NATIVE UPDATE ORIENTATION IN FROZEN PROTEIN STRUCTURE PREDICTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study whether adapter parameterization changes the effect of output-channel rotation under a fixed training budget. Our Factor head composes residue-level increments through a frozen predictor's native pair-generating operators. The generic head G+ explicitly receives the interface anchors and has a free output layer, making its function class closed under output rotation. Our four-cell design trains Native and Rotated versions of each head separately from the same zero-residual function under paired settings. At identical inputs, interface anchors and parameters, orthogonal rotations preserve local writer spectra. The interaction Ψ is the difference between the two heads' signed Native–Rotated score differences, rather than their absolute responses. On 192 new targets locked before prediction, fixed Protenix/ESMC models establish a Cα pair-lDDT interaction of +0.02515 [+0.01832, +0.03213]; OpenFold's Fresh96 primary interaction remains unestablished. In an observed-panel follow-up, the Protenix interaction remains positive after a prespecified doubling of the update budget: Factor's Native–Rotated difference decreases, while G+'s point estimate becomes more negative. In an OpenFold follow-up using one fixed rotation on previously evaluated targets, first-pass-only injection yields the largest interaction but the lowest prediction scores among the tested schedules. A positive interaction need not establish a positive Factor direction effect. We then test whether adding a shared trainable orthogonal output map gives rotated Factor a larger gain than Native Factor. Rotated Factor improves in both executions, but this gain difference was not re-established in the same-seed repeat. A prespecified reconstruction diagnostic did not establish the predicted ranking of post-training Native–Rotated score differences across eight tested rotations. Together, these findings motivate reporting absolute performance, signed within-head score differences and their interaction under explicit training budgets.

1 INTRODUCTION

Protein structure predictors map sequence information to geometry through learned pairwise representations (Jumper et al., 2021; Gong et al., 2025; Seo et al., 2026). We adapt a frozen predictor by adding residual updates to these representations. Factor constructs these updates by composing predicted residue-level factor increments through native pair-generating operators. This construction inherits both a structured update form and an output-channel orientation relative to the frozen downstream network.

An orthogonal output rotation preserves a residual's norm but can change the frozen downstream network's response. We ask: *under a fixed task-training budget, does output-channel rotation affect adaptation differently across parameterizations?*

Our four-cell design trains Native and Rotated versions of Factor and G+ separately from the same zero-residual predictor under paired settings. G+ uses the same additional sequence features, receives interface anchors explicitly, and has a free affine output layer. Each Rotated arm applies a fixed orthogonal transform only to its completed residual. G+'s function class is closed under output rotation, but its AdamW trajectories need not match (Loshchilov & Hutter, 2019). For each head H, we measure the signed Native–Rotated score difference Δ_H and their interaction $\Psi = \Delta_F - \Delta_{G+}$.

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.

Positive Ψ means a larger signed difference for Factor; it implies neither a positive Factor difference nor a larger absolute response to rotation.

Fixed Protenix/ESMC models establish the primary interaction on 192 new targets locked before prediction (Fresh192). Fixed OpenFold Train96 models (Ahdritz et al., 2024) do not establish their Fresh96 primary interaction. Native Factor outperforms G+ in the tested Protenix recipes, whereas G+ wins some OpenFold comparisons: the interaction does not determine which head gives higher prediction quality.

On the now-observed Fresh192 panel, a prespecified doubling of the Protenix update budget retains a positive interaction and reduces Factor's own score difference, without establishing a paired change in Ψ. An amended OpenFold schedule study on observed targets, using one fixed rotation R1, finds the largest interaction with First-only injection, which gives the lowest absolute quality. A larger Ψ therefore need not identify a better adapter or schedule.

We also test whether a shared trainable orthogonal output map benefits rotated Factor more than Native Factor. Both selected rotated Factor arms improve in the original execution and a same-seed retraining repeat. Only the original establishes a larger compensation gain than Native's; the repeat gives a smaller positive estimate of this gain difference whose interval includes zero. A prespecified teacher-update reconstruction diagnostic did not establish the predicted ranking of post-training Native–Rotated score differences across eight tested rotations.

The new-target test and budget extension establish a Protenix/ESMC head-by-rotation interaction at both tested budgets. The OpenFold schedule study shows that a larger interaction need not yield higher prediction quality, while the mechanism studies leave its cause unresolved.

