Systematic Disagreement between ESMFold and MSA-Based Structure Predictors on Chimeric Fusion Proteins

Anonymous authors
Paper under double-blind review

Abstract

Single-sequence structure predictors such as ESMFold are widely used as computational filters for designed proteins, including multi-domain fusions absent from natural sequence databases. Whether this filter agrees with MSA-based predictors on such constructs has not been established. We assemble 1,000 chimeric domain pairs that do not occur in UniProt, held out by protein family, and score each fusion by the worse domain's TM-score against its native AlphaFold Database conformation. ESMFold exhibits a strong, internally reproducible preference for one domain order (Pearson r=0.90 under composition-matched linker swaps) that Boltz-2 does not share (r=0.07). On an unbiased subset of 300 constructs, ESMFold assigns TM<0.7 to 17.7% of fusions and Boltz-2 to 8.7%, with only ten shared failures. AlphaFold 2, on the same alignments, fails on 7/10 of those consensus failures and on 1/43 of the failures unique to ESMFold. Isolated-domain controls partition the latter set: 26 constructs contain a domain that ESMFold already fails alone—domains that isolated Boltz-2 and AlphaFold 2 both fold—and 17 fail only after fusion. The fusion-dependent failures coincide with elevated drift in the affected domain's pair block, recur across unrelated partners, and a held-out dummy fusion flags them at high precision. On the six ESMFold-only domains with deposited crystal structures, Boltz-2 and AlphaFold 2 agree with the crystal (mean TM 0.88 and 0.89) while ESMFold does not (0.22). Neither a stricter TM threshold nor an ESMFold confidence gate resolves the disagreement. MSA models recover a population-level advantage of natural structured junctions over a glycine–serine linker, but do not agree on which pairs benefit. Predictor scores are therefore not interchangeable on artificial fusions; the claim concerns the predictors, not the proteins.

1 Introduction

Designed proteins are increasingly accepted or rejected by a structure predictor rather than by a crystal or an activity assay. The shortcut is especially tempting for fusion constructs—two pre-existing domains joined by a linker—because the design space is the junction, not the folds themselves (Chen et al., 2013), and because single-sequence models such as ESMFold (Lin et al., 2023) can score thousands of candidates without a multiple-sequence alignment (MSA). That practice silently assumes that a model trained on natural chains remains interchangeable with an MSA-based verdict on sequences that nature never made.

We test the assumption. The question is not how to generate a better fusion, but whether the folding models the field already uses as fusion oracles agree with one another on chimeric constructs. If they do not, every downstream pipeline that filters designs by ESMFold pLDDT or TM-score will fit one model's private calls rather than a shared structural verdict.

Three controls make the test interpretable. First, every chimera is a pair of domains taken from different source proteins, so the fused chain is absent from UniProt and from the ESMFold training distribution in its assembled form (The UniProt Consortium, 2023; Lin et al., 2023). Second, test folds are family-disjoint, so the order-sensitivity audits in Section 4 cannot train on an evaluation family; those checks are not a classifier we pro-

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.

pose to deploy. Third, when two predictors disagree we do not declare a winner by fiat. We bring in a third model whose architecture is not an AlphaFold 3 reimplementation—AlphaFold 2 (Jumper et al., 2021), run through OpenFold (Ahdritz et al., 2024) on the same ColabFold MSAs (Mirdita et al., 2022) already used by Boltz-2 (Wohlwend et al., 2024; Passaro et al., 2025)—and we require that this third model reproduce the consensus failures before we let it arbitrate the private ones.

The paper is organised around three questions that a fusion-design pipeline actually has to answer (Figure 1). Does domain order matter? ESMFold says yes: 26.5% of 1,000 chimeras change worse-domain TM by more than 0.05 when the two domains are swapped, and the signed gap replicates at r=0.90 under neutral linker swaps. MSA models say no (r=0.07 with Boltz-2). When the models disagree about a low-q call, do they cluster? On an unbiased 300-construct sample ESMFold fails 17.7%, Boltz-2 fails 8.7%, and AF2 fails 4.0%; only 10 failures are shared, and AF2 reproduces ESMFold's 43 private failures at 2.3%. Do the tested TM cutoffs or ESMFold pLDDT gates resolve the disagreement? No: a cutoff sweep leaves recall of Boltz-2 failures stuck at 38.5%, and ESMFold pLDDT is almost collinear with its own TM.

Contributions. (i) A family-disjoint chimeric fusion benchmark of 1,000 never-observed domain pairs (1,223 source proteins, zero family-to-fold leakage). (ii) Evidence that ESMFold's order preference is model-dependent: it is internally reproducible and is not reproduced by the MSA evaluators tested here. (iii) An unbiased comparison in which ESMFold calls nine percentage points more AFDB-referenced low-q constructs than Boltz-2; AF2 agrees with Boltz-2 more than with ESMFold; the extra calls split into isolated-domain difficulty (26/43) and fusion-context loss (17/43), and a disjoint redraw gives 20/38 and 18/38. (iv) A negative result on cheap remedies (the tested TM cutoffs and ESMFold confidence gates), a held-out throwaway-chimera screen for context-fail, and an operational reading, not a validated accept/reject rule: joint MSA-model q<0.7 rejects fewer candidates than ESMFold alone; on structured gaps a natural junction beats G4S3 as a population default, but ESMFold cannot rank individual linkers.

2 Related work

Structure predictors as oracles. AlphaFold 2 (Jumper et al., 2021) and its database (Varadi et al., 2022) made high-confidence backbone prediction routine for natural proteins. Single-sequence successors—ESMFold (Lin et al., 2023), OmegaFold (Chowdhury et al., 2022)—remove the MSA so that designed sequences can be scored at library scale. Community assessments have documented that these models are less reliable away from the training distribution, including on orphan sequences and some multi-domain proteins (Akdel et al., 2022; Tunyasuvunakool et al., 2021). We isolate a narrower disagreement: assembled chimeras whose parts are individually familiar.

Fusion and chimeric design. Linker choice and domain order are classical variables in fusion-protein engineering (Chen et al., 2013; Esposito & Chatterjee, 2006). Recent generative work often conditions on function tags and emits a full chain (Madani et al., 2023; Nijkamp et al., 2023). That framing is easy to leak through homology and hard to audit for whether both source domains survived. We instead hold the two domain sequences fixed and vary only order and linker, which makes a negative call attributable to the assembled chain rather than to a missing domain sequence.

When predictors disagree. Disagreement between ESMFold and AF2 is usually summarized by global TM or pLDDT correlation on natural proteins. On designed binders and complexes the field has already learned not to trust a single model (Bryant et al., 2022; Evans et al., 2022). Chimeric sequences are a documented failure mode for MSA-based models: Vedula et al. show that AF2/AF3 lose accuracy on structured peptides fused to workhorse scaffolds because a single-search MSA drops coverage on one partner, and that concatenating separately computed alignments (a windowed MSA) restores the peptide (Vedula et al., 2025). ESMFold is the weakest of their three predictors even on isolated peptides. That setting is complementary to ours. They study short tags on four scaffolds and diagnose an

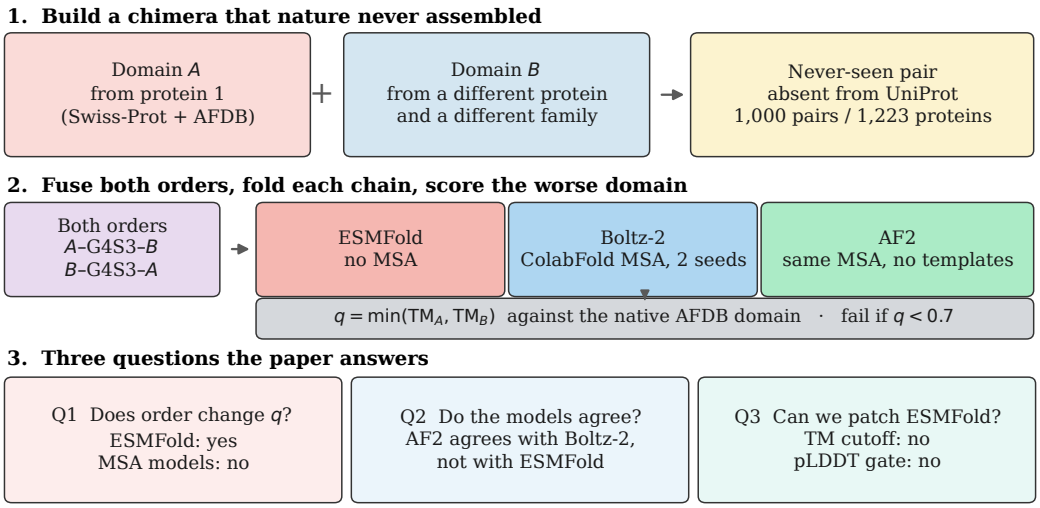

Figure 1: Study design. A chimeric pair is two domains from different proteins, fused in both orders with a G4S3 linker. Every chain is folded by ESMFold (no MSA), Boltz-2 (ColabFold MSA), and AlphaFold 2 (the same MSA), then scored by the worse domain's TM against its AFDB-predicted reference conformation. RoseTTAFold2 (Baek et al., 2023) is run on those same MSAs (Appendix S). Q1–Q3 are answered in Sections 4–6.

MSA-search artefact; we study two full domains from never-observed family pairs and find the opposite disagreement: ESMFold assigns low q to constructs that two MSA models, given the same unwindowed ColabFold alignment, score highly (Figure 6). Windowed re-folds, and a local re-search that does not call api.colabfold.com, do not collapse that private set (Appendices F and G).

3 Setup: a memory-controlled fusion test

The design is in Figure 1. We hold the two domain sequences fixed, vary only order and linker, and ask three predictors whether each domain still matches its native conformation. That removes assembled-chain memorisation as an explanation of a positive call, and it lets a negative call be split into isolated-domain difficulty versus a fusion-context loss (Appendix J).

Constructs. A chimeric pair is two domains A and B taken from different Swiss-Prot proteins, each with an AlphaFold Database structure. The two orders A-L-B and B-L-A are emitted with a fixed 15-residue glycine-serine linker L=GGGGS×3, unless a reliability experiment substitutes a length-matched alternative. Spans in the fused chain are recorded so that each domain can be sliced out after folding.

Score. Let $\mathrm{TM}(X, X^\star)$ be a known-correspondence TM-score (Zhang & Skolnick, 2004; Xu & Zhang, 2010) of a predicted domain against its AFDB-predicted reference conformation (iterative Cα Kabsch (Kabsch, 1976) on the recorded span). The construct score is the worse domain,

$$q \;=\; \min\big(\mathrm{TM}(A, A^\star),\, \mathrm{TM}(B, B^\star)\big). \tag{1}$$

A construct fails if $q < 0.7$ and fails catastrophically if $q < 0.5$. Every domain of the 300-construct panel is also scored with USalign (Zhang et al., 2022) as an implementation check (r=0.99, two verdict changes in 900 domains; Appendix U). The reference must be the AFDB multi-domain conformation, not a single-domain re-fold: a natural junction induces genuine inter-domain packing, so agreement with an isolated-domain prediction systematically favors the more flexible linker (Appendix A).

Comparing predictors. The anti-pattern is to correlate order gaps $q_{AB} - q_{BA}$. If the second model folds both orders well, the true gap is near zero and the correlation is near zero by construction; noise and "no effect" are indistinguishable. We therefore compare absolute q on matched constructs and, for rates, use a pair-clustered bootstrap (10,000 resamples of chimera_id) so that the two orders of one pair are never treated as independent observations.

Floor-effect guard. If a reference model fails more than 60% of all constructs, including ones the other model folds, agreement on failures is vacuous. Single-sequence Boltz-2 triggers this guard (96.2% of constructs at $q < 0.7$) and is never used for a claim.

