0.2602	0.2529
35M unfrozen MLP4 ST-argmax [39efc845] / 3Di-only	0.1893	0.1393	0.2119	0.2074	0.1207	0.1029	0.1443	0.1426
35M unfrozen MLP4 ST-argmax [39efc845] / AA+3Di	0.1840	0.1417	0.2082	0.2051	0.1145	0.0992	0.1340	0.1349
35M unfrozen MLP4 softmax [f7815ea1] / 3Di-only	0.2355	0.1421	0.2555	0.2433	0.1720	0.1148	0.2074	0.1983
35M unfrozen MLP4 softmax [f7815ea1] / AA+3Di	0.2539	0.1417	0.2707	0.2557	0.1885	0.1130	0.2216	0.2084
35M unfrozen MLP4 softmax+mask [929752d4] / 3Di-only	0.3326	0.1576	0.3423	0.3169	0.2765	0.1572	0.3119	0.3041
35M unfrozen MLP4 softmax+mask [929752d4] / AA+3Di	0.3397	0.1560	0.3476	0.3208	0.2873	0.1547	0.3169	0.3063

O SUPPLEMENTAL FIGURES

The figure guide at the start of the appendix identifies each comparison series. All plots use the model inventory and evaluation definitions given above.

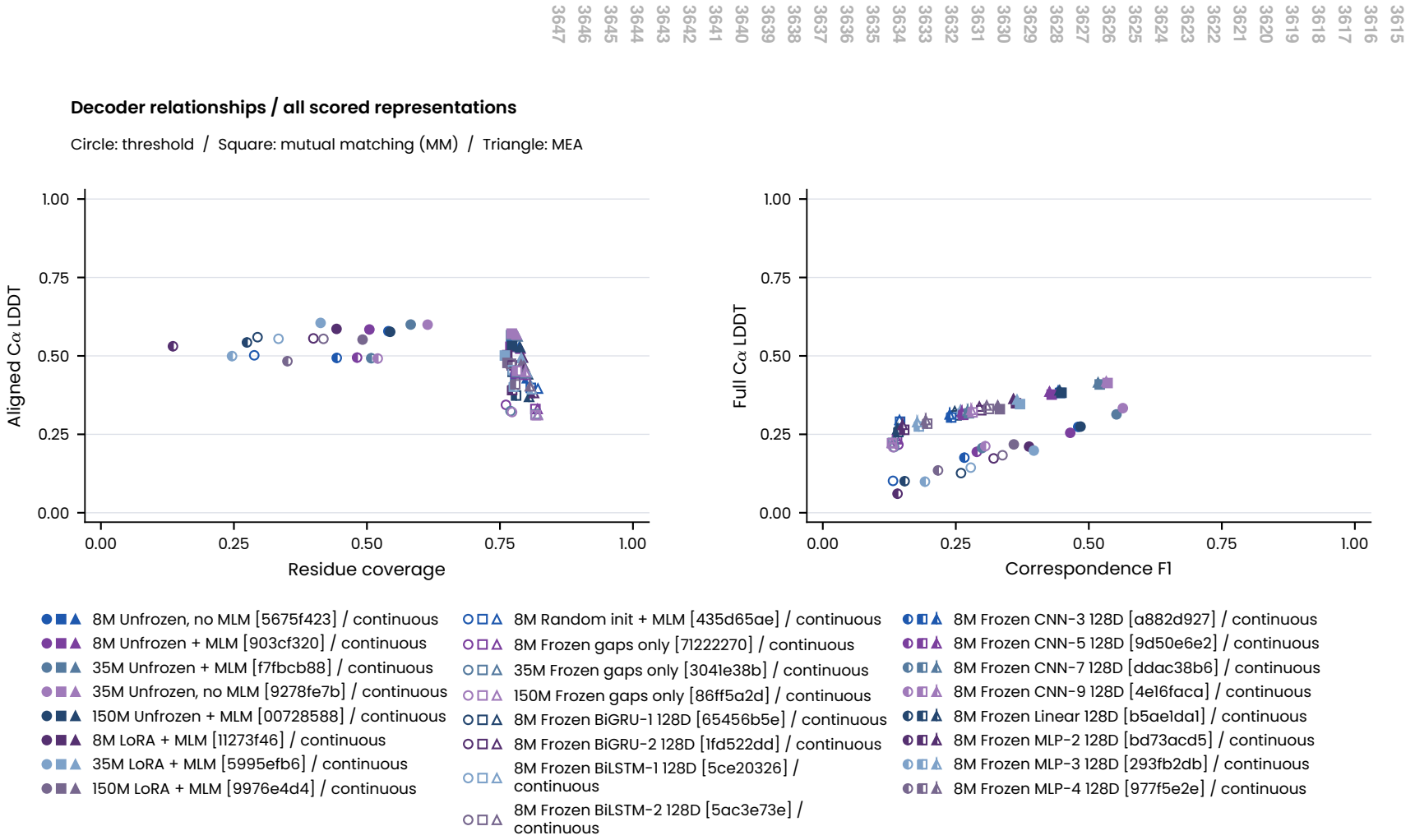


Figure S5: Coverage, correspondence and geometric agreement for every representation. Unconnected markers identify threshold (circle), mutual matching (square), and MEA (triangle). Color and fill identify the checkpoint, as shown in the key. Both axes use 0–1 scores. The companion aggregate matrices report all four LDDT endpoints, F1 and coverage under each decoder, so no separate repeated profile pages are needed.

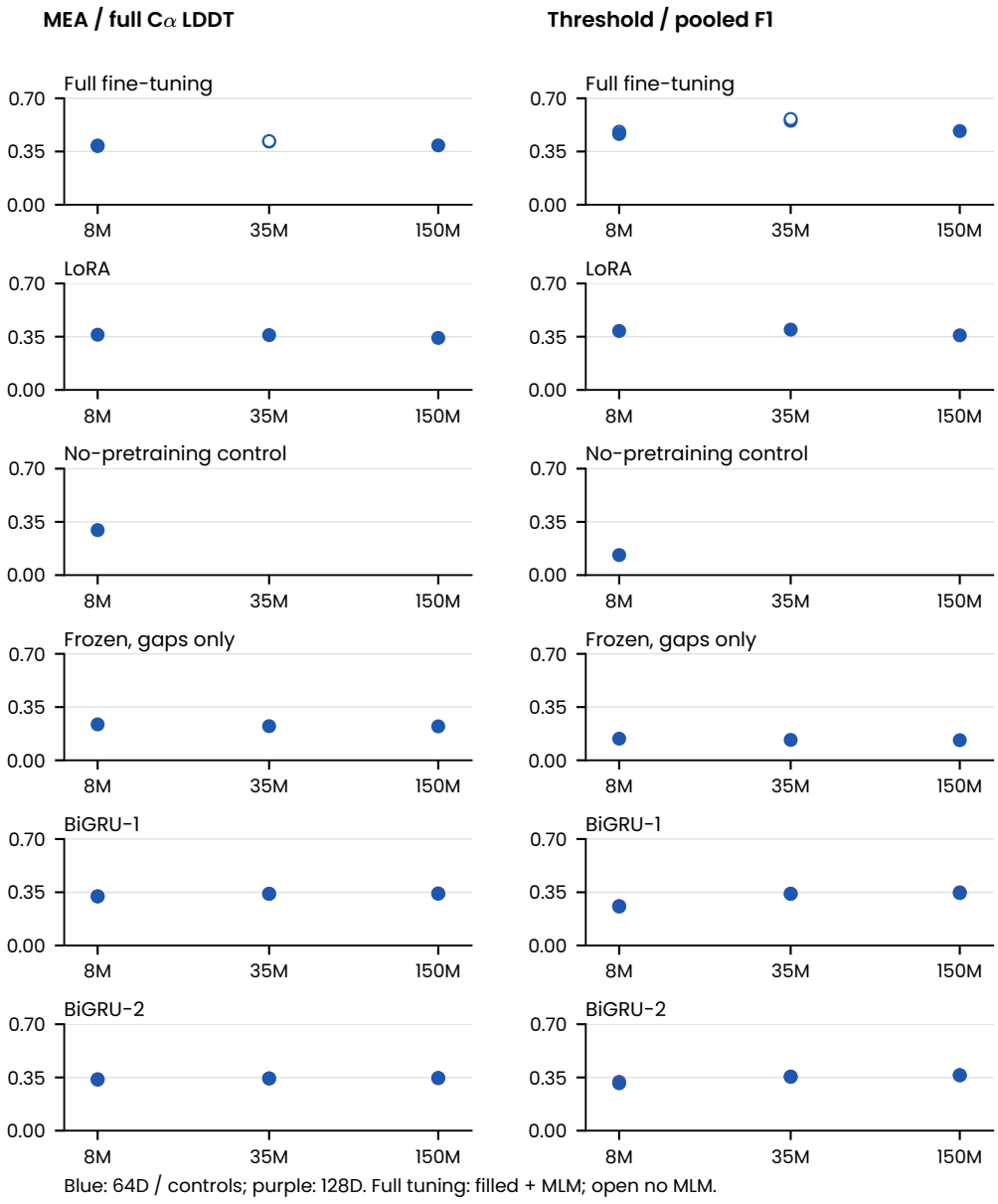


Figure S6: Backbone-size comparisons for every continuous ESM-2 checkpoint, grouped by architecture or adaptation regime. Every configuration is plotted separately, including coincident points; no best-adapter selection or averaging is performed. MLM and no-MLM use filled and open markers; exact configuration scores and checkpoint IDs appear in the aggregate matrices.

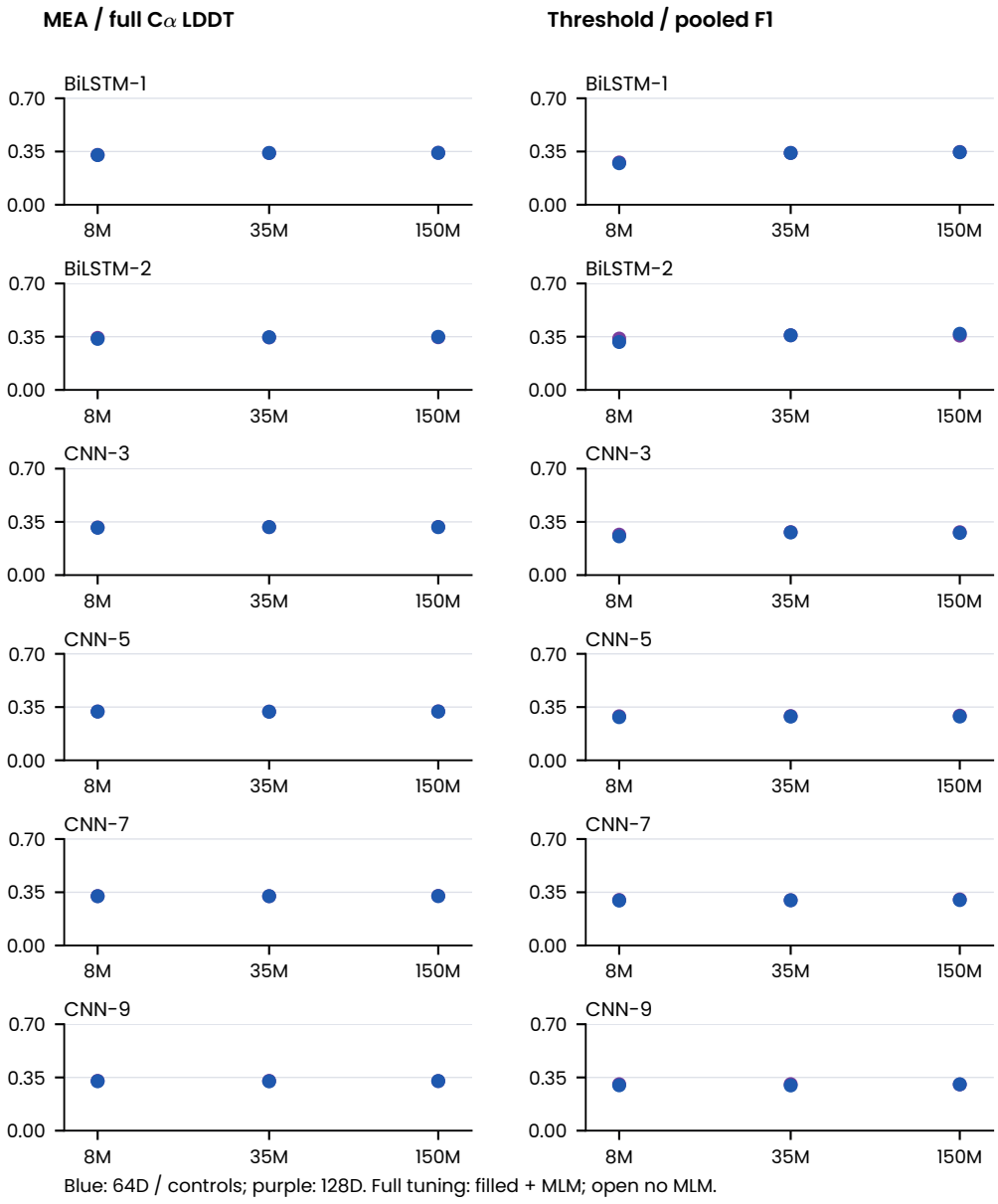


Figure S7: Backbone-size comparisons for every continuous ESM-2 checkpoint, grouped by architecture or adaptation regime. Every configuration is plotted separately, including coincident points; no best-adapter selection or averaging is performed. MLM and no-MLM use filled and open markers; exact configuration scores and checkpoint IDs appear in the aggregate matrices.

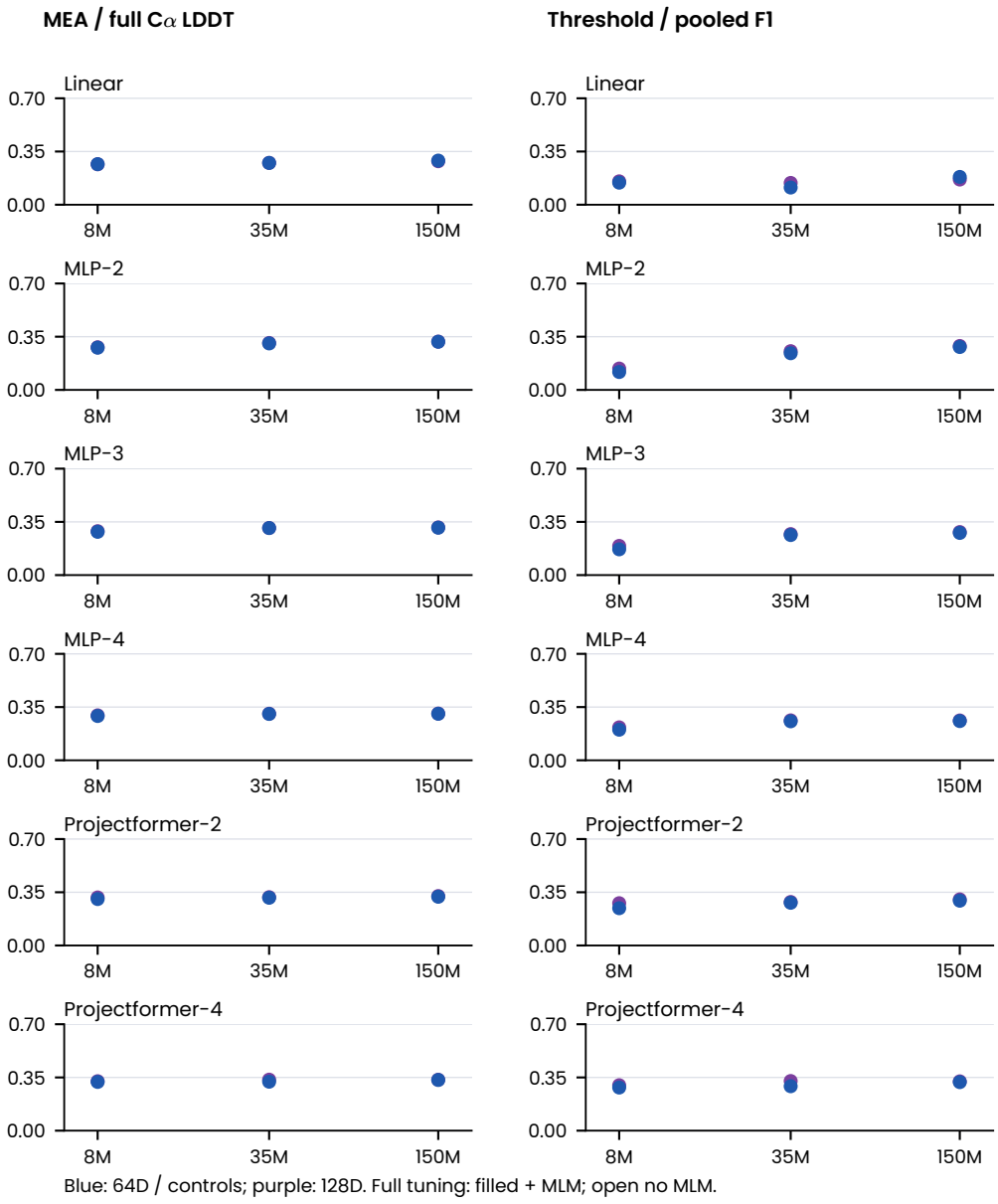


Figure S8: Backbone-size comparisons for every continuous ESM-2 checkpoint, grouped by architecture or adaptation regime. Every configuration is plotted separately, including coincident points; no best-adapter selection or averaging is performed. MLM and no-MLM use filled and open markers; exact configuration scores and checkpoint IDs appear in the aggregate matrices.

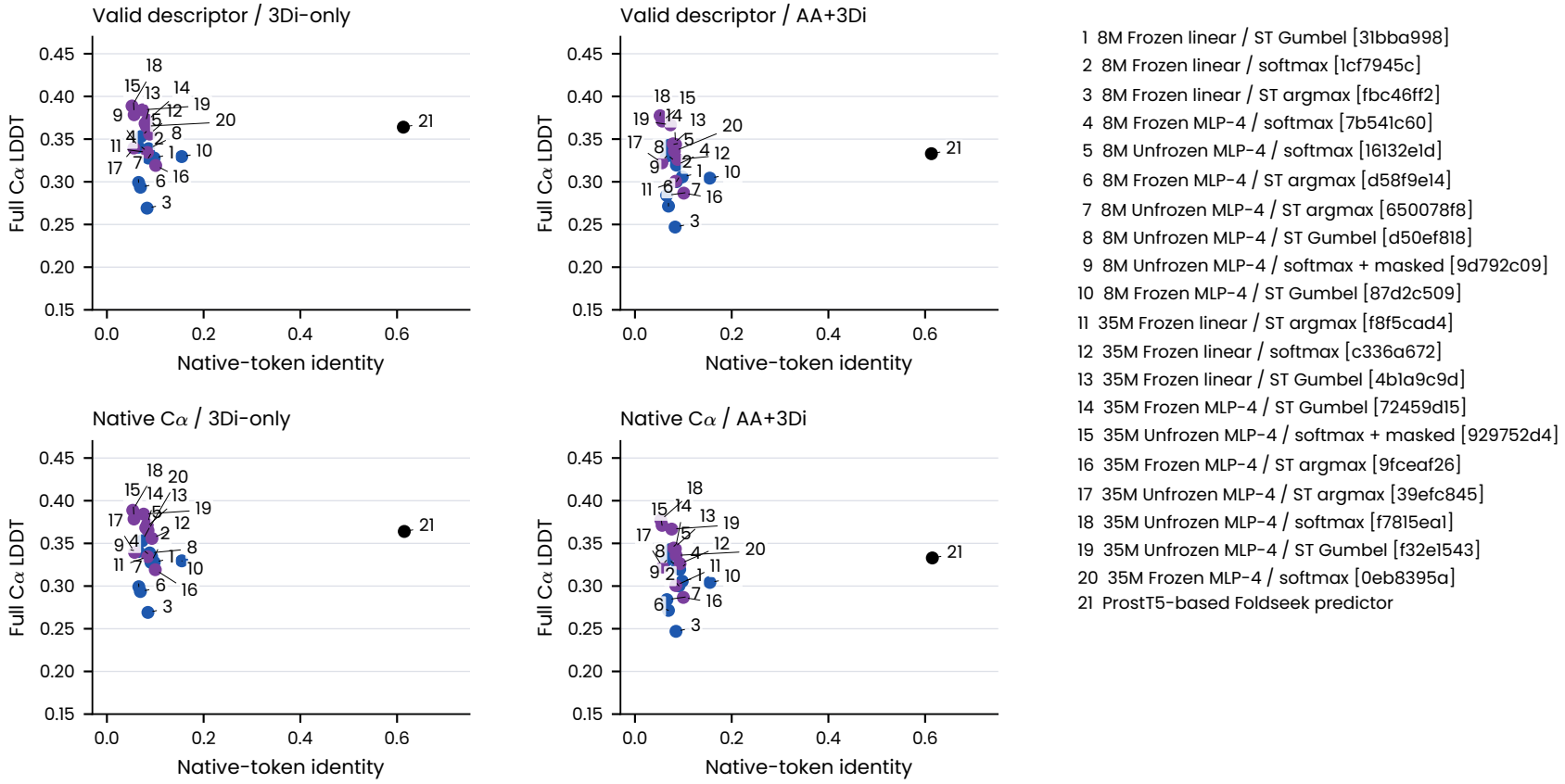


Figure S9: Native-token identity and native Foldseek LDDT for all 21 predictors, both masks and both scoring modes. Numbers identify the same predictor across panels; the adjacent key retains each checkpoint identity. Blue: 8M; purple: 35M; black: the ProstT5-based Foldseek predictor. Every original point is retained.

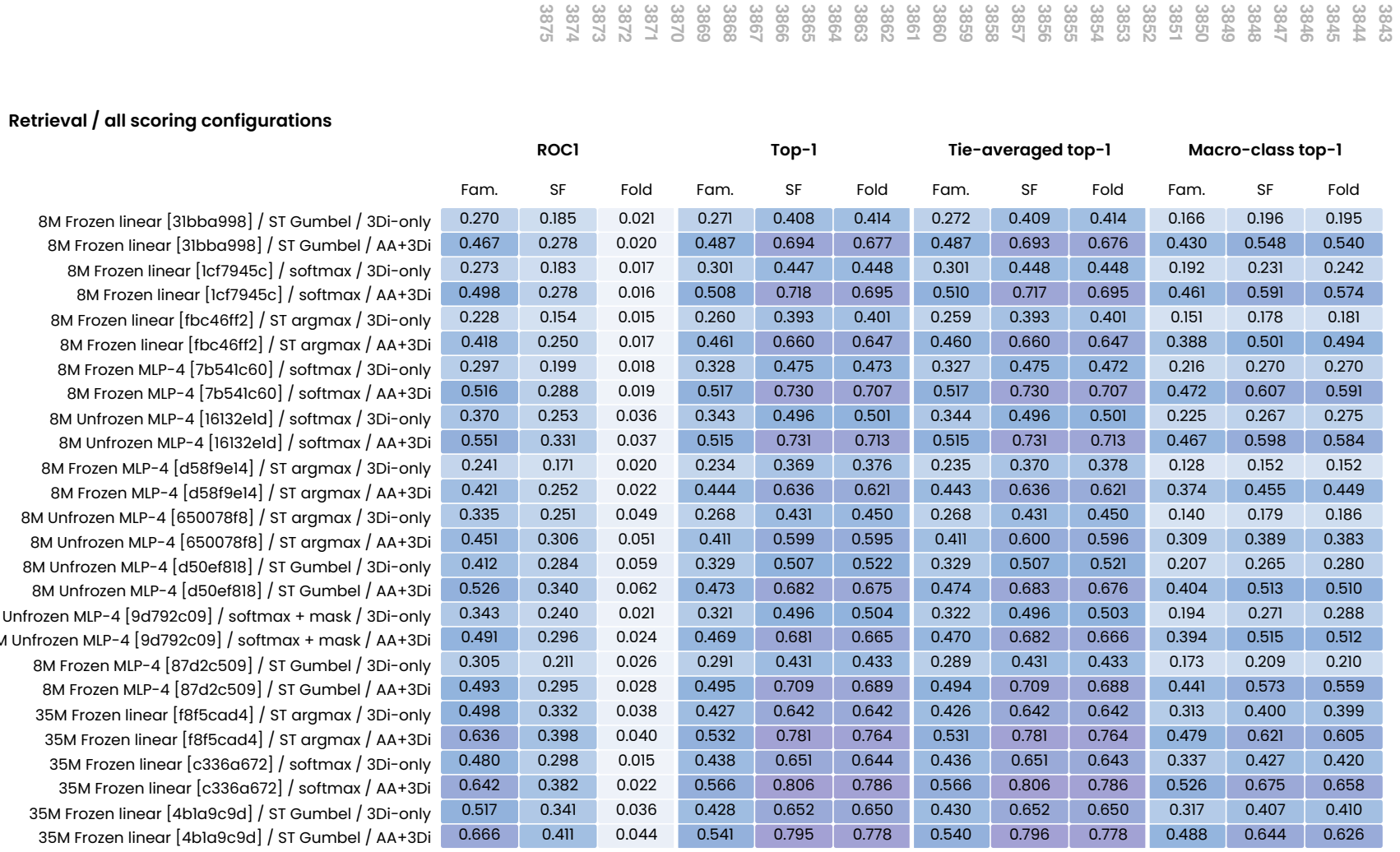


Figure S10: Complete retrieval comparison. Each metric is reported at family (Fam.), superfamily (SF), and fold levels. Row labels specify both Foldseek scoring modes for every learned, structure-based and ProstT5-based configuration. All four TM-align score normalizations and MMseqs2 are included.

3908 3907 3906 3905 3904 3903 3902 3901 3900 3899 3898 3897 3896 3895 3894 3893 3892 3891 3890 3889 3888 3887 3886 3885 3884 3883 3882 3881 3880 3879 3878 3877 3876

Retrieval / all scoring configurations

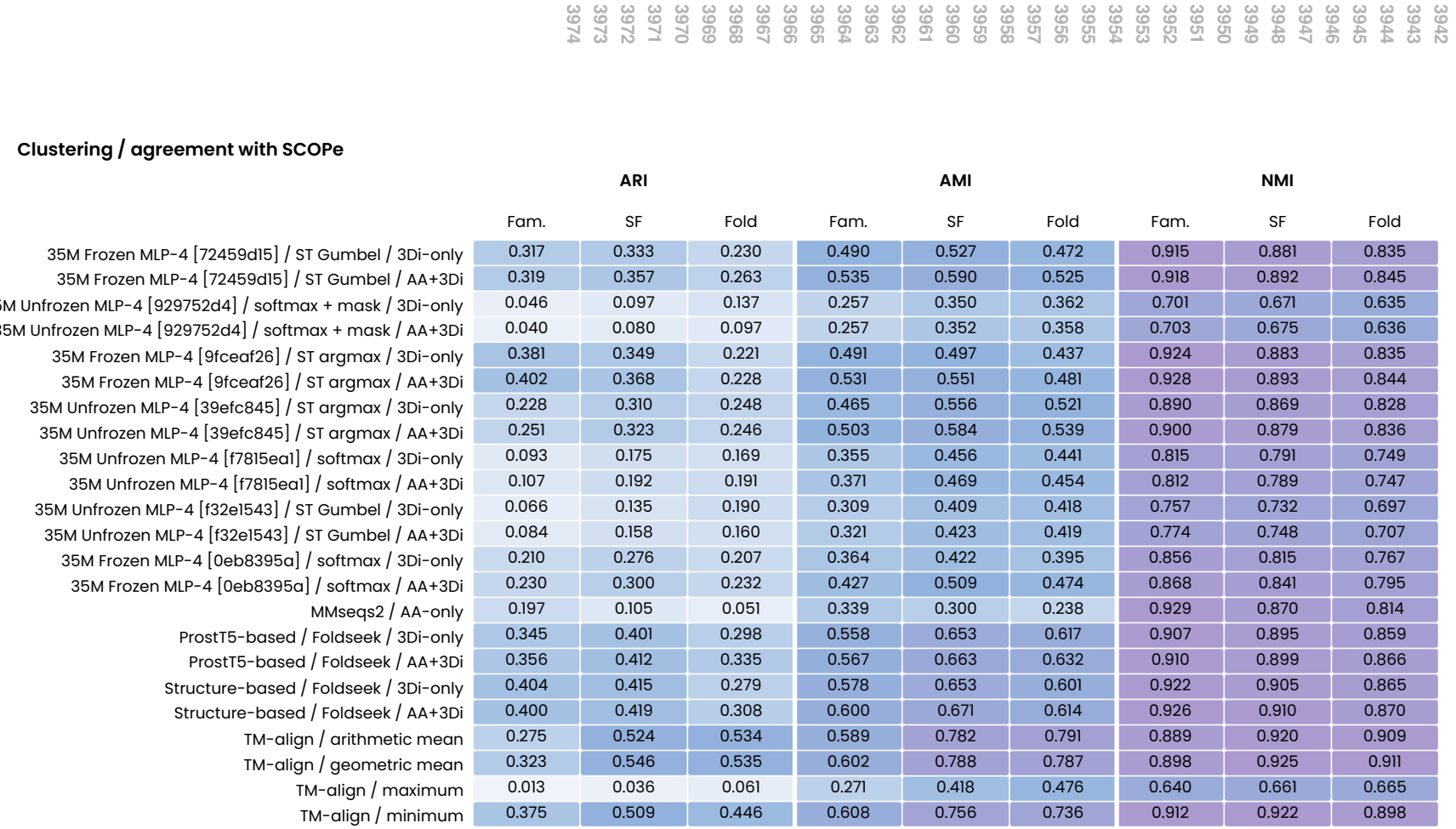
	ROC1			Top-1			Tie-averaged top-1			Macro-class top-1		
	Fam.	SF	Fold	Fam.	SF	Fold	Fam.	SF	Fold	Fam.	SF	Fold
35M Frozen MLP-4 [72459d15] / ST Gumbel / 3Di-only	0.527	0.349	0.061	0.430	0.647	0.645	0.431	0.648	0.645	0.321	0.406	0.406
35M Frozen MLP-4 [72459d15] / ST Gumbel / AA+3Di	0.679	0.420	0.066	0.547	0.800	0.782	0.547	0.801	0.782	0.502	0.656	0.636
35M Unfrozen MLP-4 [929752d4] / softmax + mask / 3Di-only	0.624	0.424	0.083	0.448	0.694	0.697	0.448	0.694	0.697	0.352	0.457	0.463
35M Unfrozen MLP-4 [929752d4] / softmax + mask / AA+3Di	0.689	0.467	0.081	0.525	0.778	0.767	0.526	0.779	0.767	0.469	0.615	0.606
35M Frozen MLP-4 [9fceaf26] / ST argmax / 3Di-only	0.457	0.315	0.045	0.371	0.585	0.584	0.371	0.585	0.584	0.245	0.324	0.320
35M Frozen MLP-4 [9fceaf26] / ST argmax / AA+3Di	0.612	0.385	0.050	0.490	0.734	0.718	0.491	0.734	0.718	0.419	0.532	0.513
35M Unfrozen MLP-4 [39efc845] / ST argmax / 3Di-only	0.589	0.416	0.086	0.414	0.660	0.665	0.414	0.660	0.665	0.295	0.397	0.409
35M Unfrozen MLP-4 [39efc845] / ST argmax / AA+3Di	0.682	0.466	0.087	0.501	0.753	0.745	0.501	0.753	0.745	0.423	0.555	0.548
35M Unfrozen MLP-4 [f7815ea1] / softmax / 3Di-only	0.625	0.425	0.071	0.466	0.713	0.712	0.465	0.713	0.712	0.371	0.487	0.497
35M Unfrozen MLP-4 [f7815ea1] / softmax / AA+3Di	0.711	0.476	0.073	0.537	0.795	0.784	0.537	0.795	0.784	0.487	0.647	0.639
35M Unfrozen MLP-4 [f32e1543] / ST Gumbel / 3Di-only	0.640	0.453	0.100	0.463	0.710	0.719	0.463	0.710	0.719	0.369	0.474	0.491
35M Unfrozen MLP-4 [f32e1543] / ST Gumbel / AA+3Di	0.726	0.498	0.098	0.538	0.805	0.795	0.538	0.805	0.795	0.491	0.650	0.640
35M Frozen MLP-4 [0eb8395a] / softmax / 3Di-only	0.519	0.338	0.034	0.437	0.665	0.661	0.436	0.665	0.662	0.332	0.428	0.432
35M Frozen MLP-4 [0eb8395a] / softmax / AA+3Di	0.656	0.407	0.038	0.547	0.802	0.785	0.547	0.802	0.785	0.500	0.659	0.645
MMseqs2 / AA-only	0.376	0.153	0.001	0.524	0.709	0.673	0.524	0.709	0.672	0.494	0.622	0.605
ProstT5-based / Foldseek / 3Di-only	0.799	0.532	0.109	0.609	0.901	0.893	0.611	0.901	0.893	0.601	0.823	0.813
ProstT5-based / Foldseek / AA+3Di	0.835	0.578	0.121	0.629	0.923	0.917	0.629	0.923	0.917	0.626	0.863	0.845
Structure-based / Foldseek / 3Di-only	0.745	0.466	0.083	0.602	0.898	0.890	0.602	0.897	0.890	0.592	0.817	0.808
Structure-based / Foldseek / AA+3Di	0.818	0.535	0.096	0.627	0.928	0.921	0.628	0.928	0.921	0.627	0.871	0.851
TM-align / arithmetic mean	0.877	0.633	0.212	0.630	0.941	0.950	0.630	0.941	0.950	0.629	0.893	0.888
TM-align / geometric mean	0.876	0.636	0.220	0.631	0.942	0.950	0.631	0.942	0.950	0.631	0.895	0.890
TM-align / maximum	0.724	0.475	0.096	0.586	0.869	0.868	0.586	0.869	0.868	0.576	0.785	0.766
TM-align / minimum	0.842	0.593	0.201	0.631	0.931	0.943	0.631	0.931	0.943	0.634	0.874	0.877

Figure S11: Complete retrieval comparison. Each metric is reported at family (Fam.), superfamily (SF), and fold levels. Row labels specify both Foldseek scoring modes for every learned, structure-based and ProstT5-based configuration. All four TM-align score normalizations and MMseqs2 are included.

Clustering / agreement with SCOPe

	ARI			AMI			NMI		
	Fam.	SF	Fold	Fam.	SF	Fold	Fam.	SF	Fold
8M Frozen linear [31bba998] / ST Gumbel / 3Di-only	0.285	0.259	0.151	0.339	0.333	0.288	0.910	0.854	0.801
8M Frozen linear [31bba998] / ST Gumbel / AA+3Di	0.336	0.296	0.170	0.442	0.444	0.381	0.917	0.870	0.818
8M Frozen linear [1cf7945c] / softmax / 3Di-only	0.195	0.169	0.099	0.231	0.238	0.211	0.871	0.807	0.750
8M Frozen linear [1cf7945c] / softmax / AA+3Di	0.232	0.227	0.136	0.357	0.380	0.332	0.887	0.837	0.782
8M Frozen linear [fbc46ff2] / ST argmax / 3Di-only	0.302	0.237	0.131	0.333	0.317	0.270	0.908	0.850	0.796
8M Frozen linear [fbc46ff2] / ST argmax / AA+3Di	0.343	0.290	0.162	0.428	0.424	0.361	0.914	0.865	0.811
8M Frozen MLP-4 [7b541c60] / softmax / 3Di-only	0.172	0.167	0.106	0.223	0.230	0.207	0.869	0.803	0.748
8M Frozen MLP-4 [7b541c60] / softmax / AA+3Di	0.225	0.221	0.134	0.336	0.353	0.310	0.881	0.827	0.772
8M Unfrozen MLP-4 [16132e1d] / softmax / 3Di-only	0.211	0.241	0.188	0.323	0.336	0.310	0.879	0.823	0.773
8M Unfrozen MLP-4 [16132e1d] / softmax / AA+3Di	0.272	0.295	0.189	0.394	0.416	0.368	0.889	0.841	0.789
8M Frozen MLP-4 [d58f9e14] / ST argmax / 3Di-only	0.263	0.226	0.134	0.315	0.321	0.286	0.893	0.837	0.784
8M Frozen MLP-4 [d58f9e14] / ST argmax / AA+3Di	0.316	0.301	0.191	0.406	0.432	0.387	0.899	0.853	0.803
8M Unfrozen MLP-4 [650078f8] / ST argmax / 3Di-only	0.143	0.208	0.184	0.339	0.383	0.368	0.863	0.817	0.773
8M Unfrozen MLP-4 [650078f8] / ST argmax / AA+3Di	0.180	0.231	0.181	0.375	0.427	0.398	0.876	0.835	0.789
8M Unfrozen MLP-4 [d50ef818] / ST Gumbel / 3Di-only	0.103	0.173	0.215	0.284	0.346	0.345	0.793	0.748	0.704
8M Unfrozen MLP-4 [d50ef818] / ST Gumbel / AA+3Di	0.102	0.167	0.182	0.276	0.344	0.346	0.788	0.743	0.699
8M Unfrozen MLP-4 [9d792c09] / softmax + mask / 3Di-only	0.231	0.261	0.200	0.355	0.394	0.366	0.879	0.833	0.785
8M Unfrozen MLP-4 [9d792c09] / softmax + mask / AA+3Di	0.255	0.289	0.201	0.418	0.473	0.432	0.887	0.850	0.802
8M Frozen MLP-4 [87d2c509] / ST Gumbel / 3Di-only	0.321	0.257	0.154	0.374	0.359	0.313	0.916	0.861	0.810
8M Frozen MLP-4 [87d2c509] / ST Gumbel / AA+3Di	0.393	0.346	0.201	0.512	0.506	0.434	0.927	0.885	0.833
35M Frozen linear [f8f5cad4] / ST argmax / 3Di-only	0.375	0.354	0.222	0.498	0.534	0.476	0.917	0.882	0.835
35M Frozen linear [f8f5cad4] / ST argmax / AA+3Di	0.406	0.386	0.232	0.534	0.583	0.520	0.920	0.892	0.846
35M Frozen linear [c336a672] / softmax / 3Di-only	0.289	0.290	0.181	0.395	0.423	0.374	0.889	0.843	0.790
35M Frozen linear [c336a672] / softmax / AA+3Di	0.338	0.372	0.237	0.497	0.556	0.490	0.905	0.877	0.826
35M Frozen linear [4b1a9c9d] / ST Gumbel / 3Di-only	0.333	0.324	0.217	0.489	0.527	0.465	0.918	0.884	0.836
35M Frozen linear [4b1a9c9d] / ST Gumbel / AA+3Di	0.396	0.394	0.241	0.553	0.599	0.527	0.924	0.897	0.849

Figure S12: Complete clustering comparison. Each metric is reported at family (Fam.), superfamily (SF), and fold levels. Row labels specify both Foldseek scoring modes for every learned, structure-based and ProstT5-based configuration. TM-align references use native set cover on score graphs at the fixed 0.5 threshold; the full threshold sweep is separate. Pair precision, recall and F1 concern co-clustered proteins, not residue alignments.



Clustering / agreement with SCOPe

	ARI			AMI			NMI		
	Fam.	SF	Fold	Fam.	SF	Fold	Fam.	SF	Fold
35M Frozen MLP-4 [72459d15] / ST Gumbel / 3Di-only	0.317	0.333	0.230	0.490	0.527	0.472	0.915	0.881	0.835
35M Frozen MLP-4 [72459d15] / ST Gumbel / AA+3Di	0.319	0.357	0.263	0.535	0.590	0.525	0.918	0.892	0.845
35M Unfrozen MLP-4 [929752d4] / softmax + mask / 3Di-only	0.046	0.097	0.137	0.257	0.350	0.362	0.701	0.671	0.635
35M Unfrozen MLP-4 [929752d4] / softmax + mask / AA+3Di	0.040	0.080	0.097	0.257	0.352	0.358	0.703	0.675	0.636
35M Frozen MLP-4 [9fceaf26] / ST argmax / 3Di-only	0.381	0.349	0.221	0.491	0.497	0.437	0.924	0.883	0.835
35M Frozen MLP-4 [9fceaf26] / ST argmax / AA+3Di	0.402	0.368	0.228	0.531	0.551	0.481	0.928	0.893	0.844
35M Unfrozen MLP-4 [39efc845] / ST argmax / 3Di-only	0.228	0.310	0.248	0.465	0.556	0.521	0.890	0.869	0.828
35M Unfrozen MLP-4 [39efc845] / ST argmax / AA+3Di	0.251	0.323	0.246	0.503	0.584	0.539	0.900	0.879	0.836
35M Unfrozen MLP-4 [f7815ea1] / softmax / 3Di-only	0.093	0.175	0.169	0.355	0.456	0.441	0.815	0.791	0.749
35M Unfrozen MLP-4 [f7815ea1] / softmax / AA+3Di	0.107	0.192	0.191	0.371	0.469	0.454	0.812	0.789	0.747
35M Unfrozen MLP-4 [f32e1543] / ST Gumbel / 3Di-only	0.066	0.135	0.190	0.309	0.409	0.418	0.757	0.732	0.697
35M Unfrozen MLP-4 [f32e1543] / ST Gumbel / AA+3Di	0.084	0.158	0.160	0.321	0.423	0.419	0.774	0.748	0.707
35M Frozen MLP-4 [0eb8395a] / softmax / 3Di-only	0.210	0.276	0.207	0.364	0.422	0.395	0.856	0.815	0.767
35M Frozen MLP-4 [0eb8395a] / softmax / AA+3Di	0.230	0.300	0.232	0.427	0.509	0.474	0.868	0.841	0.795
MMseqs2 / AA-only	0.197	0.105	0.051	0.339	0.300	0.238	0.929	0.870	0.814
ProstT5-based / Foldseek / 3Di-only	0.345	0.401	0.298	0.558	0.653	0.617	0.907	0.895	0.859
ProstT5-based / Foldseek / AA+3Di	0.356	0.412	0.335	0.567	0.663	0.632	0.910	0.899	0.866
Structure-based / Foldseek / 3Di-only	0.404	0.415	0.279	0.578	0.653	0.601	0.922	0.905	0.865
Structure-based / Foldseek / AA+3Di	0.400	0.419	0.308	0.600	0.671	0.614	0.926	0.910	0.870
TM-align / arithmetic mean	0.275	0.524	0.534	0.589	0.782	0.791	0.889	0.920	0.909
TM-align / geometric mean	0.323	0.546	0.535	0.602	0.788	0.787	0.898	0.925	0.911
TM-align / maximum	0.013	0.036	0.061	0.271	0.418	0.476	0.640	0.661	0.665
TM-align / minimum	0.375	0.509	0.446	0.608	0.756	0.736	0.912	0.922	0.898

Figure S13: Complete clustering comparison. Each metric is reported at family (Fam.), superfamily (SF), and fold levels. Row labels specify both Foldseek scoring modes for every learned, structure-based and ProstT5-based configuration. TM-align references use native set cover on score graphs at the fixed 0.5 threshold; the full threshold sweep is separate. Pair precision, recall and F1 concern co-clustered proteins, not residue alignments.

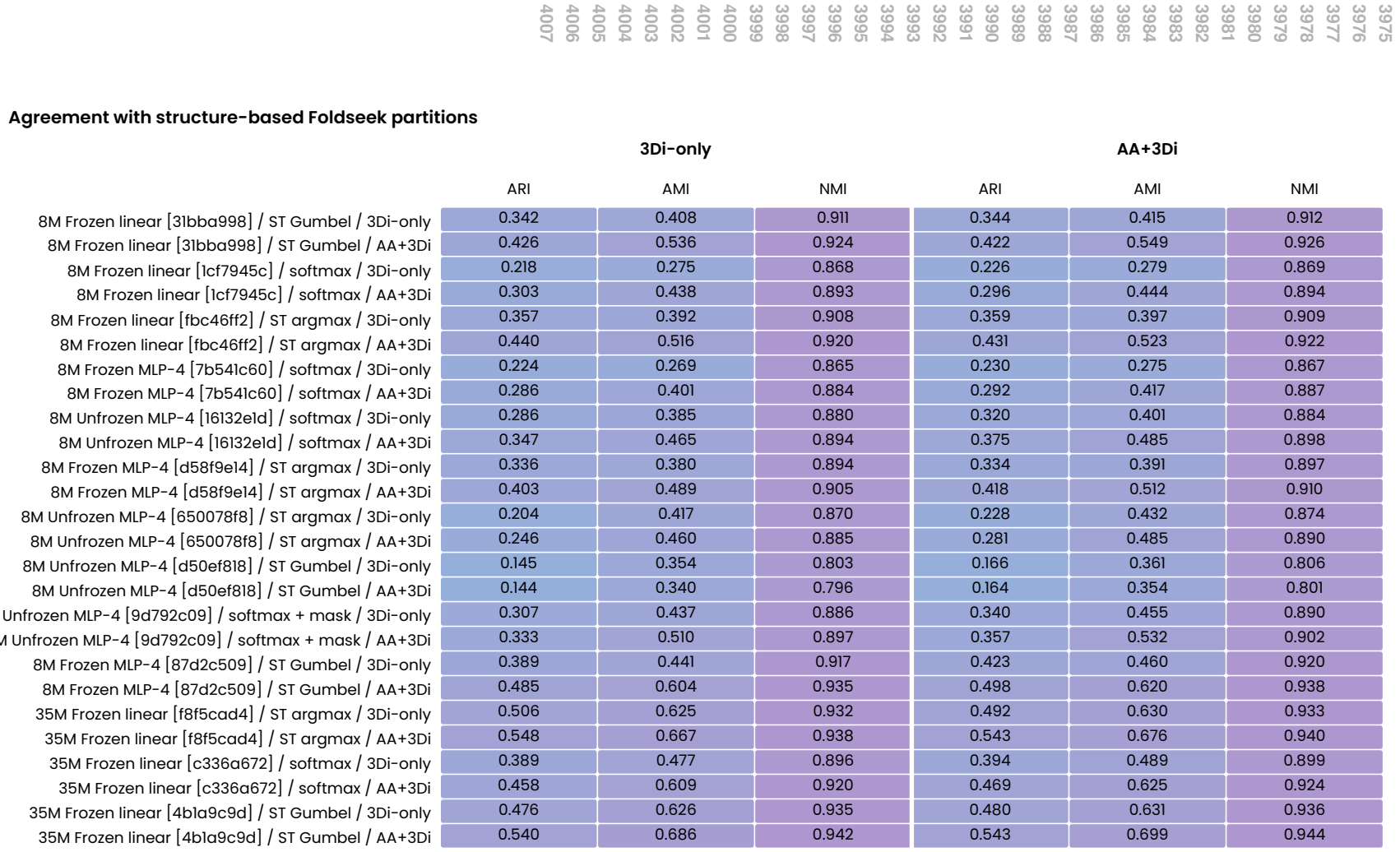


Figure S14: Every native clustering compared with both structure-based Foldseek reference partitions. Row modes identify the evaluated clustering; column groups identify the reference. The color range is $[-1, 1]$.