2 Native update construction and orientation controls

2.1 Residual adaptation with native interface anchors

Let $e(x)$ denote frozen sequence features. Each independently trained Factor adapter has its own trainable predictor f_θ of residue-level factor increments, with parameters shared across residues and interface calls. At interface k and recycle t, let $a_{k,t}, b_{k,t}$ denote the interface anchors (native factors), $G_{k,t}$ the frozen operator used to construct the residual, and $U_{k,t,0}$ the separately retained native update at the selected interface. This is an update, not the full pair state or a common query-only trajectory: OpenFold computes it on each arm's current state, while Protenix reuses its cached query update. We define

$$(\delta a, \delta b) = f_\theta(e(x)), \tag{1}$$

$$\Delta U_{k,t} = G_{k,t}(a_{k,t} + \delta a, b_{k,t} + \delta b) - G_{k,t}(a_{k,t}, b_{k,t}), \tag{2}$$

$$U^{(R)}_{k,t} = U_{k,t,0} + \mathcal{R}_k \Delta U_{k,t}. \tag{3}$$

The last predictor affine layer has zero weights and bias, giving zero residual and exactly the un-adapted initial function. The sequence encoder retains its default initialization. Its three residual convolution layers, sequence-mean projection, head dimensions, and seeded constructor order are specified in Appendix A.1. Only the adapter is optimized; gradients propagate through the frozen folding model. The interface factors anchor the composition; the increment predictor receives sequence features, not structure labels.

The operator is backbone-specific. Protenix reuses cached query anchors and the same update across passes within a prediction. OpenFold's first main Evoformer OPM derives its anchors from the current MSA representation, after pair-biased attention and other MSA updates; these anchors can therefore depend on adapted recycle history. AtlasFold's first LM difference/product hook reads the current pass's MLM-dependent features before fusion with recycled states; its anchor does not directly read adapted history. Figure 1 summarizes the intervention.

2.2 FULL AND FIRST-ORDER FACTOR UPDATES

For Protenix, let a_q, b_q denote the cached query anchors and U_q the original query update. Its frozen decoder is

$$G_D(a,b)_{ij} = \frac{DW\,\mathrm{vec}(a_i b_j^\top) + b_{\mathrm{out}}}{D+\epsilon}, \qquad \widetilde{W} = \frac{D}{D+\epsilon}W. \tag{4}$$

Here a_i and b_j are column vectors; vec flattens their outer product in row-major order (the b index varies fastest), matching W. The full residual is

$$\Delta U_{ij}^{\mathrm{full}} = \widetilde{W}\,\mathrm{vec}(\delta a_i b_{q,j}^\top + a_{q,i}\delta b_j^\top + \delta a_i \delta b_j^\top). \tag{5}$$

The first-order, or tangent, construction omits the final quadratic term while keeping the same predictor and parameter count. The output bias cancels. We use $\epsilon = 10^{-3}$ and the historical training-side scale $D = 508.5$, the median MSA depth in the original Train24 cache. Subsequent experiments retain this constant. The original depth-one query update remains the baseline; U_q is not replaced by $G_D(a_q, b_q)$. OpenFold retains its own mask and effective-row normalization; AtlasFold has no OPM depth constant (Appendix A).

2.3 FIXED ORTHOGONAL OUTPUT CONTROLS

For a pair-channel column vector, $(\mathcal{R}\Delta U)_{ij} = R\Delta U_{ij}$ with $R^\top R = I$. A single fixed matrix acts across all residue pairs, rotating only the residual, not the baseline update or protein-coordinate axes. Labels are identical across arms. Native uses $R = I$. Each rotated adapter starts from zero residual and receives task training, so the experiment does not perturb an already trained native adapter at inference time. Appendix A.2 specifies the ordinary QR sampler, its separate random stream, and the implementation's multiplication convention. With inputs and interface anchors held fixed, let J denote the Jacobian of the vectorized residual with respect to the adapter parameters. At identical parameter values, orthogonal output rotation preserves residual norms, decoder singular values, and $J^\top J$. It need not preserve its composition with the frozen downstream Jacobian, AdamW dynamics, or norms of the different updates learned after training. For state-dependent anchors, this is a local control, not a spectral equivalence of the whole recycled model.

2.4 A GENERIC HEAD THAT CAN ABSORB THE ROTATION

G+ uses the same sequence-encoder architecture, with separately fitted parameters, and explicitly receives the interface anchors a_i, b_j. Its pair input concatenates projected node features z_i, z_j, a_i, b_j, and clipped relative sequence position. A pair MLP ends in a free affine layer and produces a zero-initialized residual. Unlike the earlier Generic control, G+ has explicit factor access. For $\Delta U_G = Wh_\phi + b$ and $\Delta U_{G,R} = R(W_R h_\phi + b_R)$, the choice

$$W_R = R^\top W, \qquad b_R = R^\top b \quad \Longrightarrow \quad \mathcal{F}_G^R = \mathcal{F}_G \tag{6}$$

absorbs the rotation. This proves rotation invariance of G+'s own function class, not $\mathcal{F}_{\mathrm{Factor}} \subseteq \mathcal{F}_{G+}$. G+ is not a proven performance or expressivity upper bound. The heads also differ in residue-versus-pair organization, nonlinearities and anchor entry points. The fixed factor decoder has no corresponding general absorption guarantee. Equation 6 concerns representable functions: coordinate-wise second moments can still make AdamW sensitive to the reparameterization. We therefore measure both heads' rotation effects rather than require G+'s effect to be zero. Factor and G+ have similar, not identical, sizes: 373,824 versus 378,144 trainable parameters for Protenix and OpenFold with ESM2 inputs. The ESMC follow-up changes only the input projection width within each head (Appendix J).