3.1 The chimeric benchmark

Natural multi-domain proteins in Swiss-Prot are a contaminated test set for ESMFold: the assembled chain, or a close homologue, was likely seen during pretraining. Time holdout is infeasible—almost all AFDB-backed Swiss-Prot entries predate the ESMFold cutoff—so we build chimeras.

Domains are taken from a pool of two-domain Swiss-Prot proteins with AFDB structures. A pair is chimeric if the two domains come from different proteins and the unordered family pair is absent from the natural pool. We sample 1,000 such pairs under a cap of eight pairs per family, yielding 1,000 distinct family combinations and 1,223 source proteins. For supervised analyses we partition 297 domain families into five groups and pair only inside a group, so holding out a fold holds out every family in it. Family-to-fold leakage is zero by construction, and also at the sequence level: MMseqs2 (Steinegger & Söding, 2017) clustering (coverage 0.8) of the 1,390 source domains puts no cluster in two folds at 30%, 40%, or 70% identity.

The full 1,000-pair set is folded with ESMFold in both orders (2,000 fusions) and used for the order-sensitivity and predictability measurements in Section 4. The three-predictor comparison in Section 5 uses a uniformly random subset of 150 pairs (300 constructs), drawn with seed 20260908. Order-sensitive pairs ($|q_{AB} - q_{BA}| > 0.05$ under ESMFold) make up 25.3% of that subset versus 26.5% of the population, so the sample is not a tail.

4 Q1: ESMFold has a private opinion about domain order

On the 1,000-pair set, ESMFold's worse-domain TM changes by more than 0.05 in 26.5% of pairs, by more than 0.10 in 15.2%, and by more than 0.20 in 9.5%. The gap distribution is heavy-tailed (median 0.016, mean $|q_{AB} - q_{BA}|$=0.064, 90th percentile 0.186). Choosing the better ESMFold order raises mean q from 0.785 (A-first) or 0.786 (random order) to 0.818, a +0.032 gain, which equals mean $|q_{AB} - q_{BA}|/2$. The 0.064 figure is better-minus-worse, not the gain versus a fixed direction. Failure ($q < 0.7$) rises from 17.0% in the better order to 23.7% in the worse one; the two orientations are not globally biased (20.5% vs. 20.2%).

The gap is not noise. We re-fold 150 pairs (90 ESMFold-sensitive, 60 indifferent) with two length-matched flexible linkers, GSGSGSGSGSGSGSG and GGSGGSGGSGGSGGS. A deterministic self-check (repeating the original G4S3 linker) reproduces the original gap to numerical zero. On sensitive pairs the sign of the gap agrees on both substitutes for 77.8% of pairs, and the two substitutes correlate at r=0.90. Only 1.7% of indifferent pairs become sensitive. Whatever ESMFold is computing, it is a stable function of the domain pair.

The gap is not predictable from cheap structure. A family-disjoint classifier that predicts whether a pair is order-sensitive reaches ROC-AUC 0.716 from a single feature: the worse isolated-domain TM against native. Logistic regression on twelve symmetric features does not beat it (AUC 0.701). Predicting which order is better is worse than chance after we switch to signed, antisymmetric features and augment every pair with its swap (learned AUC 0.481 on all 1,000 pairs; 0.478 on the sensitive subset). A physics-only score comparing buriedness of the four termini in the AFDB structures reaches AUC 0.541 on sensitive pairs

Table 1: Gain in worse-domain q versus a random order on the unbiased 150 pairs. The row model chooses the orientation; the column model scores it. Diagonal cells are self-gains; off-diagonal cells are transfer. Pair-clustered bootstrap 95% CIs. No evaluator is treated as truth. Details in Appendix E.

Selector → evaluator	ESMFold	Boltz-2	AF2
ESMFold	$+0.026_{[+.017,+.036]}$	$-0.003_{[-.008,+.002]}$	$-0.001_{[-.006,+.003]}$
Boltz-2	$+0.008_{[-.002,+.018]}$	$+0.017_{[+.014,+.022]}$	$+0.003_{[-.002,+.008]}$
AF2	$+0.003_{[-.007,+.013]}$	$+0.003_{[-.002,+.008]}$	$+0.015_{[+.011,+.019]}$

Table 2: Absolute fold quality on a uniformly random sample of 150 chimeric pairs (300 constructs). Boltz-2 is the mean of two MSA-enabled seeds. Intervals are pair-clustered bootstrap 95% CIs.

	ESMFold	Boltz-2+MSA	AF2+MSA
Mean worse-domain TM	0.794	0.850	0.898
Median	0.887	0.898	0.938
Failure rate ($q < 0.7$)	17.7%	8.7%	4.0%
95% CI	[12.0, 23.7]	[4.7, 13.0]	[1.3, 7.0]
Catastrophic ($q < 0.5$)	12.7%	4.7%	2.0%
Pearson r vs. ESMFold	—	+0.39	+0.40
Pearson r vs. Boltz-2	+0.39	—	+0.83

and does not beat a constant A-first policy in realized q. The learned policy captures 4% of the oracle headroom (0.787 vs. 0.785 for A-first vs. 0.818 for the better order).

An independent model does not see the gap. A stable but unpredictable gap reads either as weak features or as a gap that belongs to ESMFold. On the unbiased 150 pairs, Boltz-2 with MSAs and two diffusion seeds is moderately self-consistent on ESMFold-sensitive pairs (r=0.64) and essentially uncorrelated with ESMFold's signed gap (r=0.07, sign agreement 57.9%). Picking that order is a self-gain of +0.026 q under ESMFold on the unbiased 150 pairs (versus a random order; pair-clustered 95% CI $[+0.017, +0.036]$) and shows no clear positive transfer under Boltz-2 or AF2 (Table 1). The three pairs where ESMFold reports one intact order ($q \geq 0.8$) and one destroyed order ($q < 0.7$) are decisive: ESMFold's bad-order mean is 0.27, Boltz-2's mean on those same chains is 0.95, and Boltz-2's own gap is +0.009. We treat order as an ESMFold-specific preference from this point on and do not build a direction model. The rest of the paper asks the more useful question: when ESMFold rejects a fusion outright, do the MSA models agree?

5 Q2: ESMFold's extra failures are not reproduced

Table 2 reports absolute quality on the unbiased 300 constructs. ESMFold's failure rate exceeds Boltz-2's by +9.0 percentage points, CI $[+2.3, +15.7]$, which excludes zero. Absolute q still correlates at r=0.39, but the two models do not agree on the decision boundary. AF2 is more optimistic still (4.0% fail, CI $[1.3, 7.0]$; median TM 0.938).

Of 53 ESMFold failures, Boltz-2 reproduces 10 (18.9%, CI $[5.9\%, 33.9\%]$). Of 26 Boltz-2 failures, ESMFold reproduces 10 (38.5%). The remaining 43 and 16 failures are private to the ESMFold/Boltz-2 split. Boltz-2's 16 private failures survive both diffusion seeds (27 at one seed, 26 after averaging), so they are not sampling noise. A first reading—"ESMFold is harsher, Boltz-2 is truth"—assumes the second model cannot itself be wrong; Section 5.1 supplies the missing vote.

A biased subset inflates the same contrast. An earlier 80-pair panel oversampled large ESMFold order gaps and gave 42.5% vs. 17.5% low-q on the 60 pairs with MSAs; that

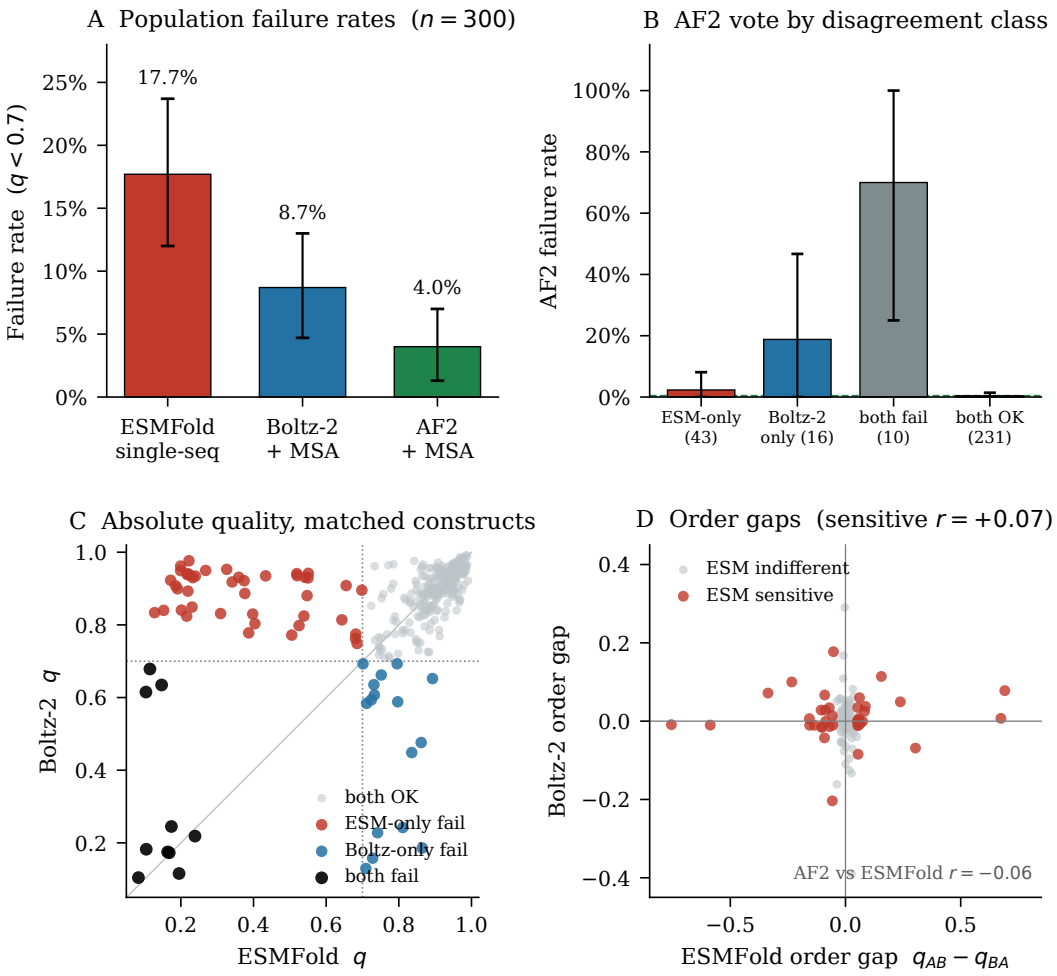

Figure 2: Unbiased 300-construct comparison. (A) Failure rates ($q < 0.7$) with pair-clustered 95% CIs. (B) AF2's failure rate inside each ESMFold/Boltz-2 class; the dashed line is AF2's rate where the other two already agree the construct is intact. (C) Matched absolute quality. Red points are ESMFold-only failures: Boltz-2 folds them. (D) Signed order gaps. ESMFold-sensitive pairs (red) do not align with Boltz-2 (r=0.07).

rate is selection, not prevalence (Appendix C). We report 17.7% vs. 8.7% as the population estimate. A second uniform 150-pair draw, disjoint from that sample (seed 20260916), reproduces the excess and the AF2 arbitration (Appendix D).

5.1 A third predictor breaks the tie

Table 3: AF2 failure rate inside each ESMFold/Boltz-2 disagreement class, on the same 300 constructs and the same MSAs. CIs are pair-clustered bootstrap.