Figure S15: Every native clustering compared with both structure-based Foldseek reference partitions. Row modes identify the evaluated clustering; column groups identify the reference. The color range is $[-1, 1]$.

4073 4072 4071 4070 4069 4068 4067 4066 4065 4064 4063 4062 4061 4060 4059 4058 4057 4056 4055 4054 4053 4052 4051 4050 4049 4048 4047 4046 4045 4044 4043 4042 4041

Native score rankings / query-wise Spearman

	Mean				Median			
	Min	Max	Geom.	Arith.	Min	Max	Geom.	Arith.
8M Frozen linear [31bba998] / ST Gumbel / 3Di-only	0.144	0.060	0.140	0.131	0.106	0.046	0.105	0.100
8M Frozen linear [31bba998] / ST Gumbel / AA+3Di	0.157	0.071	0.157	0.147	0.105	0.051	0.111	0.109
8M Frozen linear [1cf7945c] / softmax / 3Di-only	0.112	0.036	0.103	0.094	0.095	0.020	0.079	0.070
8M Frozen linear [1cf7945c] / softmax / AA+3Di	0.132	0.045	0.123	0.112	0.103	0.023	0.088	0.078
8M Frozen linear [fbc46ff2] / ST argmax / 3Di-only	0.129	0.050	0.124	0.114	0.107	0.036	0.099	0.093
8M Frozen linear [fbc46ff2] / ST argmax / AA+3Di	0.139	0.064	0.138	0.128	0.103	0.044	0.102	0.096
8M Frozen MLP-4 [7b541c60] / softmax / 3Di-only	0.138	0.055	0.132	0.120	0.099	0.030	0.086	0.076
8M Frozen MLP-4 [7b541c60] / softmax / AA+3Di	0.151	0.060	0.145	0.132	0.102	0.030	0.091	0.080
8M Unfrozen MLP-4 [16132e1d] / softmax / 3Di-only	0.151	0.094	0.161	0.153	0.095	0.067	0.109	0.102
8M Unfrozen MLP-4 [16132e1d] / softmax / AA+3Di	0.173	0.100	0.182	0.172	0.106	0.070	0.120	0.115
8M Frozen MLP-4 [d58f9e14] / ST argmax / 3Di-only	0.104	0.055	0.104	0.099	0.071	0.031	0.069	0.064
8M Frozen MLP-4 [d58f9e14] / ST argmax / AA+3Di	0.114	0.061	0.115	0.109	0.076	0.033	0.073	0.068
8M Unfrozen MLP-4 [650078f8] / ST argmax / 3Di-only	0.202	0.095	0.201	0.185	0.159	0.077	0.155	0.142
8M Unfrozen MLP-4 [650078f8] / ST argmax / AA+3Di	0.220	0.104	0.221	0.203	0.187	0.086	0.177	0.164
8M Unfrozen MLP-4 [d50ef818] / ST Gumbel / 3Di-only	0.292	0.147	0.302	0.284	0.281	0.139	0.276	0.259
8M Unfrozen MLP-4 [d50ef818] / ST Gumbel / AA+3Di	0.296	0.145	0.305	0.286	0.287	0.133	0.273	0.258
8M Unfrozen MLP-4 [9d792c09] / softmax + mask / 3Di-only	0.158	0.077	0.160	0.151	0.132	0.060	0.138	0.128
8M Unfrozen MLP-4 [9d792c09] / softmax + mask / AA+3Di	0.167	0.080	0.170	0.159	0.134	0.060	0.143	0.132
8M Frozen MLP-4 [87d2c509] / ST Gumbel / 3Di-only	0.151	0.074	0.152	0.144	0.118	0.054	0.120	0.114
8M Frozen MLP-4 [87d2c509] / ST Gumbel / AA+3Di	0.161	0.078	0.163	0.153	0.121	0.056	0.123	0.118
35M Frozen linear [f8f5cad4] / ST argmax / 3Di-only	0.162	0.086	0.165	0.155	0.132	0.061	0.127	0.119
35M Frozen linear [f8f5cad4] / ST argmax / AA+3Di	0.156	0.094	0.163	0.155	0.120	0.066	0.120	0.114
35M Frozen linear [c336a672] / softmax / 3Di-only	0.113	0.056	0.112	0.105	0.076	0.033	0.069	0.063
35M Frozen linear [c336a672] / softmax / AA+3Di	0.130	0.063	0.129	0.120	0.080	0.040	0.078	0.073
35M Frozen linear [4b1a9c9d] / ST Gumbel / 3Di-only	0.150	0.082	0.153	0.145	0.110	0.063	0.110	0.106
35M Frozen linear [4b1a9c9d] / ST Gumbel / AA+3Di	0.165	0.088	0.169	0.159	0.122	0.067	0.125	0.117

Figure S16: Mean and median query-wise Spearman correlations for every native scoring configuration and all four TM-score normalizations. Constant queries retain the saved exclusion policy. The color range is $[-1, 1]$.

4106 4105 4104 4103 4102 4101 4100 4099 4098 4097 4096 4095 4094 4093 4092 4091 4090 4089 4088 4087 4086 4085 4084 4083 4082 4081 4080 4079 4078 4077 4076 4075 4074

Native score rankings / query-wise Spearman

	Mean				Median			
	Min	Max	Geom.	Arith.	Min	Max	Geom.	Arith.
35M Frozen MLP-4 [72459d15] / ST Gumbel / 3Di-only	0.177	0.101	0.184	0.174	0.136	0.067	0.132	0.119
35M Frozen MLP-4 [72459d15] / ST Gumbel / AA+3Di	0.193	0.108	0.200	0.189	0.144	0.070	0.143	0.133
35M Unfrozen MLP-4 [929752d4] / softmax + mask / 3Di-only	0.333	0.158	0.342	0.317	0.276	0.157	0.312	0.304
35M Unfrozen MLP-4 [929752d4] / softmax + mask / AA+3Di	0.340	0.156	0.348	0.321	0.287	0.155	0.317	0.306
35M Frozen MLP-4 [9fceaf26] / ST argmax / 3Di-only	0.100	0.076	0.109	0.107	0.052	0.039	0.053	0.052
35M Frozen MLP-4 [9fceaf26] / ST argmax / AA+3Di	0.104	0.081	0.115	0.113	0.055	0.040	0.058	0.056
35M Unfrozen MLP-4 [39efc845] / ST argmax / 3Di-only	0.189	0.139	0.212	0.207	0.121	0.103	0.144	0.143
35M Unfrozen MLP-4 [39efc845] / ST argmax / AA+3Di	0.184	0.142	0.208	0.205	0.114	0.099	0.134	0.135
35M Unfrozen MLP-4 [f7815ea1] / softmax / 3Di-only	0.235	0.142	0.255	0.243	0.172	0.115	0.207	0.198
35M Unfrozen MLP-4 [f7815ea1] / softmax / AA+3Di	0.254	0.142	0.271	0.256	0.189	0.113	0.222	0.208
35M Unfrozen MLP-4 [f32e1543] / ST Gumbel / 3Di-only	0.275	0.164	0.296	0.282	0.227	0.160	0.267	0.255
35M Unfrozen MLP-4 [f32e1543] / ST Gumbel / AA+3Di	0.273	0.168	0.295	0.283	0.226	0.165	0.260	0.253
35M Frozen MLP-4 [0eb8395a] / softmax / 3Di-only	0.132	0.067	0.133	0.125	0.104	0.047	0.100	0.092
35M Frozen MLP-4 [0eb8395a] / softmax / AA+3Di	0.145	0.074	0.146	0.137	0.103	0.052	0.102	0.095
MMseqs2 / AA-only	0.043	0.032	0.041	0.033	-0.005	0.064	0.005	0.011
ProstT5-based / Foldseek / 3Di-only	0.257	0.213	0.297	0.296	0.210	0.202	0.262	0.273
ProstT5-based / Foldseek / AA+3Di	0.221	0.215	0.267	0.271	0.170	0.197	0.220	0.236
Structure-based / Foldseek / 3Di-only	0.227	0.199	0.266	0.266	0.188	0.184	0.237	0.248
Structure-based / Foldseek / AA+3Di	0.195	0.196	0.236	0.241	0.151	0.174	0.194	0.208

Figure S17: Mean and median query-wise Spearman correlations for every native scoring configuration and all four TM-score normalizations. Constant queries retain the saved exclusion policy. The color range is $[-1, 1]$.

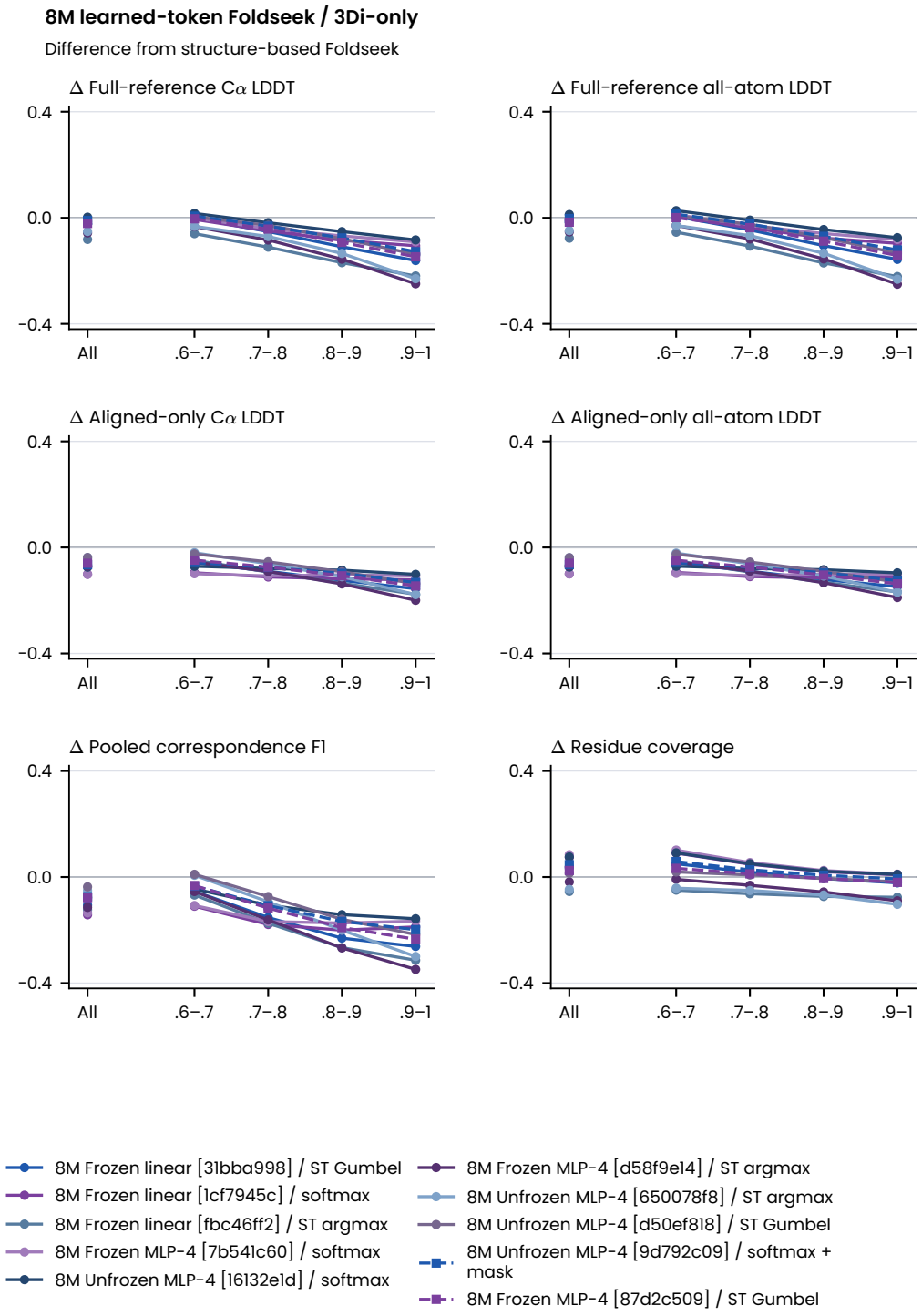


Figure S18: 8M learned-token Foldseek, 3Di-only, compared with structure-based Foldseek in the same mode. Model curves are overlaid on six metric axes. Positive differences favor learned tokens; the horizontal zero line is the structure-based reference. The separate left point is the overall aggregate; four connected points follow the reference TM-score bins $[0.6, 0.7)$ through $[0.9, 1.0]$. Both pipelines use native Foldseek correspondence maps. Differences are on the 0–1 metric scale.

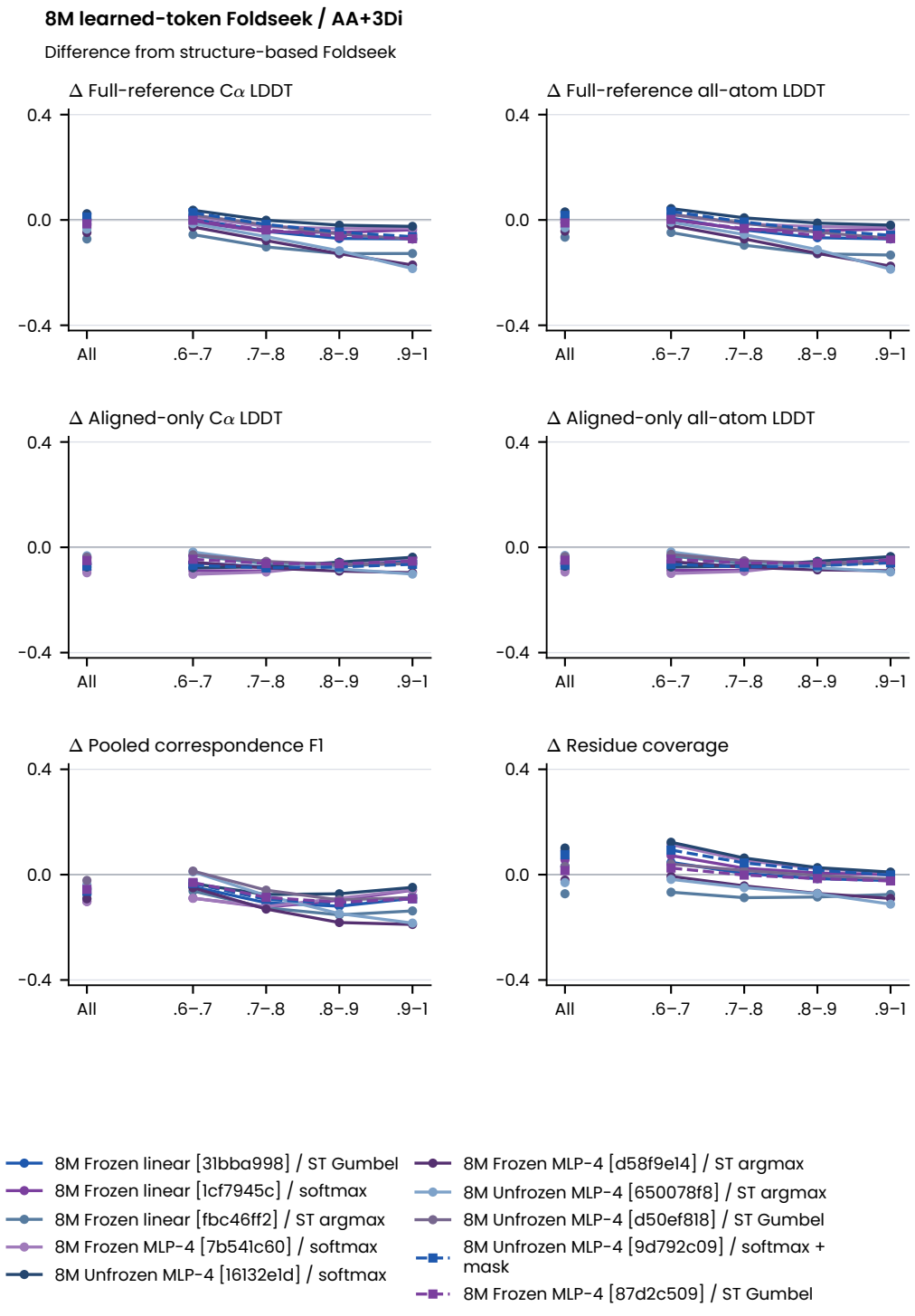


Figure S19: 8M learned-token Foldseek, AA+3Di, compared with structure-based Foldseek in the same mode. Model curves are overlaid on six metric axes. Positive differences favor learned tokens; the horizontal zero line is the structure-based reference. The separate left point is the overall aggregate; four connected points follow the reference TM-score bins $[0.6, 0.7)$ through $[0.9, 1.0]$. Both pipelines use native Foldseek correspondence maps. Differences are on the 0–1 metric scale.

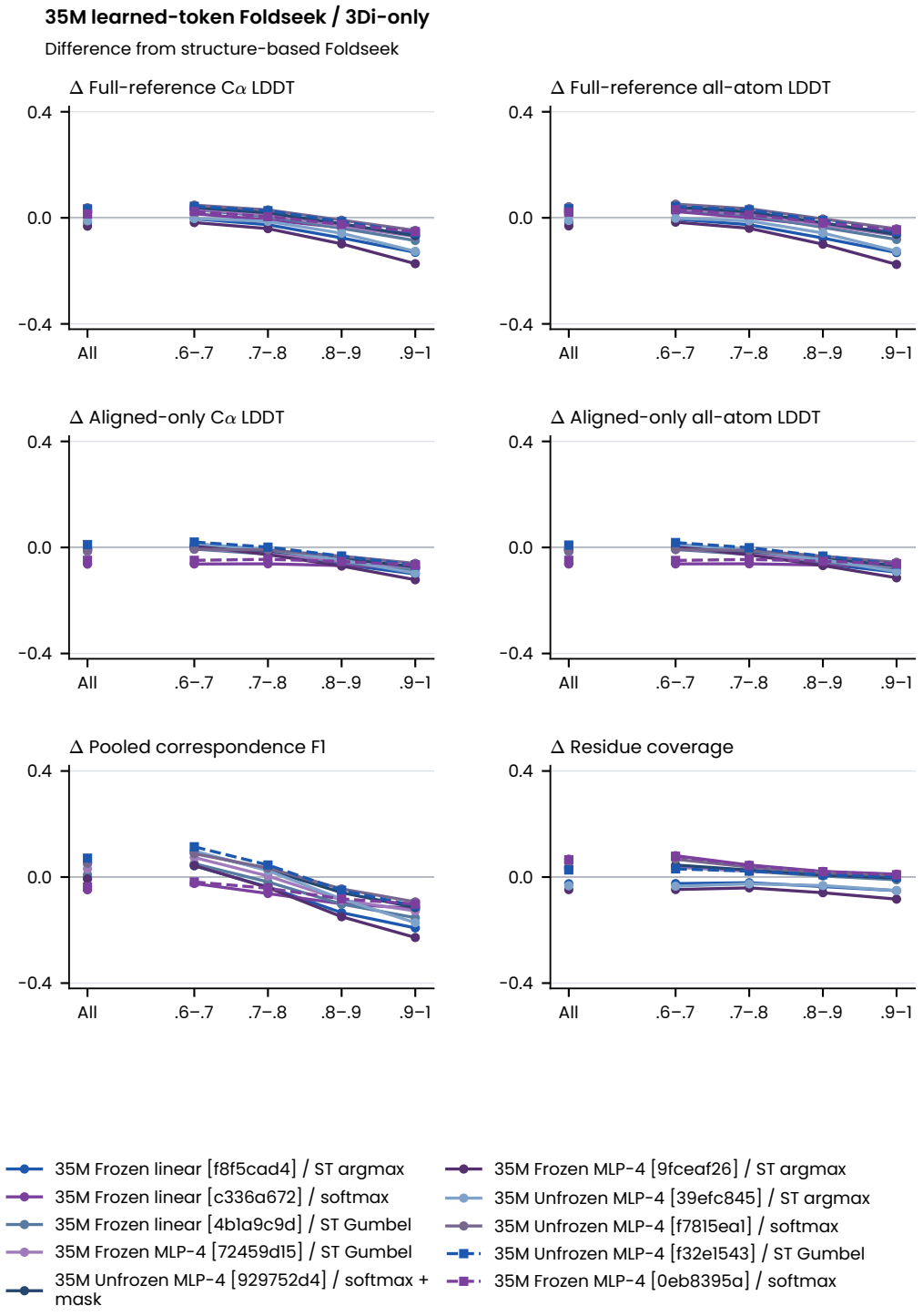


Figure S20: 35M learned-token Foldseek, 3Di-only, compared with structure-based Foldseek in the same mode. Model curves are overlaid on six metric axes. Positive differences favor learned tokens; the horizontal zero line is the structure-based reference. The separate left point is the overall aggregate; four connected points follow the reference TM-score bins $[0.6, 0.7)$ through $[0.9, 1.0]$. Both pipelines use native Foldseek correspondence maps. Differences are on the 0–1 metric scale.

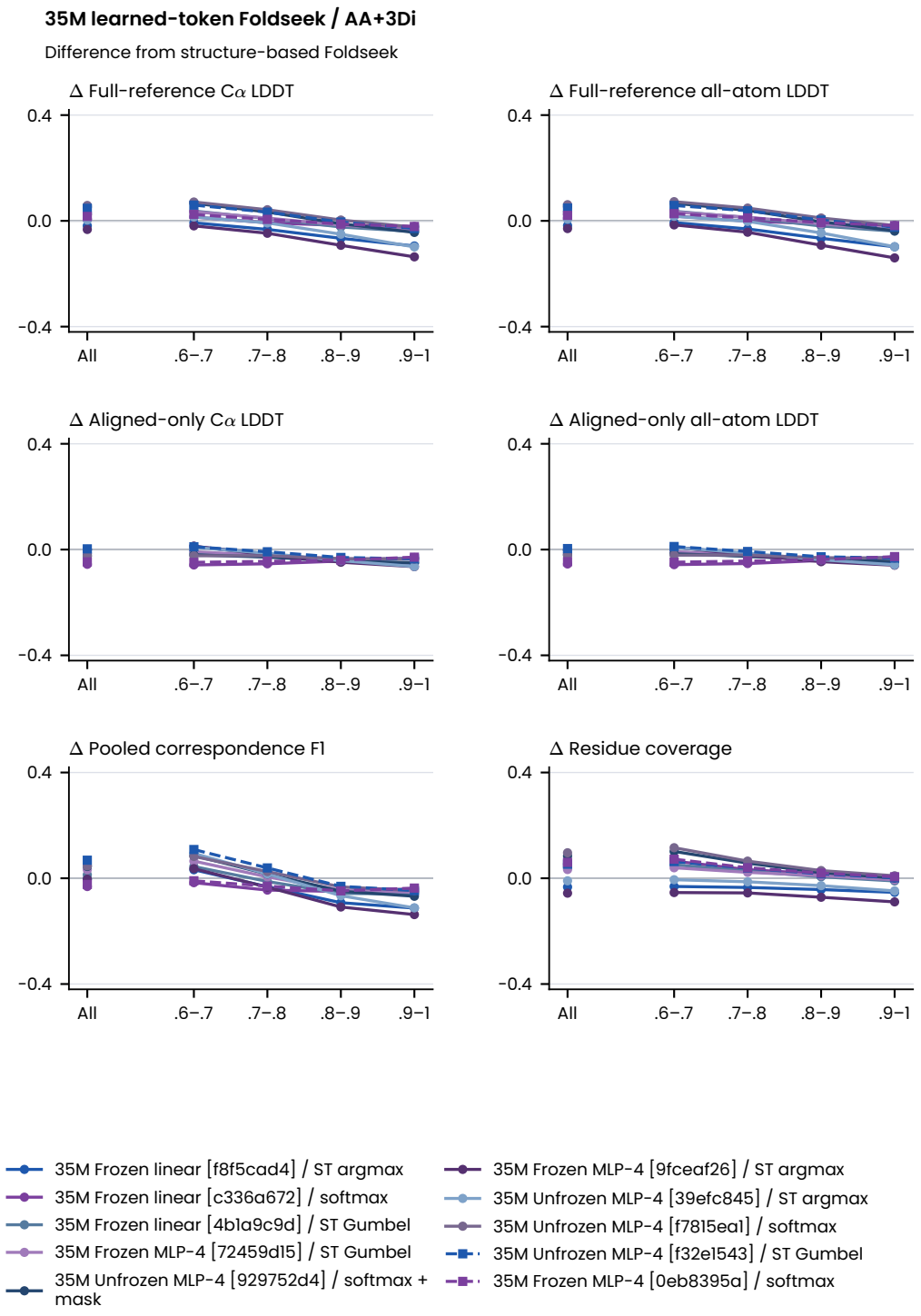


Figure S21: 35M learned-token Foldseek, AA+3Di, compared with structure-based Foldseek in the same mode. Model curves are overlaid on six metric axes. Positive differences favor learned tokens; the horizontal zero line is the structure-based reference. The separate left point is the overall aggregate; four connected points follow the reference TM-score bins $[0.6, 0.7)$ through $[0.9, 1.0]$. Both pipelines use native Foldseek correspondence maps. Differences are on the 0–1 metric scale.

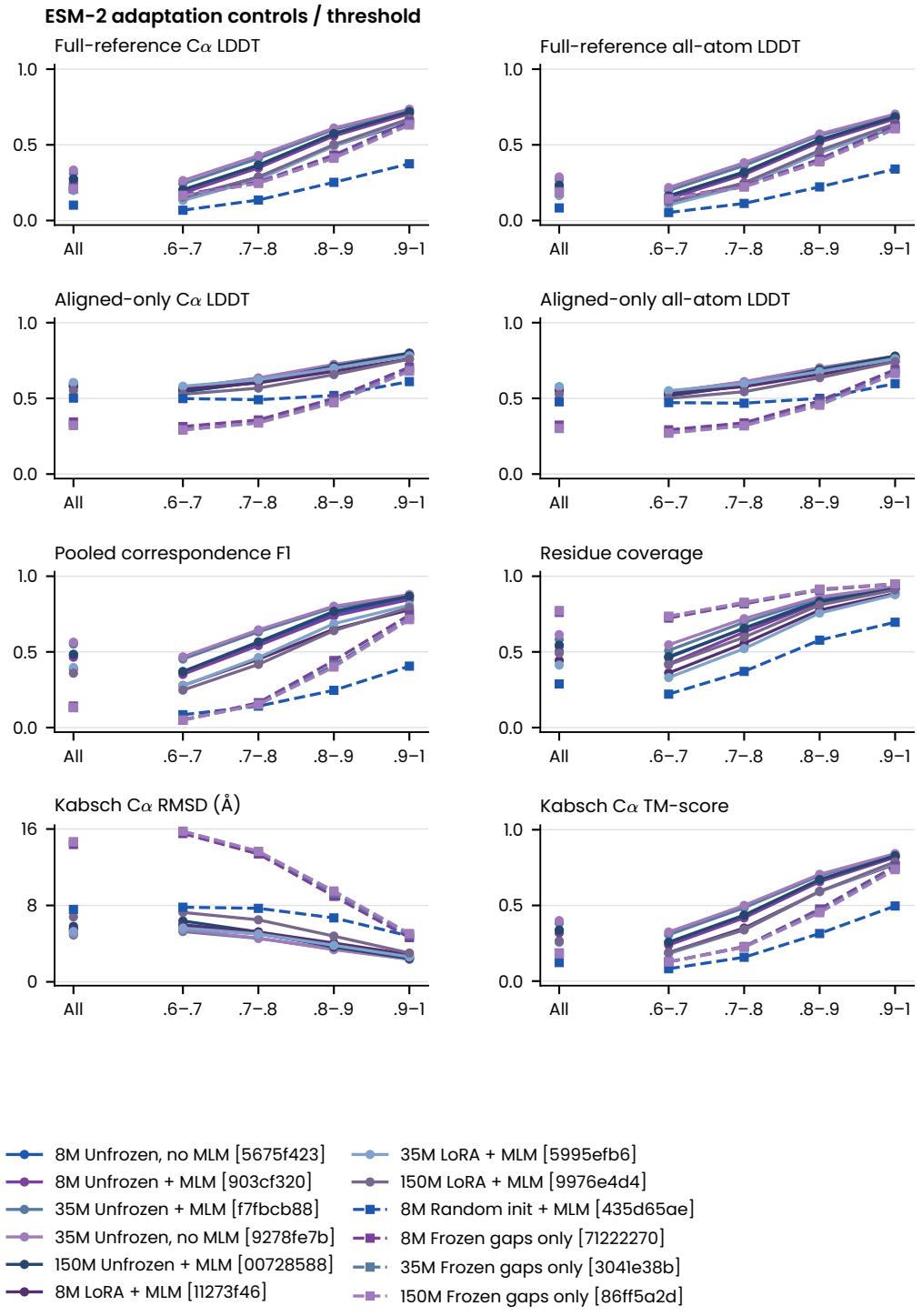


Figure S22: ESM-2 adaptation controls / threshold. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

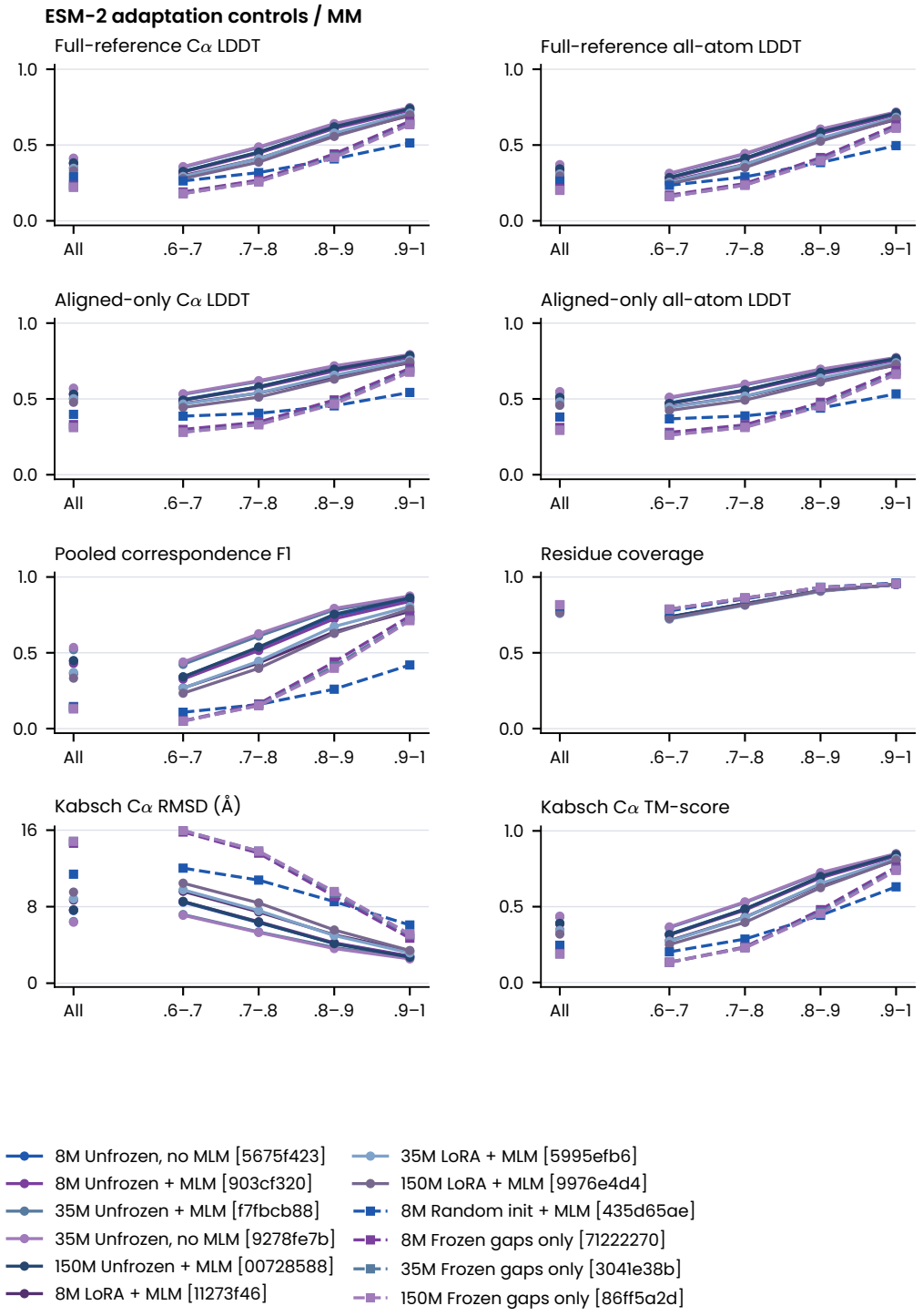


Figure S23: ESM-2 adaptation controls / MM. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

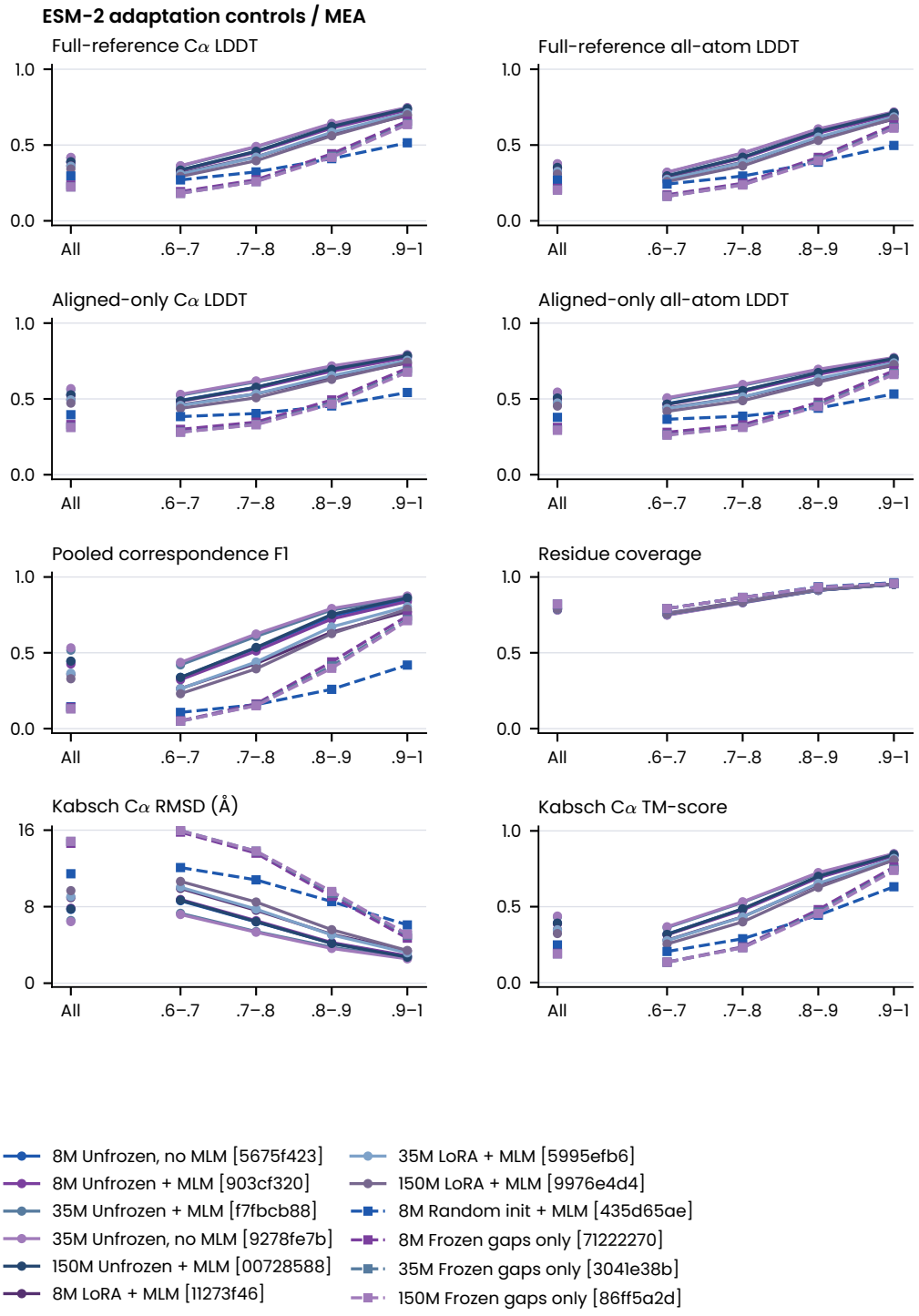


Figure S24: ESM-2 adaptation controls / MEA. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

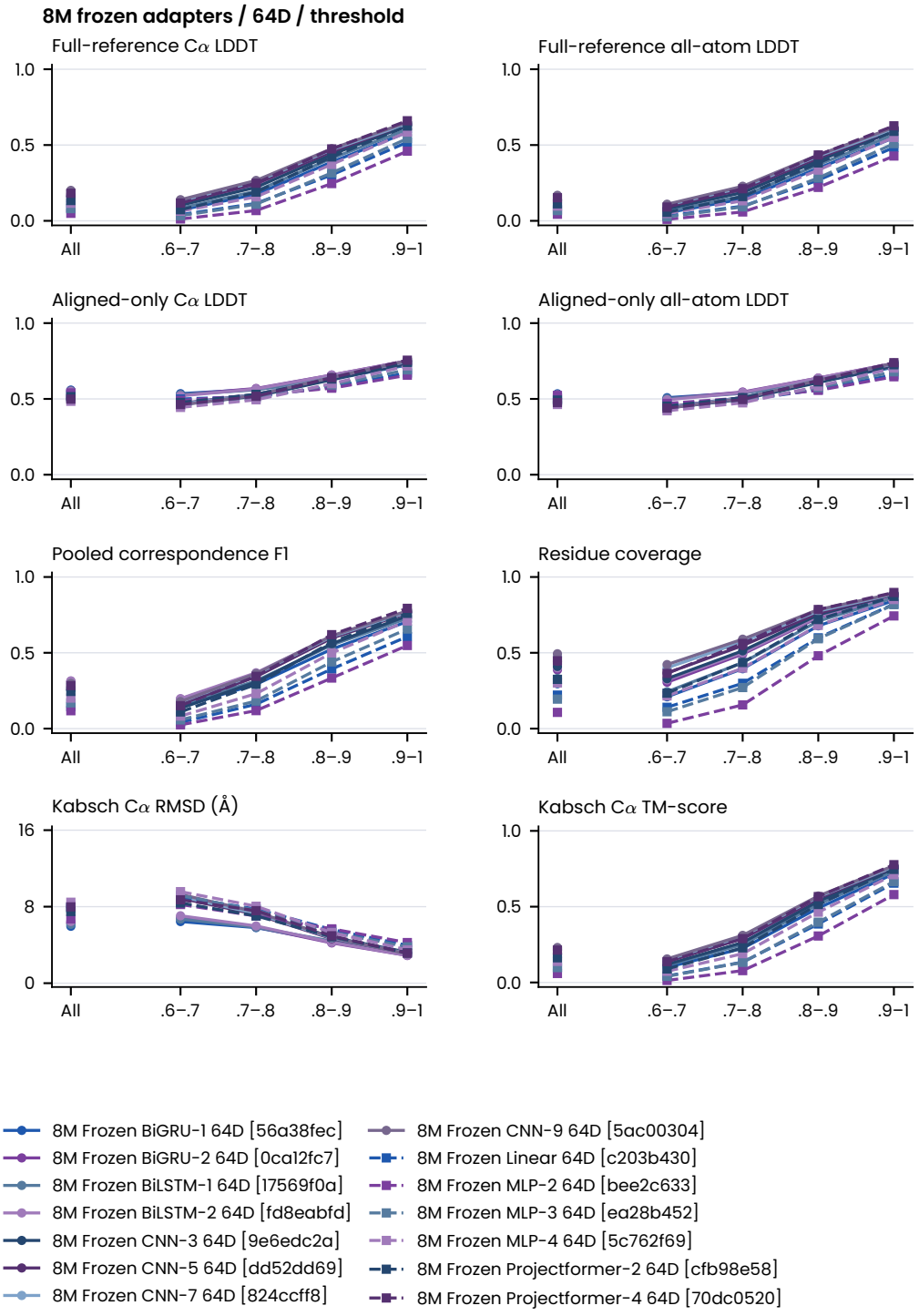


Figure S25: 8M frozen adapters / 64D / threshold. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

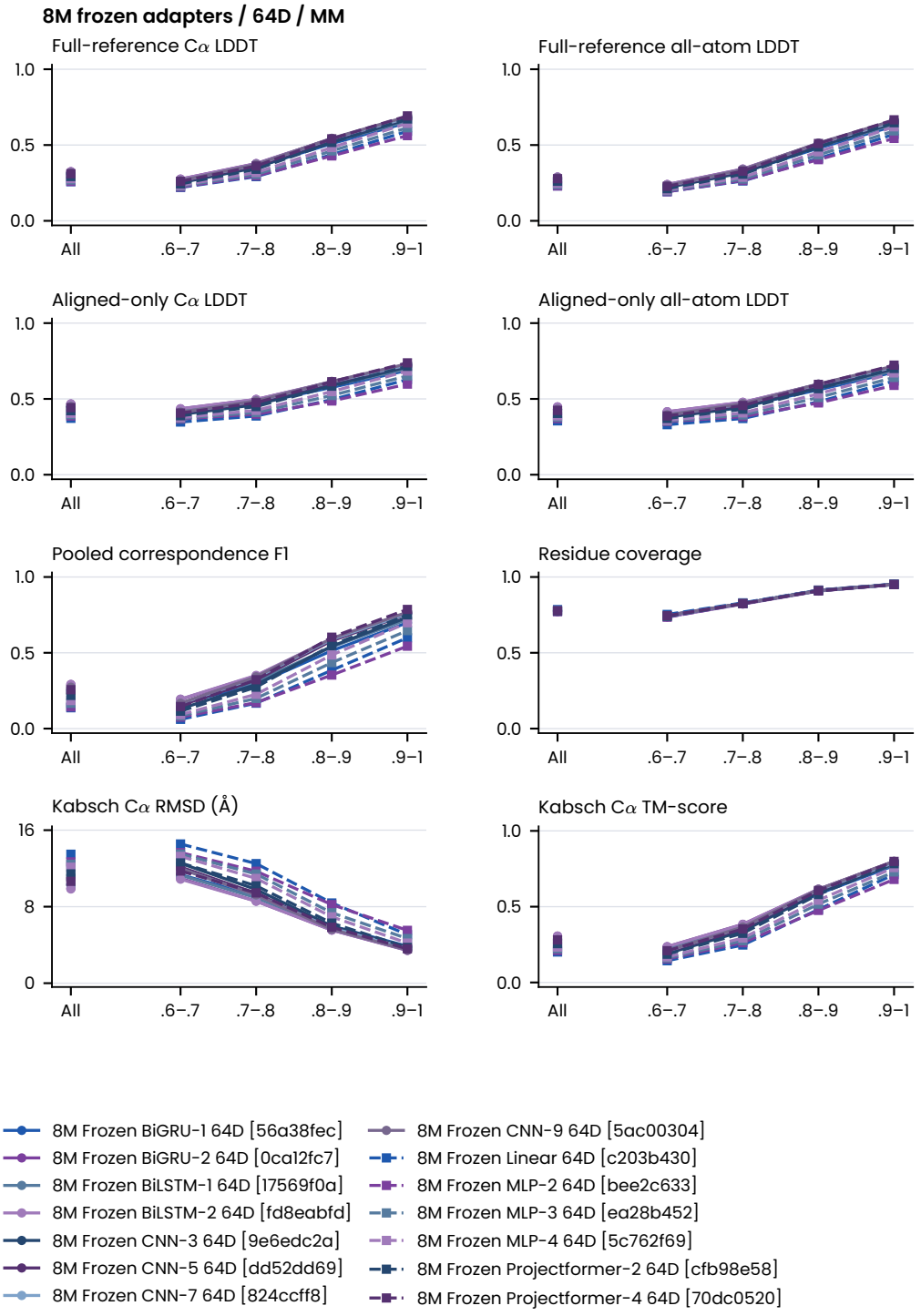


Figure S26: 8M frozen adapters / 64D / MM. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