3 EXPERIMENTAL DESIGN

Predictors and information. We freeze Protenix-Mini/Tiny default checkpoints, official AF2 `model_3_ptm` weights through OpenFold, and AtlasFold with native AtlasLM-3B retained. Evaluation inputs exclude homologous sequences, templates and reference structures; adapter training uses structural supervision. All four arms within each recipe use the same frozen ESM2-35M features (Lin et al., 2023) or ESMC-600M features (Candido et al., 2026). Unadapted baselines lack

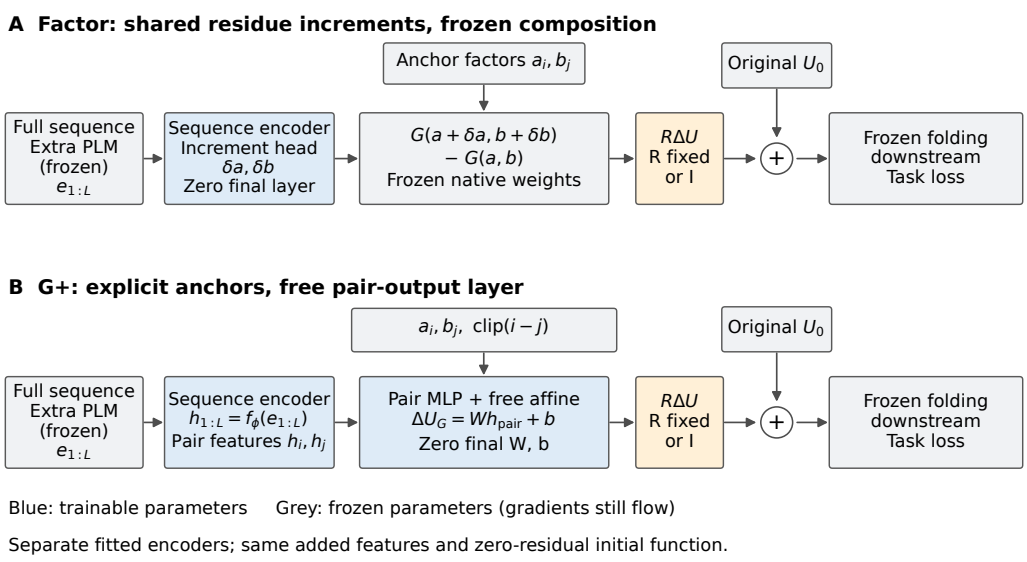

Figure 1: Actual adapter paths. Blue modules are trainable; grey modules have frozen parameters but transmit gradients. Factor's full-sequence predictor outputs residue increments; anchors enter only the frozen composition. G+ combines encoded residue features, explicit anchors and clipped relative position in a pair MLP with a free output affine. Only final output layers are zero-initialized. Fixed R (or I) acts on the complete residual before adding the separate original update U_0; $G(a, b)$ need not equal U_0. The two head families use separate fitted encoders. Factor is shown in Full form; the bilinear tangent version omits the quadratic increment term.

these added features, so their comparison with adapted models combines feature access and adaptation. Native–Rotated holds feature access fixed within each head. Appendix Table 3 gives the backbone-specific losses and inference settings.

Training and pairing. Core four-cell recipes use 1536 updates, three training seeds and three fixed rotations. Each head has three Native and nine Rotated fits. Within a recipe, arms share data order, seed choices, random-condition rules and optimizer settings. We evaluate the final checkpoint with one inference sample per complete target, without selection using evaluation labels. Task-loss gradients propagate only through the final recycle pass; the separate injection-schedule study retains gradients across four passes. The historical Train24 tangent study and intervention-specific training settings are detailed in Appendices B, R and W.

Panels and study identity. Train24 $\subset$ Train96 $\subset$ Train384, with no target-ID overlap between Train384 and Dev8 or either Confirm96 panel. Confirm96-A/B each contain 96 short-chain PDB targets (Berman et al., 2000); Length48 contains 48 complete sequences of 392–750 residues, beyond the Train384 adapter training range of 152–383. These panels were selected before their original evaluations; later reuse is reported as observed-panel follow-up. Fresh96 and Fresh192 test fixed OpenFold and Protenix/ESMC models, respectively, on targets locked before prediction, each with pair-lDDT Ψ as its sole primary endpoint. Eligibility criteria, exposure exclusions and study-specific endpoint roles are in Appendix C and Table 4. Sequence screening does not establish family or foundation-model pretraining isolation.