Class	n	ESM TM	Boltz TM	AF2 fail
ESMFold-only failure	43	0.378	0.883	2.3% [0.0, 8.1]
Boltz-2-only failure	16	0.774	0.473	18.8% [0.0, 46.7]
Consensus failure	10	0.149	0.314	70.0% [25.0, 100]
Consensus OK	231	0.901	0.893	0.4% [0.0, 1.4]

We run AlphaFold 2 model_1_ptm through OpenFold with templates disabled at inference, using each construct's Boltz-retrieved unpaired MSA. AF2 is a different architecture from Boltz-2 (Evoformer plus structure module, versus an AF3-style diffusion model), so agreement is not explained by a shared inductive bias at the network level. The MSAs are shared, deliberately: a difference in evolutionary input would confound the comparison.

Two controls decide whether AF2 is allowed to vote (Table 3). On the 10 constructs that ESMFold and Boltz-2 both call $q<0.7$, AF2 fails 7/10—it is measuring the same AFDB-referenced event. On the 231 constructs both others call $q\geq 0.7$, AF2 fails 1/231 (0.4%)—it is not a universally harsh model. Those two numbers justify reading the private sets the same way.

AF2 fails 2.3% of ESMFold's 43 private failures (mean TM 0.899) and 18.8% of Boltz-2's 16 private failures. The first rate is indistinguishable from the 0.4% consensus-OK baseline; the second is an order of magnitude above it, with a wide interval because n=16. Pairwise correlations of absolute q tell the same story: Boltz-2–AF2 r=0.83, ESMFold–AF2 r=0.40, ESMFold–Boltz-2 r=0.39. The two MSA models form a cluster; ESMFold is the outlier (Figure 3). RoseTTAFold2 on the same MSAs, no templates, fails 6/43 of the ESMFold-private set and tracks Boltz-2/AF2 (r=0.65), not ESMFold (r=0.26; Appendix S).

Deposited structures side with the MSA models. AFDB is itself an AF2 model, so the reference could be doing the work. We therefore re-score against deposited coordinates (Berman et al., 2000) wherever SIFTS (Dana et al., 2019) gives a residue-level map that survives a 95% identity check (72 domains in 33 pairs; Appendix V). On the six ESMFold-private domains with an experimental reference, ESMFold scores 0.22 against the crystal while Boltz-2 scores 0.88 and AF2 0.89 on the same residues. Boltz-2's private failures stay low against the crystal (0.62, 3/3): neither private set is a reference artefact. Dropping the reference instead—scoring each fused domain against ESMFold's own isolated fold—leaves 34/43 private calls low and the order preference at chance (Appendix W).

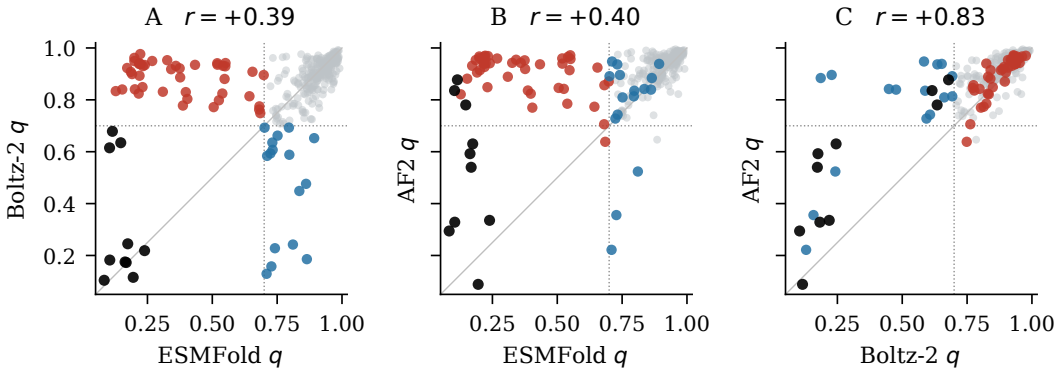

Figure 3: Pairwise absolute quality on the same 300 constructs, coloured as in Figure 2C. Boltz-2 and AF2 track each other (r=0.83); each is only moderately correlated with ESMFold. ESMFold-only failures (red) sit at high MSA-model q.

We therefore state the claim that a two-model comparison could not support: relative to two AFDB-referenced MSA models, ESMFold reports more low-q calls on artificial fusions. The magnitude remains the unbiased one (17.7% vs. 8.7% vs. 4.0%), not the biased-panel one. One construct from the private set is in Figure 6 (Appendix X): ESMFold assigns TFIIS q=0.22 while folding the CS domain (q=0.99); Boltz-2 and AF2 recover both.

The extra low-q calls are two ESMFold regimes. Of the 43 ESMFold/Boltz-2-private constructs, 26 already fail one isolated domain (one-sided; fused worse side is that domain in 24/26) and 17 lose a fold only in the chimera, with mean $\Delta\min = -0.43$ (Table 8). Isolated Boltz-2 and AF2 fold both domains in all 43 (Table 9), so already-hard is ESMFold-specific isolated difficulty, not a domain the MSA models cannot fold. Isolated difficulty, not chain

length, tracks the private set (40/43 of it sits in the hardest isolated-TM tercile; Appendix J). Pair-block drift then separates the two low-q regimes at matched q (Appendix X). A fused-score drop is a model-output change, not evidence that the protein unfolded.

6 Q3: the tested cutoffs do not resolve the disagreement

A tempting repair is that ESMFold is ranking constructs correctly and only the conventional $q<0.7$ cut is too aggressive. ESMFold's q predicts Boltz-2's binary verdict at AUC 0.86, which looks like support—until the cutoff is swept. Recall of Boltz-2's 26 failures is 38.5% at every threshold in $[0.30, 0.70]$. The 10 consensus failures all sit below q=0.30; the 16 Boltz-2-private failures all sit at ESMFold $q\geq 0.70$. There is no operating point in between (Table 10). Precision versus Boltz-2 never exceeds 36%. The high AUC is a ranking statement over the whole 300, carried by the large jointly high-q mass, not a true-error rate and not a statement that a cutoff in $[0.30, 0.70]$ recovers Boltz-2's private failures (Figure 4A).

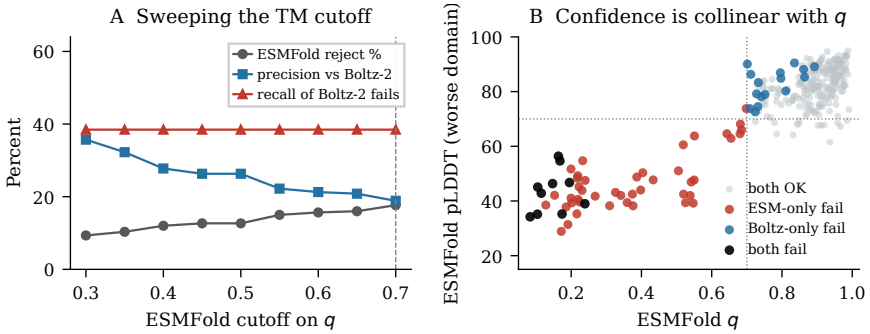

Figure 4: Two failed repairs of the ESMFold oracle. (A) Sweeping the TM cutoff; recall of Boltz-2 failures is flat over $[0.30, 0.70]$. (B) Worse-domain pLDDT against q, near-collinear with the $q<0.7$ call.

A second repair is to reject only when ESMFold is confident. Of 53 ESMFold failures, 52 have domain pLDDT<70 (median 43.9); of 247 successes, 94.3% have pLDDT≥70 (median 82.8). Confidence is real, but it is almost collinear with ESMFold's own TM call rather than an independent filter on other models' verdicts: imposing pLDDT≥70 on top of $q<0.7$ rejects one construct out of 300, and Boltz-2 folds that one. Inside the failure set it does not separate private from shared failures (AUC 0.40, CI $[0.13, 0.67]$).

A throwaway chimera is a cheap screen for the context-fail regime. Isolated q already flags already-hard domains for free. Context-fail needs a chain: fuse the candidate to one fixed dummy partner that shares no family with the panel and ask whether ESMFold still folds it (Appendix Y). On the disjoint second panel that dummy score outranks isolated q (paired $\Delta\text{AUC} = +0.068$, p=0.012). At $q_{\text{dummy}}<0.7$, every MSA-private slot it flags is a true fusion loss (20/20 across both panels, 7/7 held-out; Figure 5). That screens ESMFold, not function.

7 Implications for fusion design

Do not accept or reject a chimera with ESMFold q or pLDDT alone; joint MSA-model $q<0.7$ rejects 10 of 300 here (53 for ESMFold alone) and is not a validated optimum or a substitute for function. Enumerate both orders: an ESMFold-better orientation is worth +0.026 under ESMFold and does not transfer (Table 1). A throwaway chimera predicts fusion-context loss better than isolated q on a held-out panel (Figure 5); on structured gaps, prefer a natural junction over G4S3 as a population default (Appendix B).

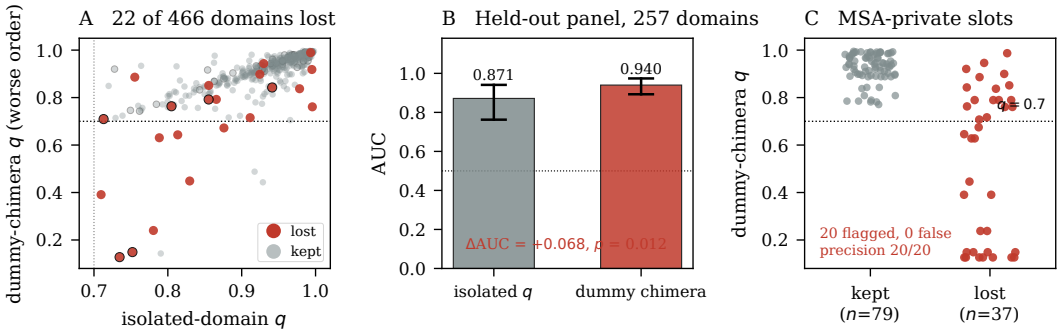

Figure 5: Throwaway-chimera screen. (A) Isolated q against worse-order dummy q for every domain ESMFold folds alone. Red: lost in a real panel construct. (B) Held-out panel: dummy score vs. isolated q, domain-clustered 95% CIs. (C) MSA-private slots. Every slot below 0.7 is a true loss (20/20).

8 Limitations

Windowed and local ColabFold re-folds of the 69-construct low-q union (43+16+10) keep the private set at high MSA-model q (Appendices F and G), but both still use ColabFold software and databases. G4S×1 leaves 41/43 ESMFold-private low-q calls intact (Appendix I). Source-protein bootstrap (105 components) keeps the excess above zero ([+1.6, +16.6] pp). OmegaFold's overlap holds at its default 10 recycles (38/43) and it is harsher on the jointly high-q set (Appendix H). The Boltz-2-private set is small (n=16; n=10 on the second panel, Appendix D). AFDB is an AF2-predicted reference; deposited coordinates reproduce the split on the 33 pairs that have them (Appendix V), and most pairs have none.

What would move a claim about proteins. (i) A dual-function assay on the 43 ESMFold-private constructs with per-domain controls (inactivity is not automatically an MSA miss). (ii) Native versus G4S3 activity on the structured natural subset.

9 Conclusion

On chimeric fusions that nature did not assemble, ESMFold is a reproducible oracle that two MSA-based predictors do not corroborate: it has an order preference they do not share, and it assigns low q to constructs that Boltz-2, AF2 and—on the six private domains with a crystal—the deposited structure all score highly. The tested cutoffs do not resolve that disagreement, and a population nat−G4S3 advantage does not license per-pair ESMFold ranking. Do not accept or reject a chimera with ESMFold q alone; joint MSA-model q<0.7 is a stricter rule on this sample, not a validated optimum or a substitute for function.