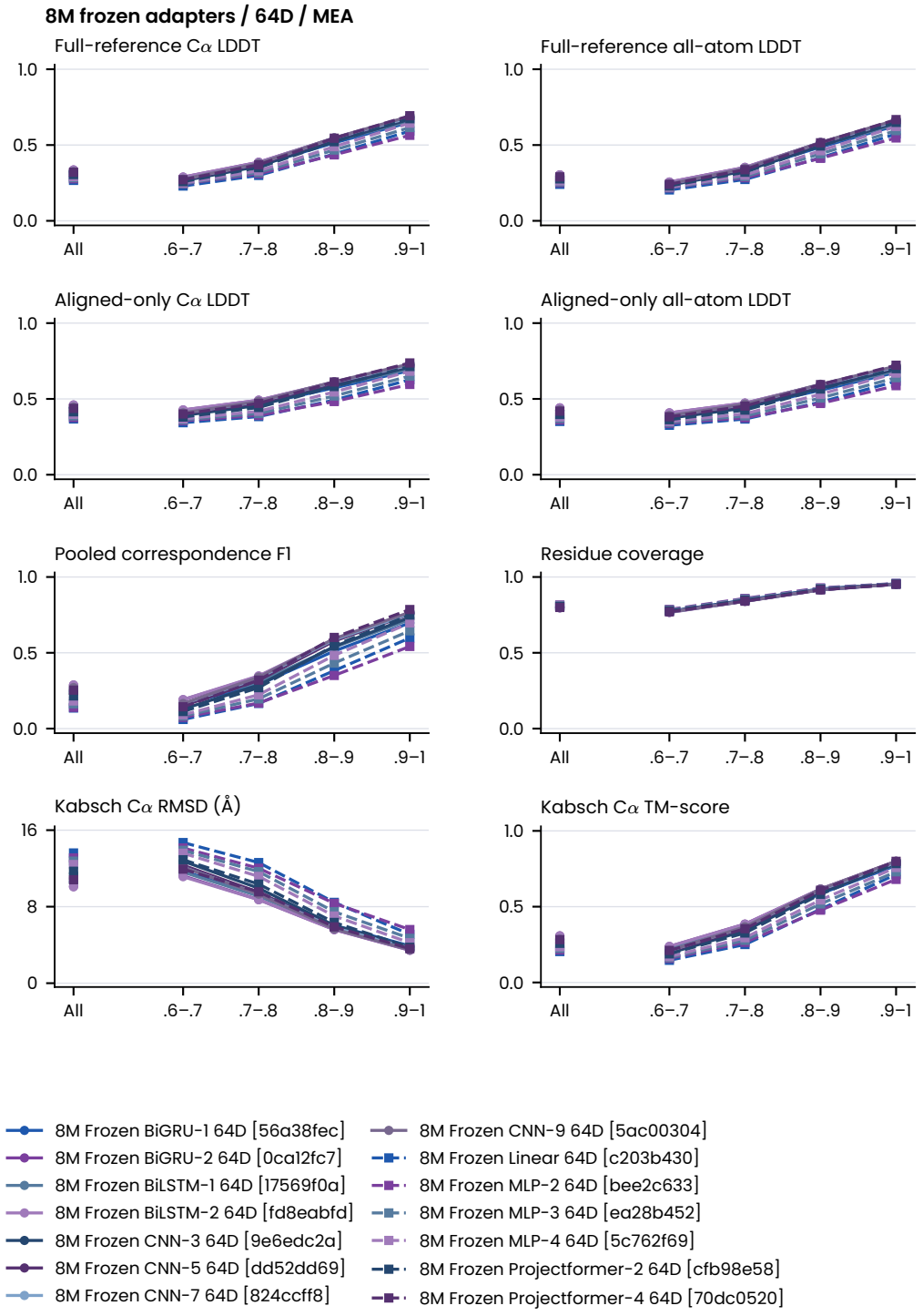


Figure S27: 8M frozen adapters / 64D / MEA. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

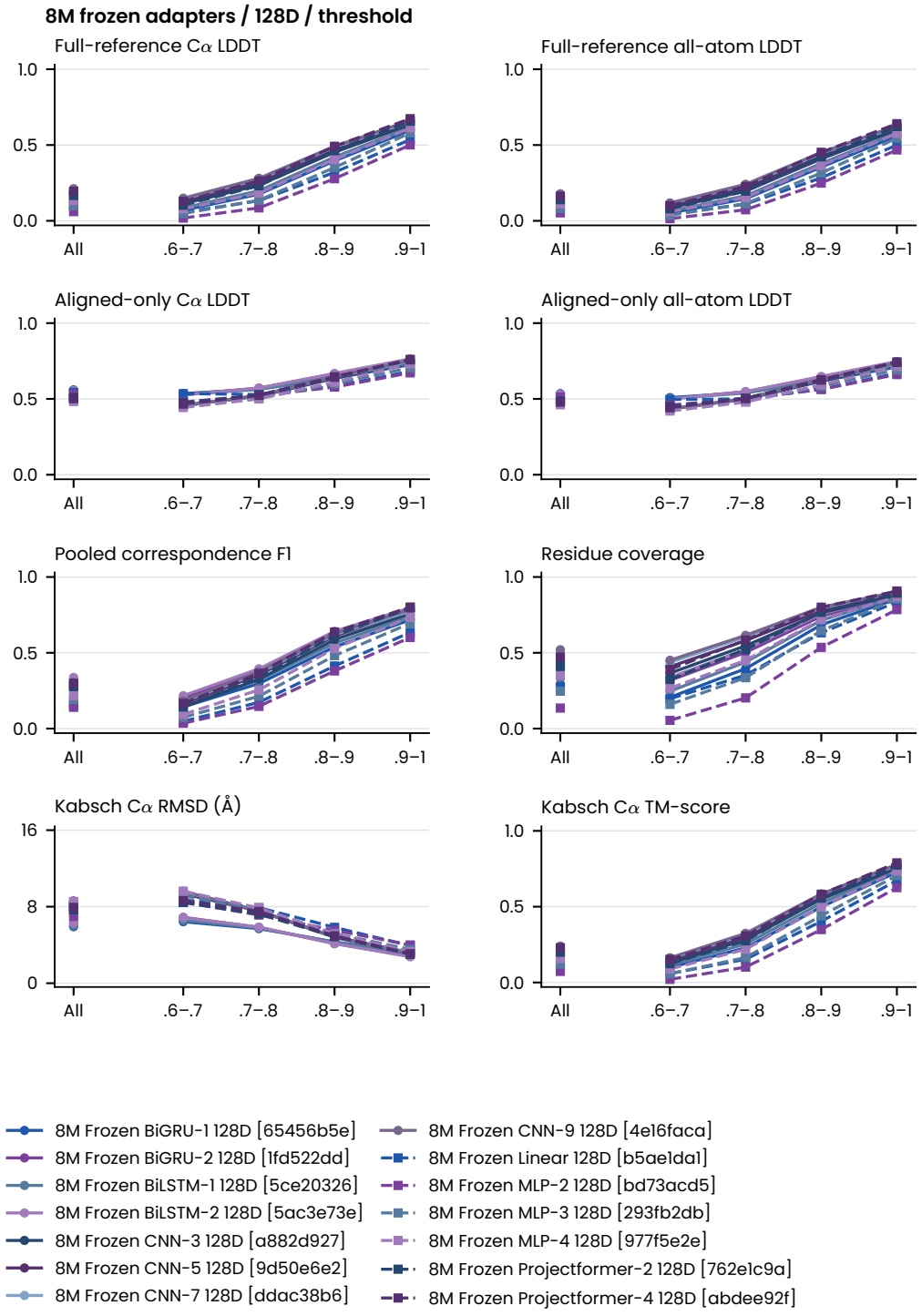


Figure S28: 8M frozen adapters / 128D / threshold. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

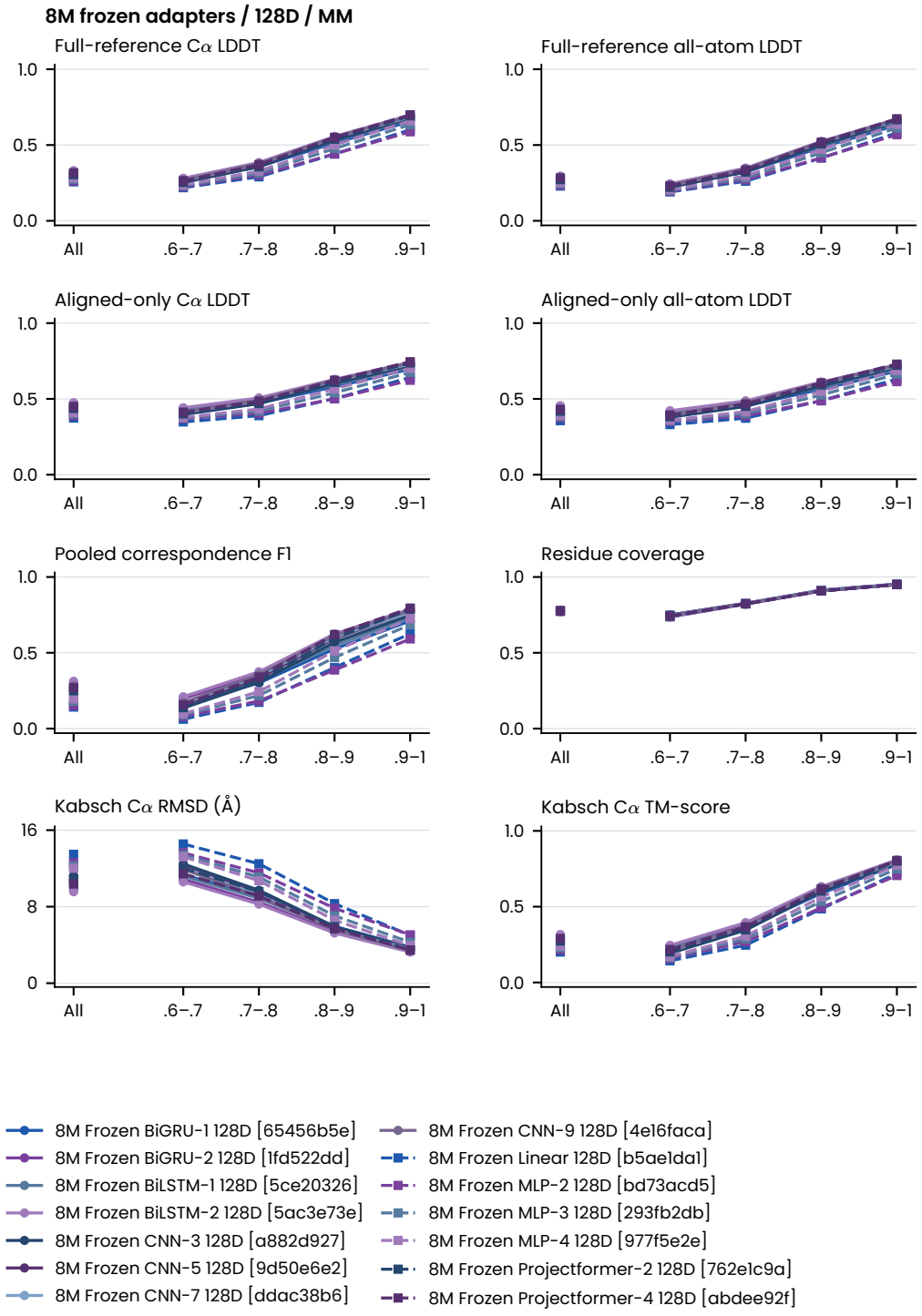


Figure S29: 8M frozen adapters / 128D / MM. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

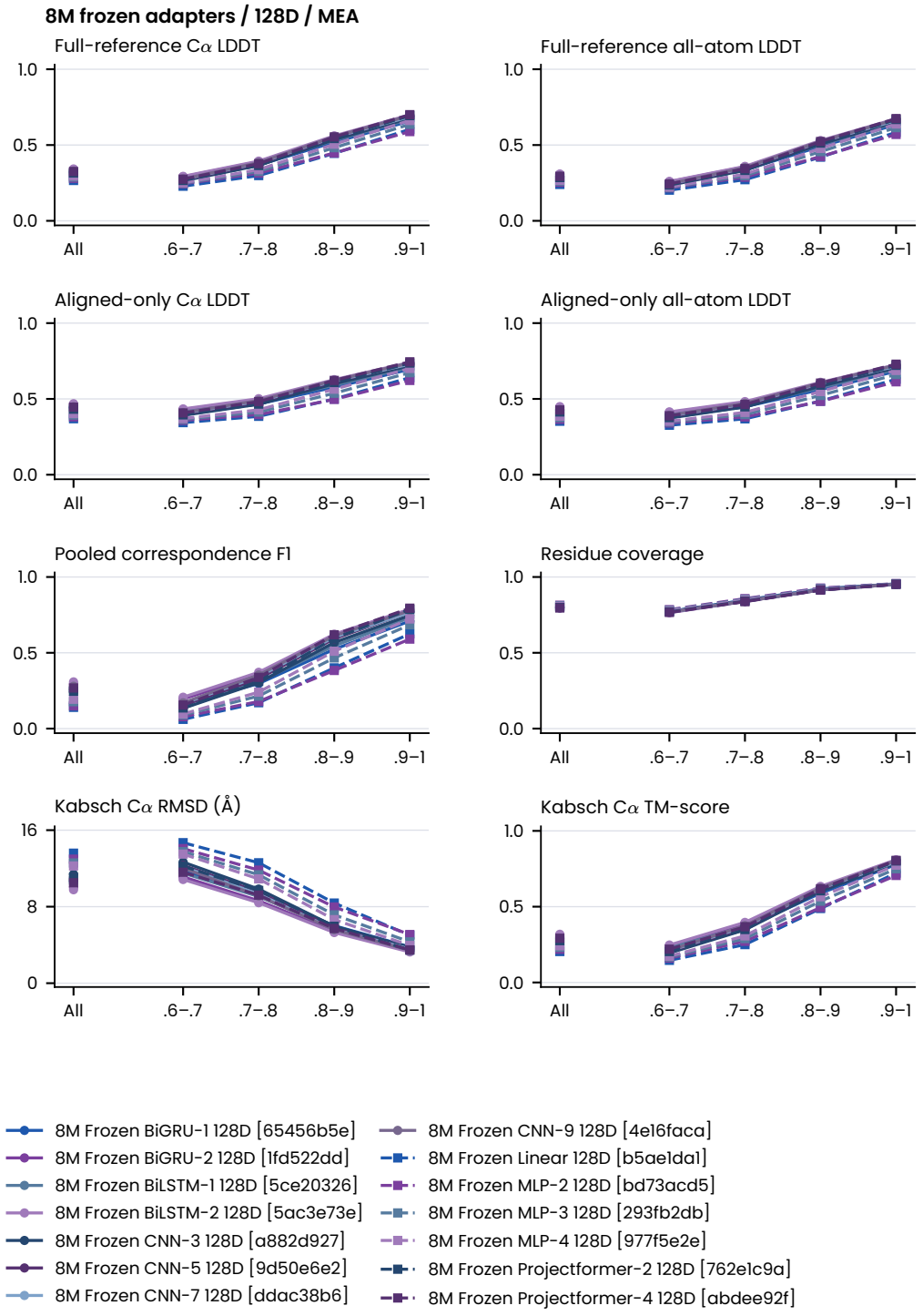


Figure S30: 8M frozen adapters / 128D / MEA. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

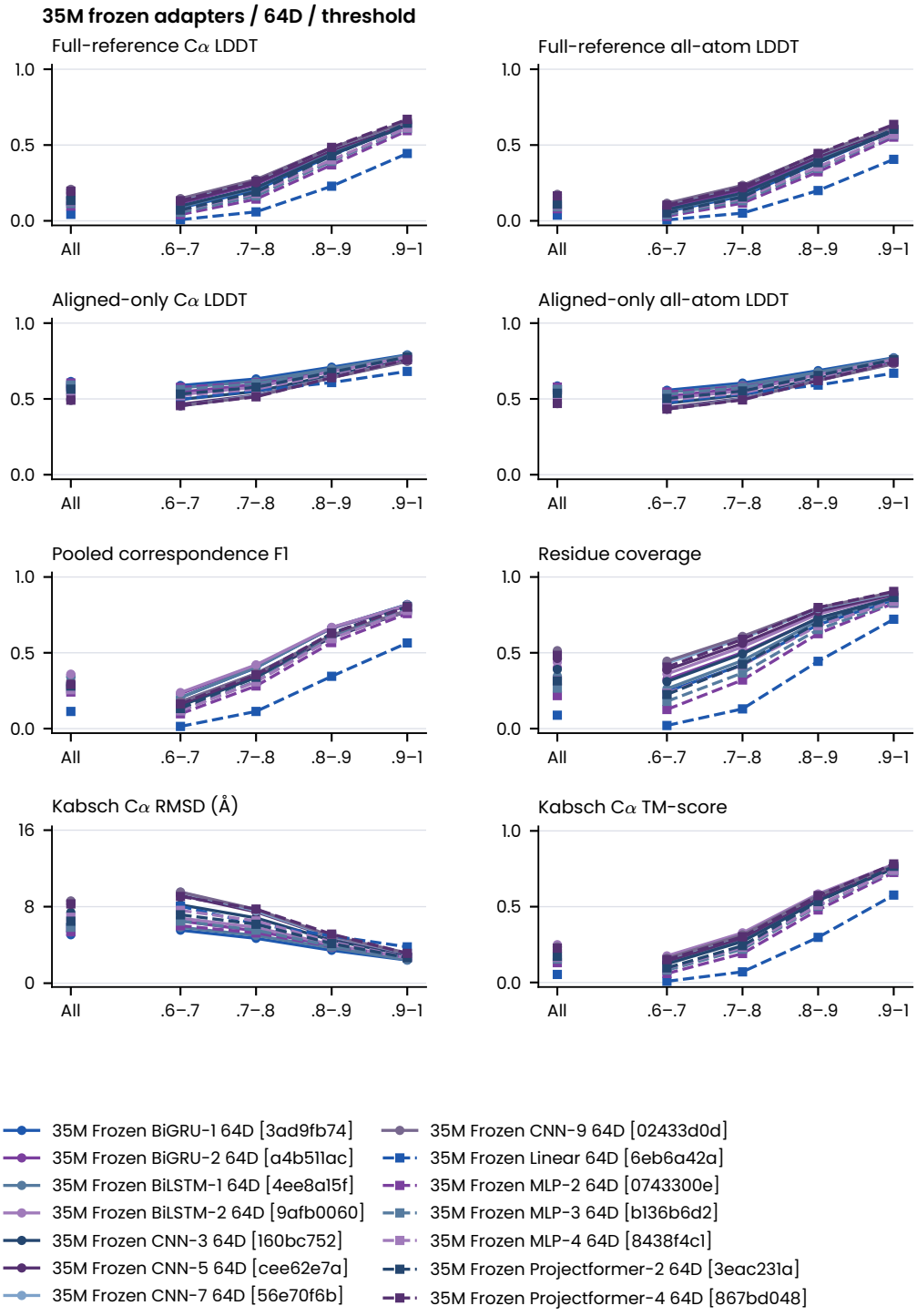


Figure S31: 35M frozen adapters / 64D / threshold. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

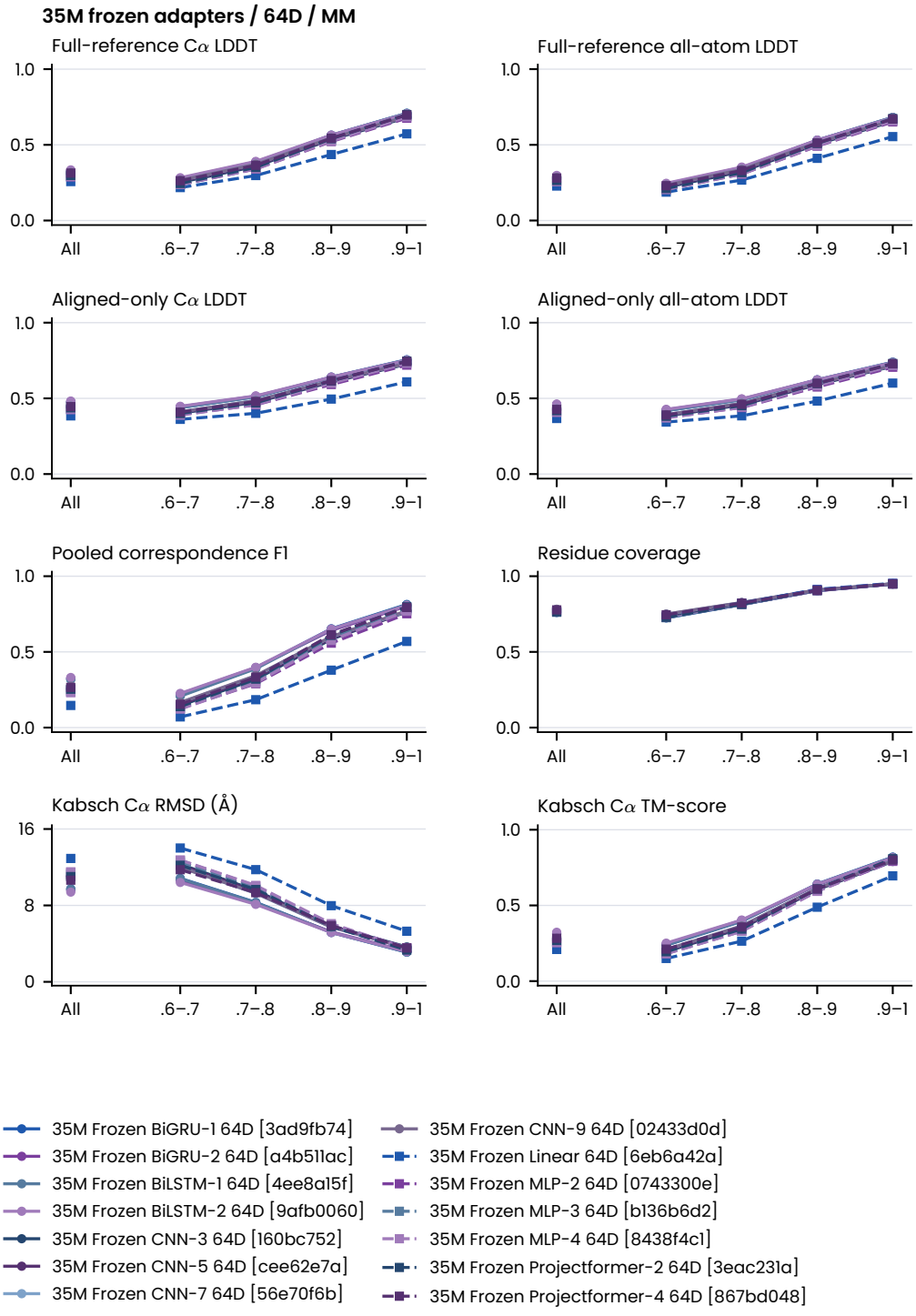


Figure S32: 35M frozen adapters / 64D / MM. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

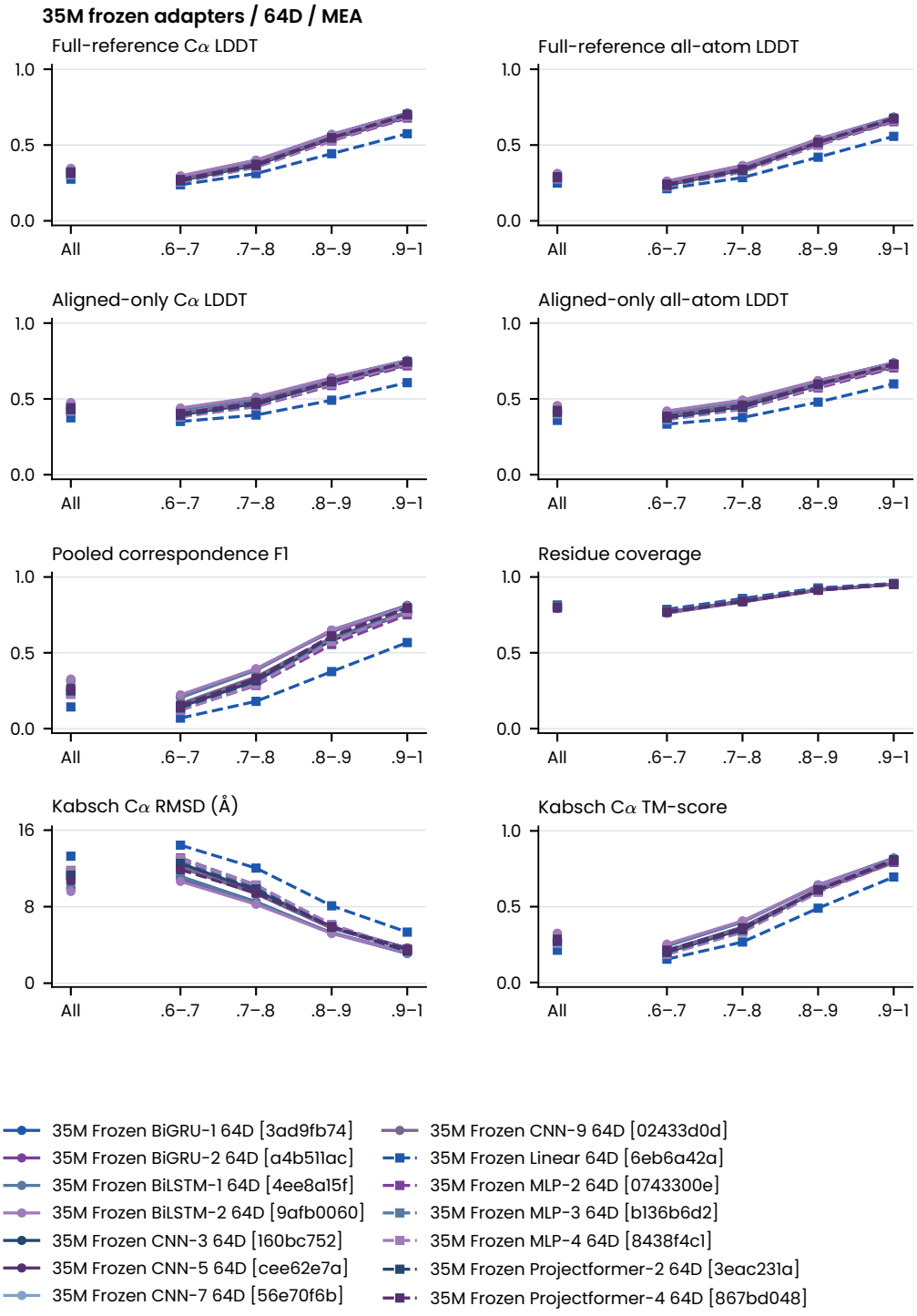


Figure S33: 35M frozen adapters / 64D / MEA. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

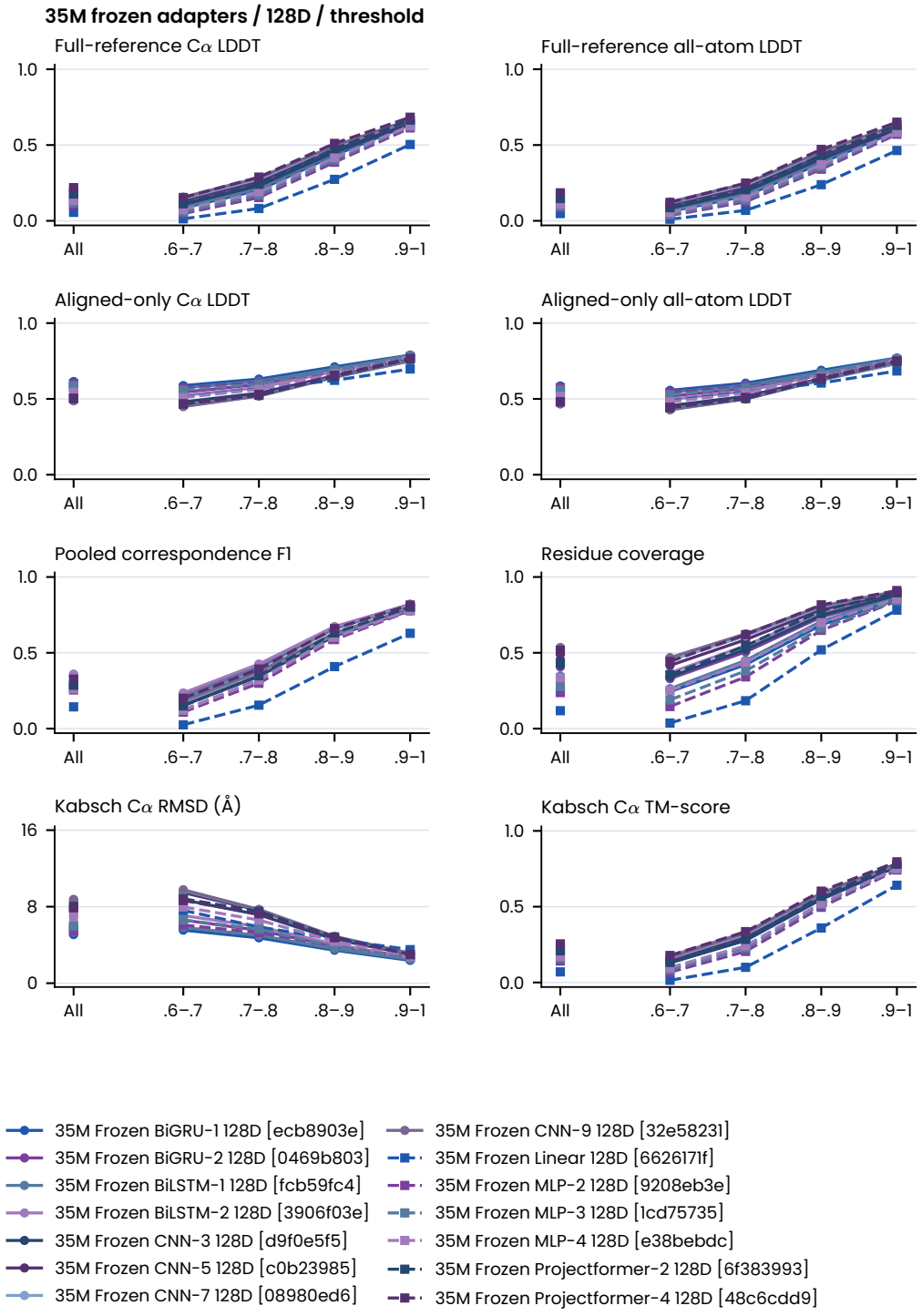


Figure S34: 35M frozen adapters / 128D / threshold. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

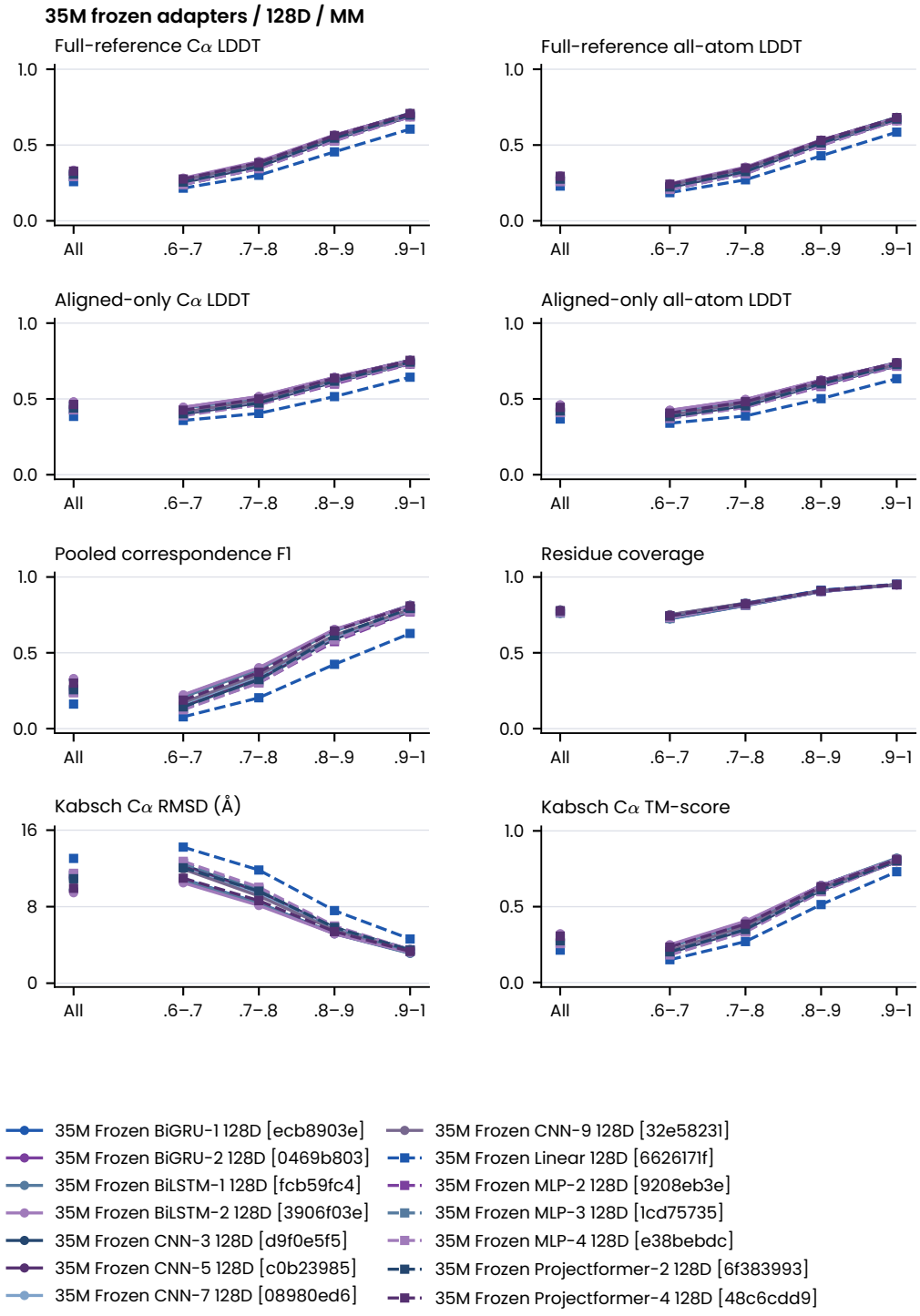


Figure S35: 35M frozen adapters / 128D / MM. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

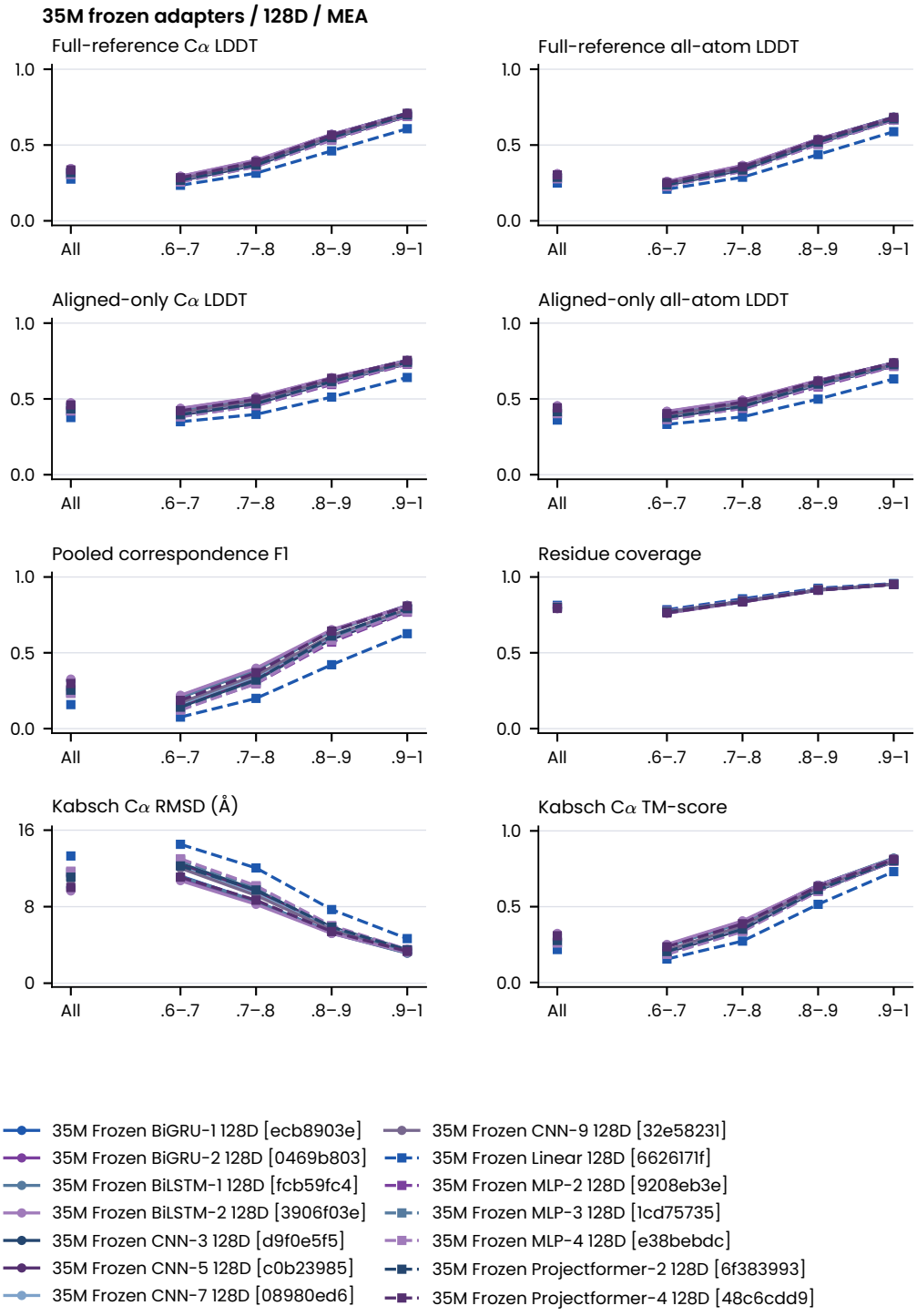


Figure S36: 35M frozen adapters / 128D / MEA. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

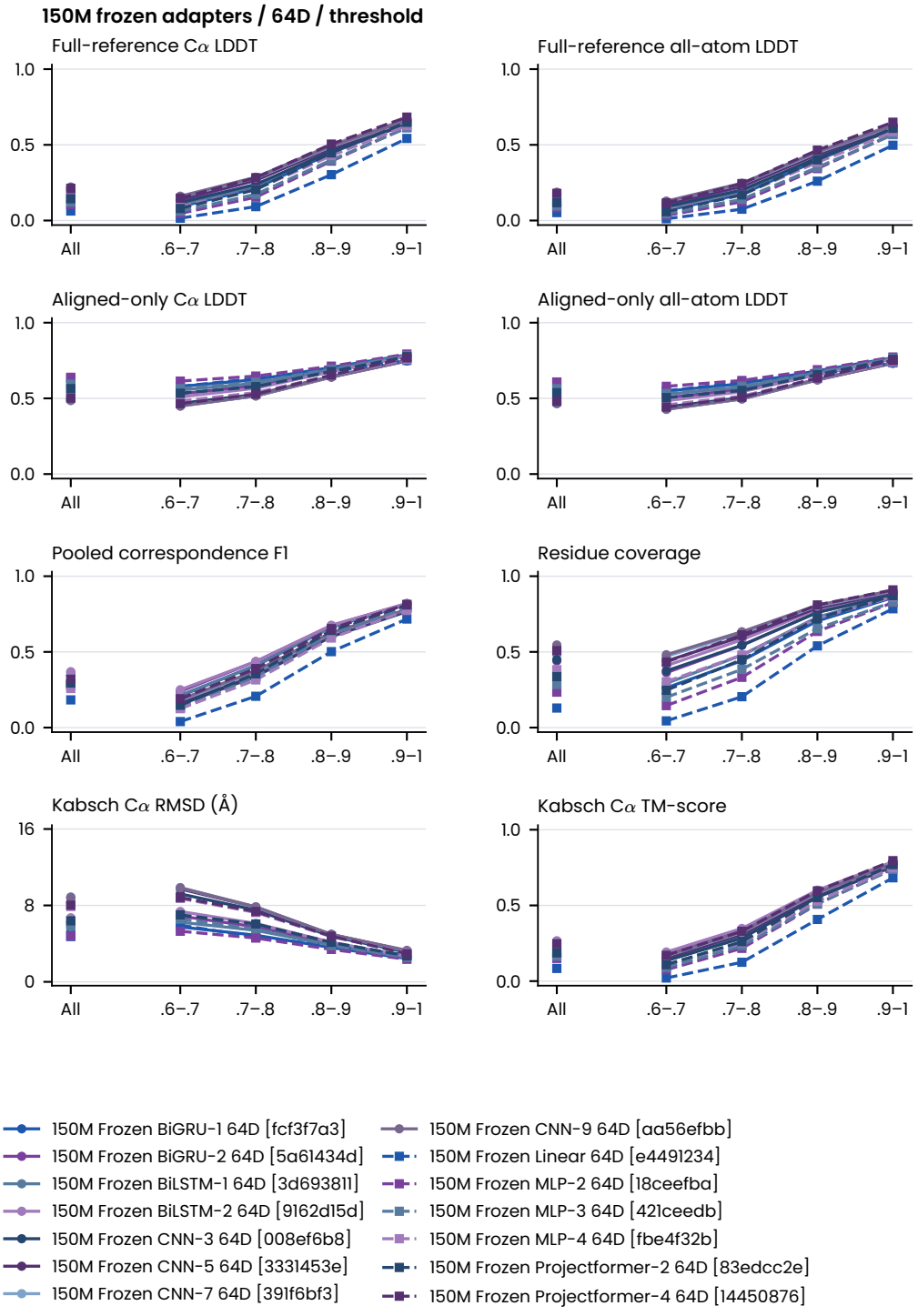


Figure S37: 150M frozen adapters / 64D / threshold. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

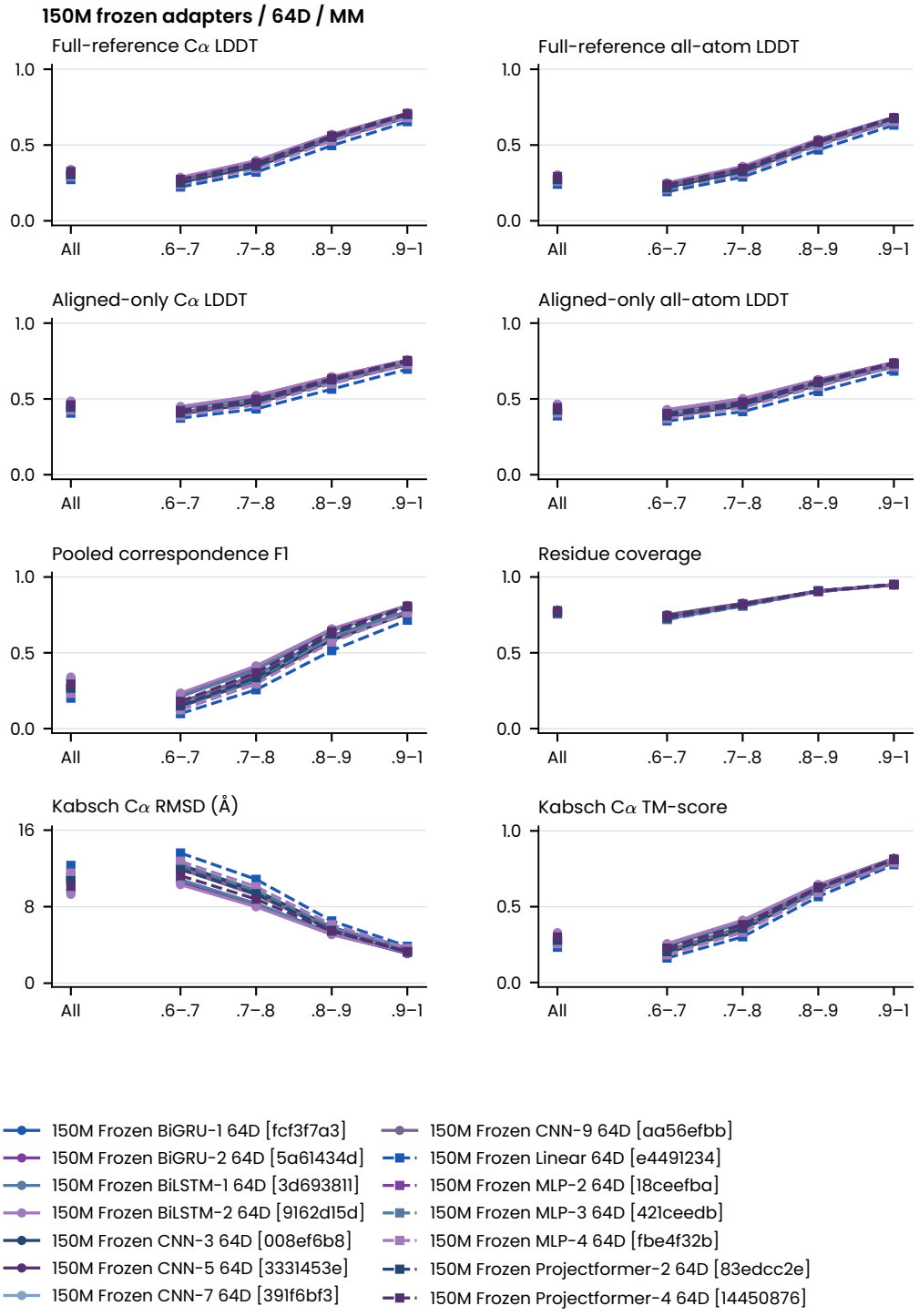


Figure S38: 150M frozen adapters / 64D / MM. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