Scores and uncertainty. Our primary score is Cα pair-lDDT: for fixed reference pairs within 15 Å, average distance-preservation credit at thresholds 0.5, 1, 2 and 4 Å, then weight targets equally. Missing prediction pairs receive zero credit with the reference denominator unchanged. This pair-averaged lDDT variant (Mariani et al., 2013) differs from supplementary residue-average Cα lDDT; fixed-correspondence TM-score (Zhang & Skolnick, 2004) is also supplementary. For head $H \in \{F, G+\}$, $S_{H,i,s}$ and $S_{H,R_r,i,s}$ denote Native and rotated scores on target i under training seed s,

respectively. Here R_r denotes the rth fixed output rotation. With N targets in the panel, define

$$d_{H,i} = \frac{1}{3}\sum_s \left(S_{H,i,s} - \frac{1}{3}\sum_r S_{H,R_r,i,s} \right), \quad \Psi_i = d_{F,i} - d_{G+,i}, \quad \Psi = \frac{1}{N}\sum_i \Psi_i. \tag{7}$$

The single-rotation schedule study uses R1 in place of the inner average. With $\Delta_H = N^{-1}\sum_i d_{H,i}$, $\Psi = \Delta_F - \Delta_{G+}$ compares signed score differences. Positive Ψ alone implies neither $\Delta_F > 0$ nor $|\Delta_F| > |\Delta_{G+}|$ and does not rank absolute prediction quality. For structure-score contrasts, we use 20,000 paired target-bootstrap draws, retaining complete within-target seed/rotation records and resampling within length strata where specified by the protocol. Fresh192 and its budget extension resample within four fixed length strata. Intervals condition on fitted models and rotations rather than quantifying variability over new training runs. Secondary and follow-up confidence intervals are unadjusted; the OpenFold feature study applies Holm correction to the p values of three prespecified secondary tests (Holm, 1979) (Appendix J).

4 Results

4.1 Do Native–Rotated score differences vary between heads?

Prospective tests of fixed models. For Protenix, we selected the Train384/ESMC recipe from observed evidence, then evaluated all 24 fixed adapters and Query on Fresh192 without further training. The sole primary interaction is $\Psi = +0.02515$ [+0.01832, +0.03213]. Factor's Native–Rotated score difference is +0.02392, while G+'s is -0.00124; its interval includes zero without establishing equivalent Native and Rotated G+ scores. Fresh192 establishes a larger signed score difference for Factor than for G+ under this fixed recipe on screened short-chain targets (Appendix T); Table 2 retains all five means.

For OpenFold, we evaluated the 24 fixed Train96/ESM2 adapters and Query on Fresh96, locked before prediction, without further training (Appendix M). Factor's score difference remains positive: +0.01240 [+0.00278, +0.02218], but the sole primary interaction is +0.01014 [-0.00318, +0.02347]. It is not established on new targets; a positive Factor difference cannot replace this endpoint. The studies differ in predictor, training scale, features and panel, so their significance outcomes do not estimate a between-predictor difference in Ψ.

Observed-panel comparisons. The OpenFold Train96/ESM2 interaction is positive on Confirm96-B and Length48; Protenix Train96 also supports it on Confirm96-B (Figure 2; Appendix Table 14). These observed-panel results do not replace new-target tests. Seed/rotation marginals and supplementary metrics are reported in Appendices E and K. Under the OpenFold Train96/ESM2 recipe, signed channel permutations retain a positive Confirm96-B interaction, while the Length48 interval includes zero. The effect is therefore not confined to dense mixing, although ideal AdamW coordinate compatibility does not establish equivalent FP32 trajectories (Appendix P).

4.2 What changes with training budget and injection schedule?

A fixed budget extension. After Fresh192 was observed, we continued all 24 Protenix/ESMC models from 1536 to 3072 updates with their AdamW states and unchanged learning rate. All four cells improve (Table 1). Factor's Native–Rotated score difference decreases, with a paired change of -0.01241 [-0.01736, -0.00754]. The prespecified final interaction remains positive: $\Psi_{3072} = +0.01820$ [+0.01176, +0.02504]. G+'s difference becomes more negative in point estimate, with a paired change interval including zero. The paired interaction change is $K = \Psi_{3072} - \Psi_{1536} = -0.00695$ [-0.01423, +0.00046], which does not establish a change in Ψ. These results extend beyond the original stopping point without establishing convergence, budget invariance or another new-target confirmation (Appendix V).

Injection schedule and absolute quality. A separate OpenFold Train96 study retrains First-only, All and Last-only injection with gradients across four passes and a common Train/Dev-calibrated learning rate. Computational-budget constraints limited the amended scope to R1 and 36 fits; this amendment preceded unified scoring (Appendix W). All yields higher mean Factor scores than First-only, but a smaller interaction: $\Psi_{\mathrm{All}} - \Psi_{\mathrm{First}} = -0.01546$ [-0.03047, -0.00089]. Among the three