AI use statement

In this work, we used a coding assistant (Cursor; Grok and Claude models) to implement the analysis and folding scripts, to design follow-up controls (recycling depth, linker length, pair-representation probes, and the throwaway-chimera screen), to interpret the resulting score tables, and to draft the manuscript. We did not use generative AI to invent amino-acid sequences, to choose the reported statistics, or to write proofs. The scientific question and the primary q<0.7 rule were fixed before those controls. Every number in the text was checked against the script outputs. We take responsibility for the final content, including text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

This work uses only public protein sequences (UniProt Swiss-Prot) and public predicted structures (AlphaFold Database). No human subjects, personal data, or pathogen-enhancement experiments are involved. The chimeric benchmark contains assembled sequences that do not occur in nature; the supplementary archive releases them as amino-acid strings and domain coordinates, not as biological materials. The paper argues against treating a single computational oracle as a substitute for experiment, which reduces rather than increases the risk of acting on a false structural claim.

Reproducibility statement

The anonymous supplementary archive contains experimental results only: the chimeric pair list (amino-acid strings and domain spans), per-construct scores for the two unbiased panels, the three-predictor summaries, the throwaway-chimera screen, the recurrence counts, and the crystal-referenced domain scores. Analysis scripts are not in that archive. Weights are the public ESMFold v1 checkpoint, Boltz-2 via its official cache, and AF2 params_model_1_ptm; they are not redistributed. Random seeds for pair sampling (20260908, 20260916) and Boltz-2 (101, 202) are fixed. Each control's protocol is in the appendix section that reports it.

References

Gustaf Ahdritz, Nazim Bouatta, Christina Floristean, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, et al. OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. Nature Methods, 21:1514–1524, 2024.

Mehmet Akdel, Douglas E V Pires, Eduard P Pardo, Jürgen Jänes, Arthur O Zalevsky, Bálint Mészáros, Patrick Bryant, Lydia L Good, Roman A Laskowski, Gabriele Pozzati, et al. A structural biology community assessment of AlphaFold2 applications. Nature Structural & Molecular Biology, 29(11):1056–1067, 2022.

Minkyung Baek, Ivan Anishchenko, Ian R Humphreys, Qian Cong, David Baker, and Frank DiMaio. Efficient and accurate prediction of protein structure using RoseTTAFold2. bioRxiv, pp. 2023.05.24.542179, 2023.

Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The Protein Data Bank. Nucleic Acids Research, 28(1):235–242, 2000.

Patrick Bryant, Gabriele Pozzati, and Arne Elofsson. Improved prediction of protein-protein interactions using AlphaFold2. Nature Communications, 13(1):1265, 2022.

Xiaoying Chen, Jennica L Zaro, and Wei-Chiang Shen. Fusion protein linkers: property, design and functionality. Advanced Drug Delivery Reviews, 65(10):1357–1369, 2013.

Ratul Chowdhury, Nazim Bouatta, Surojit Biswas, Christina Floristean, Anant Kharkar, Koushik Roy, Charlotte Rochereau, Gustaf Ahdritz, Jiarui Zhang, George M Church, et al. Single-sequence protein structure prediction using a language model and deep learning. Nature Biotechnology, 40(11):1617–1623, 2022.

Jose M Dana, Aleksandras Gutmanas, Nidhi Tyagi, Guoying Qi, Claire O'Donovan, Maria Martin, and Sameer Velankar. SIFTS: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. Nucleic Acids Research, 47(D1):D482–D489, 2019.

Dominic Esposito and Deb K Chatterjee. Fusion tags for protein solubility and expression. Current Opinion in Biotechnology, 17(4):353–358, 2006.

Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Timothy Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with AlphaFold-Multimer. bioRxiv, 2022.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873):583–589, 2021.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. Acta Crystallographica Section A, 32(5):922–923, 1976.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637):1123–1130, 2023.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. Nature Biotechnology, 41(8):1099–1106, 2023.

Giosue Migliorini, Aristofanis Rontogiannis, Grigori Guitchounts, Nicholas Franklin, Axel Elaldi, and Olivia Viessmann. PairSAE: Mechanistic interpretability from pair representations in protein co-folding. arXiv preprint arXiv:2606.27440, 2026.

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. ColabFold: making protein folding accessible to all. Nature Methods, 19(6):679–682, 2022.

Erik Nijkamp, Jeffrey A Ruffolo, Eli N Weinstein, Nikhil Naik, and Ali Madani. ProGen2: exploring the boundaries of protein language models. Cell Systems, 14(11):968–978, 2023.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Seth Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. bioRxiv, 2025. doi: 10.1101/2025.06.14.659707.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology, 35(11):1026–1028, 2017.

The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research, 51(D1):D523–D531, 2023.

Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. Highly accurate protein structure prediction for the human proteome. Nature, 596(7873):590–596, 2021.

Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research, 50(D1):D439–D444, 2022.

Sanketh Vedula, Alex Bronstein, and Ailie Marx. Improving prediction accuracy in chimeric proteins with windowed multiple sequence alignment. Computational and Structural Biotechnology Journal, 27, 2025. doi: 10.1016/j.csbj.2025.07.039.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, et al. Boltz-1: Democratizing biomolecular interaction modeling. bioRxiv, 2024. doi: 10.1101/2024.11.19.624167.

Jinrui Xu and Yang Zhang. How significant is a protein structure similarity with TM-score=0.5? Bioinformatics, 26(7):889–895, 2010.

Chengxin Zhang, Morgan Shine, Anna Marie Pyle, and Yang Zhang. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. Nature Methods, 19(9):1109–1115, 2022.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. Proteins, 57(4):702–710, 2004.

A Reference-frame trap

On a 300-pair natural two-domain pilot, scoring each domain against a single-domain ESMFold re-prediction reverses the ranking of conditions: the native junction looks worse than a G4S3 replacement (paired $\Delta q = -0.0058$, native better on only 39.7% of pairs). The native linker induces real inter-domain packing, which must move each domain away from its isolated conformation; G4S3 leaves the domains more independent and therefore closer to the isolated prediction. All numbers in the main text use the AFDB native conformation as $X^\star$.

B Natural-pair pilot and structured-linker MSA re-check

The project began as a fusion-design study. Three hundred natural two-domain proteins were folded in four conditions (native linker, G4S3, composition-shuffled native linker, domain-order swap). Native beat G4S3 by +0.0097 mean TM ($p = 4\times10^{-15}$, Cohen $d_z = 0.25$) and beat the swap by +0.028. Failure ($q<0.7$) was 10.7% / 11.7% / 11.0% / 19.0% for native / G4S3 / shuffle / swap. The swap effect did not correlate with Cα–Cα distance (Spearman −0.01). A structured-spacer subset ($n=110$) showed a larger native-vs-G4S3 gap (+0.0153 mean of the two domain TMs) than truly flexible linkers (+0.0064, $n=140$). Those measurements motivated the chimera study and the order hypothesis; the swap and failure-rate numbers are not used as evidence for Sections 4–6, because they cannot exclude ESMFold memorization of the assembled natural chains.

The structured contrast is the exception we re-measured. The same 110 pairs were folded as native versus G4S3 with Boltz-2+MSA and with AF2 on those MSAs, scored against the same AFDB natives. One native construct (s045_nat, UniProt Q0TV31) returned no MSA, leaving a paired $n=109$ (Table 4). All three 95% CIs exclude zero, and both MSA deltas are larger than ESMFold's. The population default—prefer a real junction over G4S3 on structured gaps—is therefore not an ESMFold artefact. Pair-level deltas still disagree (ESMFold–Boltz-2 $r=0.025$, ESMFold–AF2 $r=0.259$, Boltz-2–AF2 $r=0.204$ on mean TM; −0.048, +0.213, +0.168 on worst-domain TM, common $n=109$), so ESMFold cannot rank which natural linker to keep. The chains are natural two-domain proteins, so the result supports that default, not a memorization-proof per-pair design rule.

Table 4: Paired native-minus-G4S3 TM on structured-gap natural pairs. ESMFold uses all 110 pairs; Boltz-2 and AF2 use the 109 pairs for which an MSA was returned (s045_nat had none). mean is $(\mathrm{TM}_A + \mathrm{TM}_B)/2$, the metric behind the original +0.0153; worst is $\min(\mathrm{TM}_A, \mathrm{TM}_B)$, the score used on chimeras. Intervals are pair-level bootstrap 95% CIs.

	ESMFold	Boltz-2+MSA	AF2
Δ mean TM	+0.0153	+0.0231	+0.0303
95% CI	[+0.0094, +0.0220]	[+0.0095, +0.0389]	[+0.0194, +0.0436]
Native better	78.2%	67.0%	78.0%
Δ worst-domain TM	+0.0229	+0.0360	+0.0467

C Selection bias on the first Boltz-2 panel

The first MSA-enabled Boltz-2 panel was designed as 50 largest-gap ESMFold-sensitive pairs plus 30 low-gap controls (80 pairs, 160 constructs). Sensitivity is $|q_{AB} - q_{BA}| > 0.05$ under ESMFold, not a prior q<0.7 call. MSA Boltz-2 completed 60 pairs / 120 constructs (40/50 large-gap and 20/30 controls; names p000–p039 and p050–p069). The remaining 20 pairs (10+10) were not submitted to the MSA run. On those 120 constructs the low-q rates were ESMFold 42.5% vs. Boltz-2 17.5% (excess +25 pp, CI $[+15, +35]$). The 50 largest-gap pairs all have one failing order, so the rate is partly the selection rule. The panel was the right design for asking whether order gaps replicate, and the qualitative conclusion of Section 4 replicates on the unbiased sample. It was the wrong design for a population rate. We re-drew 150 pairs uniformly and treat only that re-draw as a prevalence estimate.

D A second uniform panel

The 150-pair comparison in Section 5 is one uniform draw (seed 20260908). A second draw of 150 pairs (seed 20260916) is taken from the 850 pairs that draw did not take, so the two panels are disjoint and their union is still a uniform sample of the 1,000-pair population. Construct names are prefixed x so nothing collides. MSAs, Boltz-2 (seeds 101 and 202) and AF2 model_1_ptm follow the same protocol as the first panel.

Table 5: Absolute fold quality on the second 150-pair panel (300 constructs), same protocol as Table 2. Intervals are pair-clustered bootstrap 95% CIs.

	ESMFold	Boltz-2+MSA	AF2+MSA
Mean worse-domain TM	0.782	0.841	0.889
Median	0.898	0.905	0.930
Failure rate ($q < 0.7$)	20.7%	11.3%	7.0%
95% CI	[14.7, 26.7]	[6.7, 16.7]	[3.3, 11.3]
Catastrophic ($q < 0.5$)	15.7%	9.0%	4.3%
Pearson r vs. ESMFold	—	+0.57	+0.62
Pearson r vs. Boltz-2	+0.57	—	+0.80

ESMFold's failure rate exceeds Boltz-2's by +9.3 percentage points, CI $[+4.0, +14.7]$, which again excludes zero (Table 5). The first panel's excess was +9.0, CI $[+2.3, +15.7]$; the second interval sits inside the first. Order-sensitive pairs ($|q_{AB} - q_{BA}| > 0.05$ under ESMFold) are 28.0% of the second draw against 26.5% of the population and 25.3% of the first, so this sample is also not a tail.

Table 6: AF2 failure rate inside each ESMFold/Boltz-2 class on the second panel, same protocol as Table 3. Zero-event cells are counts: a bootstrap over zeros has no width.

Class	n	ESM TM	Boltz TM	AF2 fail
ESMFold-only failure	38	0.359	0.877	2.6% [0.0, 8.6]
Boltz-2-only failure	10	0.835	0.445	0/10
Consensus failure	24	0.258	0.318	83.3% [61.9, 100]
Consensus OK	228	0.906	0.908	0/228

AF2 is again allowed to vote, and again sides with the MSA cluster (Table 6). It fails 20/24 of the constructs both others call low-q and 0/228 of the constructs both others call high-q. It fails 1/38 of the ESMFold-private set (mean TM 0.895), indistinguishable from the consensus-OK baseline, and 0/10 of the Boltz-2-private set. The first panel's n=16 Boltz-2-private cell was the one AF2 did not cleanly refuse; the second draw does not enlarge that cell and does not have AF2 backing it.