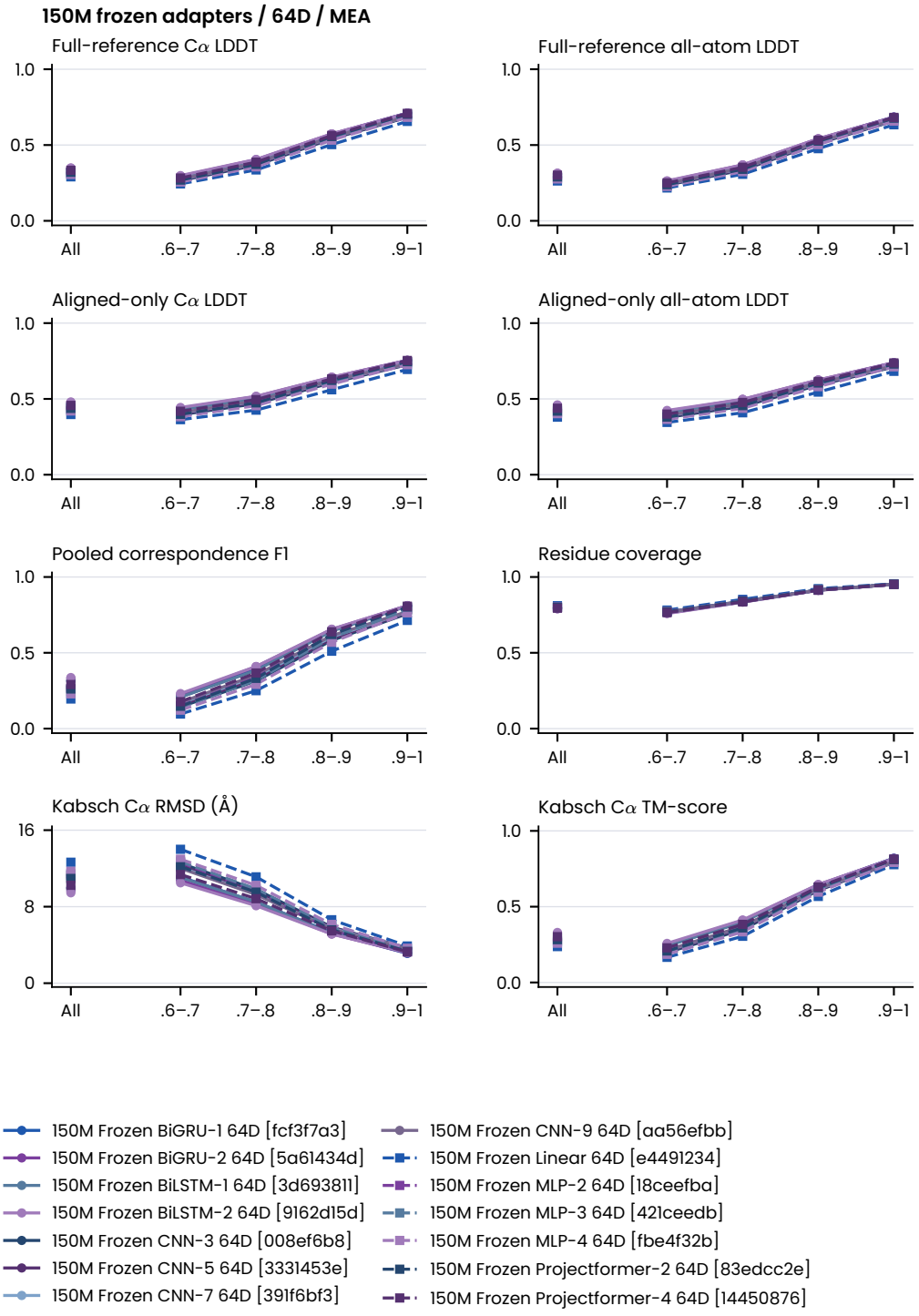


Figure S39: 150M frozen adapters / 64D / MEA. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

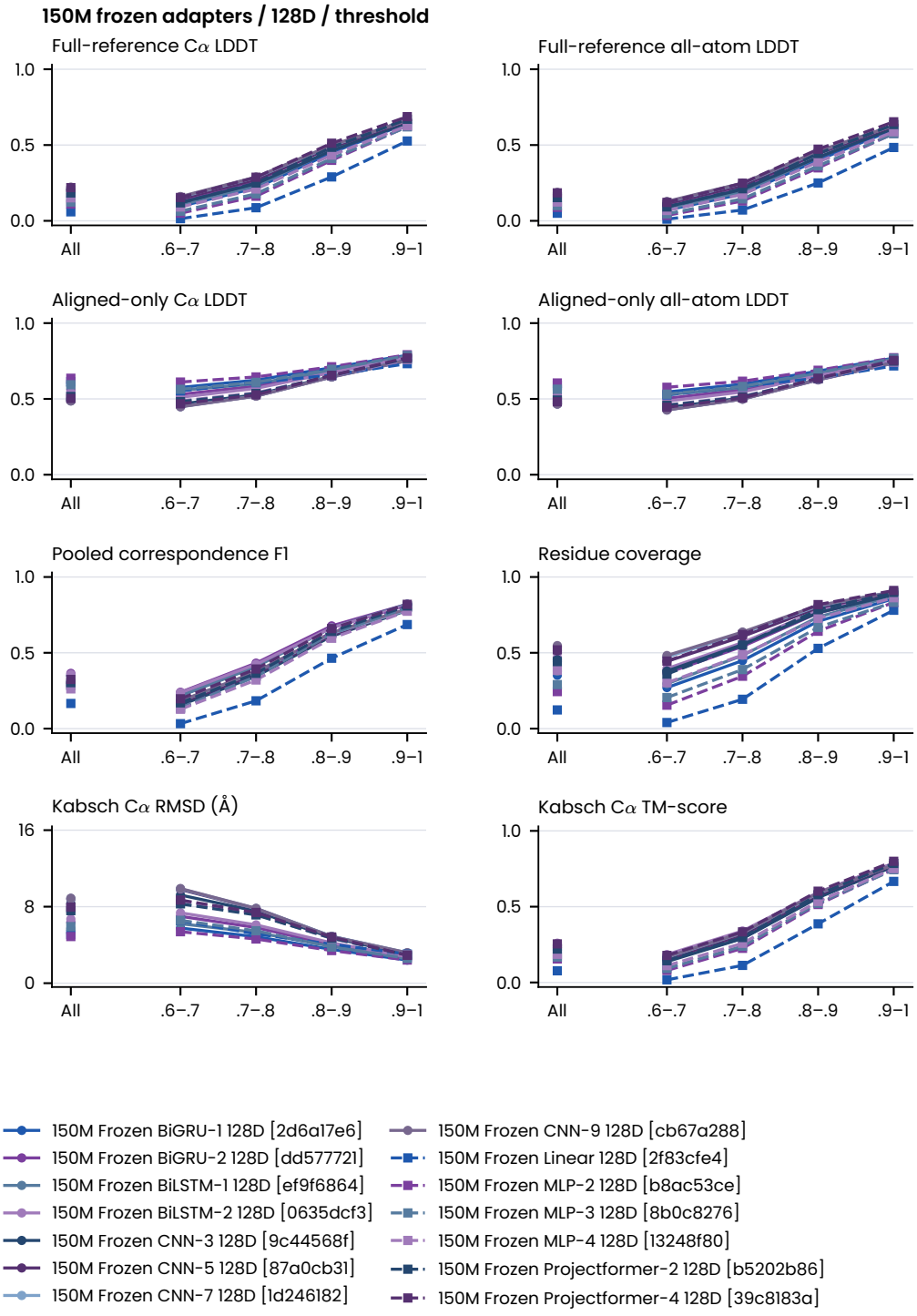


Figure S40: 150M frozen adapters / 128D / threshold. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

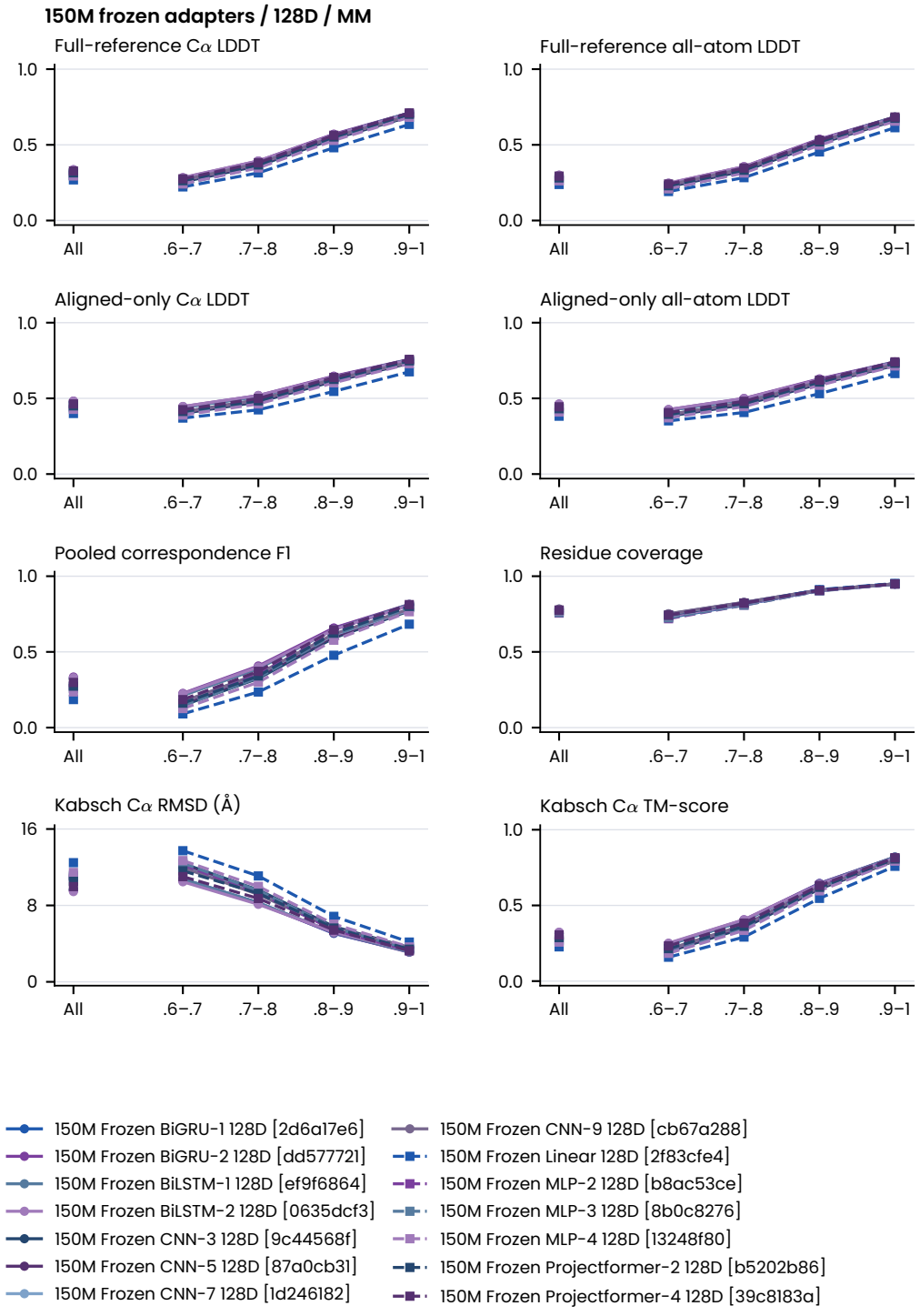


Figure S41: 150M frozen adapters / 128D / MM. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

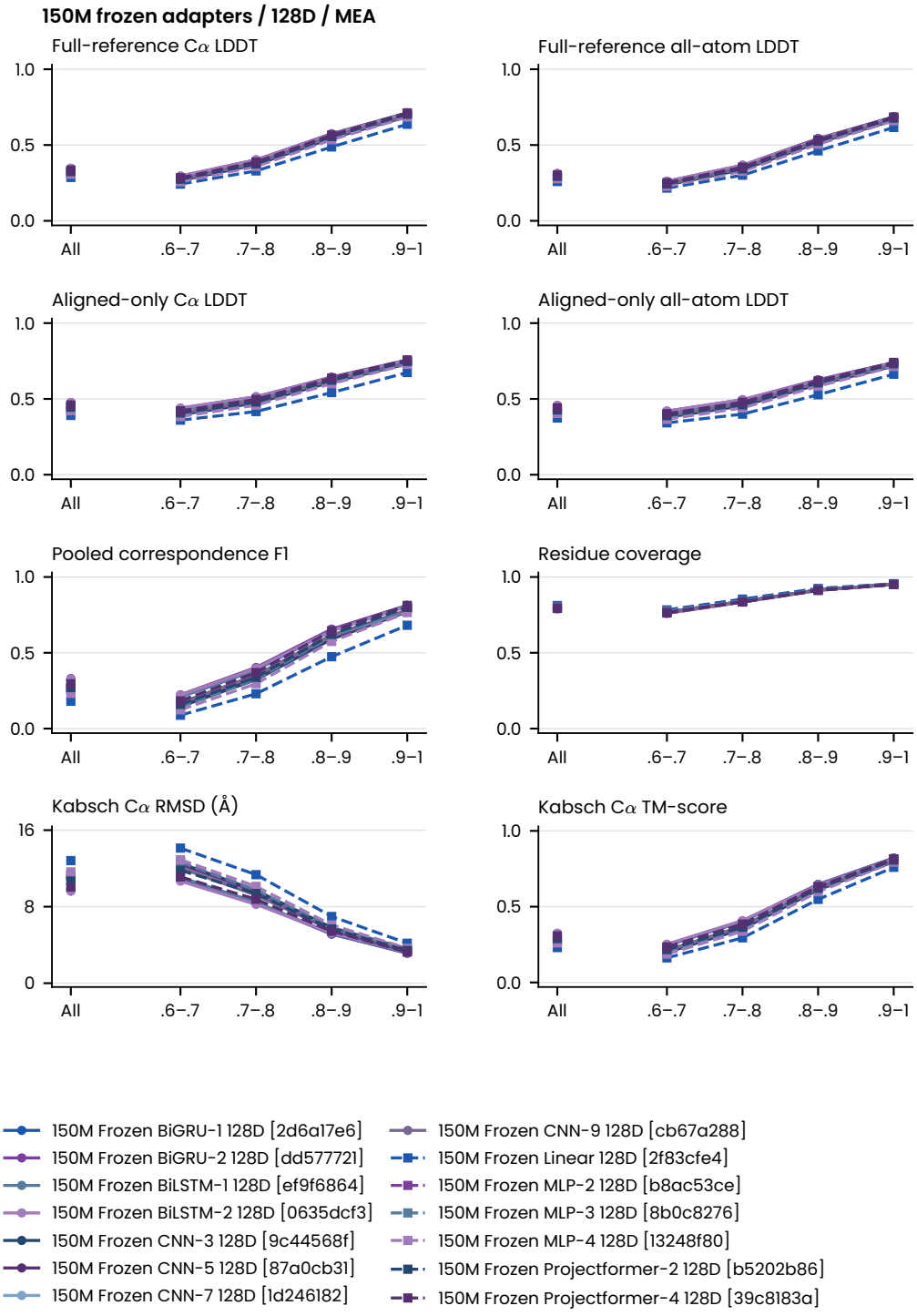


Figure S42: 150M frozen adapters / 128D / MEA. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

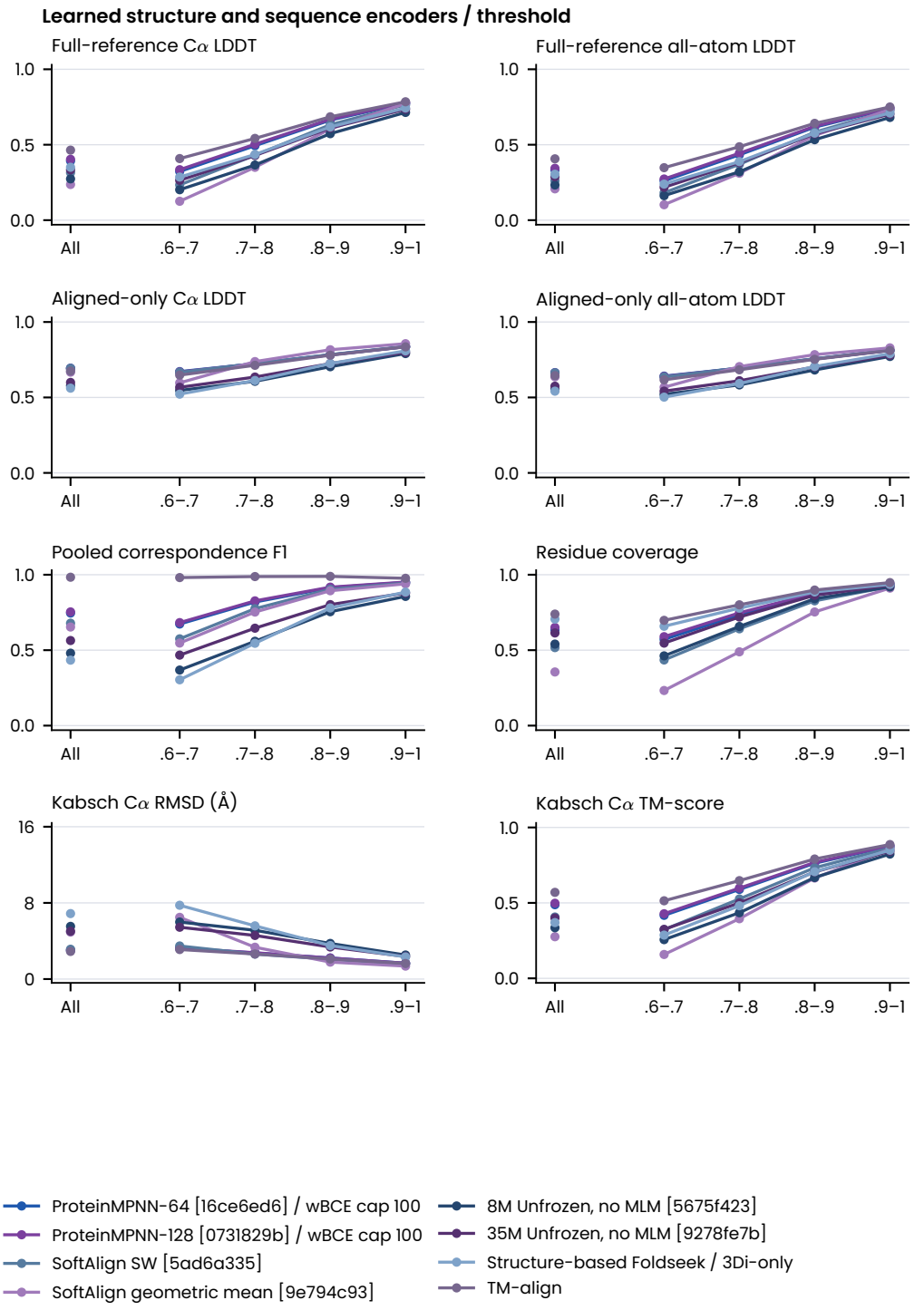


Figure S43: Learned structure and sequence encoders / threshold. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å. ProteinMPNN uses wBCE cap 100.

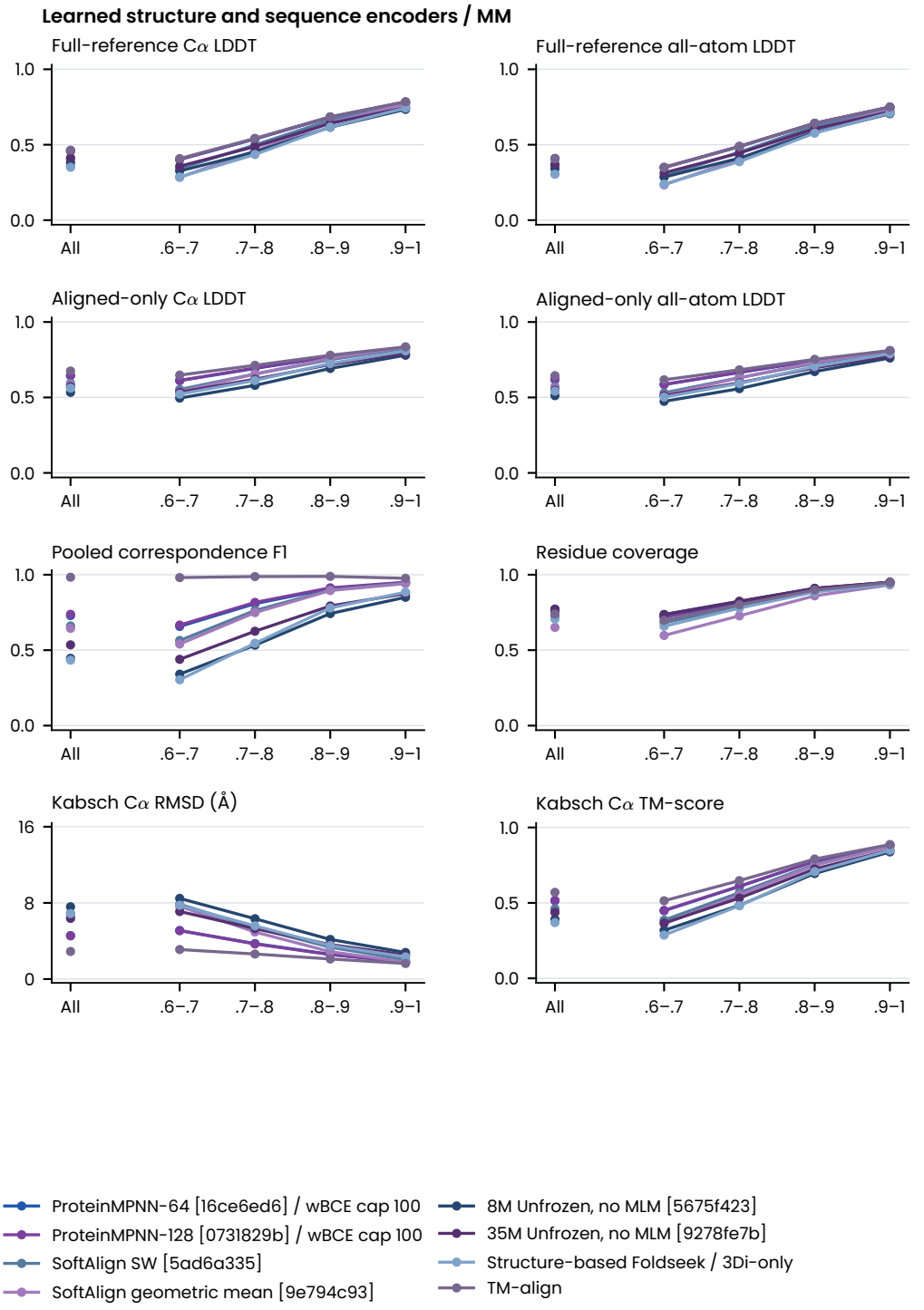


Figure S44: Learned structure and sequence encoders / MM. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å. ProteinMPNN uses wBCE cap 100.

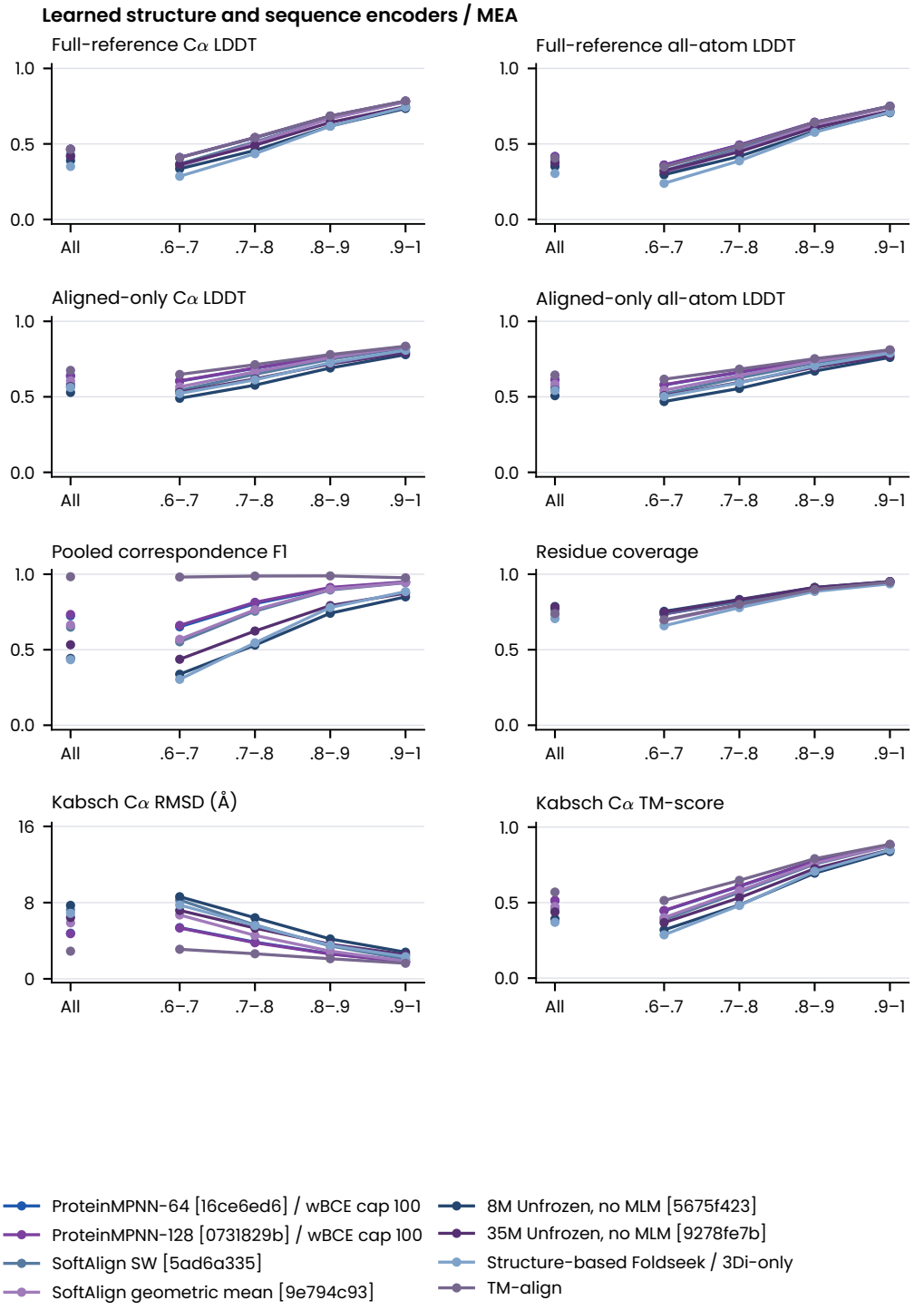


Figure S45: Learned structure and sequence encoders / MEA. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å. ProteinMPNN uses wBCE cap 100.

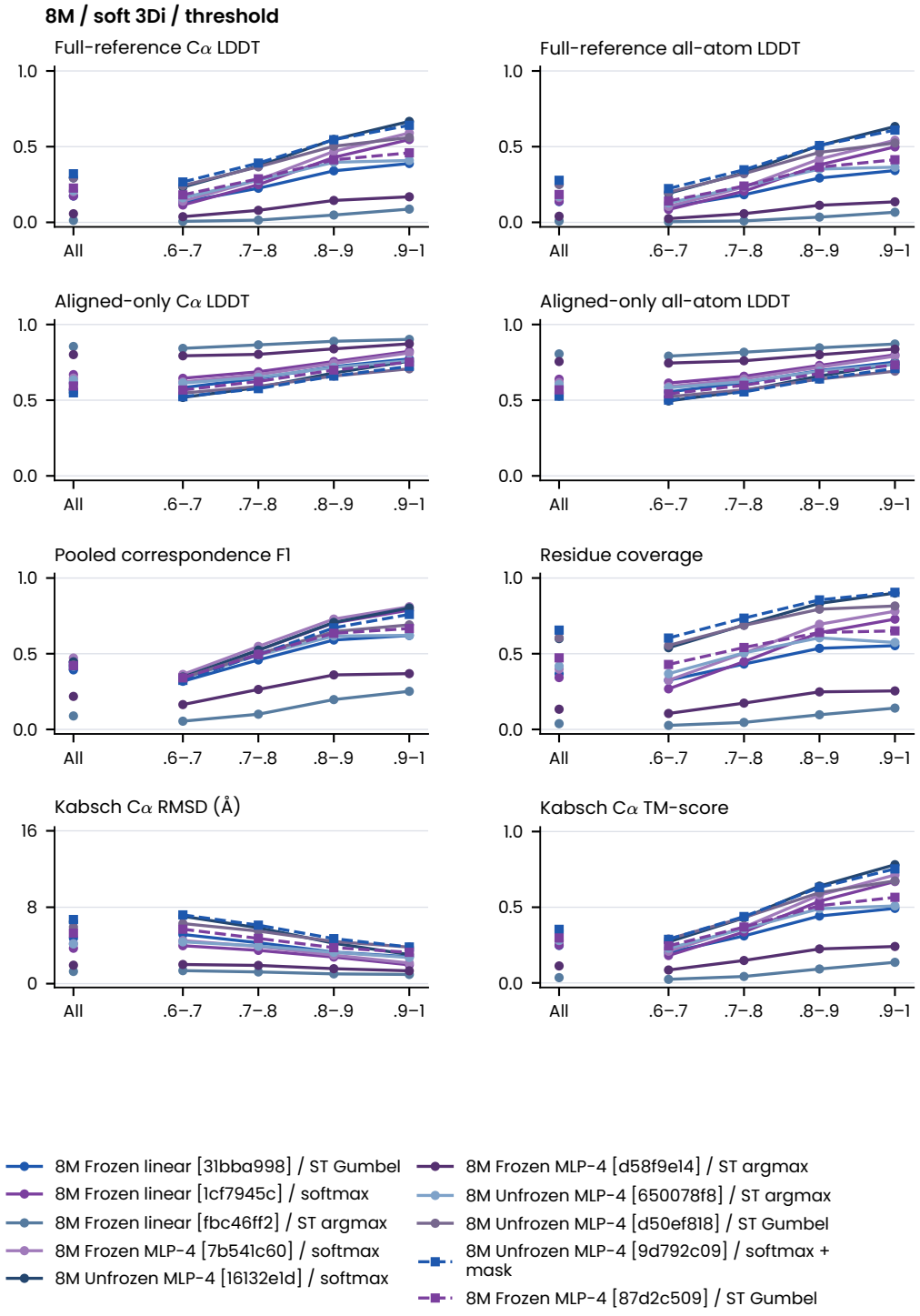


Figure S46: 8M / soft 3Di / threshold. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

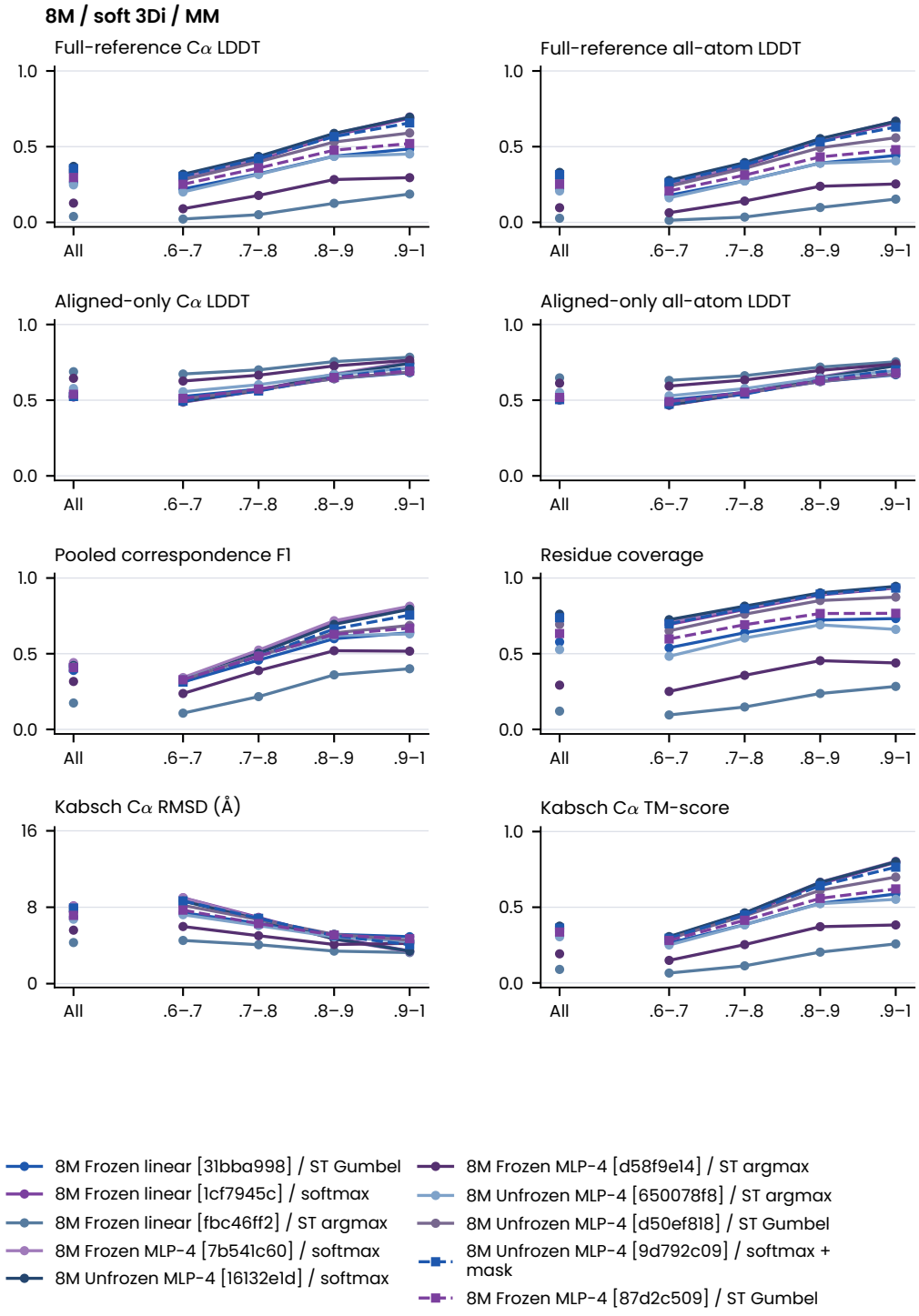


Figure S47: 8M / soft 3Di / MM. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

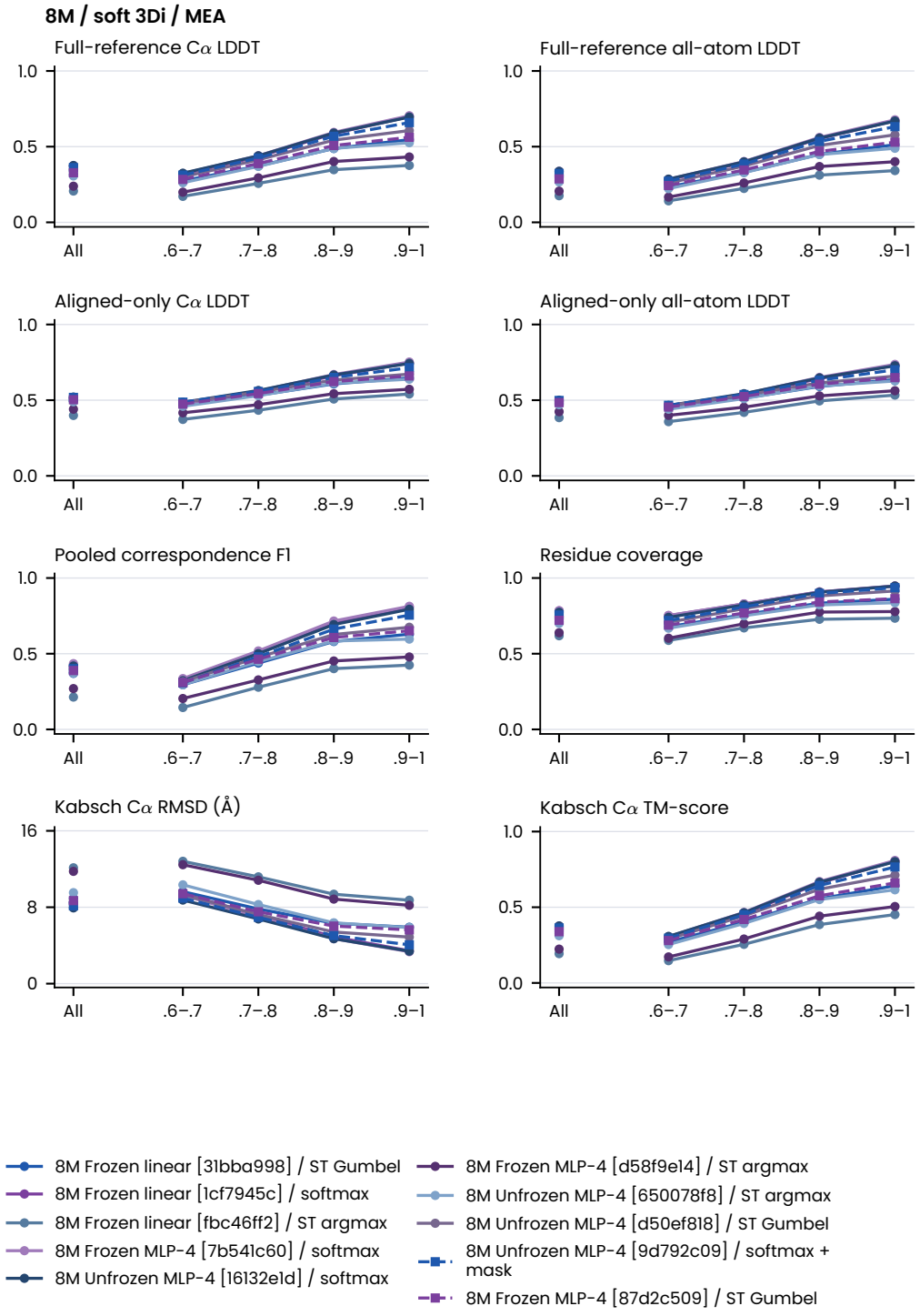


Figure S48: 8M / soft 3Di / MEA. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

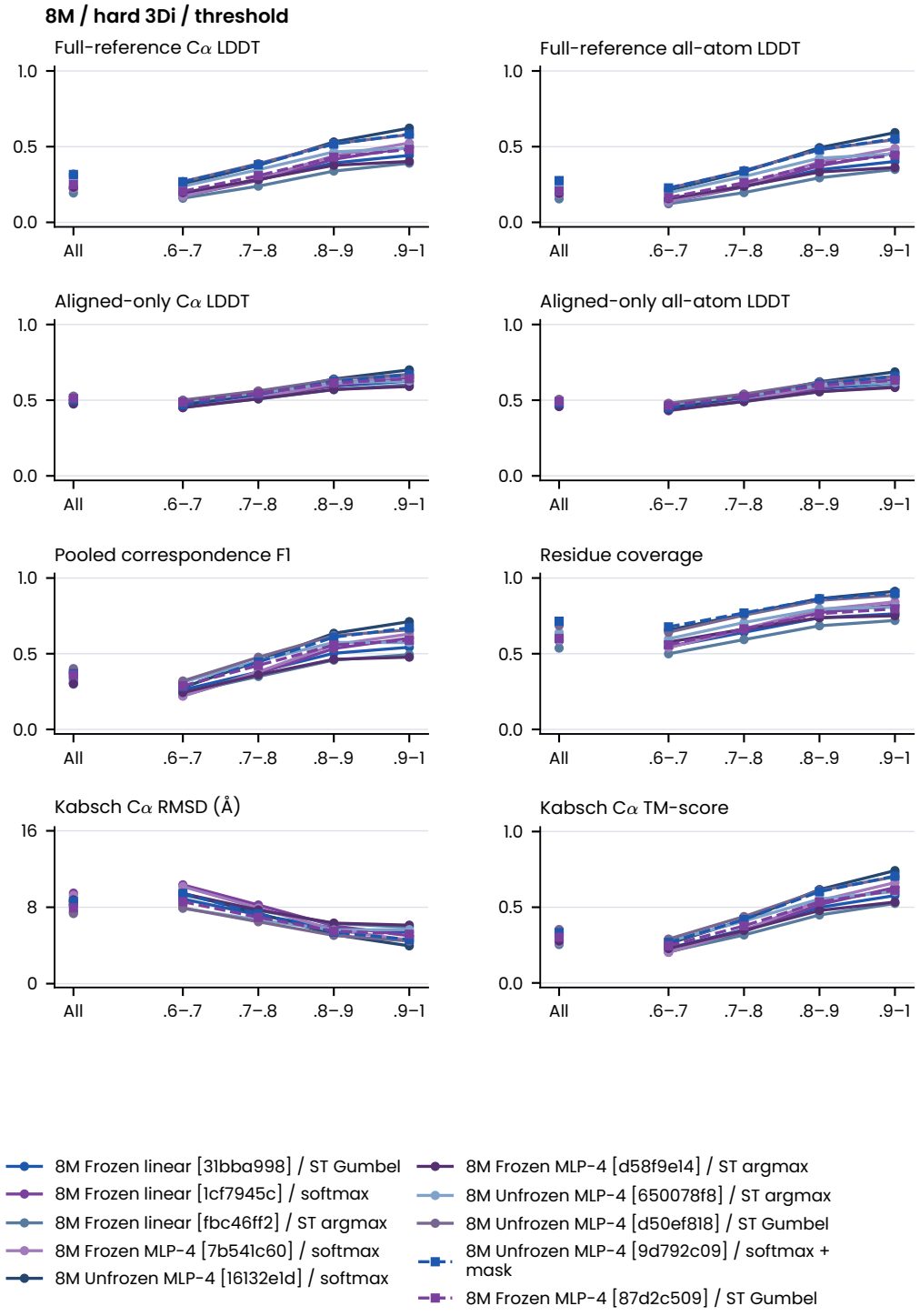


Figure S49: 8M / hard 3Di / threshold. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

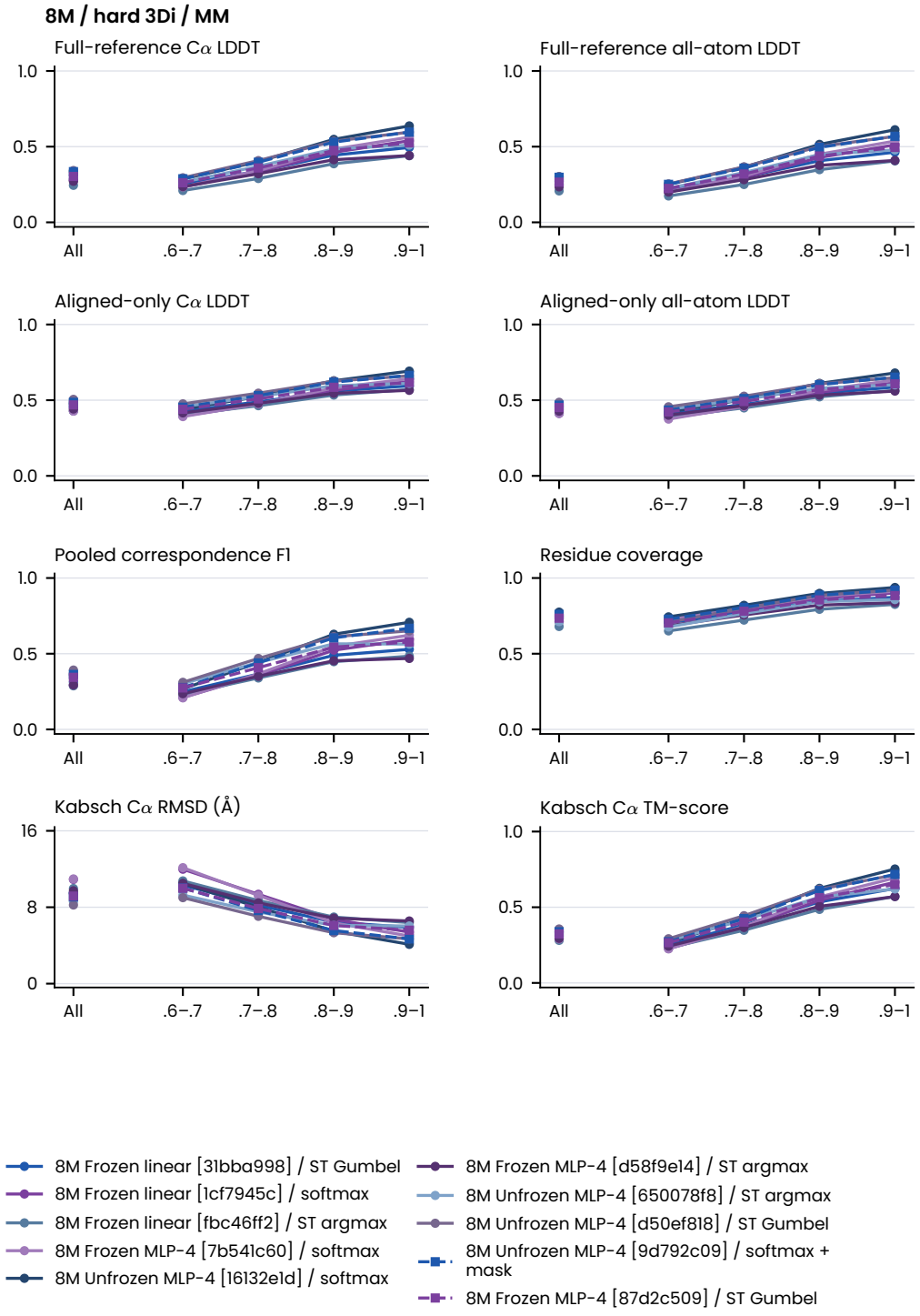


Figure S50: 8M / hard 3Di / MM. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