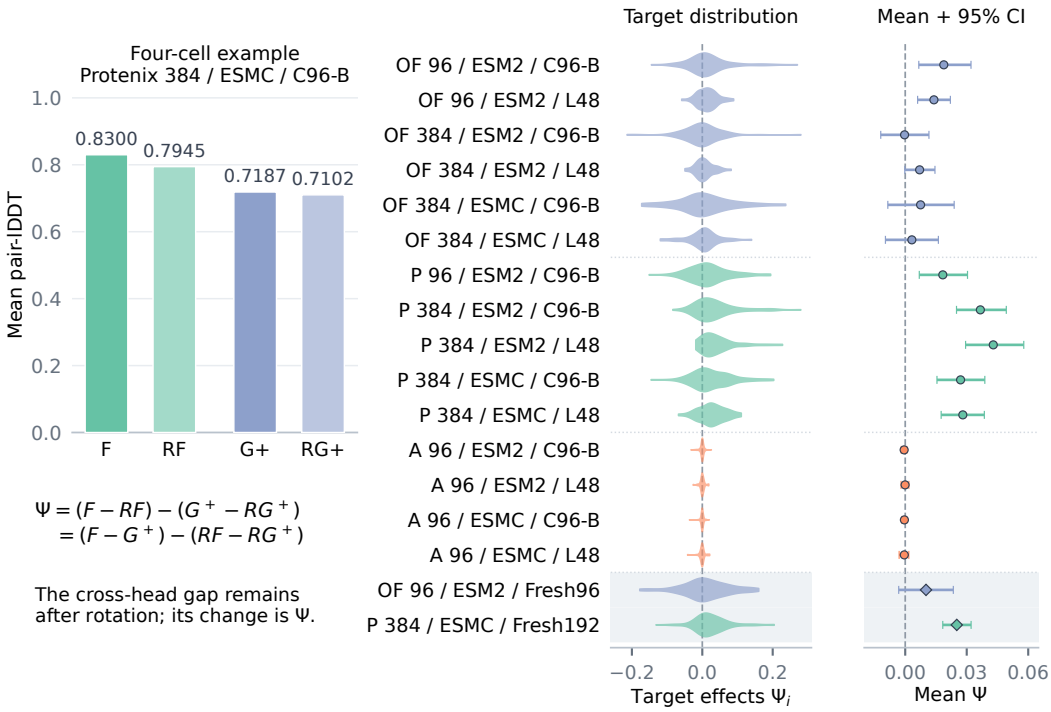

Figure 2: Head competitiveness and signed score differences. Left: Protenix four-cell means. Right: target-level Ψ_i, averaging seeds and rotations within each panel's 48, 96 or 192 targets. Violins span observed extrema (bandwidth 0.35, equal width); the separate, narrower axis shows Ψ and its original conditional 95% bootstrap interval. Shaded diamonds mark the new-target primary tests: Fresh96's interval includes zero; Fresh192's is positive. Other rows reuse observed panels and are not independent replications (Table 4). Feature-study multiplicity and backend limits are in Appendix K.

Table 1: Absolute quality and signed score differences in the budget and schedule follow-ups. All scores are Cα pair-lDDT; $\Delta_H = H_N - H_R$ and $\Psi = \Delta_F - \Delta_{G+}$. Each row averages three training seeds; Protenix also averages three rotations, while OpenFold uses R1 only. Query means are 0.3912 and 0.3019, respectively. The OpenFold schedules share the new cross-pass-gradient recipe and 1536 updates. These are separate observed-panel studies. Full conditional intervals and paired changes are in Appendices V–W; contrasts use unrounded scores.

Condition	F_N	F_R	G^+_N	G^+_R	Δ_F	Δ_{G+}	Ψ
Protenix/ESMC, Train384; observed Fresh192							
1536 updates	0.8281	0.8042	0.7280	0.7293	+0.02392	-0.00124	+0.02515
3072 updates	0.8332	0.8217	0.7633	0.7700	+0.01150	-0.00670	+0.01820
OpenFold/ESM2, Train96; observed Confirm96-B, R1 only							
First-only	0.3365	0.3063	0.3024	0.3026	+0.03013	-0.00018	+0.03031
All	0.4391	0.4371	0.3338	0.3466	+0.00206	-0.01279	+0.01485
Last-only	0.3907	0.3792	0.3898	0.3917	+0.01150	-0.00190	+0.01340

tested schedules, First-only has the largest estimated interaction and the lowest mean pair-lDDT in each of the four cells (Table 1). Under All, the interaction interval lies above zero, whereas Factor's difference interval includes zero and G+'s lies below zero. A larger interaction need not identify

Table 2: Four-cell Cα pair-lDDT means at 1536 updates. F/G+: three seeds; RF/RG+: three seeds $\times$ three rotations. Query lacks the added PLM features; Atlas retains AtlasLM-3B. Blank recipe cells inherit preceding entries within each predictor. Fresh96/Fresh192 use new targets and fixed models; other panels are observed (Table 4). Contrasts use unrounded scores. Protenix/96 has no L48 run.