The ESMFold-private set again splits by isolated fold: 20/38 already-hard and 18/38 context-fail, against 26/43 and 17/43 on the first panel. Pre-specified terciles on this panel

put 30/38 of the private calls in the hardest isolated-TM third. The shared-failure count is higher here (24 vs. 10), and absolute q correlates more tightly between ESMFold and Boltz-2 (r=0.57 vs. 0.39). Those are sampling differences, not a change of claim: the excess, the AF2 arbitration of the ESMFold-private set, and the already-hard / context-fail split all come back on a disjoint draw.

Pair-block drift on this panel's 18 context-fail constructs repeats the first-panel contrast (Appendix X).

E Design choices do not transfer across predictors

Table 1 is the selector×evaluator matrix on the unbiased 150 pairs. The baseline is a random order (the mean of the two q values); the ceiling is the evaluator's own oracle. No evaluator is treated as truth. Pair-clustered bootstrap 95% CIs. The +0.026 ESMFold self-gain is that 150-pair number, not the 1,000-pair +0.032 versus A-first.

ESMFold's self-gain is real; its transfer shows no clear positive effect (sign agreement with Boltz-2 48%, with AF2 53%; both CIs include 0). The reverse transfers are also consistent with no clear positive effect. Boltz-2 and AF2 agree with each other only modestly (sign 57%, gap r=0.19). On the 38 pairs where $|\text{gap}_{\text{ESM}}|>0.05$, ESMFold's pick is worth +0.082 under ESMFold, +0.001 under Boltz-2, and −0.012 under AF2.

Reject-set Jaccard on the 300 constructs is 0.145 (ESMFold–Boltz-2), 0.140 (ESMFold–AF2), and 0.357 (Boltz-2–AF2). Of the constructs ESMFold keeps, Boltz-2 keeps 93.5% and AF2 keeps 98.4%; instability is in what gets discarded, not in what gets kept. The structured-linker pair-level result is the same pattern for a different design knob: population nat−G4S3 replicates, per-pair ranking does not (Appendix B).

Table 7: Three-predictor 2×2×2 counts on the 300 constructs. Cells are (ESMFold q<0.7, Boltz-2 q<0.7, AF2 q<0.7). ESMFold-private (43=42+1) and consensus-ok (231=230+1) names are relative to the ESMFold/Boltz-2 split. The 10 ESMFold/Boltz-2 joint low-q constructs are not the same 10 as the Boltz-2/AF2 joint low-q set (3+7); all three are low-q on 7 constructs.

ESMFold / Boltz-2	AF2 high q	AF2 low q	Row	Name
high / high	230	1	231	consensus-ok
high / low	13	3	16	Boltz-2-private
low / high	42	1	43	ESMFold-private
low / low	3	7	10	consensus-fail

F Windowed MSA does not collapse AF2's private-set calls

Vedula et al. restore peptide coverage by concatenating two domain-restricted alignments. We approximate that construction without a new search: each existing ColabFold homolog is kept on the domain it covers ($\geq$ 30% of that domain's query columns) and gapped elsewhere, including the linker. All 69 constructs in the ESMFold/Boltz-2 low-q union (43+16+10) have usable depth (mean 4.5k A-hits and 4.6k B-hits; minima 51 and 154). AF2 and Boltz-2 are then re-run on those block-diagonal CSVs and scored with the same worse-domain TM against AFDB.

Windowed AF2 q tracks the original joint-MSA q at Pearson r=0.967 (mean $\Delta = -0.0015$). Of the 43 ESMFold-private failures, 41 stay at $q\geq$0.7 (one borderline collapse, 0.77→0.68; the one joint-MSA AF2 failure stays failed). Boltz-2-private and consensus-fail AF2 rates are unchanged (3/16 and 7/10). Windowed Boltz-2 is at least as clear: all 43 ESMFold-private constructs stay folded (r=0.936 vs. joint-MSA Boltz-2), and six of the 16 Boltz-2-private failures are rescued rather than deepened. Neither MSA model's refusal to reproduce ESMFold's extra failures is an artefact of letting each homolog see both domains at once.

The check still uses ColabFold homologs and an AFDB reference; it does not replace the search. A local re-search is Appendix G.

G A local ColabFold re-search does not collapse the private set

The windowed check re-uses server homologs. We next search the same 69-construct low-q union locally with colabfold_search 1.6.2 against the official uniref30_2302 and colabfold_envdb_202108 databases (–prefilter-mode 1, no templates, no api.colabfold.com). All 69 CSVs have usable depth (min 1,233, median 8,564, mean 9,177). Boltz-2 and AF2 are re-run on those alignments and scored with the same worse-domain TM against AFDB.

Local Boltz-2 q tracks the server-MSA q at Pearson r=0.924 (mean $\Delta = +0.009$). Of the 43 ESMFold-private failures, 42 stay at $q \geq 0.7$ (one collapse, r062_ba 0.82→0.64). Six server-MSA Boltz-2 failures are rescued rather than deepened. Local AF2 is as stable: r=0.937 (mean $\Delta = +0.005$), fail rate on the 69-construct union unchanged at 15.9% (11/69), and all 43 ESMFold-private constructs stay folded (the one server-MSA AF2 failure on that set is rescued). The private set is therefore not a cached api.colabfold.com alignment. The search is still ColabFold software and ColabFold databases; it does not replace a JackHMMER or HHblits pipeline, and it remains AFDB-referenced.

H OmegaFold shares ESMFold's private set and is harsher

OmegaFold model 1 (official weights, 4 recycles, no MSA) was run on the same 300 constructs and scored with the same worse-domain TM against AFDB. Failure is 36.7% (pair-clustered 95% CI [29.3, 43.7]), mean q=0.696. That is harsher than ESMFold (17.7%), Boltz-2 (8.7%), and AF2 (4.0%). Pearson r is 0.62 vs. ESMFold and only 0.34/0.33 vs. the two MSA models.

Of the 43 ESMFold-private failures, OmegaFold fails 39 (mean q=0.419, vs. ESMFold 0.378 on that set). It also fails all 10 ESMFold/Boltz-2 joint low-q constructs, 10/16 Boltz-2-private constructs, and 51/231 (22%) of the ESMFold/Boltz-2 jointly high-q set (that 231 includes one AF2 low-q call; 230 constructs are high-q under all three). The ESMFold/Boltz-2-private set is therefore not an ESMFold idiosyncrasy: a second language-model folder rejects almost the same constructs. The decision boundary is still not shared—OmegaFold assigns low q to many constructs the other three call intact—and the comparison remains AFDB-referenced.

The package default is 10 recycles. Re-running all 300 constructs at 10 recycles leaves the claim intact: fail 35.7% (CI [28.3, 43.0]), mean q=0.710, Pearson r=0.63 vs. ESMFold. The ESMFold/Boltz-2-private overlap is 38/43 (one rescue, r145_ab 0.67→0.86); jointly high-q fails fall to 48/231. We keep the 4-recycle numbers as the originally specified run and treat 10 recycles as the package-default sensitivity; 4 vs. 10 does not move the qualitative comparison. OmegaFold's extra rejects relative to the ESMFold/Boltz-2 split are not a biological private set.

I Private failures survive a shorter G4S linker

The chimeric panel uses one linker (GGGGS×3). We rebuilt the 43 ESMFold-private failures with GGGGS×1, keeping both domains and the original order, and re-scored all three predictors against the same AFDB references. ESMFold still fails 41/43 (mean q 0.378→0.363, r=0.88). Boltz-2 still folds every construct (mean 0.883→0.879). AF2's fail rate stays 2.3% (one rescue, one new collapse). The private set is not a G4S3-length artefact.

Pair-clustered intervals still treat two chimeras that share a UniProt as independent. The 150 pairs form 105 connected components on source proteins (31 proteins appear in more than one pair). Re-bootstrapping at that level leaves the excess-failure interval excluding zero ([+1.6%, +16.6%] vs. pair-clustered [+2.3%, +15.7%]).

Table 8: ESMFold isolated-domain TM versus fused q on the same 300 constructs. Already-hard: at least one isolated domain $q<0.7$. Context-fail: both isolated $q\geq 0.7$ but fused min $q<0.7$. Names follow the ESMFold/Boltz-2 split (Table 7). A drop is a change in ESMFold's output, not a demonstrated unfolding.

Class	n	Both iso ≥ 0.7	Already-hard	Context-fail
ESMFold-private	43	17	26	17
Boltz-2-private	16	16	0	0
Consensus-fail	10	2	8	2
Consensus-ok	231	231	0	0

J ESMFold private low-q is partly isolated-domain difficulty

ESMFold already folded every constituent domain in isolation during the 1,000-pair run. On the unbiased 300 constructs, isolated-domain TM against the same AFDB reference averages 0.904 (6.0% below 0.7). Fused-domain TM averages 0.876 (mean $\Delta = -0.028$). Both isolated domains are at $q\geq 0.7$ for 266/300 constructs; 19 of those 266 then fall below 0.7 after fusion.

Split by the ESMFold/Boltz-2 classes: of 43 ESMFold-private low-q constructs, 26 already have at least one isolated domain below 0.7 (already hard), and the other 17 lose the fold only in the fusion (mean $\Delta\min = -0.43$ on that subset). Of 10 joint low-q constructs, 8 are already hard. All 231 jointly high-q constructs have both isolated domains above 0.7 (mean $\Delta\min = -0.011$). So a majority of ESMFold's extra low-q calls are domains ESMFold already scores poorly alone; a minority are fusion-context losses. Pre-specified terciles on the 300 constructs put 40/43 ESMFold-private calls in the hardest isolated-TM third (isolated $\min q\leq 0.86$), against 2 and 1 in the other thirds. Longer domains and larger length gaps do not enrich the private set (ESMFold-private rates 18%/15%/10% by max-domain-length tercile; 21% on length-balanced pairs vs. 12% on the most asymmetric third).

The two subsets are different model behaviors. On the 26 already-hard constructs the isolated failure is always one-sided, and the fused worse domain is that side in 24/26 (mean isolated TM of the worse side 0.45). Boltz-2 and AF2 still score those chimeras at mean q 0.87 and 0.89; OmegaFold fails 25/26. On the 17 context-loss constructs both isolated TMs are high (means 0.87/0.87), yet the fused worse domain is far from ESMFold's own isolated structure (mean TM-vs-iso 0.39, 15/17 below 0.7). Five pairs are context-loss in both orders, and the same source domain is the worse side in all five; seven constructs are context-loss in only one order (the Figure 6 TFIIS example is one of them). Order-sensitive pairs ($|\text{gap}|\geq 0.05$) are enriched relative to consensus-ok (52% of ESMFold-private pairs vs. 21%), but N- versus C-terminal placement of the worse domain is 53%, and linker pLDDT and interface contact counts overlap the easy set. OmegaFold fails 14/17 at 4 recycles and 13/17 at 10. The split is therefore: (i) a domain the single-sequence folder already cannot recover; (ii) a domain it recovers alone and then destroys in one or both chimeric orders, while the MSA models do not.

Isolated Boltz-2 and AF2 folds of all 266 unique domains close (i) as ESMFold-specific (Table 9). The 26 already-hard constructs use 13 unique isolated-fail domains (always one-sided; both orders of 13 pairs). Those 13 have mean ESMFold isolated TM 0.44 and mean Boltz-2/AF2 isolated TM 0.91/0.91 (min 0.80/0.75; 0/13 below 0.7). On the 43 ESMFold-private constructs, Boltz-2 already-hard is 0 and AF2 already-hard is 0. The same isolated run does not make Boltz-2's 16 private failures ESMFold-like: 10 of 16 already fail a Boltz-2 isolated domain. So regime (i) is "the language-model folder misses a domain the MSA folders recover alone," not "that domain is unfoldable." Regime (ii) remains a fusion-context loss under ESMFold/OmegaFold only.