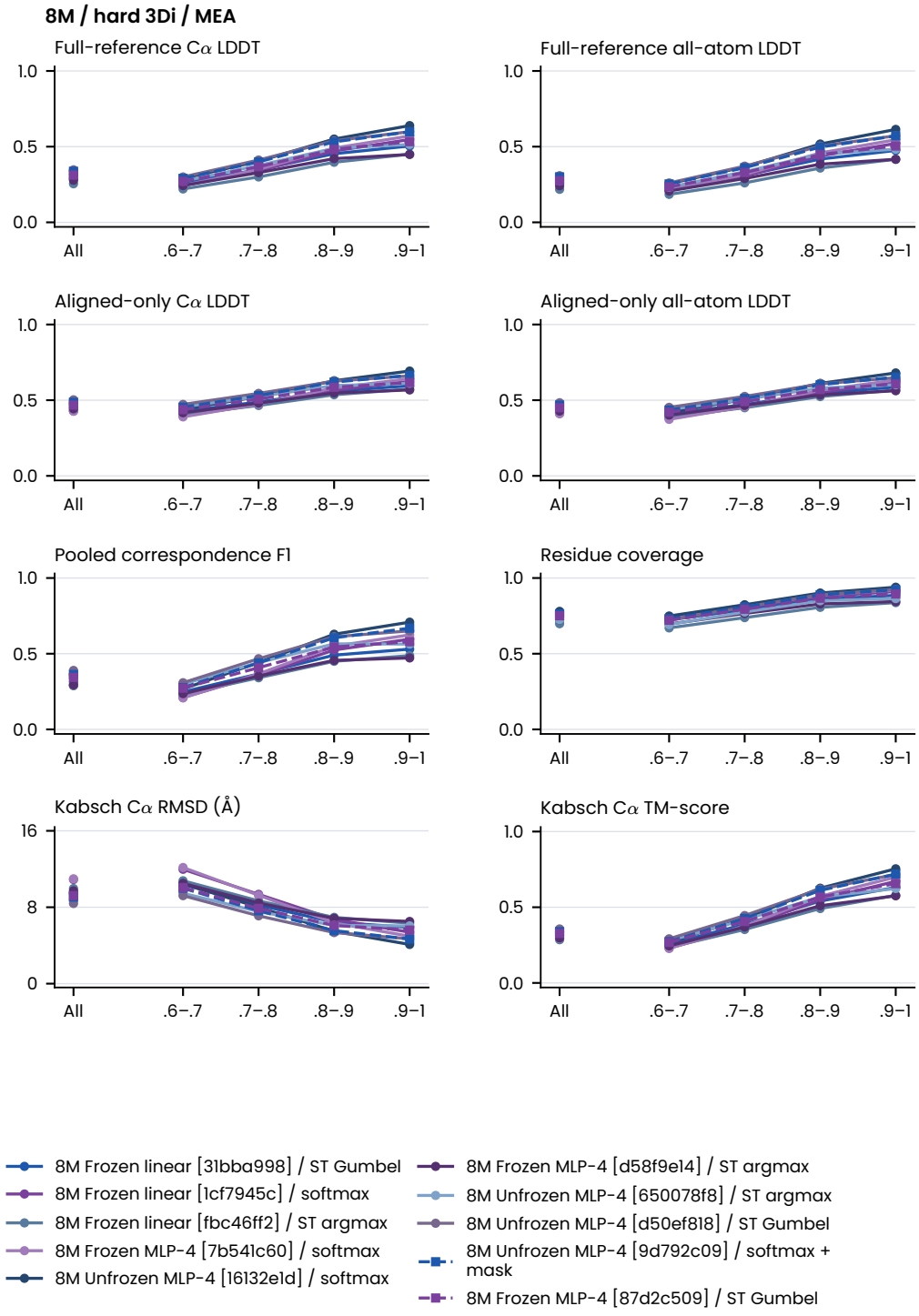


Figure S51: 8M / hard 3Di / MEA. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

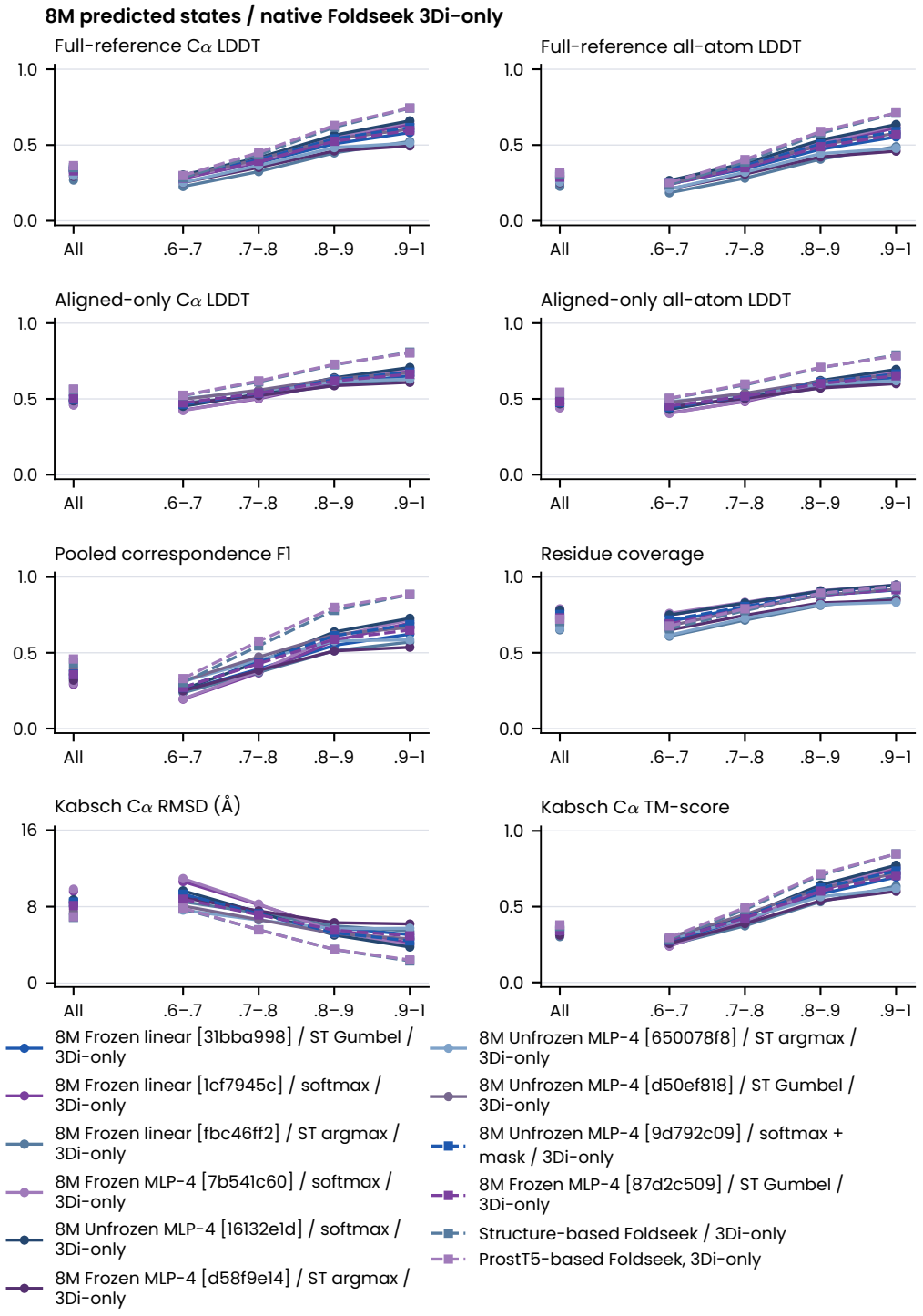


Figure S52: 8M predicted states / native Foldseek 3Di-only. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

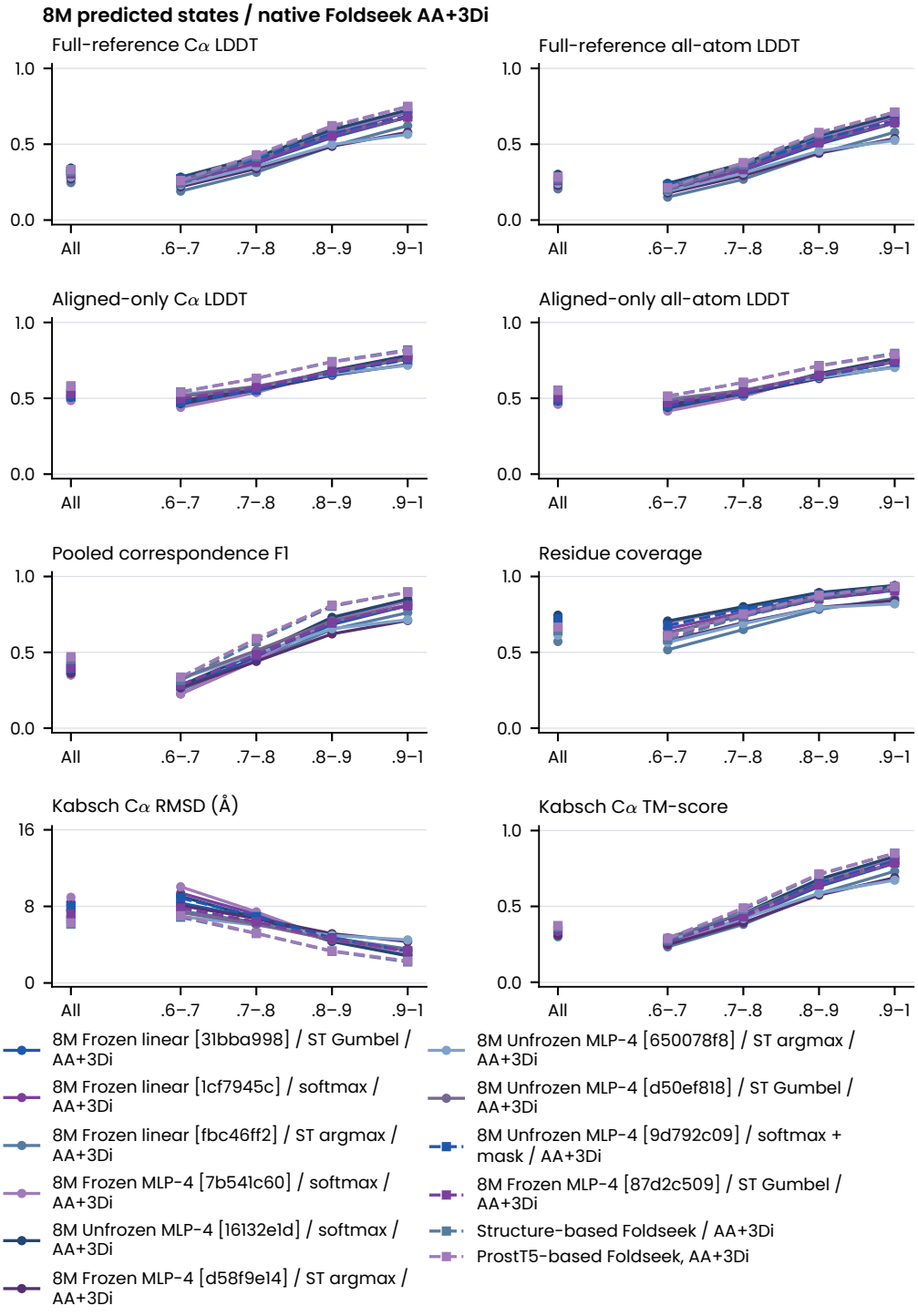


Figure S53: 8M predicted states / native Foldseek AA+3Di. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

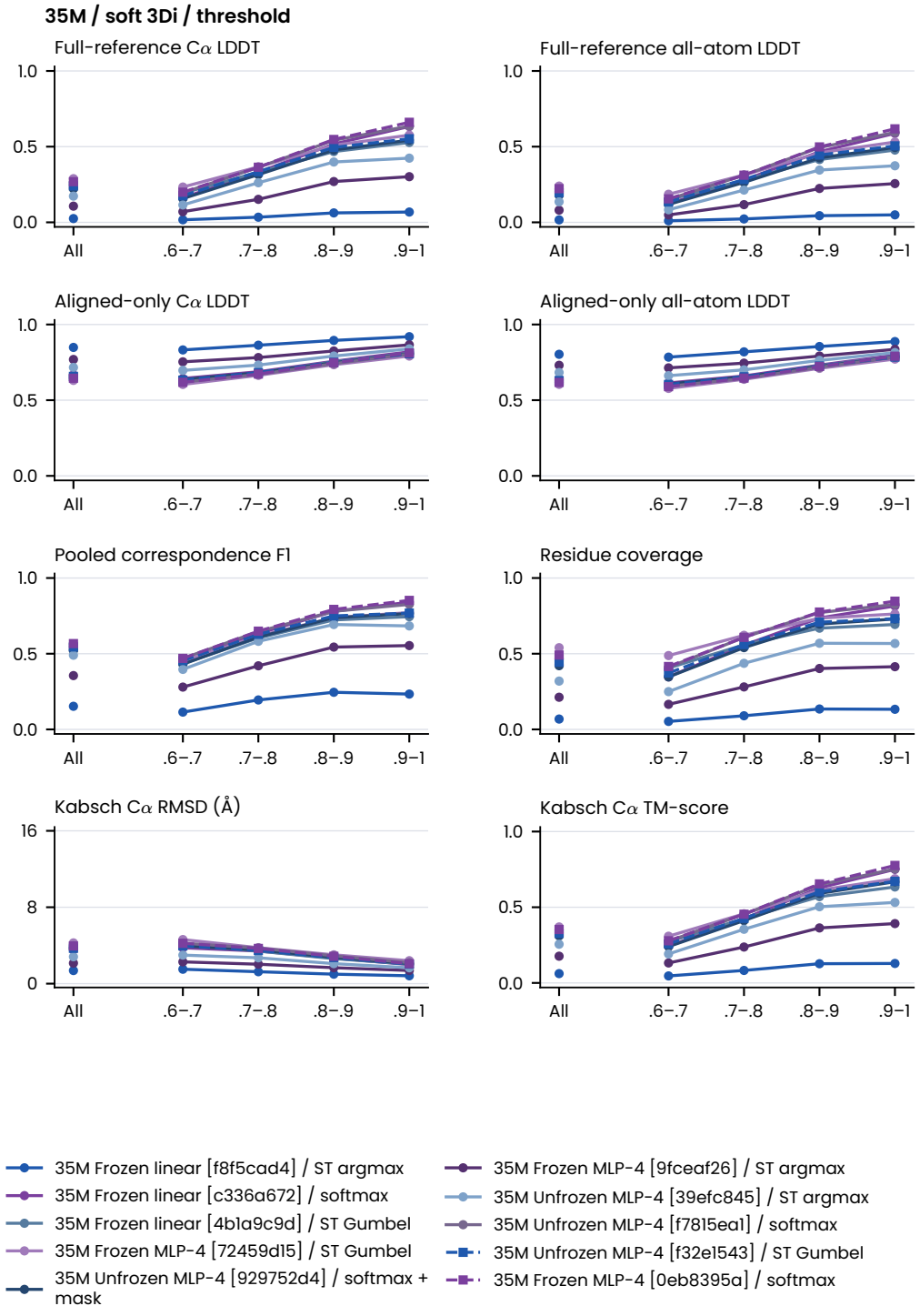


Figure S54: 35M / soft 3Di / threshold. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

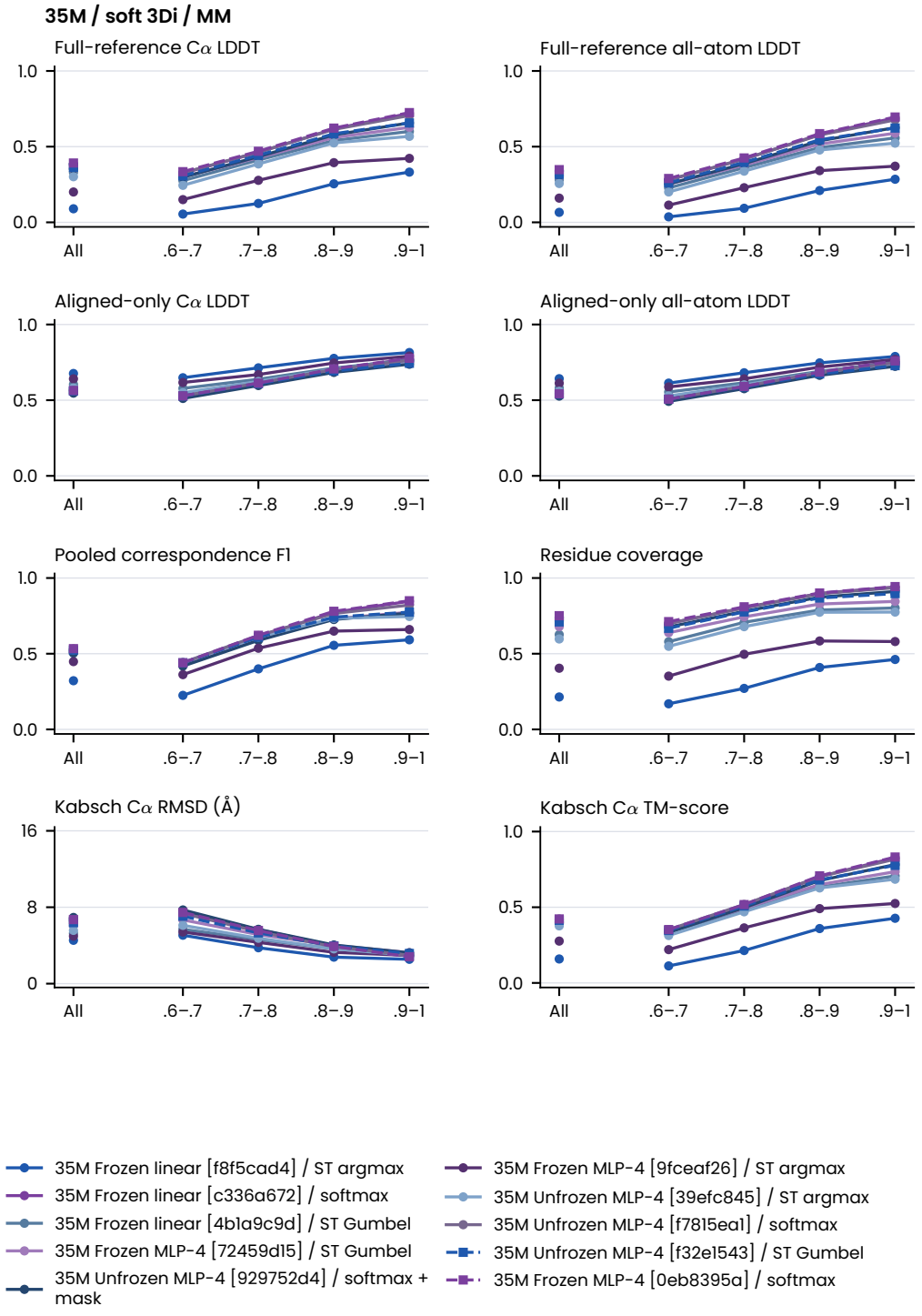


Figure S55: 35M / soft 3Di / MM. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

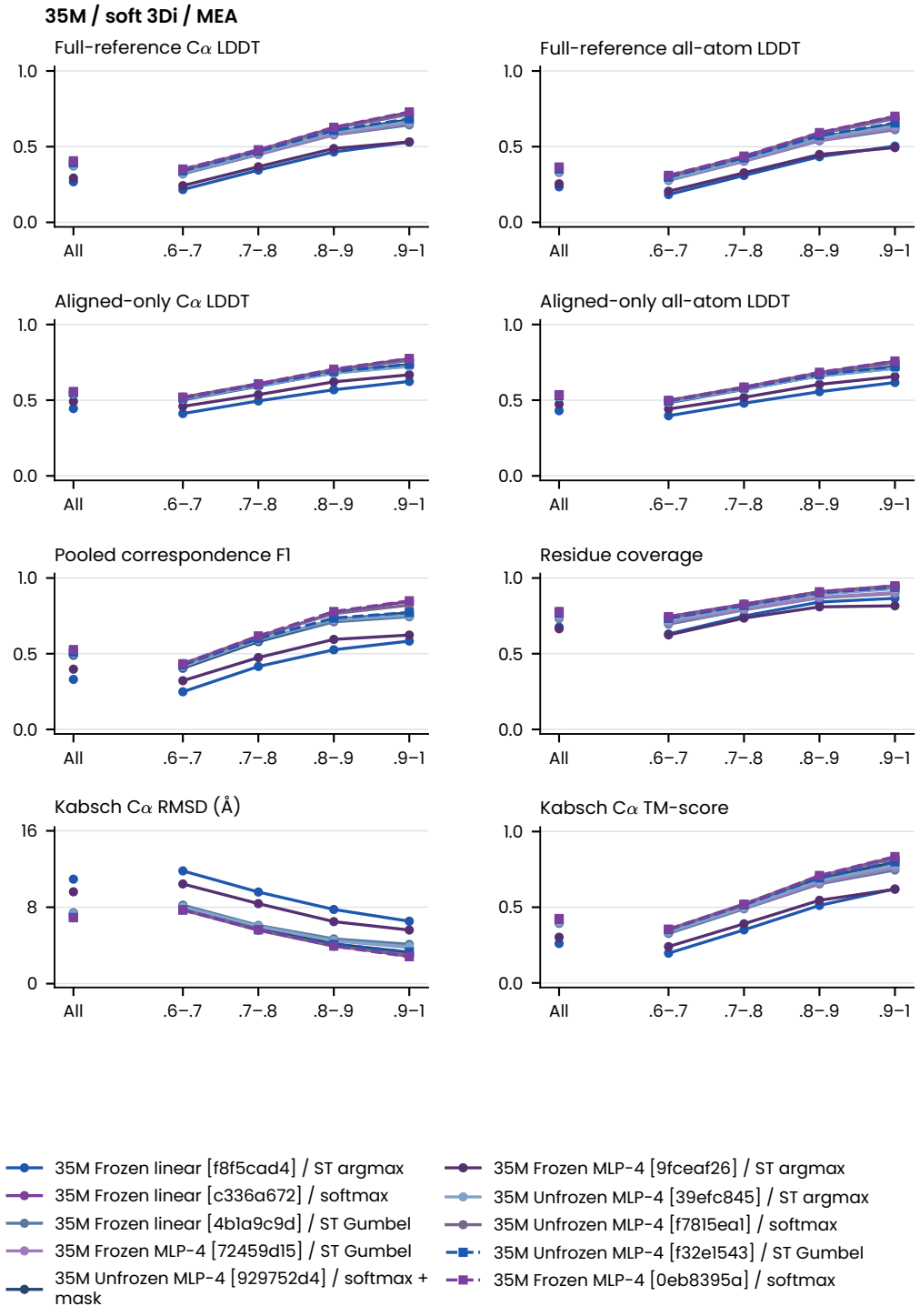


Figure S56: 35M / soft 3Di / MEA. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

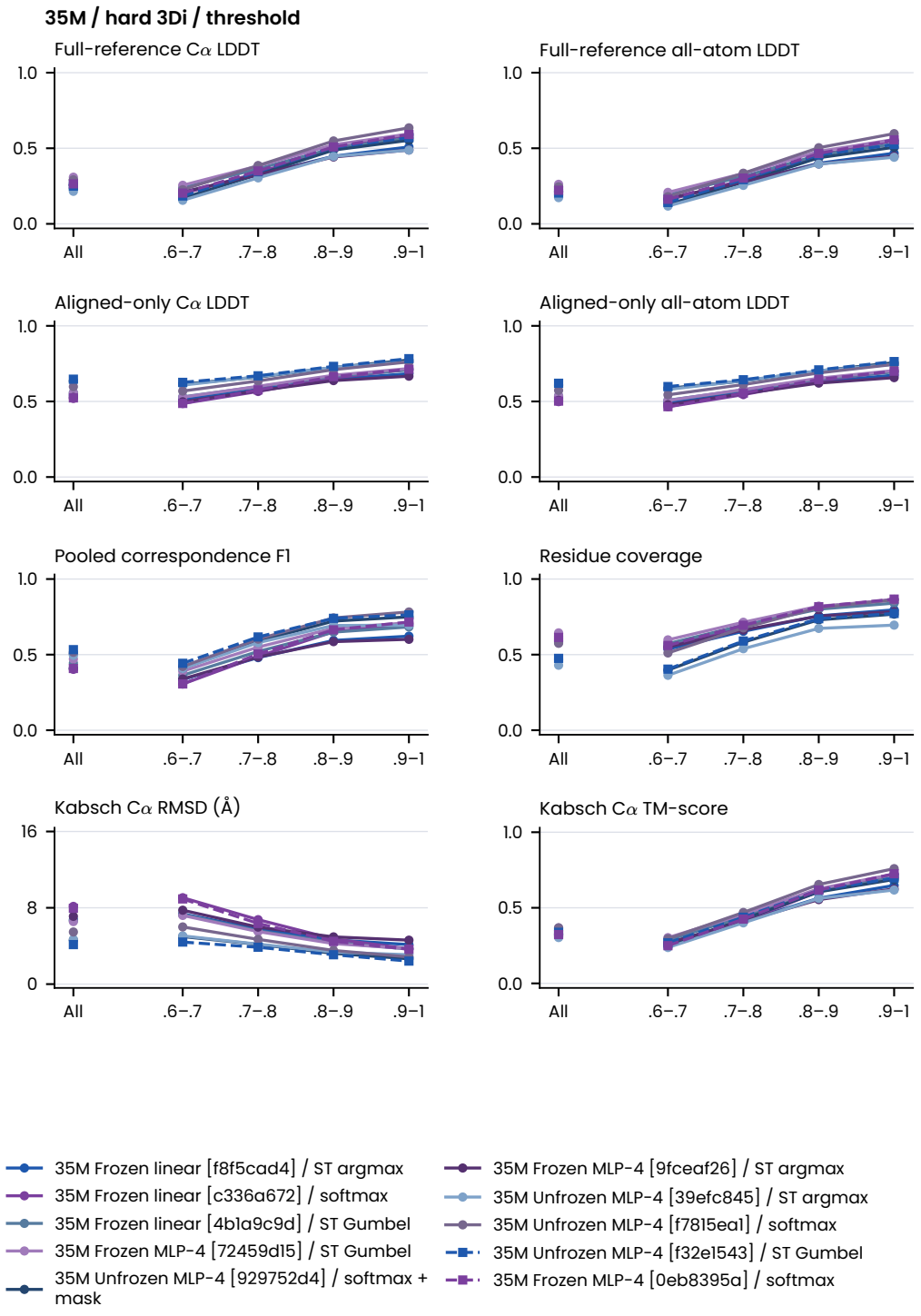


Figure S57: 35M / hard 3Di / threshold. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

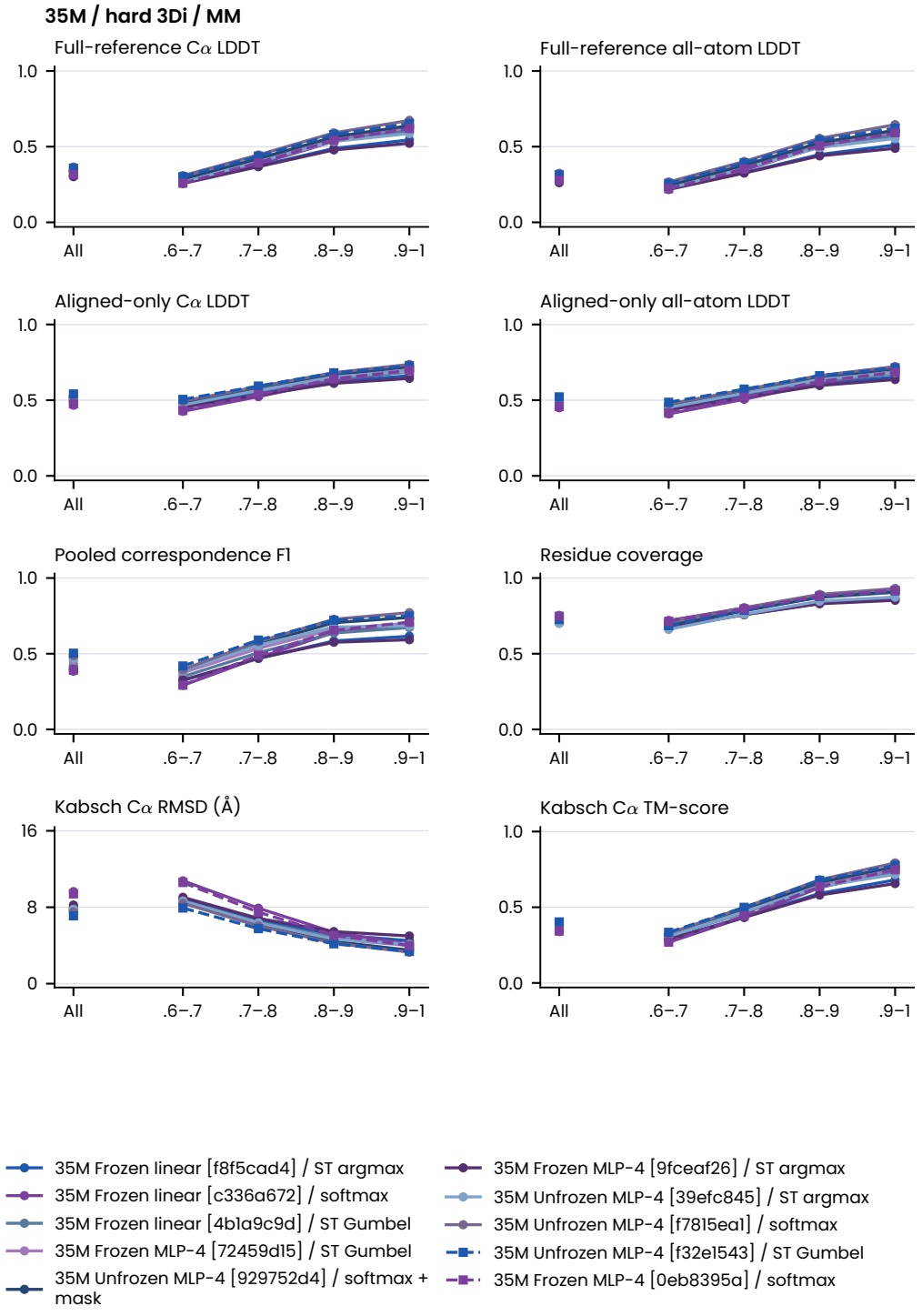


Figure S58: 35M / hard 3Di / MM. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

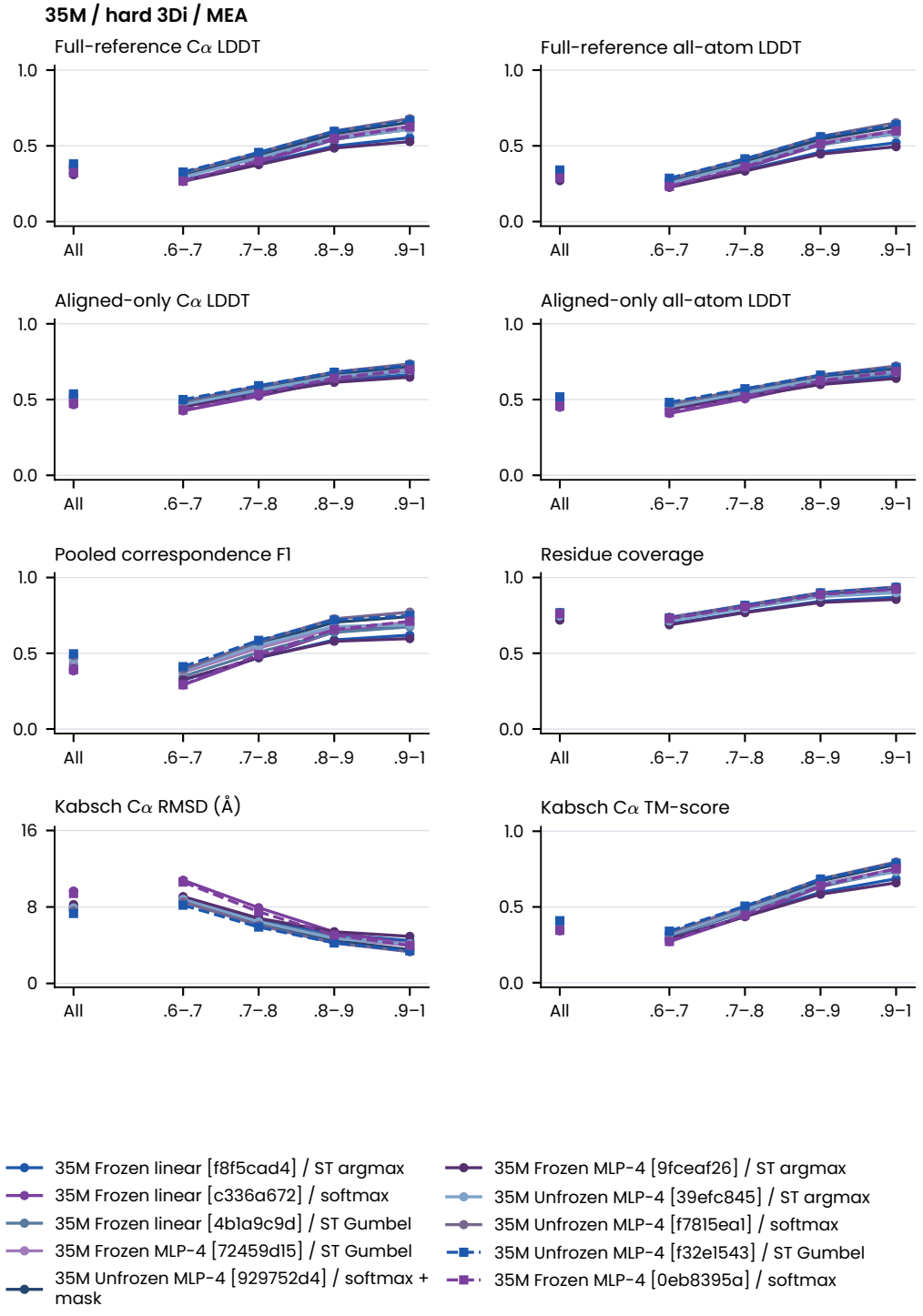


Figure S59: 35M / hard 3Di / MEA. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

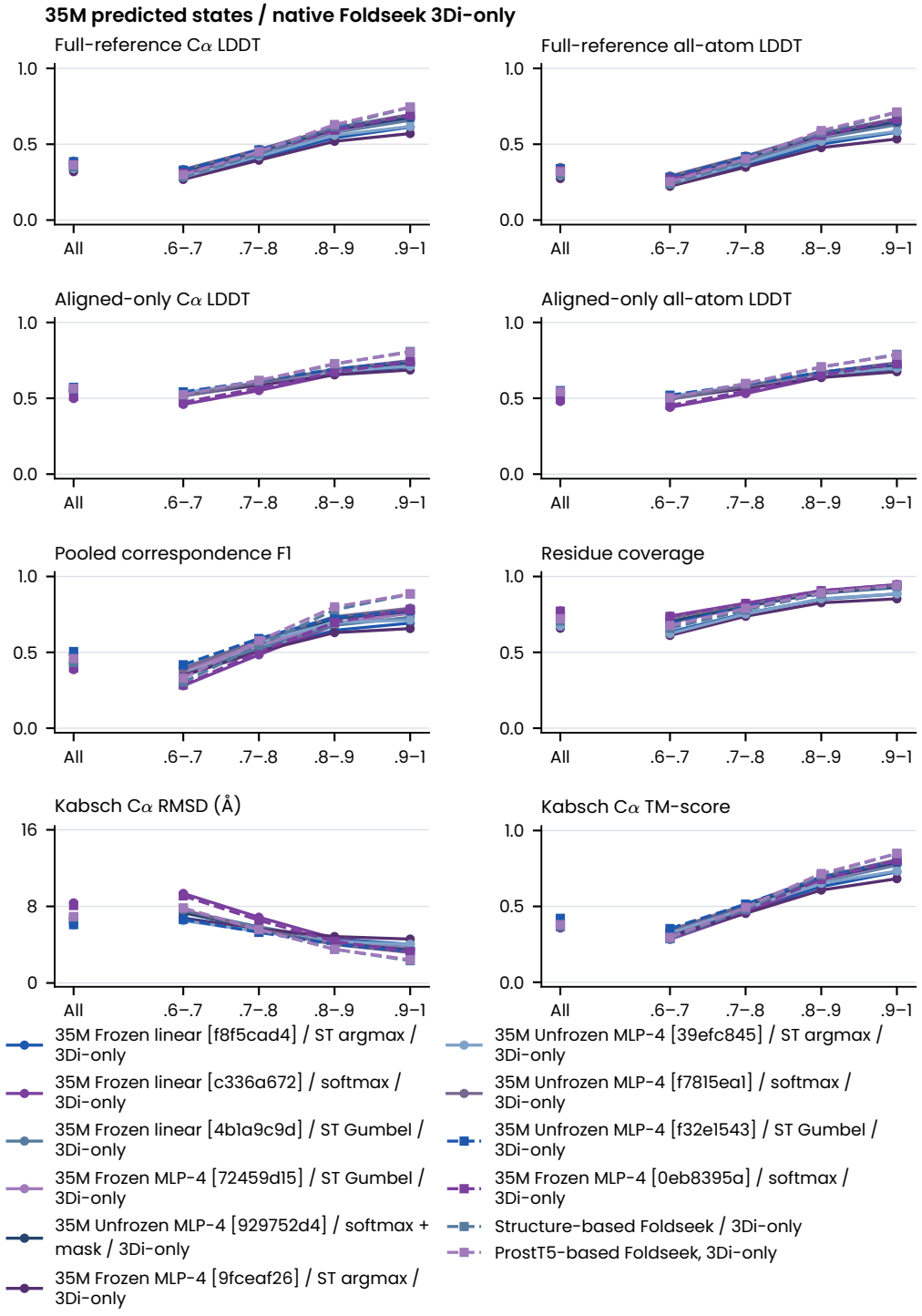


Figure S60: 35M predicted states / native Foldseek 3Di-only. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

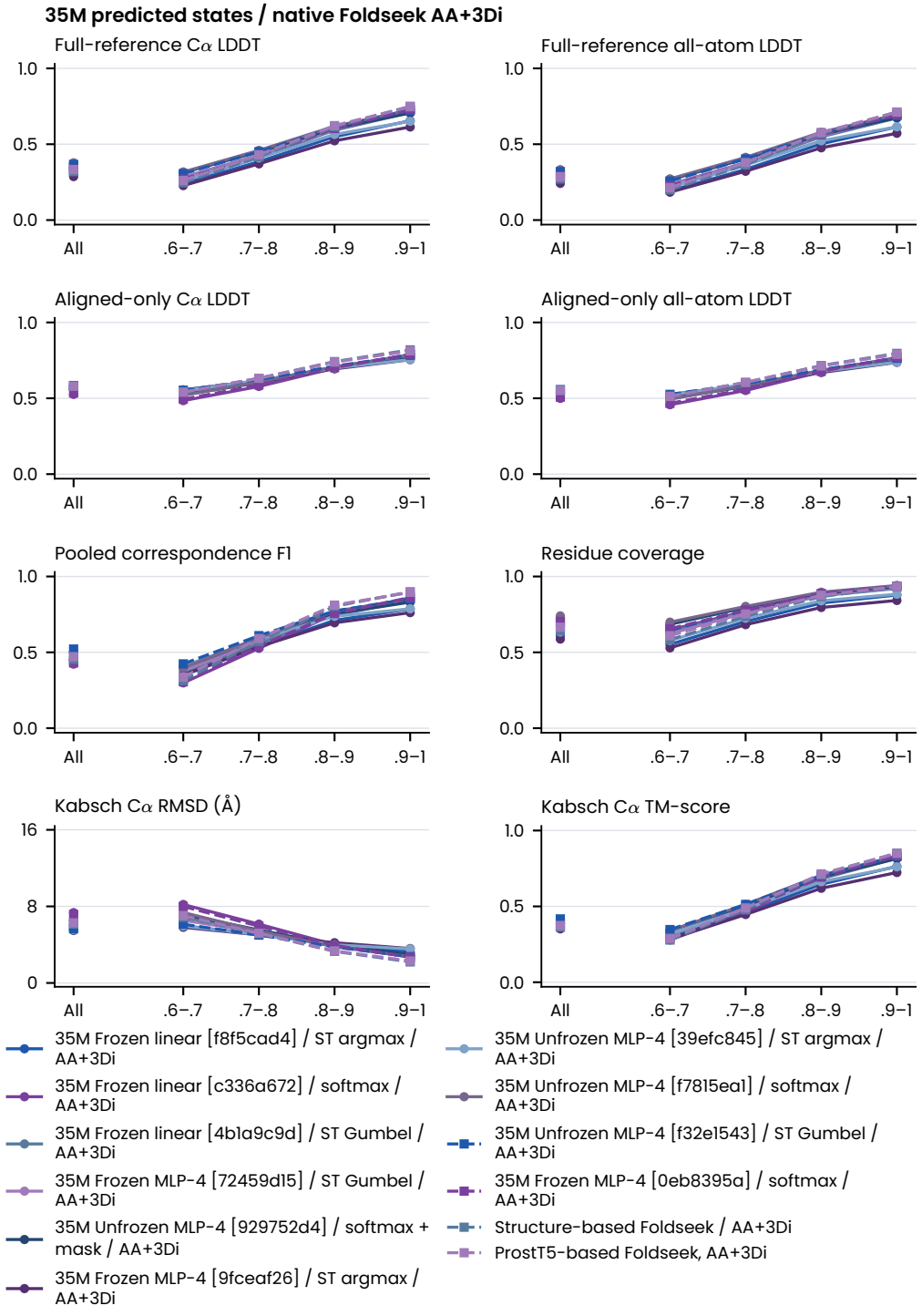


Figure S61: 35M predicted states / native Foldseek AA+3Di. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

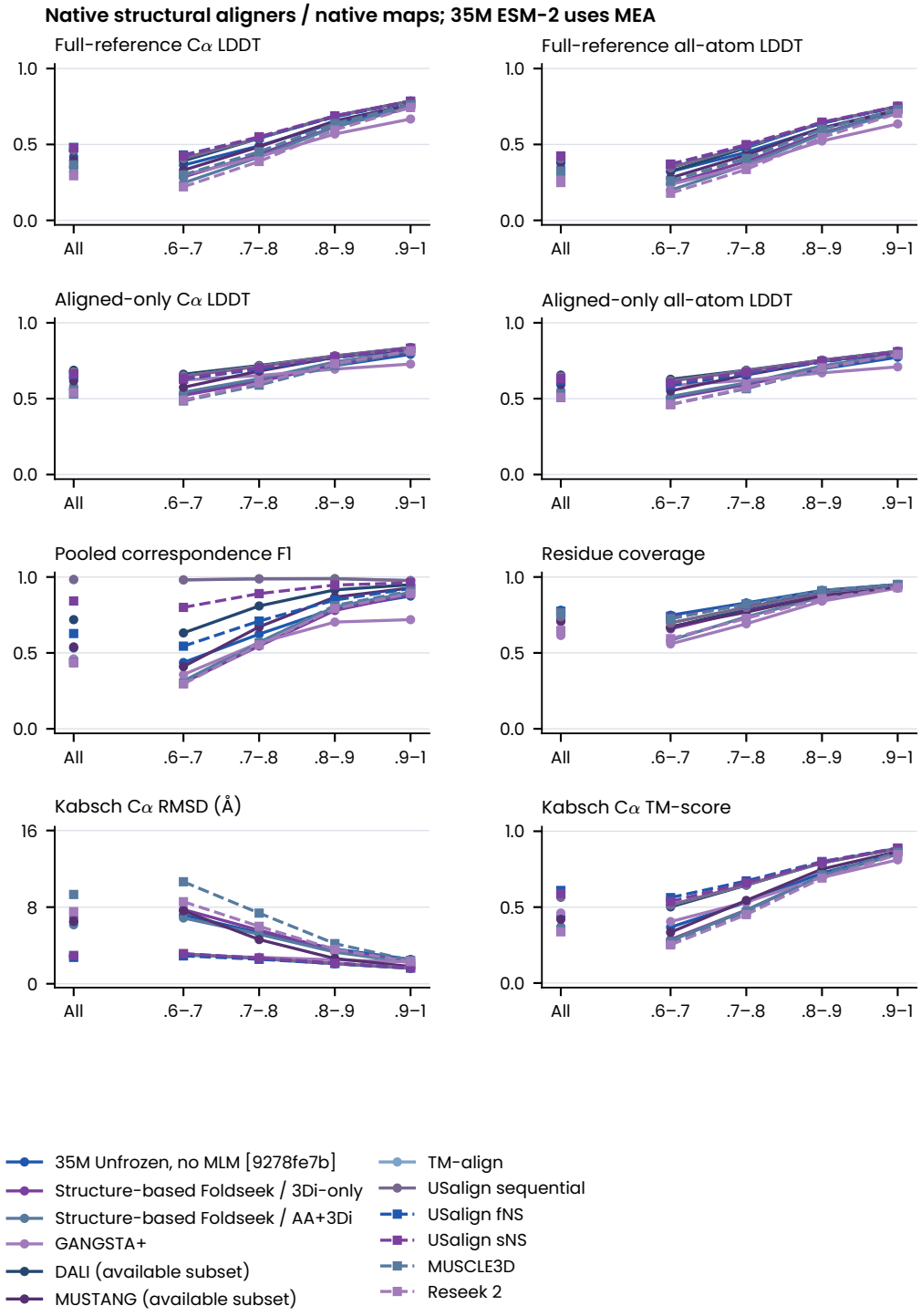


Figure S62: Native structural aligners / native maps; 35M ESM-2 uses MEA. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