Predictor	Train	PLM	Panel	Query	F	RF	G+	RG+
Protenix	96	ESM2	C96-B	0.3793	0.6165	0.5915	0.5557	0.5489
	384	ESM2	C96-B	0.3793	0.6257	0.5862	0.5494	0.5466
			L48[†]	0.3212	0.5292	0.4867	0.4478	0.4482
		ESMC	C96-B	0.3793	0.8300	0.7945	0.7187	0.7102
			L48	0.3212	0.7565	0.7184	0.5926	0.5826
OpenFold	96	ESM2	C96-B	0.3019	0.5019	0.4821	0.4953	0.4943
			L48	0.2608	0.3887	0.3783	0.3985	0.4021
	384	ESM2	C96-B	0.3019	0.5068	0.5038	0.5133	0.5100
			L48	0.2608	0.4017	0.3904	0.4087	0.4044
		ESMC	C96-B	0.3019	0.6703	0.6578	0.6762	0.6712
			L48	0.2608	0.4888	0.4816	0.5053	0.5013
AtlasFold	96	ESM2	C96-B	0.9568	0.9563	0.9566	0.9571	0.9568
			L48	0.9505	0.9479	0.9485	0.9484	0.9489
		ESMC	C96-B[‡]	0.9568	0.9567	0.9565	0.9571	0.9565
			L48[‡]	0.9505	0.9469	0.9478	0.9485	0.9490
New targets, fixed models								
OpenFold	96	ESM2	Fresh96	0.3330	0.5312	0.5188	0.5327	0.5304
Protenix	384	ESMC	Fresh192	0.3912	0.8281	0.8042	0.7280	0.7293

[†] Query/F/RF on H100; G+/RG+ on MI250. [‡] All adapted arms for seed 20260925 on H100, other seeds and Query on MI250. These mix backends across heads or across seeds, respectively (Appendix K).

a better schedule; a positive interaction need not establish a positive Factor difference. Jointly retraining each schedule does not isolate a feedback mechanism.

4.3 Where do effects transfer, and do they imply a better head?

Protenix Factor retains Native–Rotated advantages across the tested residual forms, checkpoints and mean-preserving rotations, and on complete sequences in Length48 (Appendix Figure 3). Length transfer does not establish family or pretraining novelty. At Train384, Protenix interactions are positive with both ESM2 and ESMC on the observed panels, while OpenFold establishes them with neither feature source (Appendix Table 14). In these Train384 comparisons, ESMC improves prediction quality over ESM2 in both predictors without establishing changes in Δ_F or Ψ. Feature family, width, precision and projection size change together (Appendix K); model size is not isolated.

Direction effects do not determine head competitiveness. Native Factor exceeds Native G+ in the matched Protenix comparison, whereas Native G+ outperforms Native Factor on Length48 under both OpenFold/ESM2 training scales (Appendix Table 23). Neither adaptation benefit nor interaction is established for AtlasFold with either feature source; high baseline and native AtlasLM remain unisolated explanations (Appendix K). Training-set changes also alter data composition and repetition counts, preventing a pure data-size attribution (Appendix N). Saved Dev8 checkpoint curves establish neither convergence nor a uniform learning delay (Appendix Q). Table 2 reports all four-cell means and backend differences.

4.4 Does shared output freedom preferentially help rotated Factor?

Using OpenFold Train96/ESM2 and evaluating on observed Confirm96-B, we jointly train each Factor head with its own mean-preserving orthogonal map C in Native and two mean-preserving Rotated conditions selected for low and high reconstruction error before their task training. Initialized at identity, C is shared across targets, pairs and recycles within that model (Appendix R). Let

$B_N = S(N + C) - S(N)$ and $B_R^C = S(R + C) - S(R)$, averaging the latter over both rotations; $T_C = B_R^C - B_N$ compares their compensation gains.

Rotated Factor improves in both the original execution and a complete same-seed repeat using the same fixed no-C baselines (Appendix Table 35). Only the original establishes a larger compensation gain than Native: $T_C = +0.01684$ [+0.00497, +0.02923], versus +0.00065 [-0.01223, +0.01338] in the repeat. The repeat adds no independent seeds and is not pooled with the original; its T_C interval includes zero. Differing interval significance does not establish a between-execution population difference (Appendix R.1). Across eight rotations, the prespecified positive association between held-out reconstruction error and post-training Native–Rotated score differences was not established ($\rho = +0.02381$, exact one-sided $p = 0.48839$; Appendix O). The useful rotated-arm gains do not validate that prediction or identify output freedom as the unique explanation of the four-cell interaction; output freedom and joint optimization remain coupled.

5 Discussion and limitations

For the fixed Protenix/ESMC recipe, the interaction remains positive at both tested budgets. In the OpenFold R1 schedule study, First-only has the largest interaction and lowest absolute scores. These results motivate joint reporting of Ψ, absolute quality and both within-head differences under explicit training budgets.

In OpenFold, compensation improves rotated Factor in both executions, but its additional gain over Native is not re-established in the repeat. The reconstruction and residual-anchor tests also fail to establish their predicted effects (Appendices O and S). Output freedom and joint optimization remain coupled, and schedule retraining changes training and inference together.

On 24 observed Protenix targets, the native residual source retains an advantage at matched global norms (Appendix G). This source contrast includes learned content and relative pair/channel magnitudes, so it does not isolate a causal effect of channel orientation.

Target intervals condition on fitted models and rotations, excluding variability across new training runs. The same-seed repeat adds no independent seeds or retrained no-C controls; the two executions are not pooled. Different significance outcomes do not establish different population effects between executions. There is no joint multiplicity correction across follow-up studies; the OpenFold feature study retains its prespecified Holm correction of p values.