The same 300 ColabFold MSAs do not explain the split by depth or A/B asymmetry (Appendix K). Residue pLDDT, Boltz-2 PAE, and chain geometry likewise fail to mark a linker-local collapse unique to the 17 (Appendices L and M). N- or C-terminal G4S tails of

Table 9: Isolated-domain already-hard counts on the same 300 constructs after folding all 266 unique domains with Boltz-2 and AF2. A construct is already-hard if that predictor's isolated TM is below 0.7 on at least one domain.

Class	n	ESMFold	Boltz-2	AF2
ESMFold-private	43	26	0	0
Boltz-2-private	16	0	10	2
Consensus-fail	10	8	7	6
Consensus-ok	231	0	5	0

linker length do not rescue that set (Appendix N). Already-hard worse-side families rarely appear as the worse side of a consensus-ok construct (1/9); context-loss families usually do (7/10), including TFIIS (Appendix O). Trimming or expanding three linker-facing residues does not rescue the 17 (Appendix P). Exploratory enzyme / antibody-like tags do not enrich the private rate (Appendix Q). Mean-pooled ESMFold language-model embeddings of the worse domain do not track the TM-vs-iso collapse (Appendix R). RoseTTAFold2 on the same MSAs fails 6/43 of the private set (Appendix S).

K Joint MSA depth does not separate ESMFold-private low-q

Each Boltz-2 CSV is scored per domain: a homolog is an A-hit or B-hit if it covers $\geq$30% of that domain's query columns. Occupancy is the fraction of homolog residues present on those columns. All 300 constructs have usable depth (minimum 35 hits on the shallower domain; median min-hits 1,365 on consensus-ok and 1,199 on ESMFold-private). ESMFold-private constructs are slightly shallower than consensus-ok (mean min-hits 1,595 vs. 1,993) and no more asymmetric (mean $|\log((n_A{+}1)/(n_B{+}1))|$ 1.63 vs. 1.61). Pearson of min-hits or of that ratio against $|q_{\mathrm{ESM}} - q_{\mathrm{Boltz}}|$ is $\approx$0 (-0.00 and -0.06); against an ESMFold-private indicator, -0.08 and $+0.01$. Linker occupancy is 0.24 vs. 0.22. The already-hard and context-loss subsets are likewise not separated by hit asymmetry. MSA models can still be using homology that ESMFold lacks; the negative result is only that among these ColabFold alignments, depth and A/B imbalance do not mark the constructs ESMFold rejects and Boltz-2/AF2 keep. Windowed and local re-folds (Appendices F and G) already showed that rearranging or re-searching the same databases does not collapse that set.

L Residue confidence and PAE do not localize a junction collapse

AF2, Boltz-2, and OmegaFold already wrote per-residue pLDDT on the 300 constructs; Boltz-2 also wrote PAE. The eight domain residues adjacent to the linker are lower-confidence than the domain core in every class (AF2 edge$-$core $-17/-10/-12$ on already-hard / context-fail / consensus-ok). That is a terminus effect, not a private-set marker. Worse-domain AF2 pLDDT stays high on both ESMFold-private cuts (82 and 78) and on consensus-ok (82); linker pLDDT is $\sim$38 in all three. Boltz-2 intra-domain PAE on the ESMFold-worse side is 6.3 / 5.4 / 5.3 Å (already-hard / context-fail / consensus-ok). Inter-domain PAE is $\sim$22 Å everywhere, as expected for a G4S3 chimera. By contrast, Boltz-2-private and consensus-fail constructs do look like domain-level MSA uncertainty: worse-domain Boltz pLDDT 56 and 54, intra-PAE 12.5 and 12.8. OmegaFold's worse-domain pLDDT is 50/52 on the two ESMFold-private cuts vs. 74 on consensus-ok, i.e. a whole-domain language-model drop, not a linker-neighborhood drop. Re-folded ESMFold maps on the padding panel agree: context-fail worse-domain pLDDT is 47 in both the eight linker-adjacent residues and the core (edge$-$core $\approx$0), against 84 core pLDDT on length-matched consensus-ok controls.

M No steric clash marks the private set, but ESMFold's own packing does

Inter-domain CA pairs closer than 2.5 Å are absent from all 43 ESMFold-private constructs under AF2, Boltz-2, and OmegaFold (two consensus-ok AF2 models have a single such pair). There are no consecutive-CA chain breaks. Linker end-to-end CA distances overlap (19–22 Å on ESMFold-private vs. 21 Å AF2 consensus-ok). Under those three predictors, interface contact counts do not separate context-fail from consensus-ok: median inter-domain pairs within 8 Å are 5 vs. 6 (AF2), 24 vs. 28 (Boltz-2) and 1 vs. 3 (OmegaFold), all null and all in the wrong direction.

ESMFold's own coordinates, available once the panel was re-folded for Appendix U, do separate them: 7 inter-domain pairs within 8 Å on context-fail against 2 on consensus-ok (permutation on medians, p=0.031, uncorrected across the 16 contrasts computed here), while already-hard is flat (3 vs. 2, p=0.51). So on the chains ESMFold loses only in the fusion, and only under ESMFold, the two domains sit closer together. Appendix X asks whether that packing is the cause; it is not. These are assembled-chain diagnostics, not a fused-chain AFDB ground truth.

N Terminal GS tails do not create the context-loss set

A 15-residue G4S tail (same length as the panel linker) was added to either terminus of the 17 context-fail constructs and 17 length-matched consensus-ok controls, leaving domain order unchanged. Already-hard constructs were re-folded without tails so ESMFold per-residue maps exist for all 43 ESMFold-private chains. On context-fail, N-pad and C-pad change mean q by -0.054 and -0.049 (median -0.002 and -0.015); 2/17 constructs change fail status in each direction. Controls move $|\Delta|$=0.015 with at most one fail flip. Pearson vs. the unpadded re-fold is 0.74/0.82 on context-fail and 0.93/0.94 on controls. A class-wide rescue under padding would have been positional sensitivity; the observed median is near zero, so the 17 are not an N/C-tail artefact. Unpadded re-folds reproduce the original q (context-fail mean 0.364).

O Already-hard families rarely succeed elsewhere

Counts are exploratory; no pairwise tests. The 26 already-hard constructs put the worse domain in 9 families and 13 source proteins. Ten of 26 are Phosphagen kinase N-terminal. Only 1/9 of those families is also the worse side of any consensus-ok construct. The 17 context-loss constructs put the worse domain in 10 families and 11 proteins; 7/10 of those families (including TFIIS central) are the worse side of at least one consensus-ok construct. So already-hard is concentrated in families ESMFold already fails alone; context-loss hits families that can succeed in another chimeric context. An EC number on either source protein in the natural-pair pool is present for 69% / 35% / 56% of already-hard / context-fail / consensus-ok constructs. That is not a balanced enzyme panel and is not a class-wise claim.

P Linker-facing boundary edits do not rescue context-loss

The 17 context-fail constructs and the same 17 length-matched controls were re-built with three residues removed from each linker-facing domain end, and separately with up to three native UniProt flanks added at those ends (AFDB sequence). Domain order is unchanged; q is scored on the remaining (trim) or original (expand) annotated span. Trim: context-fail mean $\Delta q = -0.048$ (median -0.024, 0/17 fail flips, r=0.90 vs. the unpadded re-fold); controls -0.012 (median -0.004, 0 flips). Expand: context-fail -0.024 (median -0.005, 2/17 flips, r=0.85); controls -0.001 (0 flips, r=0.99). A class-wide rescue would have implicated the InterPro cut. The 17 survive both edits.

Q Enzyme / antibody-like labels do not enrich the private set

Exploratory multi-label tags on the 300 constructs, assigned from InterPro family names plus source-protein EC numbers and a short list of annotated GO IDs in the natural-pair pool (not a GO closure). No pairwise tests. Enzyme-like n=184 (123 families, 161 proteins): ESMFold-private rate 15.2% (28/184), already-hard 22, context-fail 6, mean isolated min q=0.83. Other n=106 (83 families): private rate 14.2% (15/106), already-hard 4, context-fail 11. Antibody-like n=8 (7 families; Ig-like C2-type on 4 constructs): private 0/8. Signaling-like n=20: private 0/20. The panel background is 14.3% (43/300). Already-hard is mostly enzyme-like because Phosphagen kinase N-terminal sits in that tag (22/26); context-fail sits mostly in other (11/17). Antibody-like and signaling-like cells are too small to support a class claim. The private rate on enzyme-like constructs matches the panel, so the disagreement is not an enzyme-fusion artefact on this draw.

R Language-model embeddings stay put when the fold drops

Pre-specified tests on the ESMFold language-model trunk (no structure module, no attention maps). Each domain is mean-pooled over residues, once in isolation and once inside the chimera. H1: context-fail worse-domain isolated-vs-fused cosine is lower than on consensus-ok. H2: already-hard isolated worse-domain embeddings sit farther from the consensus-ok isolated centroid than context-fail embeddings do.

H1 holds on the mean (0.992 vs. 0.995) because one construct (r071_ba, cosine 0.934, TM-vs-iso 0.24) pulls the average; the other 16 have median cosine 0.996, matching consensus-ok (median 0.997). Pearson of cosine vs. TM-vs-iso on the 17 is 0.17. H2 holds by 0.002 cosine (0.995 vs. 0.997); already-hard isolated embeddings are not outliers (all $\geq$0.99 to the centroid). The language-model trunk therefore does not register the context loss that the folding head produces. That is a representation negative, not a claim about attention or a cellular mechanism, and it is specific to this stage: the folding trunk's pair stack does register the loss (Appendix X), so the two results locate the change after the sequence embedding rather than in it. Mean-pooling is also a coarse probe; a dictionary-learning analysis of the pair stack, as PairSAE (Migliorini et al., 2026) does for Boltz-2, would test the same question at feature resolution.

S RoseTTAFold2 joins the MSA cluster on the same alignments

RoseTTAFold2 (Baek et al., 2023) (RF2_apr23, 3 recycles, C_1, no templates, no HHblits) was run on the same 300 ColabFold MSAs used by Boltz-2 and AF2, converted from those CSVs to a3m. This asks whether a third MSA-using architecture joins the MSA cluster or ESMFold's private set. It does not test an independent search (still ColabFold software and databases).

Failure is 14.0% (pair-clustered 95% CI [8.7, 19.3]), mean q=0.829. That sits between ESMFold (17.7%) and Boltz-2 (8.7%), and above AF2 (4.0%). Pearson r is 0.65 vs. Boltz-2, 0.65 vs. AF2, and 0.26 vs. ESMFold.

Of the 43 ESMFold-private constructs, RF2 fails 6 (mean q=0.839 vs. ESMFold 0.378 on that set): 4/26 already-hard and 2/17 context-fail. The Figure 6 construct r132_ab is q=0.819. RF2 fails 8/16 Boltz-2-private constructs, 8/10 consensus-fail constructs, and 20/231 (8.7%) of the jointly high-q set. All 12 AF2 low-q calls on this panel are also RF2 low-q; RF2 adds 30 further rejects. ESMFold-low and RF2-low overlap on only 14 constructs; RF2 leaves 39 ESMFold-low calls folded. The extra RF2 rejects relative to Boltz-2/AF2 are not treated as a biological private set. A single-sequence RF2 run was not done: it would not answer the MSA-cluster question.

T The cutoff sweep in numbers

Table 10 is the sweep behind Figure 4A.

Table 10: Sweeping ESMFold's TM cutoff against Boltz-2's q<0.7 verdict on the same 300 constructs. Recall is constant: every shared failure lies below 0.30 and every Boltz-2-private failure at or above 0.70.