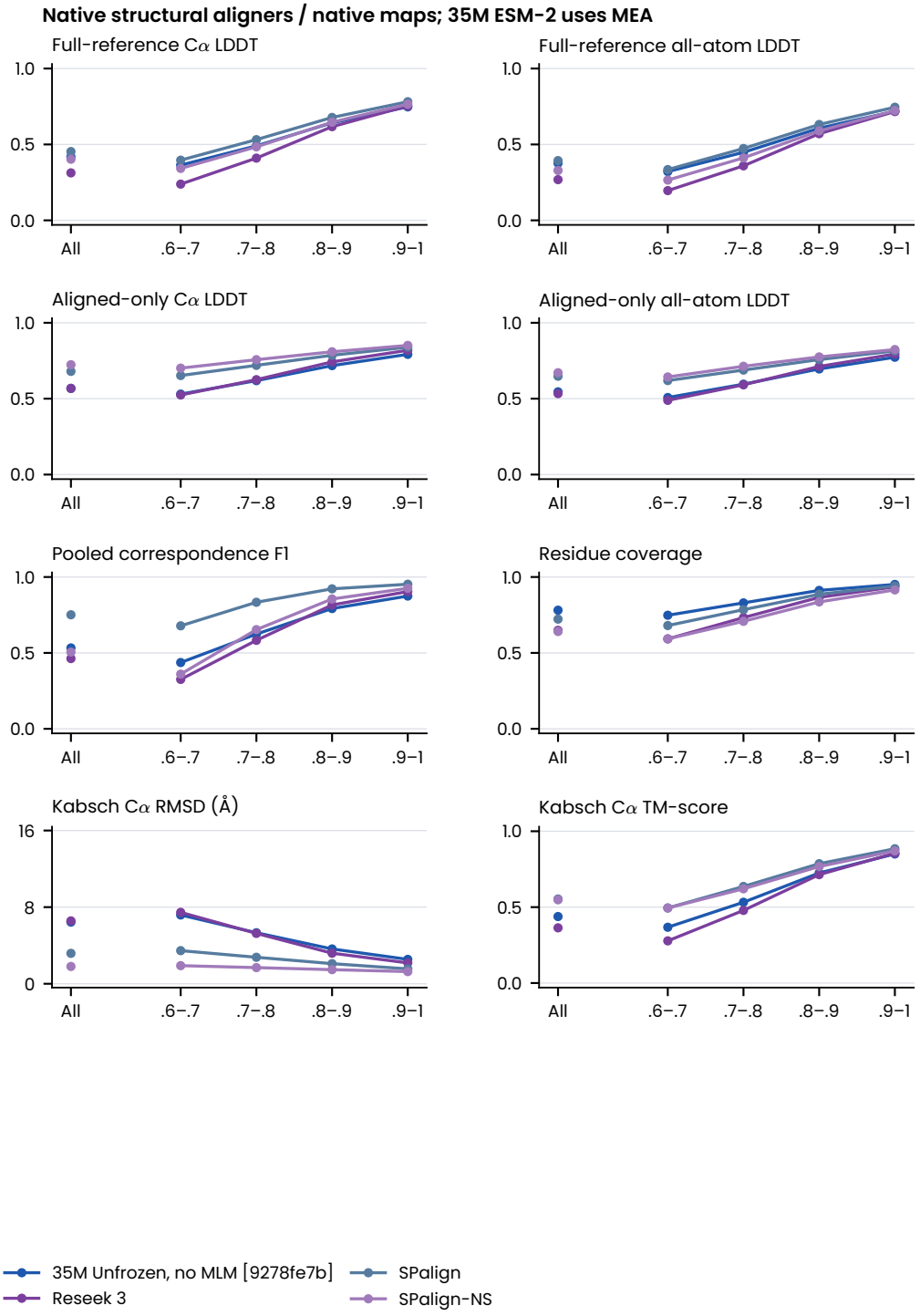


Figure S63: Native structural aligners / native maps; 35M ESM-2 uses MEA. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

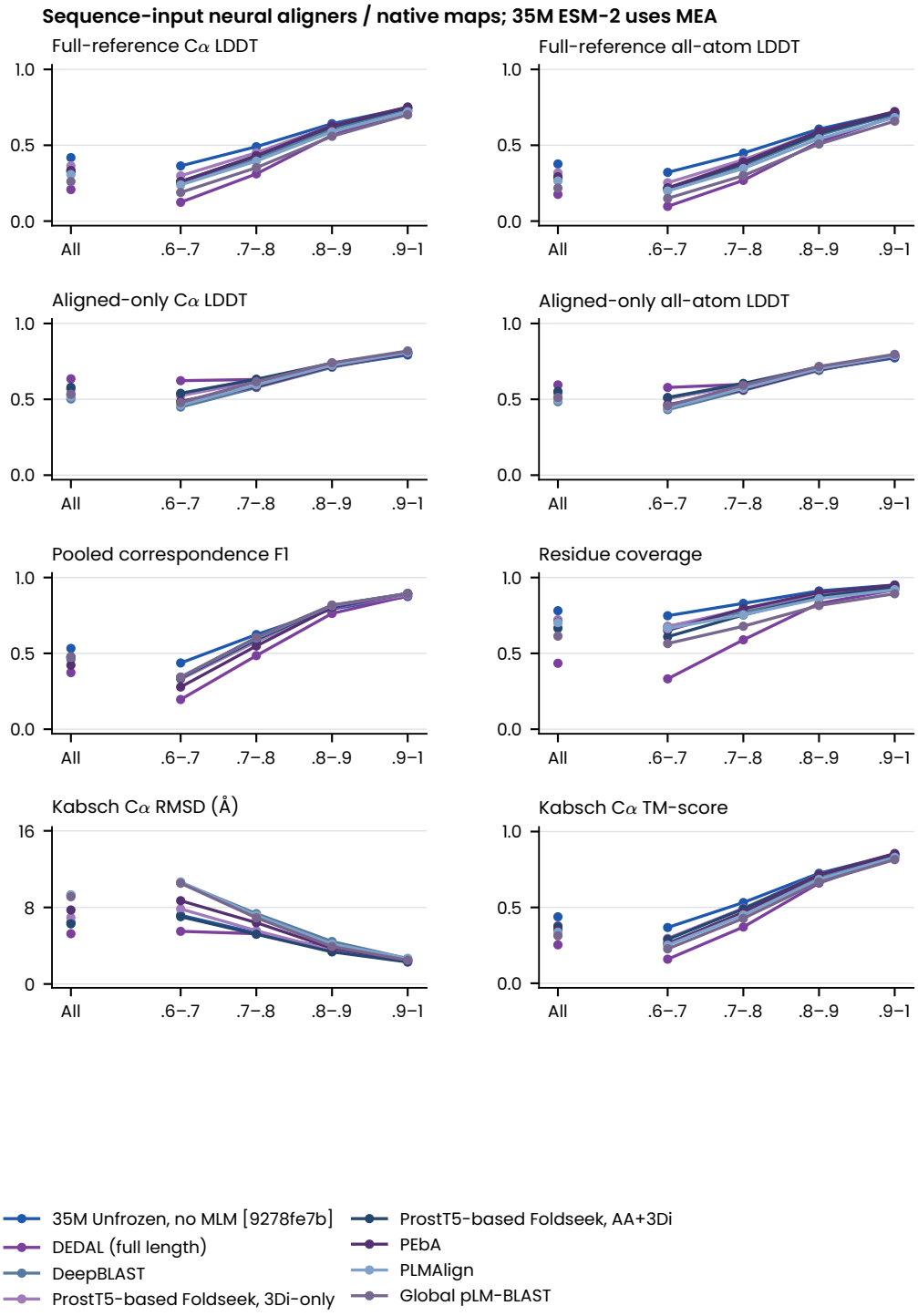


Figure S64: Sequence-input neural aligners / native maps; 35M ESM-2 uses MEA. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

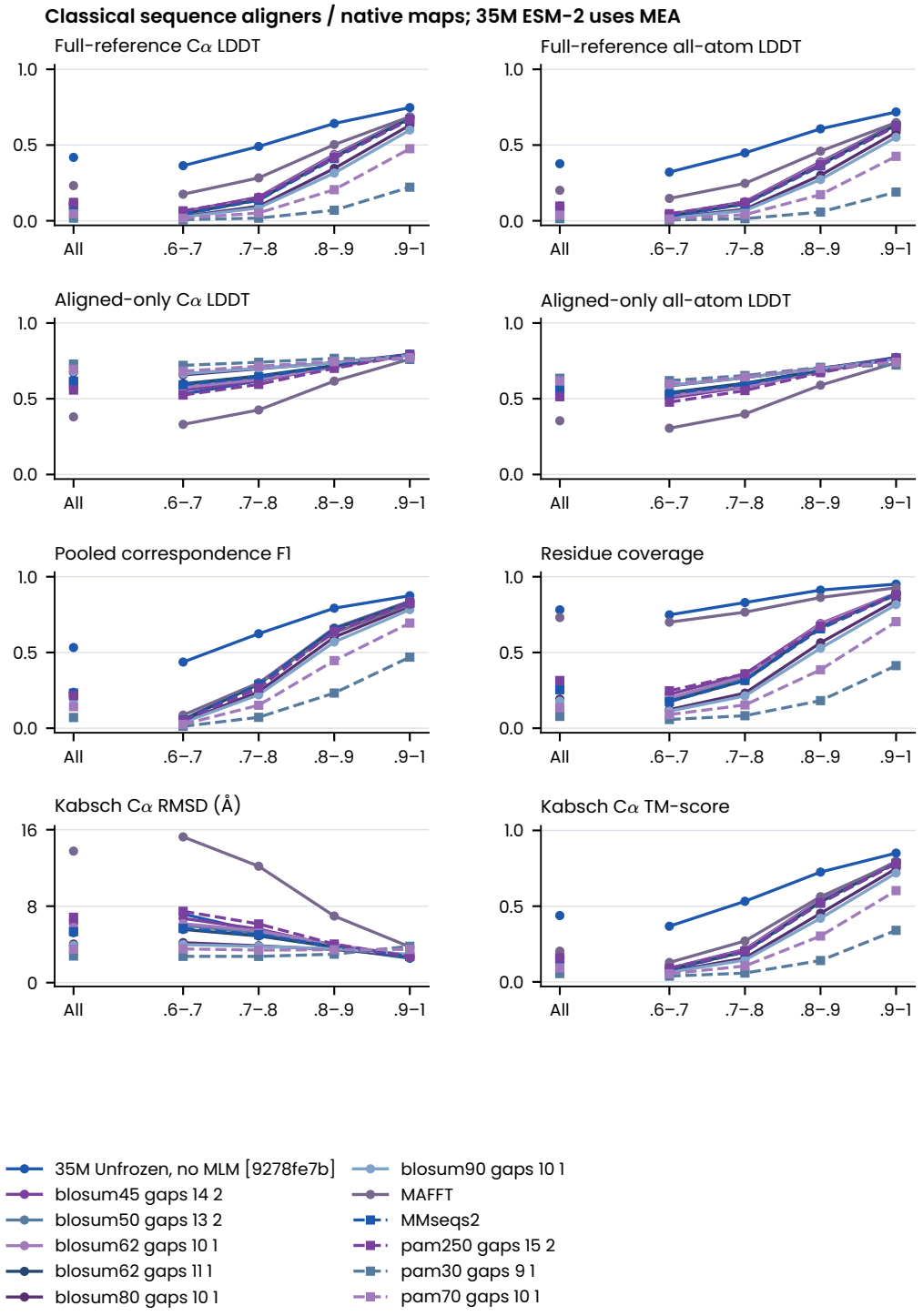


Figure S65: Classical sequence aligners / native maps; 35M ESM-2 uses MEA. Model curves are overlaid on full metric axes; the named correspondence decoder supplies all eight metrics. MM denotes mutual matching. Native reference tools retain their own alignments. The separate left point is the overall aggregate; the four connected points are reference TM-score intervals. The first three intervals exclude their upper endpoint; the last includes 1.0. All values use a 0–1 scale except RMSD in Å.

6731 6730 6729 6728 6727 6726 6725 6724 6723 6722 6721 6720 6719 6718 6717 6716 6715 6714 6713 6712 6711 6710 6709 6708 6707 6706 6705 6704 6703 6702 6701 6700 6699

Aggregate quality / continuous

	Cα full			Cα aligned			All-atom full			All-atom aligned			F1			Coverage		
	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA
8M Unfrozen, no MLM [5675f423]	0.274	0.382	0.389	0.579	0.534	0.529	0.233	0.342	0.351	0.554	0.512	0.508	0.480	0.446	0.442	0.540	0.771	0.786
8M Unfrozen + MLM [903cf320]	0.255	0.377	0.386	0.584	0.529	0.523	0.216	0.337	0.348	0.559	0.507	0.502	0.465	0.430	0.426	0.505	0.770	0.789
35M Unfrozen + MLM [f7fbcb88]	0.314	0.409	0.415	0.600	0.566	0.562	0.269	0.366	0.374	0.575	0.544	0.540	0.552	0.520	0.517	0.582	0.771	0.783
35M Unfrozen, no MLM [9278fe7b]	0.333	0.414	0.418	0.600	0.571	0.568	0.288	0.371	0.377	0.575	0.548	0.545	0.564	0.535	0.533	0.614	0.772	0.781
150M Unfrozen + MLM [00728588]	0.275	0.383	0.390	0.577	0.531	0.527	0.233	0.343	0.352	0.552	0.510	0.506	0.484	0.449	0.445	0.544	0.773	0.788
8M LoRA + MLM [11273f46]	0.211	0.349	0.363	0.586	0.502	0.494	0.177	0.311	0.328	0.560	0.481	0.474	0.388	0.363	0.358	0.443	0.764	0.793
35M LoRA + MLM [5995efb6]	0.198	0.346	0.360	0.605	0.501	0.493	0.165	0.309	0.326	0.577	0.480	0.473	0.397	0.371	0.365	0.413	0.759	0.790
150M LoRA + MLM [9976e4d4]	0.218	0.330	0.342	0.552	0.478	0.472	0.184	0.295	0.309	0.526	0.458	0.453	0.359	0.333	0.329	0.492	0.764	0.790
8M Random init + MLM [435d65ae]	0.102	0.290	0.296	0.502	0.398	0.396	0.083	0.262	0.269	0.477	0.381	0.378	0.132	0.145	0.144	0.288	0.808	0.821
8M Frozen gaps only [71222270]	0.217	0.234	0.236	0.344	0.331	0.331	0.192	0.212	0.214	0.323	0.312	0.312	0.142	0.140	0.139	0.761	0.817	0.821
35M Frozen gaps only [3041e38b]	0.209	0.223	0.224	0.324	0.314	0.314	0.187	0.202	0.204	0.304	0.295	0.296	0.134	0.132	0.132	0.770	0.817	0.821
150M Frozen gaps only [86ff5a2d]	0.209	0.222	0.223	0.321	0.311	0.311	0.186	0.201	0.203	0.301	0.293	0.293	0.132	0.130	0.130	0.773	0.817	0.821
8M Frozen BiGRU-1 128D [65456b5e]	0.126	0.310	0.324	0.560	0.449	0.441	0.104	0.276	0.293	0.535	0.431	0.423	0.260	0.252	0.248	0.294	0.774	0.803
8M Frozen BiGRU-2 128D [1fd522dd]	0.173	0.327	0.338	0.556	0.470	0.463	0.142	0.291	0.306	0.530	0.451	0.444	0.321	0.298	0.294	0.399	0.772	0.797
8M Frozen BiLSTM-1 128D [5ce20326]	0.144	0.316	0.328	0.554	0.455	0.447	0.118	0.281	0.296	0.530	0.436	0.429	0.278	0.262	0.259	0.334	0.774	0.801
8M Frozen BiLSTM-2 128D [5ac3e73e]	0.183	0.331	0.342	0.554	0.475	0.468	0.151	0.295	0.309	0.529	0.455	0.449	0.338	0.312	0.308	0.418	0.771	0.795
8M Frozen CNN-3 128D [a882d927]	0.176	0.304	0.313	0.494	0.433	0.429	0.146	0.272	0.284	0.472	0.416	0.411	0.266	0.241	0.238	0.443	0.779	0.801
8M Frozen CNN-5 128D [9d50e6e2]	0.194	0.312	0.321	0.495	0.444	0.440	0.162	0.280	0.290	0.473	0.426	0.422	0.289	0.264	0.261	0.481	0.780	0.799
8M Frozen CNN-7 128D [ddac38b6]	0.206	0.317	0.325	0.493	0.448	0.445	0.173	0.285	0.294	0.471	0.430	0.427	0.299	0.275	0.272	0.508	0.782	0.799
8M Frozen CNN-9 128D [4e16faca]	0.212	0.320	0.328	0.492	0.452	0.448	0.179	0.287	0.296	0.470	0.434	0.430	0.305	0.281	0.279	0.520	0.782	0.798
8M Frozen Linear 128D [b5ae1da1]	0.101	0.257	0.266	0.543	0.374	0.369	0.084	0.230	0.240	0.510	0.357	0.353	0.154	0.142	0.140	0.274	0.780	0.805
8M Frozen MLP-2 128D [bd73acd5]	0.061	0.264	0.280	0.531	0.391	0.382	0.053	0.234	0.254	0.513	0.374	0.365	0.140	0.153	0.150	0.135	0.773	0.814
8M Frozen MLP-3 128D [293fb2db]	0.099	0.275	0.289	0.499	0.402	0.394	0.083	0.245	0.262	0.479	0.385	0.377	0.192	0.181	0.177	0.246	0.776	0.810
8M Frozen MLP-4 128D [977f5e2e]	0.135	0.284	0.295	0.483	0.409	0.403	0.112	0.253	0.267	0.462	0.392	0.386	0.217	0.196	0.193	0.351	0.779	0.806
8M Frozen Projectformer-2 128D [762e1c9a]	0.168	0.305	0.316	0.510	0.438	0.432	0.140	0.272	0.285	0.486	0.419	0.414	0.278	0.249	0.246	0.410	0.775	0.799
8M Frozen Projectformer-4 128D [abdee92f]	0.194	0.316	0.325	0.502	0.450	0.445	0.163	0.283	0.294	0.479	0.431	0.427	0.299	0.270	0.268	0.472	0.776	0.796

Figure S66: Full-validation alignment quality for every selected checkpoint in the indicated representation. All four LDDT definitions and F1 appear together, each with threshold (T), mutual matching (MM), and MEA decoding; coverage is included alongside them. Three-decimal display values preserve the same saved endpoints; exact values remain in the source ledger.

6764 6763 6762 6761 6760 6759 6758 6757 6756 6755 6754 6753 6752 6751 6750 6749 6748 6747 6746 6745 6744 6743 6742 6741 6740 6739 6738 6737 6736 6735 6734 6733 6732

Aggregate quality / continuous

	Cα full			Cα aligned			All-atom full			All-atom aligned			F1			Coverage		
	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA
8M Frozen BiGRU-1 64D [56a38fec]	0.126	0.310	0.323	0.558	0.449	0.441	0.104	0.276	0.292	0.534	0.430	0.422	0.256	0.249	0.245	0.296	0.774	0.803
8M Frozen BiGRU-2 64D [0ca12fc7]	0.167	0.324	0.336	0.554	0.466	0.459	0.137	0.289	0.303	0.529	0.447	0.441	0.312	0.290	0.286	0.387	0.773	0.798
8M Frozen BiLSTM-1 64D [17569f0a]	0.141	0.314	0.326	0.552	0.453	0.445	0.116	0.280	0.295	0.528	0.434	0.427	0.273	0.259	0.255	0.330	0.774	0.802
8M Frozen BiLSTM-2 64D [fd8eabfd]	0.172	0.324	0.336	0.549	0.467	0.460	0.141	0.289	0.303	0.524	0.448	0.441	0.315	0.292	0.289	0.401	0.771	0.796
8M Frozen CNN-3 64D [9e6edc2a]	0.163	0.300	0.311	0.498	0.431	0.425	0.135	0.268	0.282	0.476	0.413	0.408	0.255	0.232	0.229	0.410	0.778	0.802
8M Frozen CNN-5 64D [dd52dd69]	0.180	0.310	0.320	0.506	0.443	0.438	0.150	0.277	0.290	0.484	0.425	0.420	0.284	0.257	0.254	0.444	0.778	0.800
8M Frozen CNN-7 64D [824ccff8]	0.193	0.315	0.324	0.502	0.448	0.444	0.161	0.282	0.293	0.480	0.430	0.426	0.295	0.269	0.266	0.474	0.779	0.798
8M Frozen CNN-9 64D [5ac00304]	0.201	0.317	0.325	0.497	0.450	0.446	0.168	0.284	0.294	0.475	0.432	0.428	0.298	0.273	0.271	0.494	0.780	0.798
8M Frozen Linear 64D [c203b430]	0.085	0.258	0.267	0.522	0.372	0.368	0.071	0.231	0.241	0.495	0.355	0.351	0.145	0.138	0.136	0.221	0.785	0.808
8M Frozen MLP-2 64D [bee2c633]	0.050	0.260	0.278	0.540	0.387	0.377	0.043	0.230	0.252	0.523	0.371	0.361	0.118	0.143	0.140	0.107	0.771	0.815
8M Frozen MLP-3 64D [ea28b452]	0.080	0.270	0.285	0.505	0.396	0.388	0.068	0.240	0.258	0.486	0.379	0.372	0.170	0.167	0.164	0.194	0.775	0.812
8M Frozen MLP-4 64D [5c762f69]	0.116	0.278	0.291	0.484	0.405	0.398	0.097	0.248	0.264	0.463	0.388	0.381	0.201	0.185	0.182	0.301	0.777	0.808
8M Frozen Projectformer-2 64D [cfb98e58]	0.134	0.292	0.305	0.515	0.422	0.415	0.112	0.261	0.277	0.492	0.405	0.398	0.245	0.219	0.216	0.325	0.775	0.806
8M Frozen Projectformer-4 64D [70dc0520]	0.182	0.311	0.321	0.499	0.443	0.438	0.152	0.278	0.290	0.476	0.424	0.419	0.284	0.256	0.254	0.448	0.778	0.799
35M Frozen BiGRU-1 128D [ecb8903e]	0.146	0.323	0.340	0.615	0.474	0.466	0.117	0.287	0.307	0.585	0.455	0.447	0.340	0.321	0.316	0.326	0.760	0.796
35M Frozen BiGRU-2 128D [0469b803]	0.180	0.330	0.344	0.574	0.479	0.472	0.147	0.294	0.310	0.548	0.459	0.453	0.356	0.327	0.323	0.407	0.766	0.793
35M Frozen BiLSTM-1 128D [fcb59fc4]	0.155	0.325	0.340	0.599	0.474	0.466	0.125	0.288	0.307	0.571	0.455	0.447	0.341	0.318	0.313	0.347	0.762	0.795
35M Frozen BiLSTM-2 128D [3906f03e]	0.197	0.334	0.345	0.554	0.481	0.475	0.162	0.298	0.311	0.529	0.461	0.455	0.359	0.330	0.327	0.445	0.768	0.791
35M Frozen CNN-3 128D [d9f0e5f5]	0.170	0.305	0.316	0.515	0.438	0.433	0.140	0.272	0.286	0.492	0.420	0.416	0.283	0.256	0.252	0.421	0.775	0.799
35M Frozen CNN-5 128D [c0b23985]	0.197	0.312	0.320	0.494	0.443	0.439	0.165	0.280	0.290	0.472	0.425	0.421	0.291	0.266	0.263	0.489	0.781	0.798
35M Frozen CNN-7 128D [08980ed6]	0.209	0.317	0.324	0.489	0.448	0.444	0.176	0.284	0.293	0.468	0.430	0.427	0.298	0.274	0.272	0.518	0.782	0.798
35M Frozen CNN-9 128D [32e58231]	0.218	0.321	0.328	0.488	0.452	0.448	0.184	0.288	0.297	0.467	0.433	0.430	0.306	0.283	0.280	0.535	0.783	0.798
35M Frozen Linear 128D [662617lf]	0.056	0.259	0.275	0.580	0.385	0.377	0.048	0.229	0.249	0.560	0.367	0.361	0.143	0.162	0.158	0.118	0.773	0.813
35M Frozen MLP-2 128D [9208eb3e]	0.106	0.292	0.309	0.602	0.431	0.422	0.087	0.259	0.280	0.574	0.413	0.404	0.255	0.240	0.234	0.237	0.765	0.805
35M Frozen MLP-3 128D [1cd75735]	0.118	0.293	0.310	0.588	0.433	0.425	0.096	0.260	0.281	0.560	0.415	0.407	0.270	0.248	0.242	0.278	0.762	0.802
35M Frozen MLP-4 128D [e38bebdc]	0.135	0.291	0.306	0.544	0.427	0.420	0.110	0.259	0.277	0.517	0.409	0.403	0.263	0.236	0.232	0.336	0.767	0.801

Figure S67: Full-validation alignment quality for every selected checkpoint in the indicated representation. All four LDDT definitions and F1 appear together, each with threshold (T), mutual matching (MM), and MEA decoding; coverage is included alongside them. Three-decimal display values preserve the same saved endpoints; exact values remain in the source ledger.

6797 6796 6795 6794 6793 6792 6791 6790 6789 6788 6787 6786 6785 6784 6783 6782 6781 6780 6779 6778 6777 6776 6775 6774 6773 6772 6771 6770 6769 6768 6767 6766 6765

Aggregate quality / continuous

	Cα full			Cα aligned			All-atom full			All-atom aligned			F1			Coverage		
	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA
35M Frozen Projectformer-2 128D [6f383993]	0.177	0.308	0.318	0.503	0.439	0.434	0.148	0.275	0.287	0.480	0.421	0.416	0.286	0.257	0.253	0.437	0.776	0.799
35M Frozen Projectformer-4 128D [48c6cdd9]	0.219	0.328	0.335	0.504	0.463	0.460	0.184	0.294	0.303	0.482	0.444	0.441	0.326	0.300	0.298	0.517	0.777	0.793
35M Frozen BiGRU-1 64D [3ad9fb74]	0.146	0.324	0.340	0.615	0.475	0.466	0.117	0.288	0.307	0.586	0.455	0.447	0.340	0.322	0.316	0.326	0.761	0.796
35M Frozen BiGRU-2 64D [a4b511ac]	0.178	0.330	0.344	0.575	0.479	0.472	0.145	0.294	0.310	0.548	0.460	0.453	0.354	0.327	0.322	0.402	0.765	0.793
35M Frozen BiLSTM-1 64D [4ee8a15f]	0.155	0.325	0.341	0.600	0.475	0.467	0.125	0.289	0.308	0.572	0.455	0.448	0.340	0.317	0.312	0.347	0.762	0.795
35M Frozen BiLSTM-2 64D [9afb0060]	0.194	0.334	0.347	0.559	0.482	0.476	0.159	0.298	0.313	0.533	0.463	0.457	0.359	0.331	0.327	0.437	0.767	0.792
35M Frozen CNN-3 64D [160bc752]	0.159	0.304	0.316	0.529	0.439	0.433	0.130	0.271	0.286	0.505	0.421	0.415	0.280	0.254	0.250	0.391	0.773	0.799
35M Frozen CNN-5 64D [cee62e7a]	0.186	0.310	0.319	0.500	0.443	0.438	0.155	0.278	0.289	0.478	0.425	0.421	0.287	0.261	0.259	0.463	0.778	0.798
35M Frozen CNN-7 64D [56e70f6b]	0.204	0.316	0.324	0.492	0.448	0.445	0.171	0.283	0.293	0.471	0.430	0.427	0.297	0.272	0.269	0.503	0.781	0.798
35M Frozen CNN-9 64D [02433d0d]	0.208	0.317	0.324	0.490	0.448	0.445	0.175	0.284	0.293	0.469	0.430	0.427	0.297	0.274	0.271	0.514	0.782	0.798
35M Frozen Linear 64D [6eb6a42a]	0.043	0.257	0.275	0.578	0.384	0.375	0.037	0.228	0.249	0.560	0.367	0.359	0.113	0.146	0.143	0.088	0.771	0.815
35M Frozen MLP-2 64D [0743300e]	0.097	0.289	0.306	0.605	0.427	0.418	0.080	0.256	0.277	0.578	0.409	0.401	0.241	0.232	0.227	0.218	0.765	0.806
35M Frozen MLP-3 64D [b136b6d2]	0.114	0.292	0.310	0.591	0.432	0.424	0.093	0.259	0.280	0.563	0.414	0.406	0.263	0.244	0.238	0.267	0.762	0.802
35M Frozen MLP-4 64D [8438f4c1]	0.128	0.289	0.305	0.552	0.426	0.419	0.103	0.257	0.276	0.525	0.408	0.401	0.256	0.233	0.228	0.317	0.765	0.801
35M Frozen Projectformer-2 64D [3eac231a]	0.133	0.298	0.314	0.564	0.437	0.429	0.109	0.265	0.284	0.537	0.419	0.412	0.281	0.253	0.248	0.315	0.764	0.801
35M Frozen Projectformer-4 64D [867bd048]	0.196	0.314	0.322	0.492	0.446	0.442	0.165	0.280	0.291	0.470	0.427	0.423	0.292	0.266	0.263	0.483	0.777	0.796
150M Frozen BiGRU-1 128D [2d6a17e6]	0.158	0.327	0.342	0.604	0.477	0.469	0.128	0.291	0.309	0.575	0.458	0.450	0.350	0.327	0.321	0.352	0.762	0.794
150M Frozen BiGRU-2 128D [dd577721]	0.197	0.335	0.346	0.561	0.482	0.476	0.162	0.298	0.312	0.536	0.462	0.457	0.364	0.334	0.330	0.443	0.768	0.791
150M Frozen BiLSTM-1 128D [ef9f6864]	0.169	0.328	0.342	0.584	0.476	0.469	0.138	0.292	0.309	0.557	0.456	0.449	0.347	0.320	0.315	0.379	0.765	0.794
150M Frozen BiLSTM-2 128D [0635dcf3]	0.208	0.336	0.346	0.543	0.482	0.476	0.172	0.300	0.312	0.519	0.462	0.457	0.357	0.329	0.326	0.470	0.769	0.791
150M Frozen CNN-3 128D [9c44568f]	0.182	0.307	0.317	0.502	0.439	0.435	0.151	0.275	0.286	0.480	0.421	0.417	0.282	0.256	0.254	0.455	0.777	0.798
150M Frozen CNN-5 128D [87a0cb31]	0.206	0.314	0.322	0.488	0.445	0.442	0.173	0.282	0.291	0.467	0.427	0.424	0.293	0.270	0.268	0.513	0.781	0.797
150M Frozen CNN-7 128D [1d246182]	0.219	0.319	0.326	0.487	0.450	0.447	0.184	0.286	0.295	0.466	0.432	0.429	0.303	0.280	0.278	0.539	0.782	0.797
150M Frozen CNN-9 128D [cb67a288]	0.222	0.320	0.327	0.486	0.450	0.447	0.188	0.287	0.295	0.466	0.432	0.430	0.304	0.282	0.280	0.547	0.783	0.797
150M Frozen Linear 128D [2f83cfe4]	0.059	0.269	0.286	0.614	0.400	0.391	0.050	0.238	0.259	0.591	0.382	0.374	0.166	0.185	0.180	0.123	0.768	0.812
150M Frozen MLP-2 128D [b8ac53ce]	0.110	0.298	0.318	0.636	0.444	0.435	0.089	0.265	0.288	0.605	0.425	0.417	0.289	0.274	0.267	0.244	0.757	0.801

Figure S68: Full-validation alignment quality for every selected checkpoint in the indicated representation. All four LDDT definitions and F1 appear together, each with threshold (T), mutual matching (MM), and MEA decoding; coverage is included alongside them. Three-decimal display values preserve the same saved endpoints; exact values remain in the source ledger.

6830 6829 6828 6827 6826 6825 6824 6823 6822 6821 6820 6819 6818 6817 6816 6815 6814 6813 6812 6811 6810 6809 6808 6807 6806 6805 6804 6803 6802 6801 6800 6799 6798

Aggregate quality / continuous

	Cα full			Cα aligned			All-atom full			All-atom aligned			F1			Coverage		
	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA
150M Frozen MLP-3 128D [8b0c8276]	0.124	0.296	0.314	0.594	0.438	0.430	0.100	0.263	0.284	0.564	0.420	0.412	0.283	0.259	0.254	0.290	0.760	0.800
150M Frozen MLP-4 128D [13248f80]	0.152	0.295	0.307	0.520	0.428	0.422	0.124	0.263	0.278	0.496	0.410	0.404	0.262	0.236	0.232	0.384	0.771	0.800
150M Frozen Projectformer-2 128D [b5202b86]	0.184	0.315	0.325	0.516	0.450	0.445	0.153	0.281	0.294	0.493	0.432	0.427	0.303	0.274	0.271	0.442	0.775	0.796
150M Frozen Projectformer-4 128D [39c8183a]	0.219	0.328	0.335	0.503	0.463	0.460	0.184	0.293	0.302	0.481	0.444	0.441	0.324	0.298	0.296	0.519	0.777	0.792
150M Frozen BiGRU-1 64D [fcf3f7a3]	0.154	0.325	0.341	0.605	0.476	0.468	0.124	0.288	0.308	0.577	0.456	0.449	0.345	0.322	0.317	0.344	0.761	0.794
150M Frozen BiGRU-2 64D [5a61434d]	0.197	0.335	0.346	0.561	0.482	0.476	0.162	0.298	0.313	0.536	0.463	0.457	0.365	0.335	0.331	0.441	0.767	0.791
150M Frozen BiLSTM-1 64D [3d693811]	0.167	0.327	0.341	0.583	0.475	0.467	0.136	0.291	0.308	0.556	0.455	0.448	0.345	0.319	0.314	0.375	0.765	0.794
150M Frozen BiLSTM-2 64D [9162d15d]	0.215	0.339	0.349	0.542	0.486	0.480	0.178	0.303	0.315	0.518	0.466	0.460	0.369	0.340	0.337	0.482	0.769	0.790
150M Frozen CNN-3 64D [008ef6b8]	0.179	0.305	0.315	0.502	0.437	0.433	0.148	0.273	0.285	0.480	0.420	0.415	0.277	0.252	0.249	0.449	0.777	0.799
150M Frozen CNN-5 64D [3331453e]	0.203	0.312	0.320	0.488	0.444	0.440	0.170	0.280	0.290	0.467	0.426	0.422	0.288	0.265	0.263	0.508	0.780	0.797
150M Frozen CNN-7 64D [391f6bf3]	0.216	0.318	0.325	0.486	0.448	0.446	0.182	0.285	0.293	0.465	0.430	0.428	0.299	0.277	0.274	0.536	0.782	0.796
150M Frozen CNN-9 64D [aa56efbb]	0.221	0.320	0.326	0.487	0.451	0.448	0.187	0.287	0.295	0.466	0.433	0.430	0.305	0.283	0.281	0.546	0.782	0.796
150M Frozen Linear 64D [e4491234]	0.063	0.272	0.290	0.633	0.406	0.397	0.052	0.241	0.263	0.608	0.388	0.380	0.183	0.201	0.195	0.129	0.766	0.811
150M Frozen MLP-2 64D [18ceefba]	0.106	0.297	0.317	0.639	0.443	0.434	0.085	0.263	0.287	0.608	0.424	0.415	0.282	0.270	0.264	0.234	0.756	0.802
150M Frozen MLP-3 64D [421ceedb]	0.120	0.294	0.312	0.593	0.436	0.428	0.097	0.261	0.282	0.564	0.418	0.410	0.277	0.256	0.250	0.285	0.759	0.800
150M Frozen MLP-4 64D [fbe4f32b]	0.150	0.293	0.306	0.517	0.426	0.420	0.123	0.261	0.277	0.493	0.408	0.403	0.258	0.233	0.229	0.382	0.771	0.800
150M Frozen Projectformer-2 64D [83edcc2e]	0.144	0.305	0.320	0.565	0.445	0.438	0.118	0.271	0.290	0.538	0.427	0.420	0.294	0.266	0.261	0.337	0.765	0.800
150M Frozen Projectformer-4 64D [14450876]	0.213	0.325	0.333	0.502	0.461	0.457	0.179	0.292	0.301	0.480	0.442	0.438	0.319	0.293	0.291	0.508	0.777	0.794
ProteinMPNN-128 [0731829b] / wBCE cap 100	0.403	0.460	0.466	0.692	0.646	0.640	0.344	0.409	0.417	0.662	0.620	0.614	0.754	0.738	0.733	0.651	0.758	0.770
ProteinMPNN-64 [16ce6ed6] / wBCE cap 100	0.393	0.459	0.465	0.694	0.644	0.638	0.333	0.408	0.417	0.665	0.618	0.613	0.746	0.729	0.725	0.635	0.758	0.771
SoftAlign geometric mean [9e794c93]	0.236	0.357	0.416	0.668	0.585	0.606	0.208	0.307	0.367	0.637	0.562	0.583	0.653	0.645	0.664	0.356	0.651	0.734
SoftAlign SW [5ad6a335]	0.315	0.407	0.428	0.692	0.594	0.586	0.262	0.360	0.385	0.662	0.571	0.563	0.679	0.661	0.649	0.517	0.725	0.774

Figure S69: Full-validation alignment quality for every selected checkpoint in the indicated representation. All four LDDT definitions and F1 appear together, each with threshold (T), mutual matching (MM), and MEA decoding; coverage is included alongside them. Three-decimal display values preserve the same saved endpoints; exact values remain in the source ledger.

6863 6862 6861 6860 6859 6858 6857 6856 6855 6854 6853 6852 6851 6850 6849 6848 6847 6846 6845 6844 6843 6842 6841 6840 6839 6838 6837 6836 6835 6834 6833 6832 6831

Aggregate quality / soft 3Di

	Cα full			Cα aligned			All-atom full			All-atom aligned			F1			Coverage		
	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA
8M Frozen linear [31bba998] / ST Gumbel	0.174	0.262	0.316	0.612	0.547	0.497	0.137	0.219	0.276	0.584	0.524	0.478	0.393	0.390	0.370	0.367	0.578	0.710
8M Frozen linear [1cf7945c] / softmax	0.174	0.344	0.370	0.669	0.521	0.510	0.140	0.303	0.333	0.638	0.501	0.490	0.447	0.430	0.422	0.342	0.733	0.783
8M Frozen linear [fbc46ff2] / ST argmax	0.013	0.039	0.207	0.855	0.688	0.399	0.009	0.027	0.176	0.806	0.647	0.385	0.089	0.174	0.214	0.038	0.121	0.619
8M Frozen MLP-4 [7b541c60] / softmax	0.203	0.355	0.375	0.646	0.523	0.515	0.164	0.315	0.338	0.617	0.503	0.495	0.472	0.441	0.434	0.399	0.747	0.784
8M Unfrozen MLP-4 [16132e1d] / softmax	0.294	0.370	0.376	0.550	0.524	0.521	0.252	0.330	0.337	0.528	0.503	0.500	0.446	0.422	0.419	0.600	0.761	0.774
8M Frozen MLP-4 [d58f9e14] / ST argmax	0.057	0.127	0.239	0.801	0.645	0.440	0.040	0.097	0.206	0.755	0.612	0.424	0.218	0.317	0.270	0.133	0.292	0.639
8M Unfrozen MLP-4 [650078f8] / ST argmax	0.209	0.247	0.306	0.633	0.577	0.489	0.172	0.207	0.267	0.603	0.551	0.470	0.416	0.412	0.372	0.419	0.528	0.701
8M Unfrozen MLP-4 [d50ef818] / ST Gumbel	0.294	0.331	0.348	0.568	0.535	0.514	0.251	0.288	0.308	0.544	0.513	0.494	0.425	0.412	0.399	0.605	0.694	0.744
8M Unfrozen MLP-4 [9d792c09] / softmax + mask	0.321	0.355	0.362	0.549	0.524	0.518	0.278	0.315	0.322	0.527	0.504	0.498	0.419	0.406	0.401	0.656	0.738	0.756
8M Frozen MLP-4 [87d2c509] / ST Gumbel	0.227	0.296	0.328	0.594	0.539	0.503	0.184	0.252	0.287	0.568	0.518	0.484	0.422	0.409	0.389	0.473	0.633	0.721
35M Frozen linear [f8f5cad4] / ST argmax	0.025	0.090	0.268	0.849	0.676	0.445	0.016	0.066	0.235	0.803	0.642	0.430	0.153	0.322	0.330	0.068	0.214	0.676
35M Frozen linear [c336a672] / softmax	0.233	0.383	0.403	0.667	0.562	0.553	0.190	0.338	0.362	0.638	0.541	0.532	0.547	0.524	0.516	0.429	0.741	0.778
35M Frozen linear [4b1a9c9d] / ST Gumbel	0.249	0.330	0.372	0.647	0.604	0.553	0.203	0.282	0.329	0.620	0.581	0.532	0.520	0.514	0.489	0.463	0.626	0.732
35M Frozen MLP-4 [72459d15] / ST Gumbel	0.288	0.352	0.376	0.631	0.572	0.549	0.239	0.305	0.333	0.605	0.550	0.529	0.533	0.516	0.500	0.540	0.679	0.740
35M Unfrozen MLP-4 [929752d4] / softmax + mask	0.222	0.355	0.383	0.659	0.547	0.539	0.179	0.310	0.343	0.631	0.527	0.519	0.525	0.504	0.495	0.421	0.718	0.768
35M Frozen MLP-4 [9fceaf26] / ST argmax	0.107	0.201	0.293	0.769	0.641	0.492	0.080	0.160	0.255	0.731	0.612	0.474	0.355	0.448	0.398	0.212	0.404	0.665
35M Unfrozen MLP-4 [39efc845] / ST argmax	0.173	0.301	0.380	0.716	0.581	0.543	0.136	0.257	0.339	0.683	0.559	0.523	0.490	0.525	0.501	0.319	0.597	0.754
35M Unfrozen MLP-4 [f7815ea1] / softmax	0.265	0.384	0.402	0.645	0.564	0.555	0.220	0.340	0.361	0.617	0.543	0.535	0.563	0.531	0.524	0.487	0.741	0.774
35M Unfrozen MLP-4 [f32e1543] / ST Gumbel	0.241	0.363	0.394	0.657	0.561	0.550	0.196	0.317	0.353	0.629	0.541	0.530	0.545	0.524	0.512	0.444	0.710	0.768
35M Frozen MLP-4 [0eb8395a] / softmax	0.269	0.392	0.407	0.645	0.564	0.557	0.222	0.348	0.366	0.617	0.543	0.536	0.567	0.533	0.527	0.491	0.751	0.778

Figure S70: Full-validation alignment quality for every selected checkpoint in the indicated representation. All four LDDT definitions and F1 appear together, each with threshold (T), mutual matching (MM), and MEA decoding; coverage is included alongside them. Three-decimal display values preserve the same saved endpoints; exact values remain in the source ledger.

6896 6895 6894 6893 6892 6891 6890 6889 6888 6887 6886 6885 6884 6883 6882 6881 6880 6879 6878 6877 6876 6875 6874 6873 6872 6871 6870 6869 6868 6867 6866 6865 6864

Aggregate quality / hard 3Di

	Cα full			Cα aligned			All-atom full			All-atom aligned			F1			Coverage		
	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA	T	MM	MEA
8M Frozen linear [31bba998] / ST Gumbel	0.231	0.289	0.299	0.502	0.454	0.453	0.189	0.252	0.263	0.481	0.437	0.436	0.325	0.311	0.310	0.584	0.733	0.750
8M Frozen linear [1cf7945c] / softmax	0.219	0.277	0.286	0.483	0.429	0.429	0.181	0.244	0.254	0.463	0.414	0.413	0.305	0.291	0.291	0.591	0.750	0.765
8M Frozen linear [fbc46ff2] / ST argmax	0.194	0.244	0.255	0.480	0.430	0.430	0.155	0.208	0.218	0.461	0.415	0.416	0.299	0.288	0.289	0.537	0.680	0.698
8M Frozen MLP-4 [7b541c60] / softmax	0.222	0.281	0.290	0.481	0.428	0.426	0.185	0.249	0.259	0.461	0.412	0.410	0.311	0.295	0.295	0.590	0.757	0.774
8M Unfrozen MLP-4 [16132e1d] / softmax	0.306	0.337	0.340	0.500	0.477	0.477	0.266	0.302	0.305	0.480	0.458	0.458	0.372	0.361	0.361	0.698	0.775	0.780
8M Frozen MLP-4 [d58f9e14] / ST argmax	0.232	0.271	0.278	0.475	0.442	0.443	0.191	0.234	0.242	0.458	0.427	0.428	0.303	0.294	0.294	0.610	0.713	0.724
8M Unfrozen MLP-4 [650078f8] / ST argmax	0.283	0.309	0.315	0.516	0.485	0.484	0.240	0.268	0.275	0.496	0.467	0.466	0.383	0.373	0.370	0.640	0.711	0.725
8M Unfrozen MLP-4 [d50ef818] / ST Gumbel	0.319	0.342	0.347	0.527	0.505	0.502	0.275	0.302	0.308	0.506	0.486	0.483	0.402	0.392	0.390	0.686	0.745	0.757
8M Unfrozen MLP-4 [9d792c09] / softmax + mask	0.316	0.335	0.337	0.499	0.486	0.486	0.276	0.298	0.300	0.480	0.467	0.468	0.369	0.363	0.363	0.715	0.760	0.763
8M Frozen MLP-4 [87d2c509] / ST Gumbel	0.249	0.303	0.312	0.512	0.469	0.467	0.206	0.265	0.275	0.491	0.452	0.450	0.360	0.344	0.343	0.601	0.735	0.751
35M Frozen linear [f8f5cad4] / ST argmax	0.256	0.308	0.320	0.541	0.498	0.496	0.211	0.267	0.279	0.521	0.481	0.479	0.409	0.395	0.395	0.583	0.705	0.725
35M Frozen linear [c336a672] / softmax	0.258	0.312	0.321	0.520	0.467	0.466	0.215	0.276	0.286	0.499	0.450	0.449	0.403	0.384	0.384	0.607	0.753	0.768
35M Frozen linear [4b1a9c9d] / ST Gumbel	0.291	0.342	0.351	0.558	0.514	0.512	0.243	0.302	0.311	0.536	0.496	0.494	0.444	0.427	0.427	0.620	0.740	0.754
35M Frozen MLP-4 [72459d15] / ST Gumbel	0.309	0.352	0.359	0.557	0.518	0.517	0.261	0.311	0.319	0.536	0.500	0.498	0.471	0.454	0.453	0.644	0.747	0.758
35M Unfrozen MLP-4 [929752d4] / softmax + mask	0.234	0.341	0.364	0.636	0.521	0.518	0.190	0.298	0.325	0.609	0.503	0.500	0.505	0.477	0.473	0.469	0.725	0.764
35M Frozen MLP-4 [9fceaf26] / ST argmax	0.257	0.302	0.310	0.527	0.485	0.486	0.213	0.262	0.270	0.508	0.469	0.470	0.410	0.396	0.396	0.594	0.706	0.719
35M Unfrozen MLP-4 [39efc845] / ST argmax	0.215	0.320	0.346	0.632	0.507	0.507	0.173	0.278	0.307	0.605	0.490	0.490	0.495	0.469	0.465	0.430	0.700	0.746
35M Unfrozen MLP-4 [f7815ea1] / softmax	0.291	0.365	0.377	0.598	0.533	0.530	0.245	0.324	0.337	0.573	0.513	0.511	0.515	0.487	0.484	0.575	0.749	0.769
35M Unfrozen MLP-4 [f32e1543] / ST Gumbel	0.250	0.358	0.381	0.647	0.541	0.537	0.204	0.314	0.340	0.620	0.522	0.518	0.533	0.503	0.496	0.475	0.726	0.767
35M Frozen MLP-4 [0eb8395a] / softmax	0.265	0.315	0.323	0.523	0.475	0.475	0.222	0.278	0.288	0.502	0.458	0.457	0.408	0.391	0.391	0.614	0.747	0.761

Figure S71: Full-validation alignment quality for every selected checkpoint in the indicated representation. All four LDDT definitions and F1 appear together, each with threshold (T), mutual matching (MM), and MEA decoding; coverage is included alongside them. Three-decimal display values preserve the same saved endpoints; exact values remain in the source ledger.