AtlasFold showed no established adaptation benefit. Limited budgets, one inference sample, differing interfaces, losses and backends, and incomplete exposure records restrict generalization. Sequence screening establishes neither family nor foundation-model pretraining isolation. Gaps to native single-sequence systems limit deployment claims.

6 Related work

LoRA and PiSSA adapt pretrained weights (Hu et al., 2022; Meng et al., 2024), whereas ReFT adapts representations (Wu et al., 2024). Factor composes residue-level increments through frozen native pair operators. Orthogonal finetuning (OFT) and BOFT learn orthogonal transformations (Qiu et al., 2023; Liu et al., 2024); our fixed rotations act on completed residuals, while a shared map trained jointly with Factor tests whether rotated models benefit more than Native. G+'s rotation closure distinguishes representable functions from finite-budget training behavior under AdamW (Loshchilov & Hutter, 2019). This distinction complements LoRA-RITE's study of optimization invariance across equivalent low-rank factorizations (Yen et al., 2025).

7 Conclusion

Fixed Protenix/ESMC models establish the Fresh192 interaction, which remains positive after the observed-panel budget extension. OpenFold's Fresh96 interaction remains unestablished. In a separate OpenFold mechanism study, the compensation repeat improves rotated Factor without re-establishing a larger gain than Native. Budget and schedule follow-ups show why absolute prediction quality, signed within-head differences and their interaction should be reported together.

AI USE STATEMENT

AI assistants supported literature search and discovery, research ideation, feedback on methodology and experimental design, method implementation and code development, experiment orchestration, analysis checks, and result interpretation. They also assisted with mathematical derivations and proofs, manuscript drafting and editing, and the collection and organization of supplementary code. The authors checked numerical results against experimental records and take responsibility for the final text, mathematical claims, code, and experimental conclusions.

ETHICS STATEMENT

This work studies computational adaptation using protein structure data. Its results concern the stated structure-prediction settings and do not validate biological function, clinical use, or safety of designed proteins. Data and model access remain subject to their respective source terms.

REPRODUCIBILITY STATEMENT

The accompanying materials provide adapter source snapshots, training and scoring recipe summaries, per-target scores, and code to regenerate numerical results and plots from archived scores. Appendices B, C and D describe the model-specific recipes, study sequence and statistical analysis. A one-target OpenFold example includes a trained adapter, feature extraction, prediction and scoring; prediction requires separately obtained upstream software and foundation-model weights. CPU operator examples use synthetic inputs. The package does not reproduce the full training and evaluation matrix: complete training data, feature caches, the full set of trained adapters and predicted structures, and an end-to-end retraining environment are not bundled.

REFERENCES

Gustaf Ahdritz, Nazim Bouatta, Christina Floristean, et al. OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nature Methods*, 21: 1514–1524, 2024. doi: 10.1038/s41592-024-02272-z.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000. doi: 10.1093/nar/28.1.235. URL https://doi.org/10.1093/nar/28.1.235.

Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10:421, 2009. doi: 10.1186/1471-2105-10-421. URL https://doi.org/10.1186/1471-2105-10-421.

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://www.biorxiv.org/content/10.64898/2026.06.03.729735. Preprint.

Chengyue Gong, Xinshi Chen, Yuxuan Zhang, Yuxuan Song, Hao Zhou, and Wenzhi Xiao. Protenix-Mini: Efficient structure predictor via compact architecture, few-step diffusion and switchable pLM. *arXiv preprint arXiv:2507.11839*, 2025. URL https://arxiv.org/abs/2507.11839.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

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. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

John Jumper, Richard Evans, Alexander Pritzel, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Mario Lezcano-Casado and David Martínez-Rubio. Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3794–3803. PMLR, 2019. URL https://proceedings.mlr.press/v97/lezcano-casado19a.html.

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://www.science.org/doi/abs/10.1126/science.ade2574. Earlier versions as preprint: bioRxiv 2022.07.20.500902.

Weiyang Liu, Zeju Qiu, Yao Feng, Yuliang Xiu, Yuxuan Xue, Longhui Yu, Haiwen Feng, Zhen Liu, Juyeon Heo, Songyou Peng, Yandong Wen, Michael J. Black, Adrian Weller, and Bernhard Schölkopf. Parameter-efficient orthogonal finetuning via butterfly factorization. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=7NzgkEdGyr.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

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.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 121038–121072. Curran Associates, Inc., 2024. doi: 10.52202/079017-3846. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/db36f4d603cc9e3a2a5e10b93e6428f2-Abstract-Conference.html.

C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. CATH—a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, 1997. doi: 10.1016/S0969-2126(97)00260-8. URL https://doi.org/10.1016/S0969-2126(97)00260-8.

Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 79320–79362. Curran Associates, Inc., 2023. doi: 10.52202/075280-3472. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/faacb7a4827b4d51e201666b93ab5fa7-Abstract-Conference.html.