ESMFold cutoff	Rejects	Precision vs. Boltz-2	Recall of Boltz-2 fails
$q < 0.30$	9.3%	35.7%	38.5%
$q < 0.50$	12.7%	26.3%	38.5%
$q < 0.70$	17.7%	18.9%	38.5%

U The scorer agrees with the reference implementation

Our TM optimises the superposition by iterated Kabsch fits on the currently well-fitting subset, which is a lower bound on the fragment-seeded search in the reference program. To bound that gap we re-folded the 300-construct panel with ESMFold so that its coordinates are on disk (the re-fold reproduces every published q to four decimals and the same 53/300 low-q calls), and rescored all 900 domains with USalign's TMscore (Zhang et al., 2022) in sequence-dependent mode on the recorded spans.

Table 11: Our iterative-Kabsch TM versus USalign TMscore on the same coordinates and spans. Δ is reference minus ours on the construct score q. Boltz-2 here is the seed-101 run.

Model	r (q)	mean Δ	max $\lvert\Delta\rvert$	low-q ours	low-q reference
ESMFold	0.994	+0.011	0.222	53/300	51/300
Boltz-2	0.992	+0.006	0.222	27/300	27/300
AF2	0.996	+0.003	0.190	12/300	12/300

The reference implementation is uniformly the more generous of the two, as expected from the wider seed search, and the bias is small (Table 11). No Boltz-2 or AF2 verdict moves. Two ESMFold calls cross 0.7 (r023_ba, r061_ba); both are context-loss constructs, so under the reference scorer the private set is 41 rather than 43 and the two regimes read 26 already-hard versus 15 context-loss. The ESMFold-minus-Boltz-2 excess is 8.0 pp under the reference scorer against 8.7 pp here. No claim in the paper depends on that difference.

V Deposited coordinates as the reference

The main-text reference is the AFDB conformation, which is an AF2 prediction; that makes AF2 a non-independent third voter and leaves the possibility that both MSA models agree with a reference they helped define. This appendix replaces the reference with deposited coordinates.

Verified mappings only. For every source accession with a PDBe hit we take the SIFTS (Dana et al., 2019) uniprot_segments mapping to author numbering, which handles multi-segment and offset entries, and fall back to a voted single offset when SIFTS leaves both anchors null. A chain-span is then admitted only if the deposited residue names match the AFDB residue names at $\geq$ 95% of mapped positions and cover $\geq$ 80% of the domain span; 156/158 candidates pass, and the survivors give 72 domains across 33 pairs (50 mapped from SIFTS anchors, 22 from the voted offset). A single global offset taken from best_structures does not satisfy that check on three of five chains we first tried, which is why the gate is part of the protocol rather than a diagnostic. Mean TM between the AFDB conformation and the deposited one is 0.928 on the admitted residues, so the two references mostly agree.

The disagreement survives the reference swap. Table 12 is the class breakdown. Six of the ten admitted domains in the ESMFold-private class are the ones ESMFold scores low,

Table 12: Mean TM against deposited coordinates on the admitted residues, by the AFDB-referenced class of the construct. Fractions are domains below 0.7. Classes are defined by the main-text q, so a class can contain domains that the model in question scores highly.

Class (AFDB-referenced)	Domains	ESMFold	Boltz-2	AF2
ESMFold-private low-q	10	0.49 (6/10)	0.90 (0/10)	0.91 (0/10)
Boltz-2-private low-q	3	0.75 (0/3)	0.62 (3/3)	0.78 (0/3)
Jointly high-q	59	0.88 (4/59)	0.89 (4/59)	0.90 (4/59)

and on those six ESMFold averages 0.22 against the crystal (range 0.14–0.33) while Boltz-2 averages 0.88 (worst 0.79) and AF2 0.89 (worst 0.80) on the same residues. The three crystals involved are 6d6q_K, 1vrp_B and 3c4h_A. The remaining four admitted domains in that class are the partner domains, which ESMFold also folds. Boltz-2's private failures behave the same way in reverse: all three stay below 0.7 against the crystal, so neither private set is an artefact of the AFDB reference.

In the jointly high-q class the four domains below 0.7 are the same four for all three models (1bw6_A, 3i71_B), and the AFDB conformation is equally far from those crystals (TM 0.65 and 0.64): a reference-side conformational difference, not a prediction error. This is the expected floor of the experiment and the reason the admission gate matters. No consensus-failure construct has an experimental reference, so the crystal check cannot speak to that class.

W ESMFold scored against its own isolated fold

Appendix V swaps the reference for deposited coordinates. The complementary move is to remove the external reference entirely. Our q compares a fused domain with the AFDB conformation, so an ESMFold domain-level error and a fusion-context loss enter the same number. Define instead $q_{\text{self}} = \min_d \text{TM}(\text{fused } d,\ \text{ESMFold's isolated prediction of } d)$, which cancels the first term: whatever the model gets wrong about a domain in isolation it is free to get wrong again in the fusion. Both terms are already on disk from the 1,000-pair sweep, so this needs no folding.

On the panel q_{self} averages 0.842 against 0.794 for q and the two are closely related (r=0.936), with 46/300 below 0.7 against 53/300. Of the 43 ESMFold-private constructs, 34 stay below 0.7 once the model's own isolated fold is the reference (mean q_{self} 0.46) and 9 rise above it; all ten consensus-low constructs stay below (mean 0.28).

Table 13: Each score read as a screen for Boltz-2's low-q verdict on the 300-construct panel, cutoff 0.7. Intervals are pair-clustered bootstrap.

Score	Rejected	Precision	Recall	AUC
q (AFDB reference)	53/300	0.19 [0.06, 0.34]	0.38 [0.13, 0.66]	0.855
q_{self} (own isolated fold)	46/300	0.26 [0.10, 0.44]	0.46 [0.20, 0.73]	0.735

Self-reference is therefore a slightly cleaner threshold and a worse ranking (Table 13): it discards the native, so it cannot see a domain that ESMFold folds consistently and consistently wrong, and the AUC falls. Neither reading changes the design-relevant quantity. The sign of the q_{self} order gap agrees with Boltz-2's on 50% of pairs (48% for q) and the gaps correlate at r=0.055 (0.016 for q), so the order preference is a coin flip under either reference. Together with the crystal check, the disagreement is not an artefact of scoring ESMFold against an AF2-derived conformation: it survives an independent experimental reference and it survives having no external reference at all.

X Where the context loss lives

ESMFold packs the two domains closer on the 17 context-fail constructs and no other predictor does (Appendix M). This appendix tests the obvious reading of that—the model glues the domains together and one of them deforms to fit—and rejects it, then locates the loss one level in, in the folding trunk's pair representation.

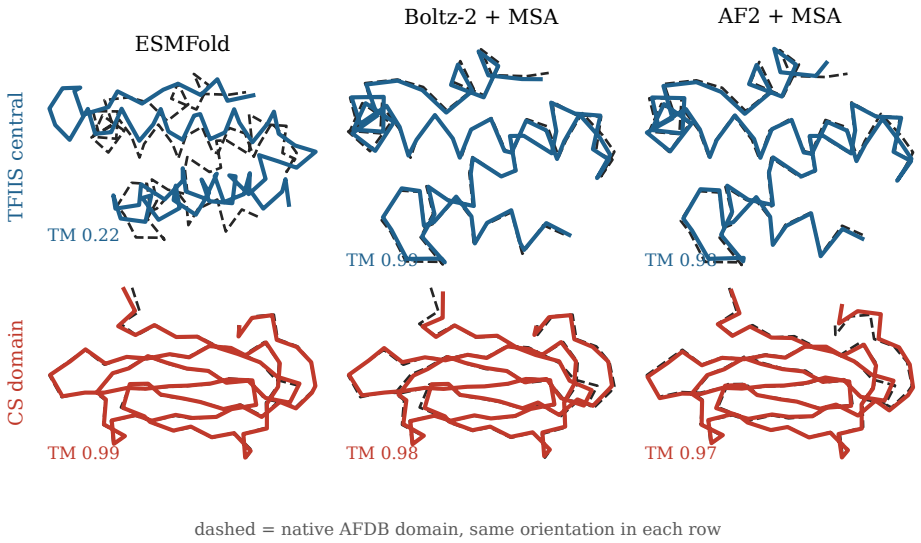

Figure 6: One ESMFold-only failure (r132_ab: TFIIS central fused to a CS domain, G4S3 linker). Each panel is the predicted domain (solid) Kabsch-superposed onto the AFDB-predicted reference domain (dashed), so a row shares one orientation. ESMFold predicts low reference similarity on TFIIS (q=0.22) and leaves CS intact (q=0.99); both MSA models recover both domains. A 60-residue linker takes ESMFold to 0.98 on the same pair.

More recycles change nothing. The panel runs ESMFold at the checkpoint default while AF2 and Boltz-2 iterate their trunks further, so the private set could be an under-converged two-domain chain. Re-folding all 300 constructs at 16 recycles moves mean q from 0.7942 to 0.7947 (r=0.988) and the low-q count from 53 to 52. The only constructs that cross are r023_ba and r061_ba out and r009_ab in; the first two are the same threshold-hugging pair that crosses under the reference scorer (Appendix U). Deep failures are untouched (r005_ab 0.219→0.216, r051_ba 0.104→0.100).

A long tether removes the packing and not the failure. If proximity costs the domain, a linker long enough to separate the domains should return it. We rebuilt the 43 private constructs and 17 length-matched consensus-ok controls with G4S×6 (30 residues) and G4S×12 (60 residues), changing nothing else.

At 60 residues the contact enrichment is gone—context-fail and the controls both sit at a median of 2 inter-domain pairs (p=1.0)—and 12 of the 17 constructs are still below 0.7 (Table 14, Figure 7C–D). The mean gain on context-fail is +0.067 and does not reach significance on 17 paired constructs (p=0.32); already-hard does not move (−0.016) and the controls stay folded (+0.006), so chain length alone neither repairs nor breaks anything. Packing accounts for at most three of the seventeen, and on inspection for fewer. Two of the three non-threshold rescues did carry elevated contacts under G4S3 (8 and 9), but the third—r132_ab, the Figure 6 construct—sat at the control median of 2 contacts and still moved 0.22→0.98. Its rescue cannot be contact relief; chain length or linker composition is doing the work.