6929 6928 6927 6926 6925 6924 6923 6922 6921 6920 6919 6918 6917 6916 6915 6914 6913 6912 6911 6910 6909 6908 6907 6906 6905 6904 6903 6902 6901 6900 6899 6898 6897

Aggregate quality / native correspondence decoders

	LDDT				Correspondence
	Cα full	Cα aligned	All-atom full	All-atom aligned	F1
blosum45 gaps 14 2	0.120	0.585	0.097	0.539	0.234
blosum50 gaps 13 2	0.112	0.606	0.091	0.556	0.236
blosum62 gaps 10 1	0.112	0.600	0.091	0.549	0.236
blosum62 gaps 11 1	0.105	0.622	0.085	0.569	0.234
blosum80 gaps 10 1	0.077	0.674	0.063	0.609	0.206
blosum90 gaps 10 1	0.070	0.680	0.057	0.611	0.192
DEDAL (full length)	0.208	0.635	0.176	0.595	0.373
DeepBLAST	0.321	0.502	0.281	0.484	0.474
35M Frozen MLP-4 [72459d15] / ST Gumbel / 3Di-only	0.372	0.549	0.328	0.529	0.465
35M Frozen MLP-4 [72459d15] / ST Gumbel / AA+3Di	0.344	0.568	0.296	0.541	0.486
35M Unfrozen MLP-4 [929752d4] / softmax + mask / 3Di-only	0.379	0.554	0.335	0.533	0.487
35M Unfrozen MLP-4 [929752d4] / softmax + mask / AA+3Di	0.371	0.557	0.325	0.533	0.498
35M Frozen linear [f8f5cad4] / ST argmax / 3Di-only	0.334	0.555	0.288	0.534	0.430
35M Frozen linear [f8f5cad4] / ST argmax / AA+3Di	0.301	0.582	0.253	0.553	0.451
35M Frozen MLP-4 [9fceaf26] / ST argmax / 3Di-only	0.319	0.553	0.274	0.531	0.427
35M Frozen MLP-4 [9fceaf26] / ST argmax / AA+3Di	0.287	0.581	0.242	0.551	0.450
35M Unfrozen MLP-4 [39efc845] / ST argmax / 3Di-only	0.339	0.565	0.294	0.543	0.485
35M Unfrozen MLP-4 [39efc845] / ST argmax / AA+3Di	0.322	0.580	0.278	0.553	0.502
35M Frozen linear [c336a672] / softmax / 3Di-only	0.356	0.499	0.318	0.479	0.386
35M Frozen linear [c336a672] / softmax / AA+3Di	0.326	0.526	0.282	0.499	0.422
35M Unfrozen MLP-4 [f7815ea1] / softmax / 3Di-only	0.389	0.551	0.347	0.530	0.486
35M Unfrozen MLP-4 [f7815ea1] / softmax / AA+3Di	0.377	0.557	0.332	0.532	0.500
35M Unfrozen MLP-4 [f32e1543] / ST Gumbel / 3Di-only	0.384	0.572	0.339	0.551	0.505
35M Unfrozen MLP-4 [f32e1543] / ST Gumbel / AA+3Di	0.367	0.584	0.319	0.557	0.522
35M Frozen MLP-4 [0eb8395a] / softmax / 3Di-only	0.365	0.513	0.326	0.493	0.397
35M Frozen MLP-4 [0eb8395a] / softmax / AA+3Di	0.336	0.535	0.292	0.508	0.430

Figure S72: Native alignment results: all four LDDT endpoints and F1 use each tool's native correspondences. Learned-token rows identify their native Foldseek scoring mode. Both structure-based and ProstT5-based Foldseek modes are retained.

6962 6961 6960 6959 6958 6957 6956 6955 6954 6953 6952 6951 6950 6949 6948 6947 6946 6945 6944 6943 6942 6941 6940 6939 6938 6937 6936 6935 6934 6933 6932 6931 6930

Aggregate quality / native correspondence decoders

	LDDT				Correspondence
	Cα full	Cα aligned	All-atom full	All-atom aligned	F1
35M Frozen linear [4b1a9c9d] / ST Gumbel / 3Di-only	0.368	0.546	0.324	0.525	0.443
35M Frozen linear [4b1a9c9d] / ST Gumbel / AA+3Di	0.344	0.564	0.297	0.537	0.467
8M Frozen MLP-4 [7b541c60] / softmax / 3Di-only	0.334	0.460	0.301	0.442	0.299
8M Frozen MLP-4 [7b541c60] / softmax / AA+3Di	0.319	0.485	0.280	0.460	0.352
8M Unfrozen MLP-4 [16132e1d] / softmax / 3Di-only	0.353	0.487	0.317	0.468	0.357
8M Unfrozen MLP-4 [16132e1d] / softmax / AA+3Di	0.343	0.507	0.302	0.482	0.399
8M Frozen MLP-4 [d58f9e14] / ST argmax / 3Di-only	0.294	0.491	0.252	0.472	0.319
8M Frozen MLP-4 [d58f9e14] / ST argmax / AA+3Di	0.271	0.517	0.229	0.492	0.362
8M Unfrozen MLP-4 [650078f8] / ST argmax / 3Di-only	0.299	0.523	0.256	0.503	0.383
8M Unfrozen MLP-4 [650078f8] / ST argmax / AA+3Di	0.284	0.549	0.242	0.523	0.415
8M Unfrozen MLP-4 [d50ef818] / ST Gumbel / 3Di-only	0.339	0.524	0.296	0.503	0.397
8M Unfrozen MLP-4 [d50ef818] / ST Gumbel / AA+3Di	0.323	0.544	0.279	0.518	0.431
8M Unfrozen MLP-4 [9d792c09] / softmax + mask / 3Di-only	0.341	0.497	0.302	0.477	0.360
8M Unfrozen MLP-4 [9d792c09] / softmax + mask / AA+3Di	0.330	0.510	0.288	0.487	0.391
8M Frozen linear [31bba998] / ST Gumbel / 3Di-only	0.328	0.491	0.288	0.471	0.328
8M Frozen linear [31bba998] / ST Gumbel / AA+3Di	0.306	0.518	0.262	0.492	0.380
8M Frozen MLP-4 [87d2c509] / ST Gumbel / 3Di-only	0.330	0.502	0.288	0.482	0.356
8M Frozen MLP-4 [87d2c509] / ST Gumbel / AA+3Di	0.304	0.531	0.260	0.504	0.397
8M Frozen linear [1cf7945c] / softmax / 3Di-only	0.328	0.461	0.293	0.442	0.291
8M Frozen linear [1cf7945c] / softmax / AA+3Di	0.300	0.495	0.260	0.469	0.351
8M Frozen linear [fbc46ff2] / ST argmax / 3Di-only	0.269	0.495	0.228	0.475	0.310
8M Frozen linear [fbc46ff2] / ST argmax / AA+3Di	0.247	0.540	0.206	0.511	0.362
GANGSTA+	0.351	0.636	0.291	0.602	0.460
DALI (available subset)	0.457	0.687	0.392	0.655	0.719
MUSTANG (available subset)	0.395	0.617	0.344	0.593	0.539
TM-align	0.464	0.675	0.407	0.645	0.983

Figure S73: Native alignment results: all four LDDT endpoints and F1 use each tool's native correspondences. Learned-token rows identify their native Foldseek scoring mode. Both structure-based and ProstT5-based Foldseek modes are retained.



Figure S74: Native alignment results: all four LDDT endpoints and F1 use each tool's native correspondences. Learned-token rows identify their native Foldseek scoring mode. Both structure-based and ProstT5-based Foldseek modes are retained.

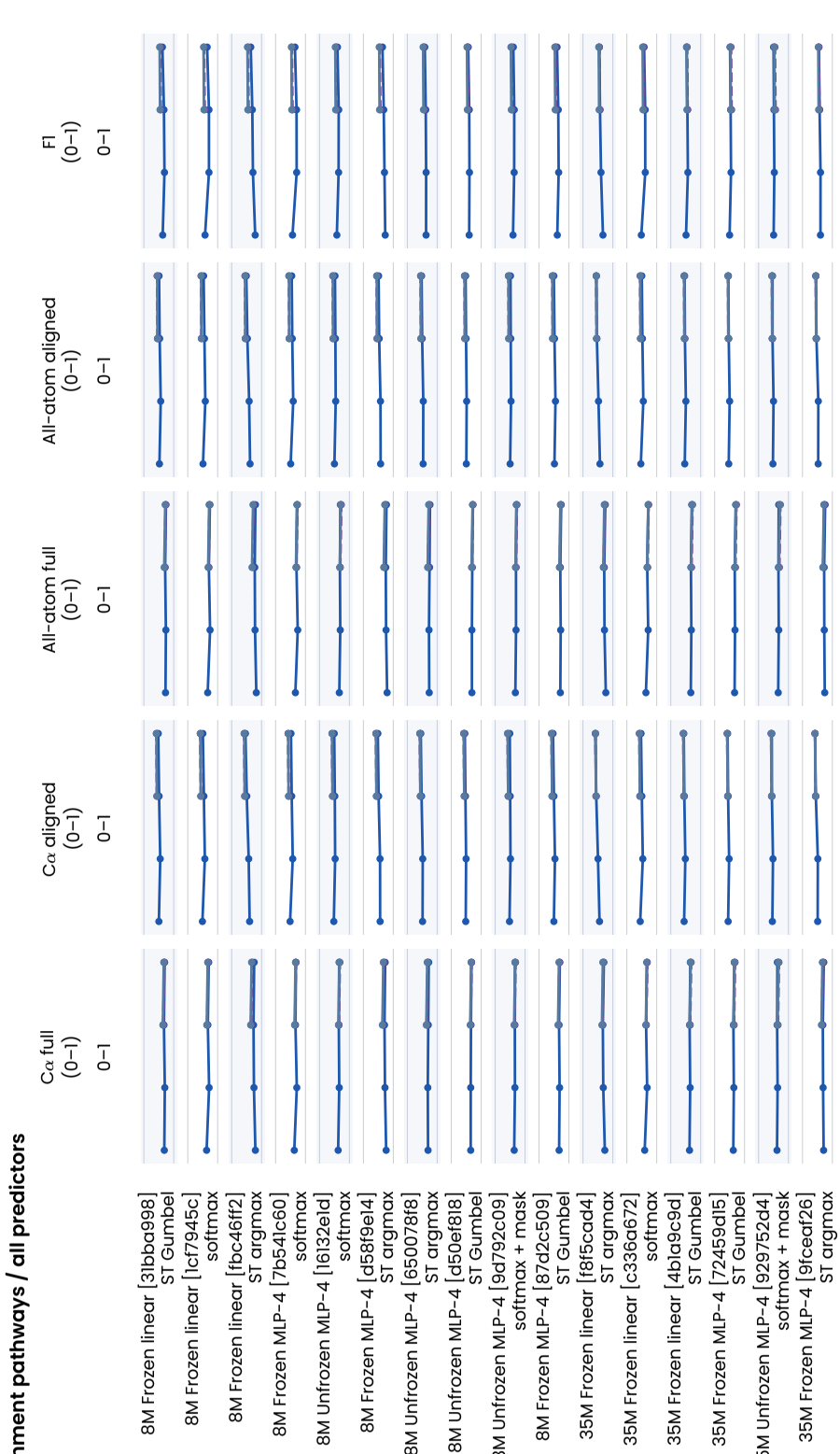


Figure S75: Every discrete predictor through four routes, left to right: soft-state Orihime MEA, hard-state Orihime MEA, native Foldseek 3Di-only, and native Foldseek AA+3Di. Blue follows the same checkpoint; dashed purple and slate segments show structure-based and ProstT5-based Foldseek in the two native modes. All four LDDT definitions and correspondence F1 are shown together.

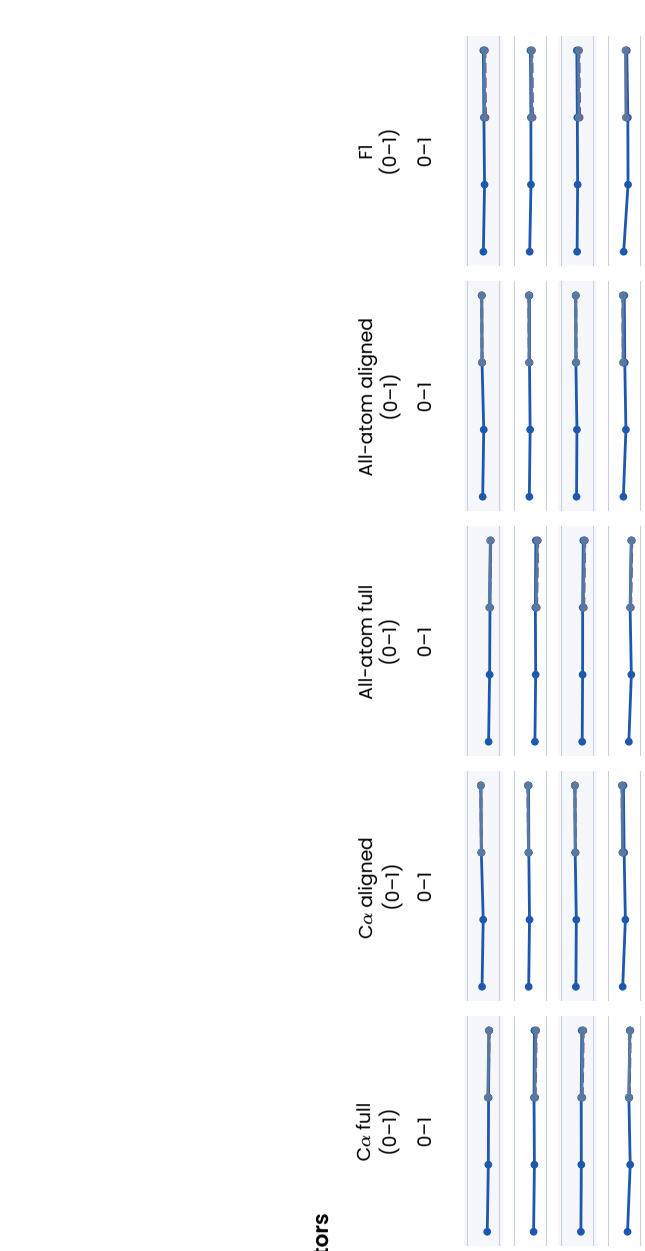


Figure S76: Every discrete predictor through four routes, left to right: soft-state Orihime MEA, hard-state Orihime MEA, native Foldseek 3Di-only, and native Foldseek AA+3Di. Blue follows the same checkpoint; dashed purple and slate segments show structure-based and ProstT5-based Foldseek in the two native modes. All four LDDT definitions and correspondence F1 are shown together.

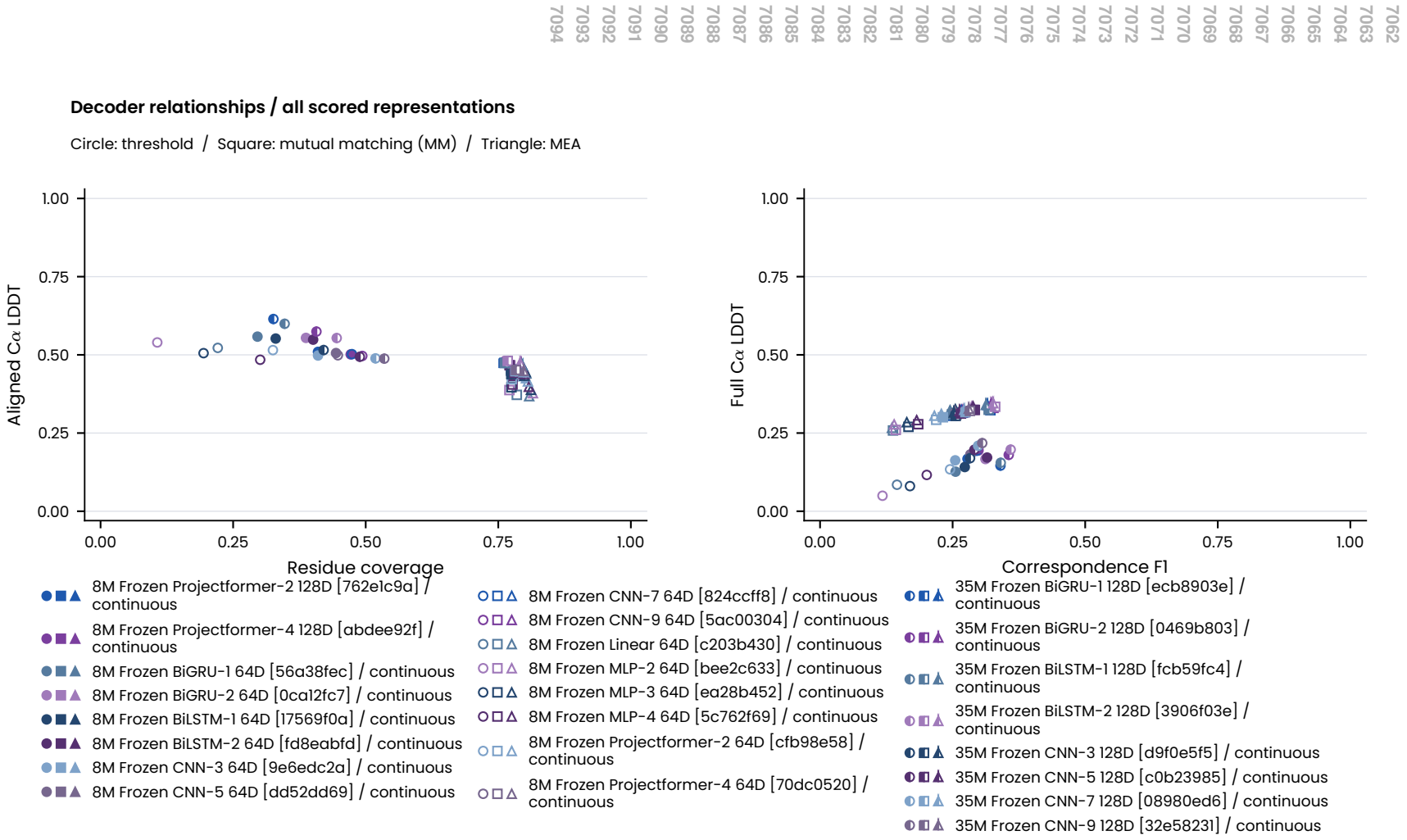


Figure S77: Coverage, correspondence and geometric agreement for every representation. Unconnected markers identify threshold (circle), mutual matching (square), and MEA (triangle). Color and fill identify the checkpoint, as shown in the key. Both axes use 0–1 scores. The companion aggregate matrices report all four LDDT endpoints, F1 and coverage under each decoder, so no separate repeated profile pages are needed.

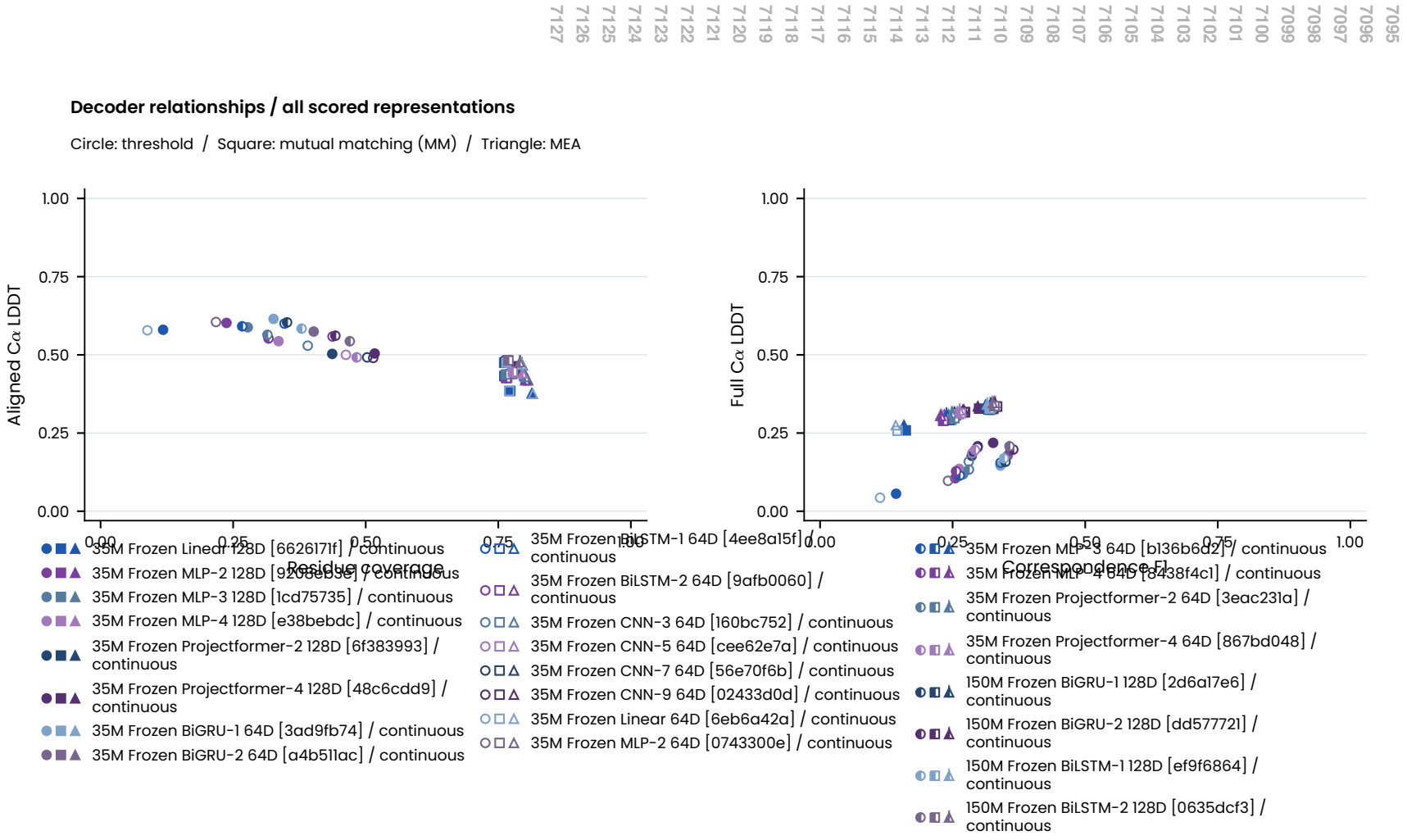


Figure S78: Coverage, correspondence and geometric agreement for every representation. Unconnected markers identify threshold (circle), mutual matching (square), and MEA (triangle). Color and fill identify the checkpoint, as shown in the key. Both axes use 0–1 scores. The companion aggregate matrices report all four LDDT endpoints, F1 and coverage under each decoder, so no separate repeated profile pages are needed.

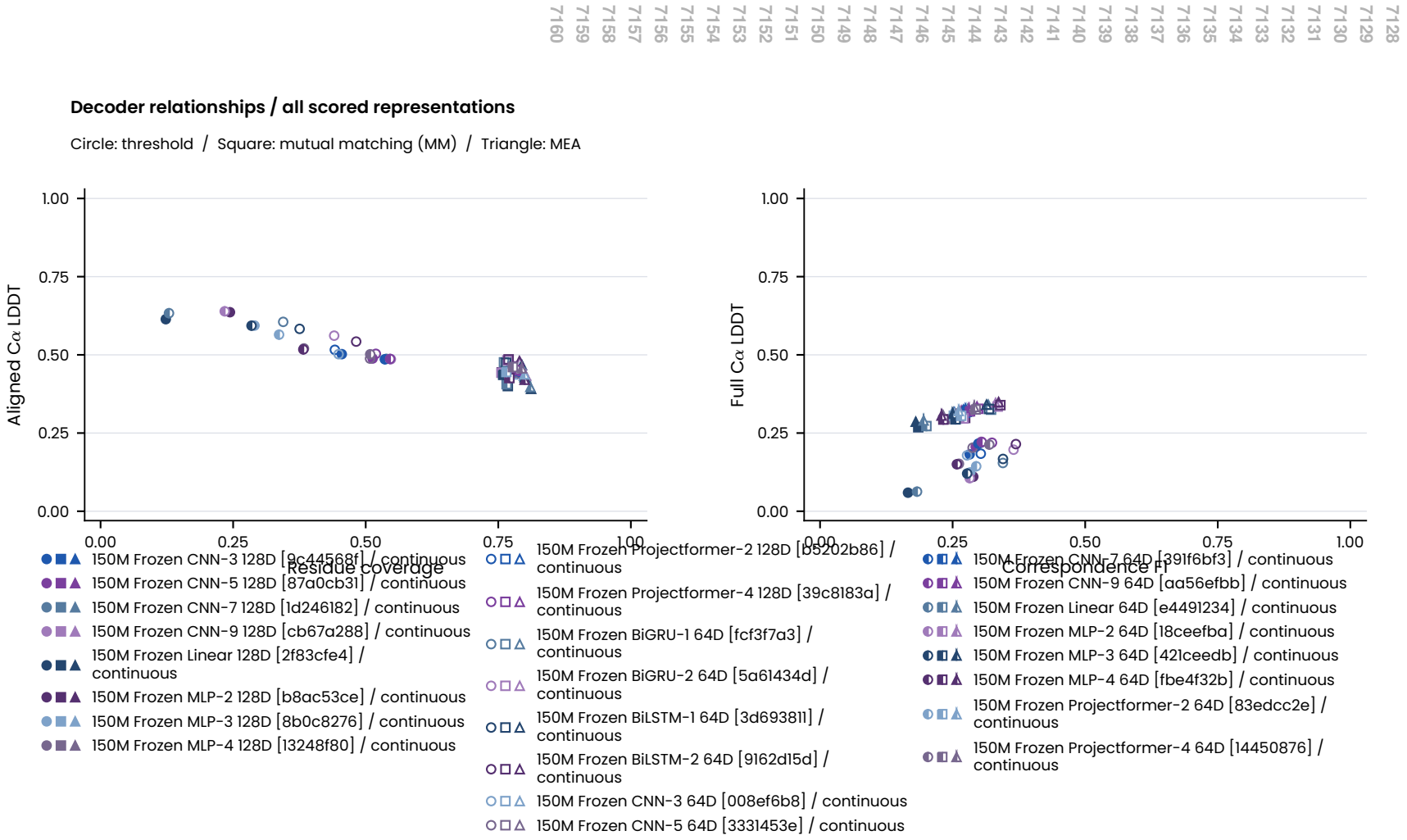


Figure S79: Coverage, correspondence and geometric agreement for every representation. Unconnected markers identify threshold (circle), mutual matching (square), and MEA (triangle). Color and fill identify the checkpoint, as shown in the key. Both axes use 0–1 scores. The companion aggregate matrices report all four LDDT endpoints, F1 and coverage under each decoder, so no separate repeated profile pages are needed.

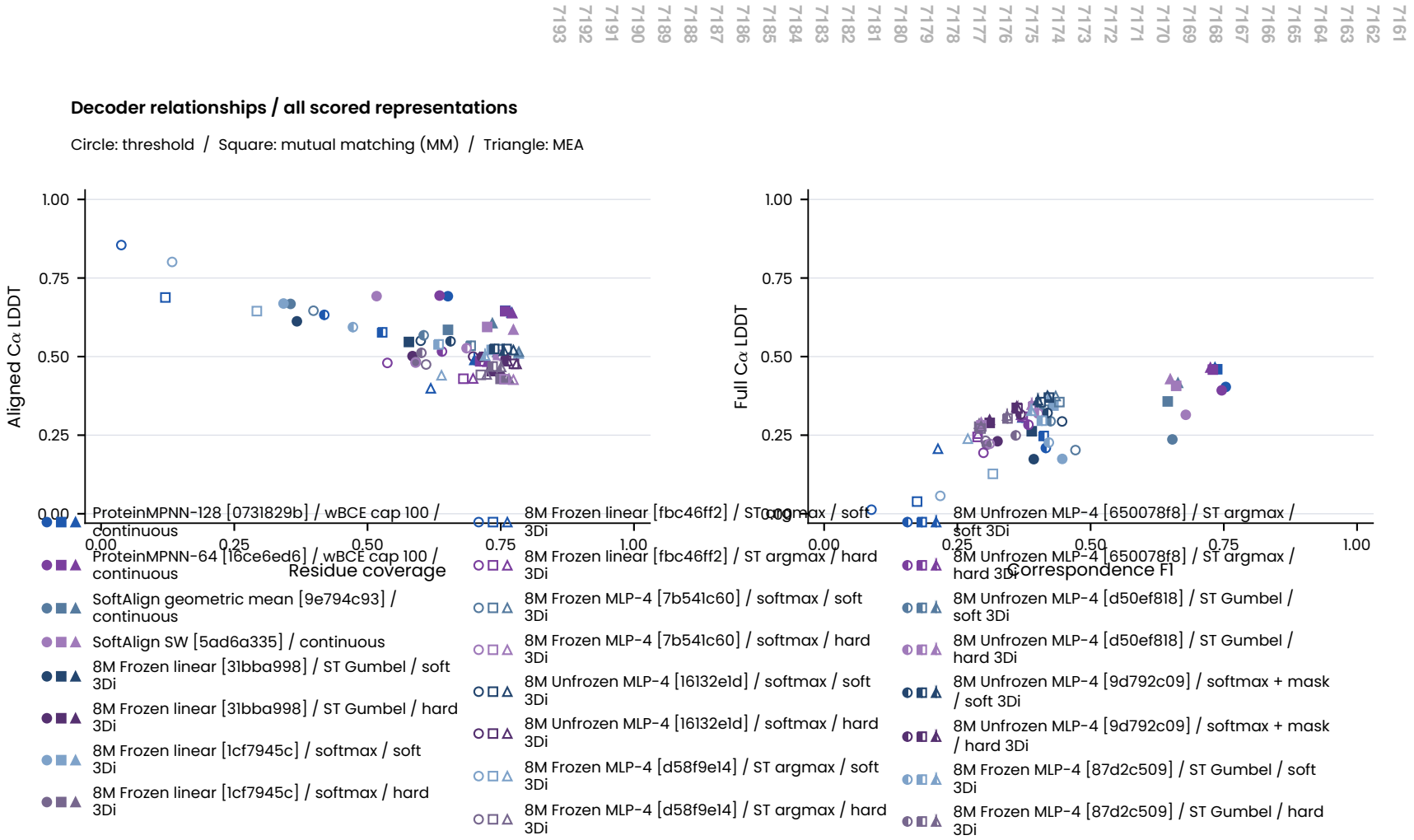


Figure S80: Coverage, correspondence and geometric agreement for every representation. Unconnected markers identify threshold (circle), mutual matching (square), and MEA (triangle). Color and fill identify the checkpoint, as shown in the key. Both axes use 0–1 scores. The companion aggregate matrices report all four LDDT endpoints, F1 and coverage under each decoder, so no separate repeated profile pages are needed.

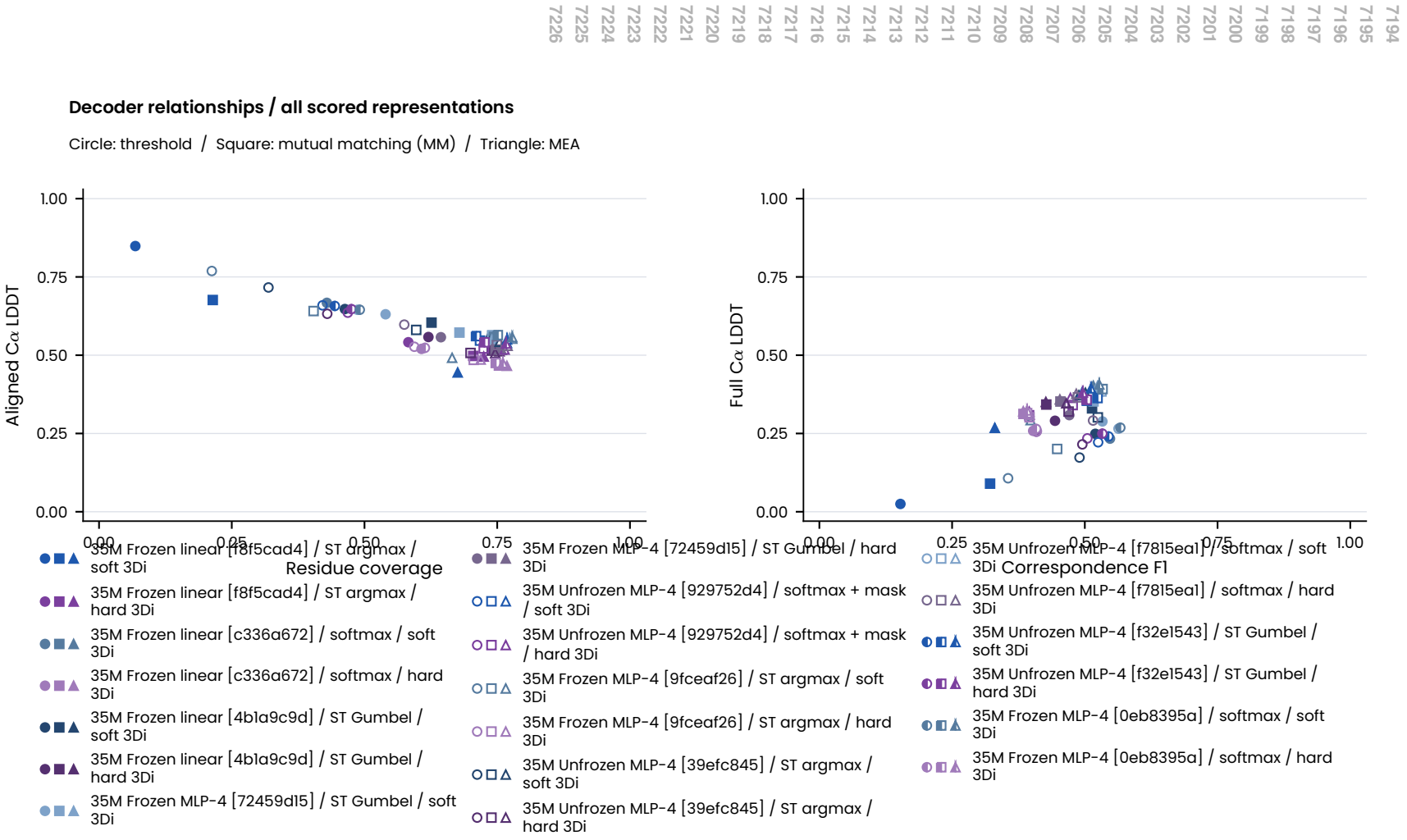


Figure S81: Coverage, correspondence and geometric agreement for every representation. Unconnected markers identify threshold (circle), mutual matching (square), and MEA (triangle). Color and fill identify the checkpoint, as shown in the key. Both axes use 0–1 scores. The companion aggregate matrices report all four LDDT endpoints, F1 and coverage under each decoder, so no separate repeated profile pages are needed.

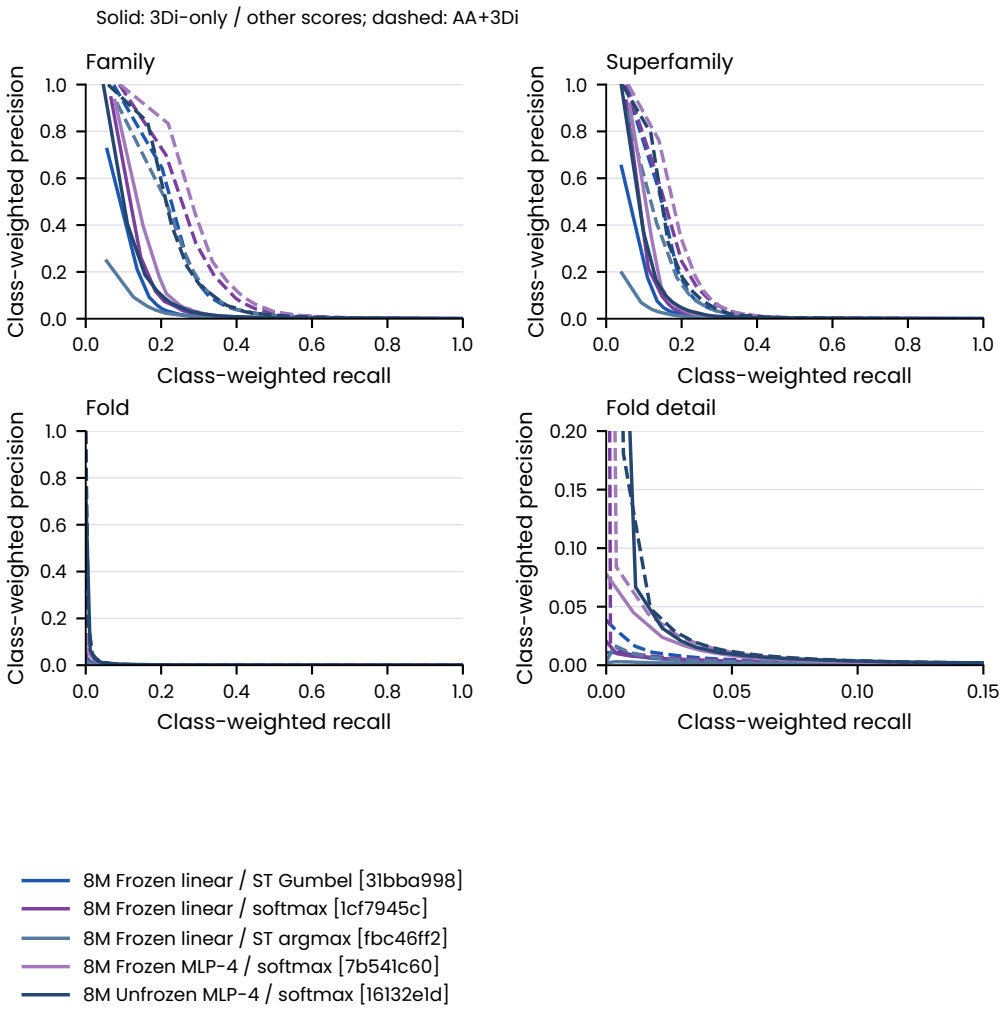


Figure S82: All supplied class-weighted precision–recall points. Colors identify predictors; both native Foldseek modes share that predictor's color, with solid 3Di-only and dashed AA+3Di lines. The reference page includes both Foldseek input types and modes, MMseqs2 and all four TM-align normalizations. The fourth panel enlarges the fold curves without changing their sampling.

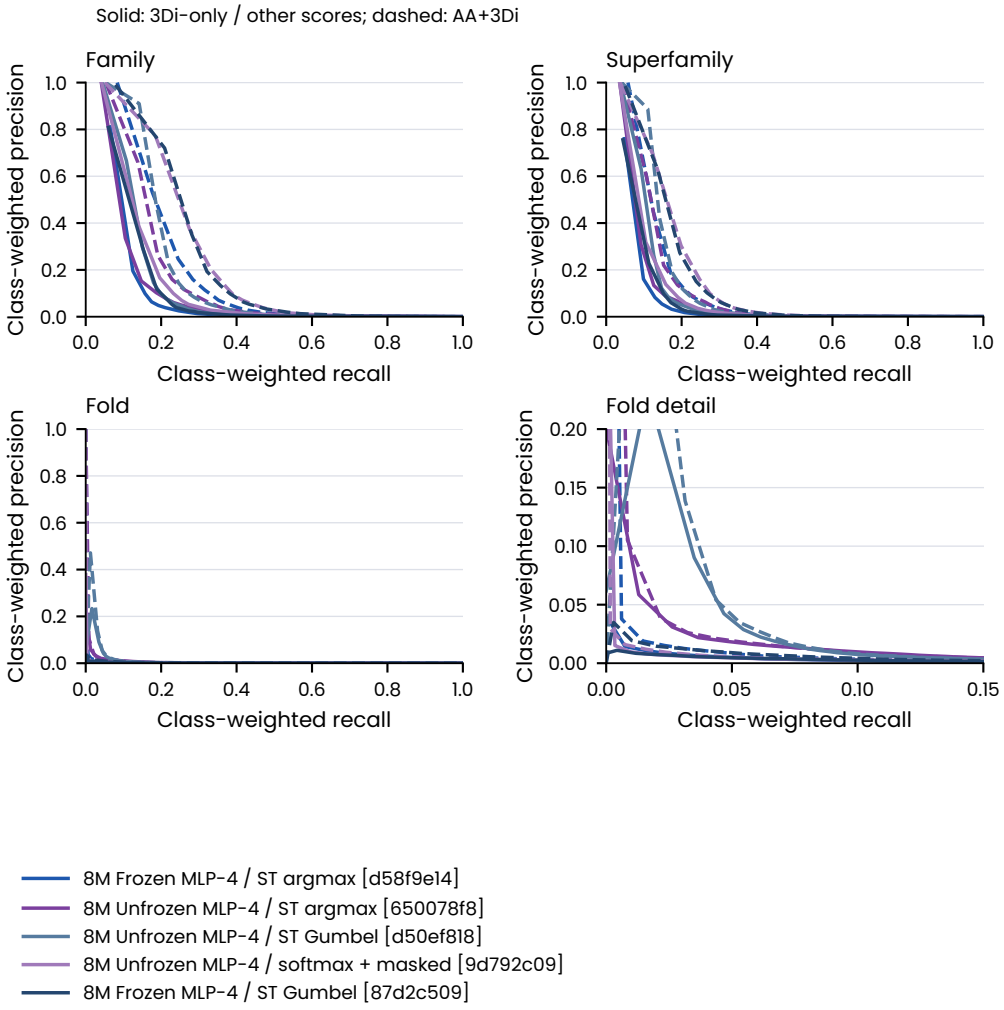


Figure S83: All supplied class-weighted precision–recall points. Colors identify predictors; both native Foldseek modes share that predictor's color, with solid 3Di-only and dashed AA+3Di lines. The reference page includes both Foldseek input types and modes, MMseqs2 and all four TM-align normalizations. The fourth panel enlarges the fold curves without changing their sampling.

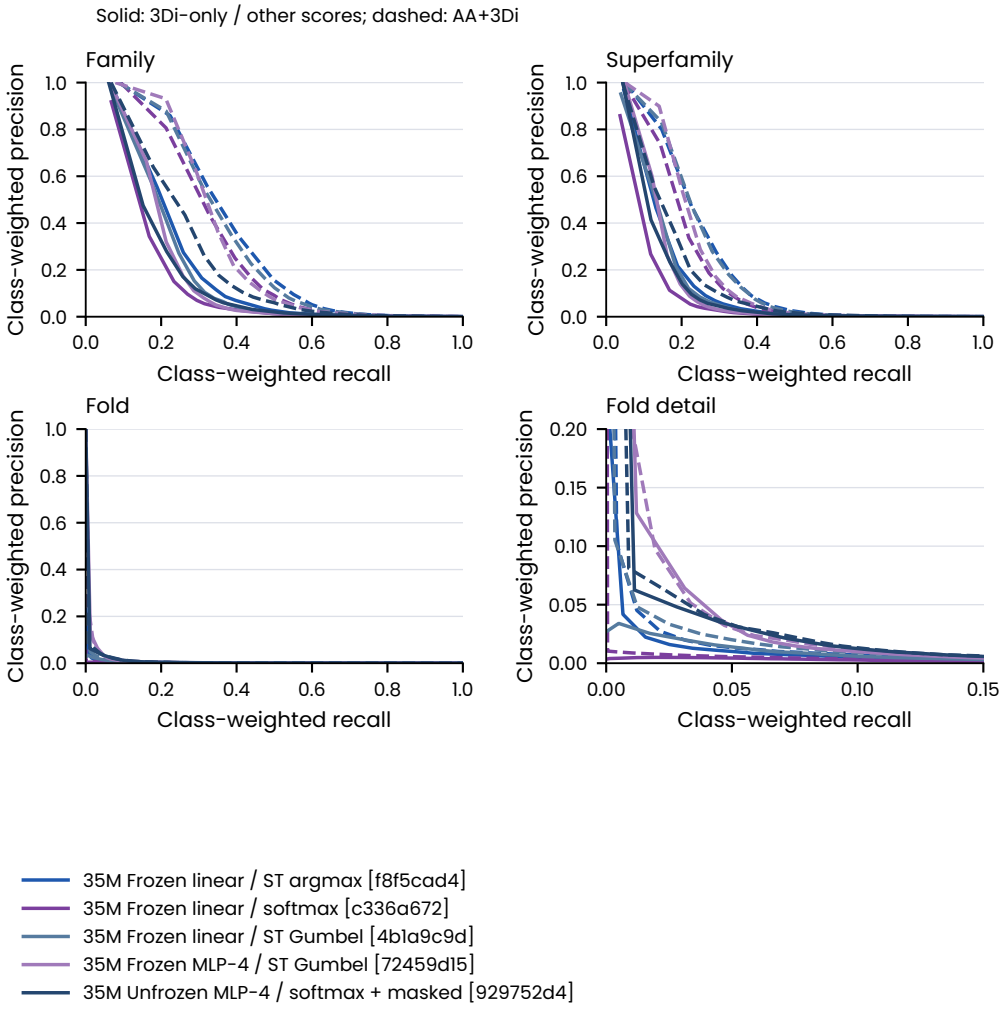


Figure S84: All supplied class-weighted precision–recall points. Colors identify predictors; both native Foldseek modes share that predictor's color, with solid 3Di-only and dashed AA+3Di lines. The reference page includes both Foldseek input types and modes, MMseqs2 and all four TM-align normalizations. The fourth panel enlarges the fold curves without changing their sampling.