The trunk does not predict the interface it builds. The contacts could still be a belief rather than a by-product, so we read the distogram head directly. Bin edges for that head are not

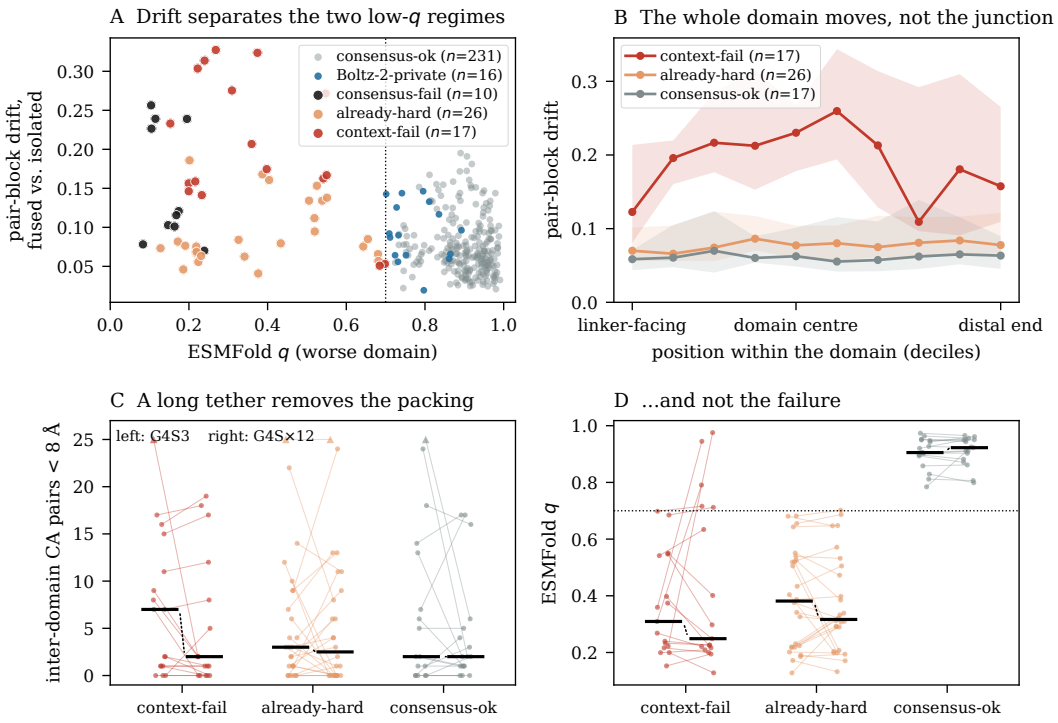

Figure 7: Where the context loss lives. (A) Pair-block drift against ESMFold q for all 300 panel constructs. Context-fail and already-hard overlap in q and separate in drift. (B) Drift by decile of position within the domain, median and interquartile band; decile 1 faces the linker. (C) Inter-domain CA pairs within 8 Å under G4S3 (left of each pair) and G4S×12 (right), thick bars are medians. Triangles are three constructs clipped at 25; one pair sits at 61/62. (D) The same constructs' q. Contacts fall to the control level while 12 of the 17 context-fail constructs stay below 0.7; the context-fail median q falls even as the mean rises +0.067, because the gain is carried by five rescues.

Table 14: ESMFold on the same constructs with a longer tether. Contacts are median inter-domain CA pairs within 8 Å; the panel linker is G4S3 (15 residues). Two of the five G4S×12 rescues are the threshold pair of the previous paragraph.

Linker	Class	n	mean q	low-q	rescued	contacts
G4S3 (panel)	context-fail	17	0.364	17	—	7
G4S×6	context-fail	17	0.347	15	2	3
G4S×12	context-fail	17	0.432	12	5	2
G4S3 (panel)	already-hard	26	0.388	26	—	3
G4S×12	already-hard	26	0.371	25	1	2
G4S3 (panel)	consensus-ok	17	0.904	0	—	2
G4S×12	consensus-ok	17	0.910	0	0	2

fixed in the inference code, so rather than assume them we fit the bin-to-distance map from each run's own coordinates on intra-domain pairs: the map is monotone at ≈0.3 Å per bin from 3.75 Å (bin 5) to 21.5 Å (bin 62), with bin 63 an overflow bin at 26.6 Å, which puts 8 Å at bin 18. Mean inter-domain probability below that bin is 1.49×10^{-3} on context-fail against 1.33×10^{-3} on consensus-ok (p=0.57), and the single most confident inter-domain contact in a construct is 0.158 against 0.169 (p=0.82). The readout has range—per-pair contact probabilities reach 0.17—so this is a negative, not a saturated statistic. The interface exists in the coordinates and not in the distogram.

What does mark the regime is the domain's own pair block. For each construct we encode the worse domain twice, alone and inside the chimera, and measure how far its intra-domain pair block moves: $\|z_{ij}^{\text{fused}}\|_2$ against $\|z_{ij}^{\text{iso}}\|_2$ on the same residue pairs, normalised by the isolated mean.

Table 15: ESMFold trunk statistics by class. Drift is the worse domain's intra-domain pair block, fused against isolated. Coupling is the inter-domain pair norm over the intra-domain pair norm. p is a 10,000-draw permutation test on medians against consensus-ok.

Class	n	mean q	pair drift	p	coupling
context-fail	17	0.364	0.175	$< 10^{-4}$	0.672
already-hard	26	0.388	0.078	0.093	0.638
consensus-fail	10	0.149	0.118	—	0.680
Boltz-2-private	16	0.774	0.091	—	0.634
consensus-ok	231	0.901	0.066	—	0.593

Drift on context-fail is 2.6× the consensus-ok median (Table 15, Figure 7A). The comparison that matters is with already-hard, whose mean q is the same to within 0.024: drift separates the two regimes at 0.175 vs. 0.078 ($p<10^{-4}$), so it is not a restatement of "q is low". That contrast also reads naturally: a domain ESMFold cannot fold alone is represented the same way alone and in the chimera—consistently wrong—whereas a context-fail domain is represented differently once the partner is present. Inter-domain coupling does not separate the two regimes (0.672 vs. 0.638, p=0.21), consistent with the linker result. Length is not behind either statistic (Pearson 0.18 and 0.13 against chain length; both classes have median length 231).

The whole domain moves, not the junction. Splitting the drift per residue with the 8-residue edge definition of Appendix L, the profile is flat: context-fail averages 0.173 over the eight residues facing the linker against 0.207 in the core, i.e. edge−core = −0.033, and the decile profile runs $0.151, 0.195, 0.214, 0.223, 0.214, 0.252, 0.214, 0.189, 0.198, 0.190$. Junction-adjacent residues move slightly less than the core (Figure 7B), which is the opposite of damage propagating inward from the linker, and it matches the per-residue pLDDT profile, where edge−core is also ≈0.

Reading. The fusion context changes how the trunk represents the domain itself, not how it represents the interface. That is why every junction-level repair in this paper fails: linker length, terminal padding, boundary edits and clash all act on the relationship between the domains, and the damage is inside one of them. Appendix Y tests that reading on the full sweep: the loss recurs on the same domains across unrelated partners. Two caveats. The drift statistic is measured on a model whose output structure has already changed, so it localises the loss rather than dating it; the already-hard contrast is what keeps it from being circular. And a pair-norm summary is a coarse instrument—dictionary learning on the pair stack (Migliorini et al., 2026) is the feature-resolution version, and it needs a larger context-fail set than the 35 constructs on the two panels before its multiple-comparison burden is payable.

The drift contrast repeats on a disjoint panel. The second 150-pair draw (Appendix D) has 18 context-fail constructs against 20 already-hard. Worse-domain pair-block drift is 0.195 vs. 0.087 (p=2×10^{-4}) against 0.062 on the 228 consensus-ok constructs—3.1× the control and 2.2× already-hard at matched q (0.343 vs. 0.375). Those are the same two ratios as Table 15 (2.6× and 2.2×). Inter-domain coupling again fails to separate the two low-q regimes (0.641 vs. 0.640). The spatial profile is again flat: edge−core = −0.017 on context-fail (linker-facing eight residues 0.175, core 0.191), against −0.033 on the first panel.

Y Context loss follows the domain, not the pairing

Appendix X says the loss lives in the domain's own pair block. If that is right, the same domain should lose in front of an unrelated partner, and the 1,000-pair sweep already contains that experiment: domains are drawn with replacement, so 260 domains that ESMFold folds in isolation appear with two or more different partners, up to eight.

We score a domain slot as a context loss when $q_{\text{iso}} - q_{\text{fused}} > 0.25$. The fixed drop replaces "crosses 0.7" on purpose—a domain scoring 0.71 alone crosses that line after an arbitrarily small drop—and if anything it favours well-folded domains, which have more room to fall. Across 3,732 eligible slots the loss rate is 3.4%, and it is graded by how well the domain folds alone: 15.5% for isolated q in $[0.70, 0.80)$ (n=238), 7.5% in $[0.80, 0.90)$ (n=584), 2.0% in $[0.90, 0.95)$ (n=896) and 1.3% above 0.95 (n=2,014; Figure 8A).

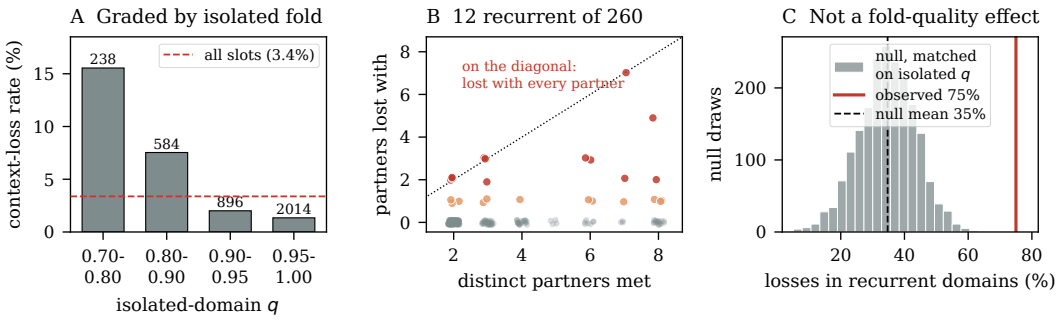

Figure 8: Context loss follows the domain. (A) Loss rate by the domain's isolated q, over all 3,732 eligible slots; the number above each bar is n. (B) The 260 reused domains: partners met against partners lost with. Grey never loses, orange loses with one partner, red with two or more; points are jittered. (C) Share of loss events sitting in recurrent domains, against 2,000 null draws that resample losses independently per slot within isolated-q strata.

Losses then concentrate on particular domains rather than particular pairings. Of the 260 reused domains, 24 lose with at least one partner and 12 lose with two or more; those 12 carry 75.0% of all 48 loss events, against 16.8% expected if each domain-partner slot failed independently ($p<10^{-4}$, 20,000 draws; Figure 8B–C). The graded risk above does not explain this: resampling within isolated-q strata still expects only 34.7% ($p<10^{-4}$). Domain identity therefore carries information beyond isolated fold quality, and the extreme cases are not the marginal ones—P37836 71–195 loses with all seven of its partners from an isolated 0.733, while A5PKE4 131–208 and Q5XIC7 130–207 both lose with both of theirs from an isolated 0.993. Those two are orthologues, and three members of one Q6R4R6/8/9 family at span ≈90–150 are all recurrent, so the behaviour tracks the sequence rather than the experiment.

Two practical consequences, now measured rather than only inferred. A domain's isolated fold is a screen but a weak one. The usable test is one throwaway chimera: fuse the candidate to a fixed dummy partner that shares no family with the panels (B5DEN9 131–227, isolated q=0.995) and ask whether ESMFold still folds it. The screen labels a domain slot as lost when isolated $q \geq 0.7$ and fused $q < 0.7$—the same gate as the rest of the paper, not the 0.25 drop used for recurrence—and the dummy was chosen by family exclusion and isolated-fold quality, not by fusion q. On the first panel that score beats isolated q only at the margin (paired $\Delta\text{AUC} = +0.127$, p=0.027). On the disjoint second panel it beats isolated q ($\Delta\text{AUC} = +0.068$, CI $[+0.007, +0.162]$, p=0.012, 257 domains). Pooling both panels, the dummy chimera's worse-order q ranks fusion-context loss at AUC 0.923 against 0.824 for isolated q ($\Delta = +0.099$, CI $[+0.021, +0.229]$, p=0.002, 466 domains). At the operating point $q_{\text{dummy}} < 0.7$, every flagged slot that the MSA models also scored as ESMFold-private is a true loss (20/20; recall 0.54), including 7/7 on the held-out panel (Figure 5). That is a predictor screen at two extra ESMFold calls per domain, not a function assay.

Restricting the recurrence test to the 600 constructs the MSA models scored—so a loss counts only when it is ESMFold-private—leaves 61 multi-partner domains and 12 private loss events. Three recurrent domains still carry 75% of those events, against 39% expected after matching on isolated q (p=0.021). The first panel alone was underpowered for that restriction (p=0.21); the direction did not move. The fragile set remains small.

This analysis of the 1,000-pair sweep is still ESMFold-internal wherever Boltz-2 and AF2 did not fold the construct. The two panels are where that separation is available, and there the recurrence reading agrees with the pair-drift one: what marks context-fail is a property of the domain that survives changing the partner.