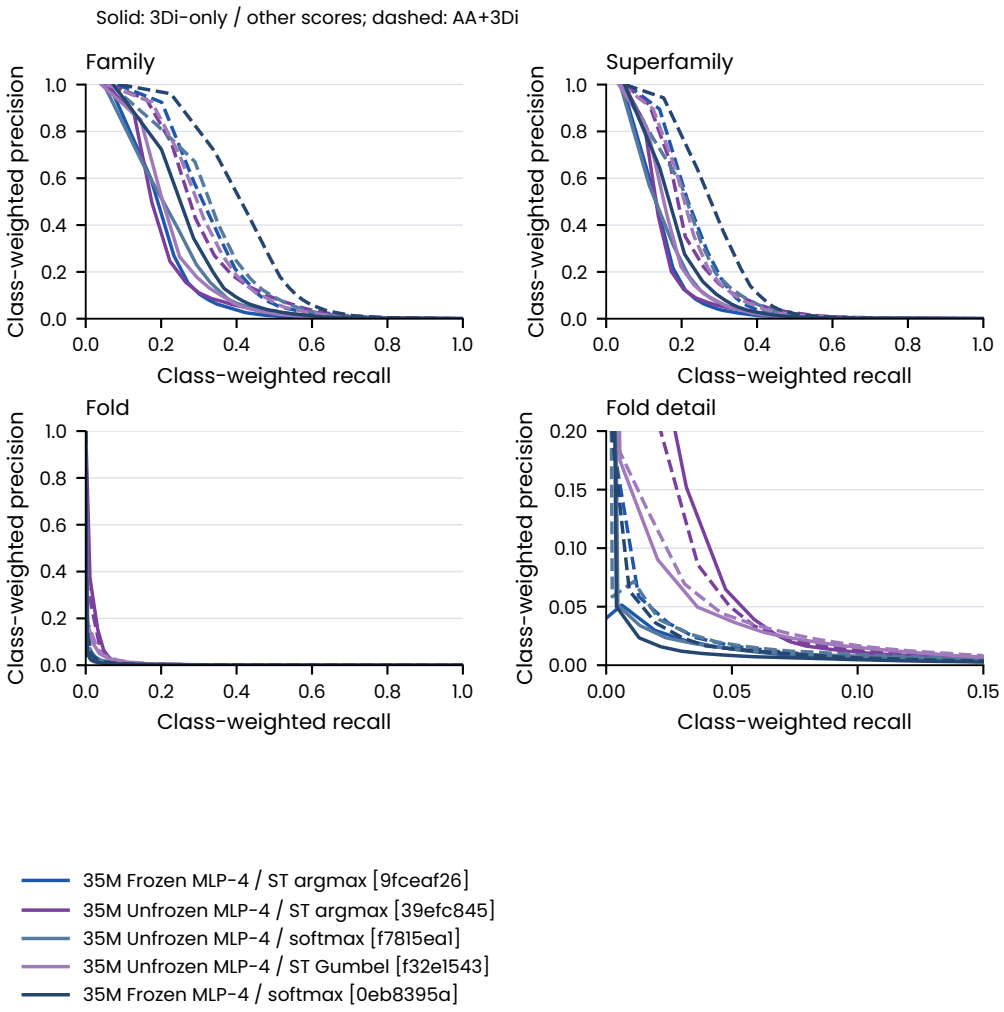


Figure S85: All supplied class-weighted precision–recall points. Colors identify predictors; both native Foldseek modes share that predictor's color, with solid 3Di-only and dashed AA+3Di lines. The reference page includes both Foldseek input types and modes, MMseqs2 and all four TM-align normalizations. The fourth panel enlarges the fold curves without changing their sampling.

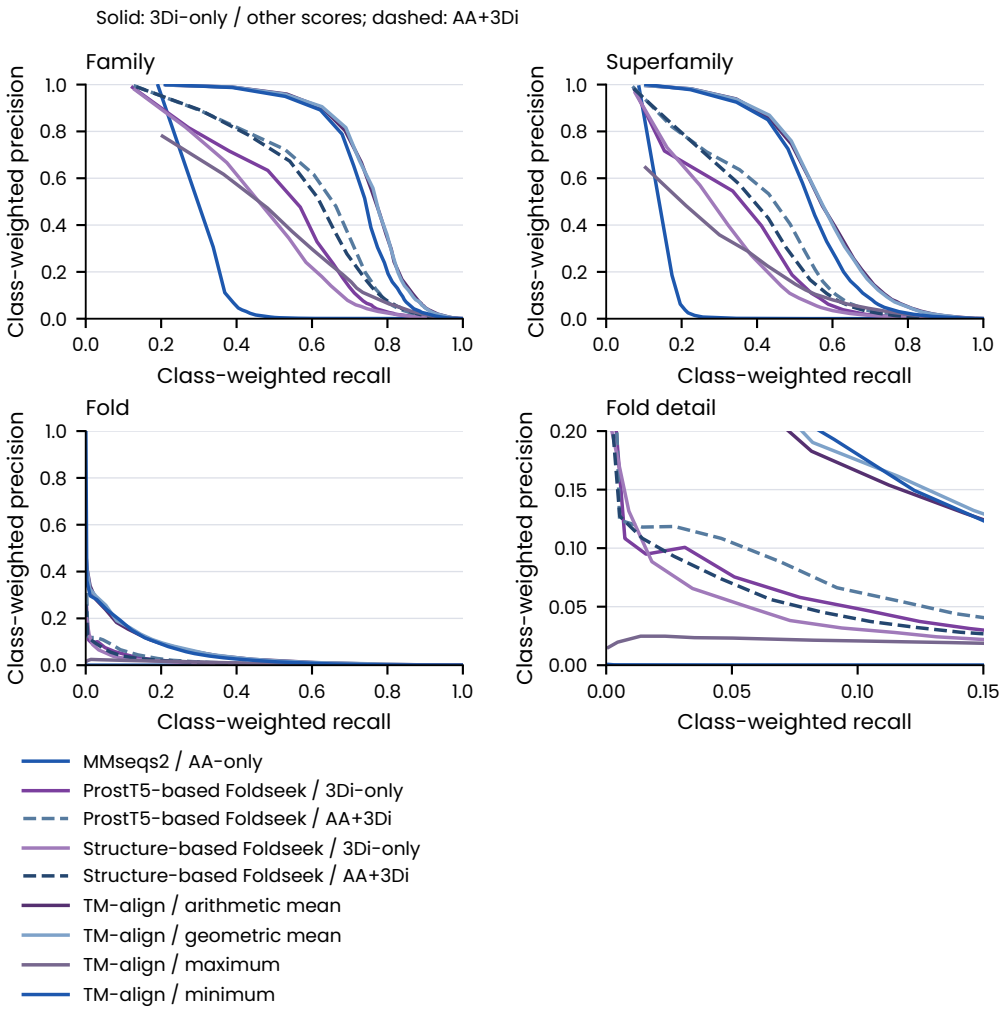


Figure S86: All supplied class-weighted precision–recall points. Colors identify predictors; both native Foldseek modes share that predictor's color, with solid 3Di-only and dashed AA+3Di lines. The reference page includes both Foldseek input types and modes, MMseqs2 and all four TM-align normalizations. The fourth panel enlarges the fold curves without changing their sampling.

Token distances / both reference masks

	Valid descriptor			Native Cα		
	Mismatch	Edit	Matrix	Mismatch	Edit	Matrix
8M Frozen linear [31bba998] ST Gumbel	0.902	0.839	0.227	0.902	0.838	0.226
8M Frozen linear [1cf7945c] softmax	0.914	0.870	0.223	0.909	0.864	0.222
8M Frozen linear [fbc46ff2] ST argmax	0.917	0.857	0.247	0.915	0.855	0.246
8M Frozen MLP-4 [7b541c60] softmax	0.914	0.874	0.236	0.908	0.867	0.234
8M Unfrozen MLP-4 [16132e1d] softmax	0.930	0.892	0.270	0.927	0.889	0.269
8M Frozen MLP-4 [d58f9e14] ST argmax	0.931	0.879	0.273	0.931	0.879	0.272
8M Unfrozen MLP-4 [650078f8] ST argmax	0.934	0.885	0.304	0.934	0.885	0.302
8M Unfrozen MLP-4 [d50ef818] ST Gumbel	0.914	0.854	0.251	0.912	0.852	0.250
8M Unfrozen MLP-4 [9d792c09] softmax + mask	0.935	0.894	0.267	0.935	0.894	0.266
8M Frozen MLP-4 [87d2c509] ST Gumbel	0.845	0.791	0.212	0.845	0.791	0.212
35M Frozen linear [f8f5cad4] ST argmax	0.916	0.864	0.269	0.916	0.863	0.268
35M Frozen linear [c336a672] softmax	0.912	0.864	0.218	0.907	0.859	0.216
35M Frozen linear [4b1a9c9d] ST Gumbel	0.921	0.854	0.253	0.920	0.852	0.251
35M Frozen MLP-4 [72459d15] ST Gumbel	0.916	0.837	0.245	0.917	0.836	0.244
35M Unfrozen MLP-4 [929752d4] softmax + mask	0.944	0.909	0.337	0.944	0.909	0.336
35M Frozen MLP-4 [9fceaf26] ST argmax	0.899	0.843	0.274	0.900	0.843	0.274
35M Unfrozen MLP-4 [39efc845] ST argmax	0.944	0.892	0.334	0.943	0.890	0.332
35M Unfrozen MLP-4 [f7815ea1] softmax	0.948	0.912	0.323	0.947	0.911	0.321
35M Unfrozen MLP-4 [f32e1543] ST Gumbel	0.926	0.872	0.286	0.924	0.870	0.284
35M Frozen MLP-4 [0eb8395a] softmax	0.919	0.874	0.229	0.915	0.870	0.227
ProstT5-based Foldseek predictor	0.387	0.376	0.069	0.385	0.374	0.069

Figure S87: All predictors under both reference masks. Mismatch is the unequal-state fraction, Edit is normalized pooled Levenshtein distance, and Matrix is normalized substitution dissimilarity. These predictor diagnostics do not use a Foldseek alignment mode.

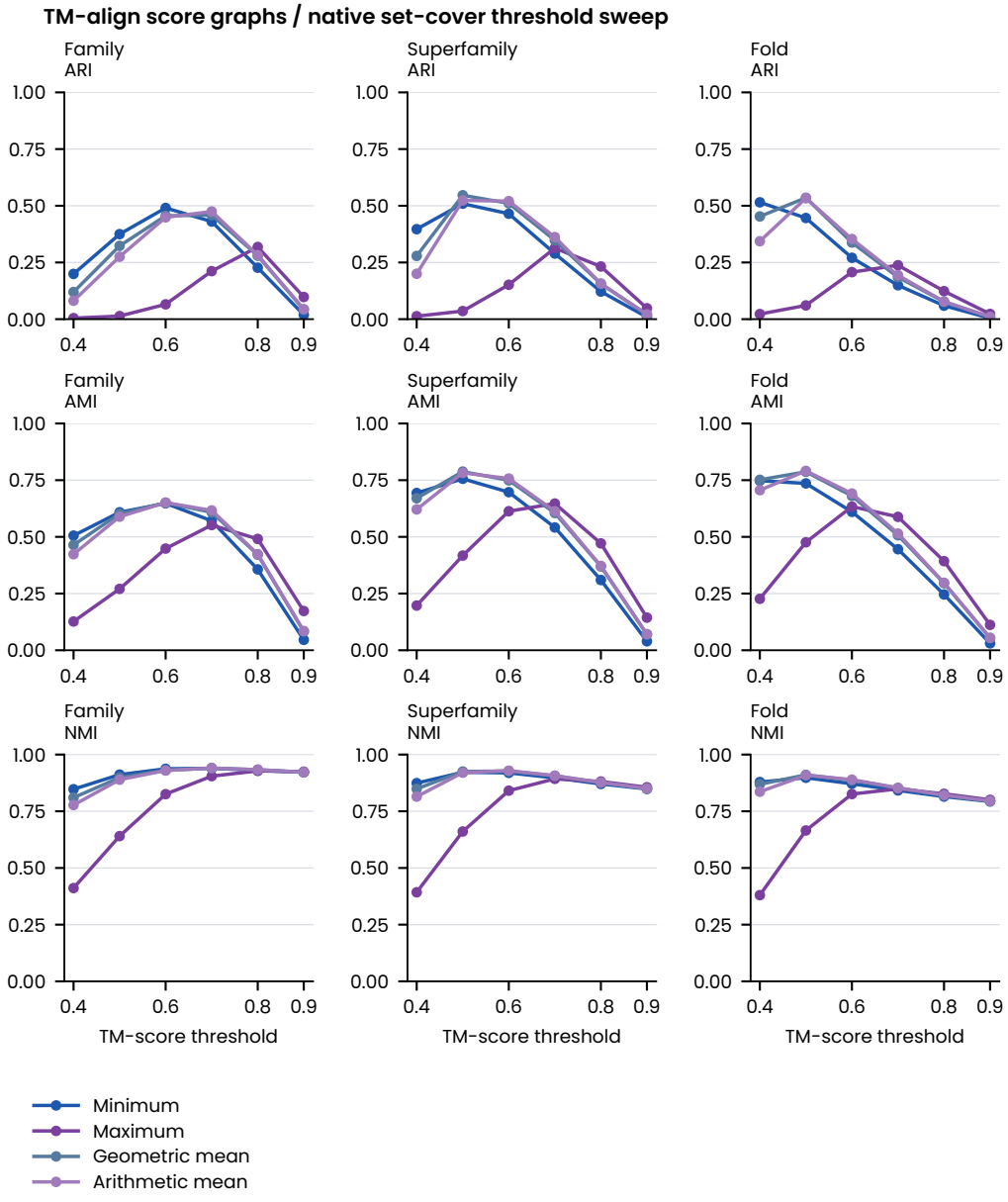


Figure S88: All four TM-score normalizations and all six fixed clustering thresholds. The unchanged Foldseek set-cover core clusters threshold graphs on the same validation proteins. The prespecified reference threshold is 0.5. Protein-pair precision, recall and F1 concern co-clustering of distinct proteins, not residue correspondences. The graph protocol does not apply the stock Foldseek cascade, E-value or coverage filters.

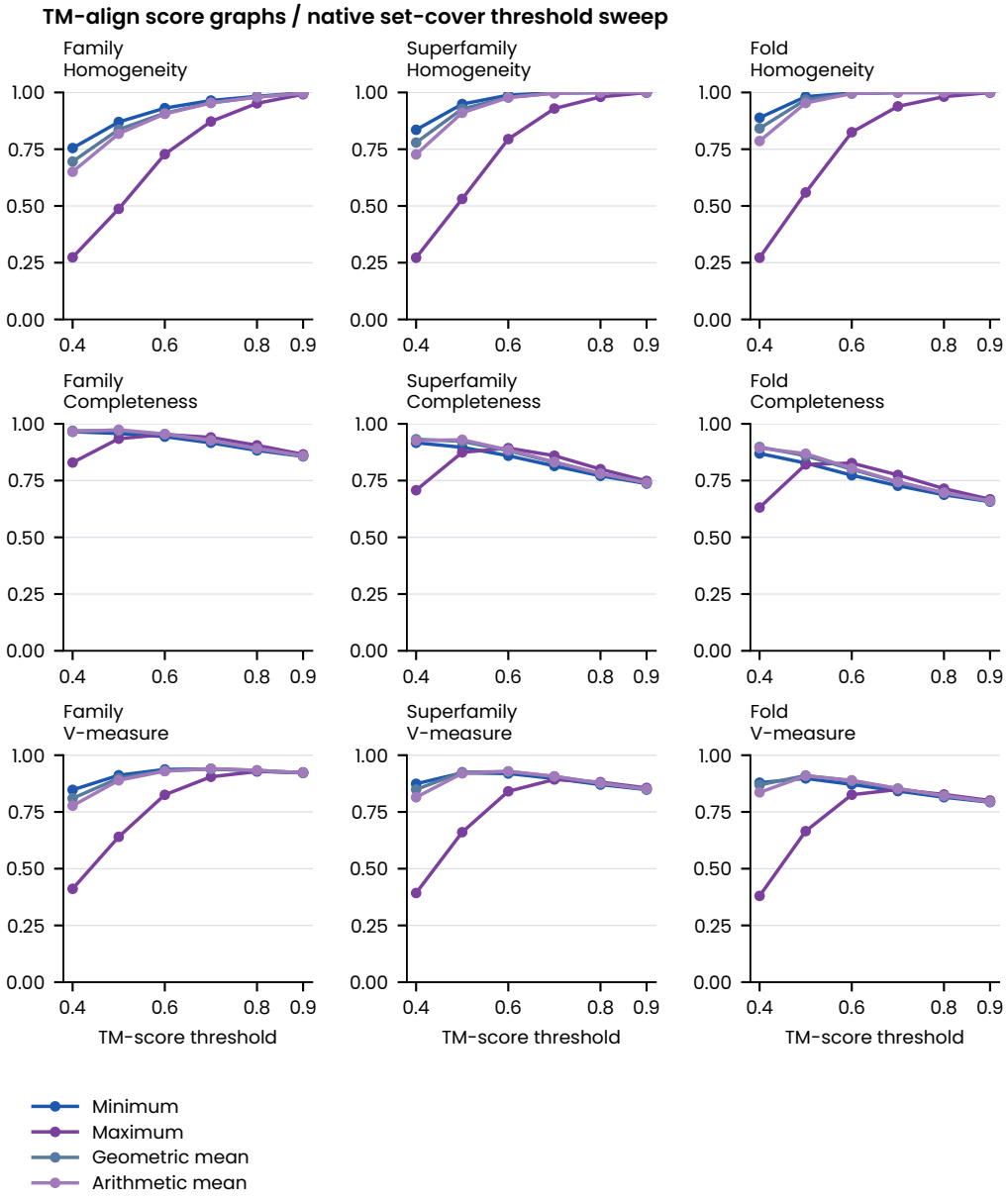


Figure S89: All four TM-score normalizations and all six fixed clustering thresholds. The unchanged Foldseek set-cover core clusters threshold graphs on the same validation proteins. The prespecified reference threshold is 0.5. Protein-pair precision, recall and F1 concern co-clustering of distinct proteins, not residue correspondences. The graph protocol does not apply the stock Foldseek cascade, E-value or coverage filters.

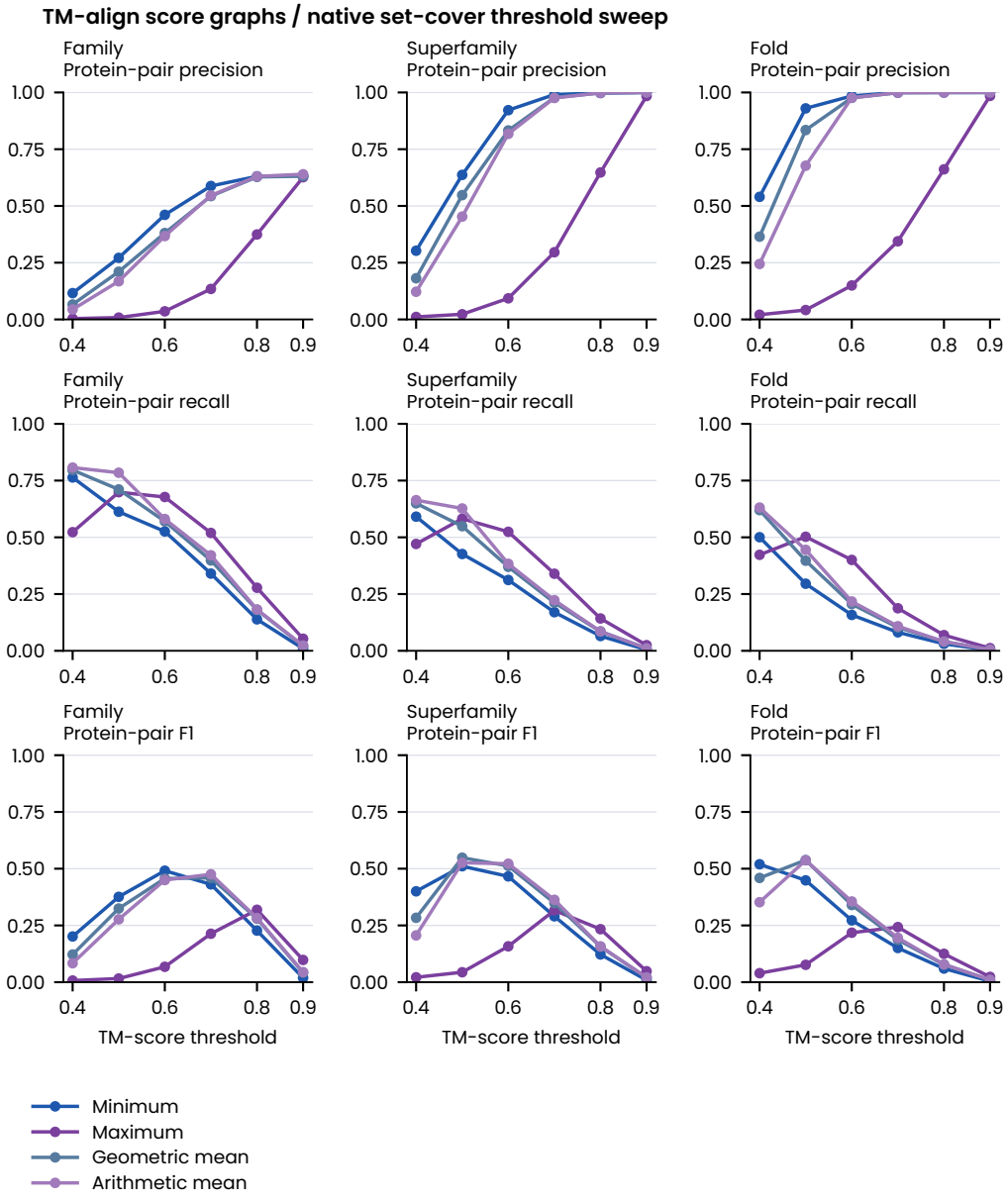


Figure S90: All four TM-score normalizations and all six fixed clustering thresholds. The unchanged Foldseek set-cover core clusters threshold graphs on the same validation proteins. The prespecified reference threshold is 0.5. Protein-pair precision, recall and F1 concern co-clustering of distinct proteins, not residue correspondences. The graph protocol does not apply the stock Foldseek cascade, E-value or coverage filters.

Table S40: TM-align graph clustering at family level. All four normalizations and all six fixed thresholds; the nominal reference is 0.5. Pair metrics concern co-clustered proteins, not residue correspondences.

Normalization	Threshold	ARI	AMI	NMI	Homogeneity	Completeness	V-measure	Pair precision	Pair recall	Pair F1
Minimum	0.4	0.1995	0.5056	0.8472	0.7550	0.9651	0.8472	0.1159	0.7636	0.2013
Minimum	0.5	0.3749	0.6084	0.9116	0.8698	0.9576	0.9116	0.2712	0.6127	0.3760
Minimum	0.6	0.4905	0.6478	0.9373	0.9310	0.9435	0.9373	0.4608	0.5258	0.4911
Minimum	0.7	0.4304	0.5707	0.9396	0.9644	0.9160	0.9396	0.5878	0.3402	0.4310
Minimum	0.8	0.2269	0.3565	0.9303	0.9830	0.8830	0.9303	0.6304	0.1386	0.2272
Minimum	0.9	0.0190	0.0458	0.9217	0.9977	0.8565	0.9217	0.6335	0.0097	0.0190
Maximum	0.4	0.0046	0.1269	0.4114	0.2735	0.8295	0.4114	0.0036	0.5226	0.0072
Maximum	0.5	0.0135	0.2707	0.6404	0.4871	0.9344	0.6404	0.0081	0.6988	0.0160
Maximum	0.6	0.0654	0.4484	0.8255	0.7279	0.9533	0.8255	0.0356	0.6776	0.0677
Maximum	0.7	0.2118	0.5520	0.9048	0.8720	0.9401	0.9048	0.1343	0.5193	0.2134
Maximum	0.8	0.3182	0.4909	0.9281	0.9520	0.9053	0.9281	0.3744	0.2778	0.3189
Maximum	0.9	0.0977	0.1729	0.9245	0.9921	0.8655	0.9245	0.6265	0.0531	0.0979
Geometric mean	0.4	0.1197	0.4643	0.8100	0.6955	0.9695	0.8100	0.0659	0.7962	0.1218
Geometric mean	0.5	0.3232	0.6017	0.8985	0.8367	0.9700	0.8985	0.2103	0.7110	0.3245
Geometric mean	0.6	0.4558	0.6506	0.9307	0.9087	0.9537	0.9307	0.3807	0.5704	0.4566
Geometric mean	0.7	0.4589	0.6082	0.9400	0.9544	0.9261	0.9400	0.5433	0.3981	0.4595
Geometric mean	0.8	0.2795	0.4214	0.9332	0.9796	0.8909	0.9332	0.6277	0.1801	0.2799
Geometric mean	0.9	0.0417	0.0827	0.9224	0.9962	0.8589	0.9224	0.6312	0.0216	0.0418
Arithmetic mean	0.4	0.0811	0.4235	0.7774	0.6506	0.9657	0.7774	0.0440	0.8072	0.0834
Arithmetic mean	0.5	0.2753	0.5889	0.8894	0.8184	0.9738	0.8894	0.1681	0.7846	0.2768
Arithmetic mean	0.6	0.4491	0.6510	0.9297	0.9055	0.9553	0.9297	0.3673	0.5806	0.4499
Arithmetic mean	0.7	0.4748	0.6167	0.9407	0.9537	0.9281	0.9407	0.5462	0.4208	0.4754
Arithmetic mean	0.8	0.2822	0.4238	0.9333	0.9795	0.8913	0.9333	0.6302	0.1821	0.2826
Arithmetic mean	0.9	0.0442	0.0849	0.9225	0.9962	0.8590	0.9225	0.6396	0.0230	0.0443

Table S41: TM-align graph clustering at superfamily level. All four normalizations and all six fixed thresholds; the nominal reference is 0.5. Pair metrics concern co-clustered proteins, not residue correspondences.

Normalization	Threshold	ARI	AMI	NMI	Homogeneity	Completeness	V-measure	Pair precision	Pair recall	Pair F1
Minimum	0.4	0.3968	0.6933	0.8741	0.8354	0.9165	0.8741	0.3026	0.5910	0.4003
Minimum	0.5	0.5092	0.7561	0.9220	0.9489	0.8966	0.9220	0.6368	0.4265	0.5109
Minimum	0.6	0.4649	0.6971	0.9193	0.9881	0.8594	0.9193	0.9220	0.3119	0.4661
Minimum	0.7	0.2893	0.5414	0.8972	0.9989	0.8143	0.8972	0.9908	0.1700	0.2902
Minimum	0.8	0.1217	0.3097	0.8705	0.9999	0.7708	0.8705	0.9978	0.0650	0.1221
Minimum	0.9	0.0090	0.0396	0.8484	1.0000	0.7368	0.8484	1.0000	0.0045	0.0090
Maximum	0.4	0.0131	0.1972	0.3928	0.2718	0.7075	0.3928	0.0110	0.4707	0.0214
Maximum	0.5	0.0357	0.4175	0.6607	0.5310	0.8743	0.6607	0.0227	0.5819	0.0438
Maximum	0.6	0.1517	0.6130	0.8409	0.7945	0.8930	0.8409	0.0930	0.5243	0.1580
Maximum	0.7	0.3132	0.6469	0.8930	0.9291	0.8597	0.8930	0.2961	0.3395	0.3163
Maximum	0.8	0.2323	0.4712	0.8812	0.9805	0.8002	0.8812	0.6474	0.1424	0.2335
Maximum	0.9	0.0480	0.1438	0.8559	0.9995	0.7483	0.8559	0.9842	0.0247	0.0482
Geometric mean	0.4	0.2787	0.6702	0.8488	0.7792	0.9321	0.8488	0.1814	0.6497	0.2836
Geometric mean	0.5	0.5463	0.7877	0.9253	0.9276	0.9230	0.9253	0.5476	0.5490	0.5483
Geometric mean	0.6	0.5109	0.7495	0.9275	0.9786	0.8815	0.9275	0.8326	0.3699	0.5122
Geometric mean	0.7	0.3478	0.6053	0.9060	0.9969	0.8302	0.9060	0.9772	0.2123	0.3488
Geometric mean	0.8	0.1560	0.3697	0.8766	0.9999	0.7804	0.8766	0.9981	0.0849	0.1565
Geometric mean	0.9	0.0200	0.0698	0.8506	1.0000	0.7400	0.8506	1.0000	0.0102	0.0201
Arithmetic mean	0.4	0.2001	0.6212	0.8146	0.7271	0.9262	0.8146	0.1219	0.6637	0.2060
Arithmetic mean	0.5	0.5236	0.7823	0.9201	0.9105	0.9299	0.9201	0.4530	0.6271	0.5260
Arithmetic mean	0.6	0.5205	0.7572	0.9290	0.9775	0.8850	0.9290	0.8177	0.3832	0.5218
Arithmetic mean	0.7	0.3618	0.6130	0.9071	0.9966	0.8323	0.9071	0.9758	0.2229	0.3628
Arithmetic mean	0.8	0.1570	0.3718	0.8769	0.9999	0.7808	0.8769	0.9981	0.0855	0.1576
Arithmetic mean	0.9	0.0210	0.0711	0.8507	1.0000	0.7401	0.8507	1.0000	0.0106	0.0211

Table S42: TM-align graph clustering at fold level. All four normalizations and all six fixed thresholds; the nominal reference is 0.5. Pair metrics concern co-clustered proteins, not residue correspondences.

Normalization	Threshold	ARI	AMI	NMI	Homogeneity	Completeness	V-measure	Pair precision	Pair recall	Pair F1
Minimum	0.4	0.5152	0.7477	0.8789	0.8886	0.8694	0.8789	0.5400	0.5005	0.5195
Minimum	0.5	0.4461	0.7357	0.8978	0.9815	0.8272	0.8978	0.9299	0.2956	0.4485
Minimum	0.6	0.2706	0.6103	0.8713	0.9972	0.7736	0.8713	0.9847	0.1581	0.2724
Minimum	0.7	0.1494	0.4455	0.8420	1.0000	0.7271	0.8420	1.0000	0.0814	0.1506
Minimum	0.8	0.0595	0.2460	0.8148	1.0000	0.6875	0.8148	1.0000	0.0309	0.0600
Minimum	0.9	0.0042	0.0305	0.7931	1.0000	0.6571	0.7931	1.0000	0.0021	0.0043
Maximum	0.4	0.0225	0.2268	0.3801	0.2719	0.6312	0.3801	0.0208	0.4229	0.0396
Maximum	0.5	0.0606	0.4760	0.6654	0.5593	0.8213	0.6654	0.0414	0.5025	0.0764
Maximum	0.6	0.2073	0.6337	0.8259	0.8249	0.8269	0.8259	0.1497	0.4004	0.2179
Maximum	0.7	0.2378	0.5882	0.8489	0.9388	0.7747	0.8489	0.3444	0.1874	0.2427
Maximum	0.8	0.1235	0.3929	0.8269	0.9814	0.7144	0.8269	0.6612	0.0690	0.1250
Maximum	0.9	0.0230	0.1120	0.8004	0.9995	0.6674	0.8004	0.9849	0.0117	0.0232
Geometric mean	0.4	0.4529	0.7516	0.8690	0.8417	0.8981	0.8690	0.3647	0.6198	0.4592
Geometric mean	0.5	0.5352	0.7871	0.9105	0.9683	0.8593	0.9105	0.8344	0.3970	0.5380
Geometric mean	0.6	0.3378	0.6806	0.8868	0.9953	0.7996	0.8868	0.9762	0.2058	0.3399
Geometric mean	0.7	0.1853	0.5070	0.8521	0.9996	0.7425	0.8521	0.9989	0.1030	0.1867
Geometric mean	0.8	0.0769	0.2957	0.8208	1.0000	0.6961	0.8208	0.9998	0.0404	0.0776
Geometric mean	0.9	0.0095	0.0538	0.7952	1.0000	0.6600	0.7952	1.0000	0.0048	0.0096
Arithmetic mean	0.4	0.3436	0.7059	0.8362	0.7861	0.8931	0.8362	0.2443	0.6312	0.3523
Arithmetic mean	0.5	0.5335	0.7910	0.9091	0.9536	0.8686	0.9091	0.6771	0.4448	0.5369
Arithmetic mean	0.6	0.3532	0.6911	0.8894	0.9954	0.8038	0.8894	0.9767	0.2172	0.3554
Arithmetic mean	0.7	0.1939	0.5147	0.8535	0.9996	0.7446	0.8535	0.9990	0.1083	0.1954
Arithmetic mean	0.8	0.0775	0.2975	0.8211	1.0000	0.6965	0.8211	0.9998	0.0407	0.0781
Arithmetic mean	0.9	0.0100	0.0548	0.7953	1.0000	0.6601	0.7953	1.0000	0.0051	0.0100

7844 7843 7842 7841 7840 7839 7838 7837 7836 7835 7834 7833 7832 7831 7830 7829 7828 7827 7826 7825 7824 7823 7822 7821 7820 7819 7818 7817 7816 7815 7814 7813 7812

Clustering / partition diagnostics

	Homogeneity			Completeness			V-measure			Pair precision			Pair recall			Pair F1		
	Fam.	SF	Fold	Fam.	SF	Fold	Fam.	SF	Fold	Fam.	SF	Fold	Fam.	SF	Fold	Fam.	SF	Fold
8M Frozen linear [31bba998] / ST Gumbel / 3Di-only	0.937	0.954	0.955	0.884	0.773	0.690	0.910	0.854	0.801	0.295	0.598	0.655	0.277	0.167	0.087	0.286	0.261	0.153
8M Frozen linear [31bba998] / ST Gumbel / AA+3Di	0.933	0.960	0.963	0.900	0.795	0.711	0.917	0.870	0.818	0.331	0.633	0.676	0.343	0.194	0.098	0.337	0.297	0.172
8M Frozen linear [1cf7945c] / softmax / 3Di-only	0.874	0.876	0.868	0.869	0.748	0.661	0.871	0.807	0.750	0.178	0.318	0.343	0.219	0.117	0.060	0.196	0.171	0.101
8M Frozen linear [1cf7945c] / softmax / AA+3Di	0.883	0.901	0.897	0.891	0.781	0.693	0.887	0.837	0.782	0.198	0.382	0.411	0.285	0.163	0.083	0.234	0.229	0.139
8M Frozen linear [fbc46ff2] / ST argmax / 3Di-only	0.934	0.948	0.948	0.883	0.769	0.686	0.908	0.850	0.796	0.328	0.586	0.612	0.282	0.150	0.074	0.303	0.238	0.132
8M Frozen linear [fbc46ff2] / ST argmax / AA+3Di	0.930	0.954	0.954	0.899	0.791	0.705	0.914	0.865	0.811	0.354	0.665	0.696	0.335	0.186	0.093	0.344	0.291	0.163
8M Frozen MLP-4 [7b541c60] / softmax / 3Di-only	0.869	0.870	0.863	0.868	0.746	0.660	0.869	0.803	0.748	0.146	0.278	0.318	0.214	0.121	0.066	0.174	0.169	0.109
8M Frozen MLP-4 [7b541c60] / softmax / AA+3Di	0.874	0.888	0.883	0.888	0.774	0.686	0.881	0.827	0.772	0.191	0.372	0.405	0.277	0.159	0.082	0.226	0.223	0.137
8M Unfrozen MLP-4 [16132e1d] / softmax / 3Di-only	0.872	0.883	0.883	0.886	0.771	0.687	0.879	0.823	0.773	0.160	0.332	0.445	0.312	0.192	0.122	0.212	0.243	0.191
8M Unfrozen MLP-4 [16132e1d] / softmax / AA+3Di	0.879	0.900	0.898	0.899	0.790	0.703	0.889	0.841	0.789	0.214	0.432	0.482	0.379	0.227	0.120	0.274	0.297	0.192
8M Frozen MLP-4 [d58f9e14] / ST argmax / 3Di-only	0.905	0.918	0.918	0.882	0.768	0.685	0.893	0.837	0.784	0.267	0.505	0.562	0.261	0.147	0.077	0.264	0.227	0.136
8M Frozen MLP-4 [d58f9e14] / ST argmax / AA+3Di	0.898	0.923	0.926	0.899	0.793	0.709	0.899	0.853	0.803	0.297	0.597	0.695	0.340	0.203	0.112	0.317	0.303	0.193
8M Unfrozen MLP-4 [650078f8] / ST argmax / 3Di-only	0.835	0.854	0.859	0.893	0.783	0.703	0.863	0.817	0.773	0.091	0.198	0.268	0.354	0.228	0.146	0.145	0.212	0.189
8M Unfrozen MLP-4 [650078f8] / ST argmax / AA+3Di	0.855	0.881	0.885	0.898	0.794	0.712	0.876	0.835	0.789	0.125	0.266	0.340	0.332	0.209	0.127	0.182	0.234	0.185
8M Unfrozen MLP-4 [d50ef818] / ST Gumbel / 3Di-only	0.718	0.727	0.724	0.887	0.770	0.684	0.793	0.748	0.704	0.060	0.129	0.216	0.451	0.289	0.230	0.105	0.178	0.223
8M Unfrozen MLP-4 [d50ef818] / ST Gumbel / AA+3Di	0.712	0.721	0.718	0.883	0.767	0.682	0.788	0.743	0.699	0.061	0.134	0.207	0.363	0.238	0.174	0.104	0.172	0.189
8M Unfrozen MLP-4 [9d792c09] / softmax + mask / 3Di-only	0.865	0.886	0.890	0.893	0.785	0.702	0.879	0.833	0.785	0.173	0.347	0.452	0.357	0.213	0.131	0.233	0.264	0.204
8M Unfrozen MLP-4 [9d792c09] / softmax + mask / AA+3Di	0.869	0.900	0.904	0.906	0.806	0.721	0.887	0.850	0.802	0.193	0.394	0.469	0.382	0.231	0.131	0.256	0.292	0.204
8M Frozen MLP-4 [87d2c509] / ST Gumbel / 3Di-only	0.945	0.964	0.968	0.888	0.778	0.697	0.916	0.861	0.810	0.364	0.679	0.775	0.288	0.159	0.086	0.322	0.258	0.155
8M Frozen MLP-4 [87d2c509] / ST Gumbel / AA+3Di	0.944	0.977	0.981	0.911	0.809	0.724	0.927	0.885	0.833	0.401	0.780	0.847	0.387	0.223	0.115	0.394	0.347	0.203
35M Frozen linear [f8f5cad4] / ST argmax / 3Di-only	0.920	0.959	0.967	0.913	0.817	0.735	0.917	0.882	0.835	0.348	0.686	0.787	0.410	0.240	0.131	0.376	0.356	0.224
35M Frozen linear [f8f5cad4] / ST argmax / AA+3Di	0.919	0.965	0.974	0.921	0.830	0.747	0.920	0.892	0.846	0.385	0.778	0.861	0.430	0.258	0.135	0.407	0.387	0.234
35M Frozen linear [c336a672] / softmax / 3Di-only	0.879	0.901	0.899	0.899	0.792	0.704	0.889	0.843	0.790	0.245	0.487	0.543	0.354	0.208	0.110	0.290	0.292	0.183
35M Frozen linear [c336a672] / softmax / AA+3Di	0.891	0.933	0.936	0.920	0.827	0.739	0.905	0.877	0.826	0.271	0.563	0.630	0.454	0.280	0.148	0.340	0.374	0.240
35M Frozen linear [4b1a9c9d] / ST Gumbel / 3Di-only	0.927	0.967	0.974	0.910	0.815	0.732	0.918	0.884	0.836	0.313	0.642	0.789	0.358	0.218	0.127	0.334	0.325	0.219
35M Frozen linear [4b1a9c9d] / ST Gumbel / AA+3Di	0.924	0.971	0.979	0.924	0.833	0.749	0.924	0.897	0.849	0.361	0.742	0.828	0.441	0.269	0.143	0.397	0.395	0.243

Figure S91: Complete cluster diagnostics comparison. Each metric is reported at family (Fam.), superfamily (SF), and fold levels. Row labels specify both Foldseek scoring modes for every learned, structure-based and ProstT5-based configuration. TM-align references use native set cover on score graphs at the fixed 0.5 threshold; the full threshold sweep is separate. Pair precision, recall and F1 concern co-clustered proteins, not residue alignments.

Clustering / partition diagnostics

	Homogeneity			Completeness			V-measure			Pair precision			Pair recall			Pair F1		
	Fam.	SF	Fold	Fam.	SF	Fold	Fam.	SF	Fold	Fam.	SF	Fold	Fam.	SF	Fold	Fam.	SF	Fold
35M Frozen MLP-4 [72459d15] / ST Gumbel / 3Di-only	0.919	0.958	0.966	0.912	0.816	0.734	0.915	0.881	0.835	0.264	0.542	0.662	0.399	0.243	0.141	0.318	0.335	0.232
35M Frozen MLP-4 [72459d15] / ST Gumbel / AA+3Di	0.913	0.960	0.969	0.923	0.833	0.750	0.918	0.892	0.845	0.252	0.526	0.673	0.440	0.273	0.165	0.321	0.359	0.266
35M Unfrozen MLP-4 [929752d4] / softmax + mask / 3Di-only	0.580	0.591	0.589	0.887	0.775	0.689	0.701	0.671	0.635	0.025	0.062	0.104	0.442	0.316	0.254	0.048	0.103	0.148
35M Unfrozen MLP-4 [929752d4] / softmax + mask / AA+3Di	0.582	0.595	0.590	0.888	0.779	0.689	0.703	0.675	0.636	0.022	0.051	0.073	0.465	0.312	0.214	0.043	0.087	0.109
35M Frozen MLP-4 [9fceaf26] / ST argmax / 3Di-only	0.942	0.975	0.983	0.907	0.807	0.725	0.924	0.883	0.835	0.368	0.722	0.844	0.398	0.231	0.128	0.382	0.350	0.223
35M Frozen MLP-4 [9fceaf26] / ST argmax / AA+3Di	0.941	0.981	0.989	0.915	0.819	0.736	0.928	0.893	0.844	0.395	0.784	0.900	0.411	0.242	0.132	0.403	0.370	0.230
35M Unfrozen MLP-4 [39efc845] / ST argmax / 3Di-only	0.864	0.910	0.922	0.918	0.831	0.751	0.890	0.869	0.828	0.157	0.352	0.457	0.424	0.281	0.173	0.230	0.313	0.251
35M Unfrozen MLP-4 [39efc845] / ST argmax / AA+3Di	0.876	0.924	0.935	0.925	0.838	0.756	0.900	0.879	0.836	0.176	0.378	0.472	0.449	0.286	0.170	0.252	0.326	0.250
35M Unfrozen MLP-4 [f7815ea1] / softmax / 3Di-only	0.738	0.769	0.771	0.910	0.814	0.728	0.815	0.791	0.749	0.053	0.127	0.165	0.435	0.308	0.190	0.095	0.180	0.177
35M Unfrozen MLP-4 [f7815ea1] / softmax / AA+3Di	0.729	0.760	0.761	0.917	0.821	0.733	0.812	0.789	0.747	0.061	0.138	0.184	0.511	0.343	0.217	0.109	0.197	0.199
35M Unfrozen MLP-4 [f32e1543] / ST Gumbel / 3Di-only	0.651	0.673	0.676	0.904	0.802	0.719	0.757	0.732	0.697	0.036	0.088	0.152	0.498	0.357	0.292	0.068	0.141	0.200
35M Unfrozen MLP-4 [f32e1543] / ST Gumbel / AA+3Di	0.676	0.700	0.699	0.904	0.804	0.716	0.774	0.748	0.707	0.048	0.113	0.153	0.423	0.295	0.189	0.086	0.164	0.169
35M Frozen MLP-4 [0eb8395a] / softmax / 3Di-only	0.815	0.838	0.837	0.900	0.794	0.707	0.856	0.815	0.767	0.143	0.305	0.369	0.406	0.257	0.148	0.211	0.279	0.211
35M Frozen MLP-4 [0eb8395a] / softmax / AA+3Di	0.825	0.861	0.865	0.916	0.821	0.735	0.868	0.841	0.795	0.157	0.332	0.416	0.443	0.278	0.165	0.232	0.302	0.236
MMseqs2 / AA-only	0.983	1.000	1.000	0.881	0.769	0.686	0.929	0.870	0.814	0.625	1.000	1.000	0.117	0.056	0.026	0.198	0.106	0.051
ProstT5-based / Foldseek / 3Di-only	0.875	0.933	0.952	0.940	0.860	0.782	0.907	0.895	0.859	0.255	0.523	0.657	0.538	0.327	0.195	0.346	0.403	0.301
ProstT5-based / Foldseek / AA+3Di	0.883	0.942	0.964	0.940	0.861	0.786	0.910	0.899	0.866	0.271	0.570	0.794	0.521	0.325	0.214	0.357	0.414	0.338
Structure-based / Foldseek / 3Di-only	0.910	0.966	0.983	0.934	0.851	0.772	0.922	0.905	0.865	0.353	0.732	0.884	0.474	0.291	0.167	0.405	0.417	0.281
Structure-based / Foldseek / AA+3Di	0.915	0.972	0.989	0.938	0.855	0.776	0.926	0.910	0.870	0.337	0.694	0.906	0.494	0.301	0.187	0.401	0.420	0.310
TM-align / arithmetic mean	0.818	0.911	0.954	0.974	0.930	0.869	0.889	0.920	0.909	0.168	0.453	0.677	0.785	0.627	0.445	0.277	0.526	0.537
TM-align / geometric mean	0.837	0.928	0.968	0.970	0.923	0.859	0.898	0.925	0.911	0.210	0.548	0.834	0.711	0.549	0.397	0.325	0.548	0.538
TM-align / maximum	0.487	0.531	0.559	0.934	0.874	0.821	0.640	0.661	0.665	0.008	0.023	0.041	0.699	0.582	0.503	0.016	0.044	0.076
TM-align / minimum	0.870	0.949	0.982	0.958	0.897	0.827	0.912	0.922	0.898	0.271	0.637	0.930	0.613	0.427	0.296	0.376	0.511	0.449

Figure S92: Complete cluster diagnostics comparison. Each metric is reported at family (Fam.), superfamily (SF), and fold levels. Row labels specify both Foldseek scoring modes for every learned, structure-based and ProstT5-based configuration. TM-align references use native set cover on score graphs at the fixed 0.5 threshold; the full threshold sweep is separate. Pair precision, recall and F1 concern co-clustered proteins, not residue alignments.