Seonghwan Seo, Hyeongwoo Kim, Seokhyun Moon, Woo Youn Kim, and Team KAIST. AtlasFold: Protein structure prediction with metagenomic-scale language models. *bioRxiv*, 2026. doi: 10.64898/2026.09.04.749352. URL https://www.biorxiv.org/content/10.64898/2026.09.04.749352v2.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 63908–63962. Curran Associates, Inc., 2024. doi: 10.52202/079017-2041. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/75008a0fba53bf13b0bb3b7bff986e0e-Abstract-Conference.html.

Jui-Nan Yen, Si Si, Zhao Meng, Felix Yu, Sai Surya Duvvuri, Inderjit S. Dhillon, Cho-Jui Hsieh, and Sanjiv Kumar. LoRA done RITE: Robust invariant transformation equilibration for LoRA optimization. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.20625.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004. doi: 10.1002/prot.20264.

A NATIVE OPERATORS AND RESIDUAL DERIVATIONS

The default predictor maps 480-dimensional frozen ESM2 features through a shared 128-dimensional sequence encoder to residue-level increments. For Protenix and OpenFold each of the two factor increments has width 32, and the pair residual has 128 channels. The Protenix residual is computed by the bilinear expansion in Eq. 5, avoiding subtraction of two large decoder outputs. For an output weight reshaped as W_{cuv}, the three terms contract $\delta a_{iu} W_{cuv} b_{jv}$, $a_{iu} W_{cuv} \delta b_{jv}$, and $\delta a_{iu} W_{cuv} \delta b_{jv}$. The affine bias cancels in the difference.

The fixed Protenix scale originates from these 24 training-side teacher depths:

114, 489, 490, 492, 493, 502, 502, 504, 504, 506, 507, 508,
509, 510, 511, 511, 512, 512, 513, 513, 513, 513, 513, 513.

Their median is $(508 + 509)/2 = 508.5$. It was not estimated from test proteins or reselected for Train96/384. This historical MSA metadata is disclosed even though task-only prediction does not take homologs as input. The residual's $D/(D+10^{-3})$ multiplier never replaces the depth-one query baseline.

OpenFold recomputes its first-OPM anchors after pair-biased MSA updates on each pass, so they can depend on previous adapted states. The residual retains that OPM's mask and effective-row denominator; later blocks are unchanged. AtlasFold uses its first LM difference/product update, with factor width 128 and the native projection order. Its anchor depends on the current pass's MLM mask, before fusion with recycled states, and does not directly read adapted history. Native AtlasLM-3B is retained. Factor increments depend on complete sequence features and shared predictor parameters, not on the anchor or pass index. Neither interface inherits Protenix's depth constant.

G+ concatenates $[z_i, z_j, a_i, b_j, \mathrm{clip}(i-j, -32, 32)/32]$ and uses a Linear–GELU–Linear pair head. The final output layer has no activation or extra output constraint. Its zero weight and bias recover the baseline. For Protenix/OpenFold its hidden pair width is 192. AtlasFold's Factor/G+ parameter counts are 423,168/415,008, versus 373,824/378,144 in the OPM studies. These are approximate size controls, not exact parameter matches.

A.1 ENCODER, HEADS, AND INITIALIZATION

For one complete sequence of length L, let $e_i \in \mathbb{R}^d$ be the aligned frozen feature at residue i, with $d = 480$ for ESM2 and $d = 1152$ for ESMC. Each fitted head has its own trainable encoder; parameters are shared over residues and interface calls within that head. The encoder is

$$h_i^{(0)} = P e_i + p, \qquad P \in \mathbb{R}^{128 \times d}, \tag{8}$$

$$h^{(\ell)} = h^{(\ell-1)} + \mathrm{GELU}\Big(\mathrm{Conv1d}_\ell(h^{(\ell-1)})\Big), \qquad \ell = 1, 2, 3, \tag{9}$$

$$h_i = h_i^{(3)} + V\left(\frac{1}{L}\sum_{j=1}^{L} h_j^{(3)}\right), \qquad V \in \mathbb{R}^{128 \times 128}. \tag{10}$$

Each convolution has 128 input and output channels, kernel size 5, stride and dilation 1, groups 1, zero padding of 2, and a bias. The global projection V has no bias; its output is broadcast to all residues. There is no encoder normalization, dropout, or positional embedding. The sequence mean uses all L aligned residues, including when residual pairs are evaluated in chunks. GELU uses the exact, non-tanh implementation.

Factor maps each h_i through $\mathrm{LayerNorm}_{128} \to \mathrm{Linear}_{128\to256} \to \mathrm{GELU} \to \mathrm{Linear}_{256\to2r}$, then splits the output into $(\delta a_i, \delta b_i)$. Here $r = 32$ for Protenix/OpenFold and $r = 128$ for AtlasFold. G+ instead forms z_i using $\mathrm{LayerNorm}_{128} \to \mathrm{Linear}_{128\to32}$. Its pair input $[z_i, z_j, a_i, b_j, \mathrm{clip}(i-j, -32, 32)/32]$ has width $65 + 2r$, and its head is $\mathrm{Linear}_{65+2r\to192} \to \mathrm{GELU} \to \mathrm{Linear}_{192\to128}$, with no final activation. All LayerNorms act on the last 128 channels, use $\epsilon = 10^{-5}$, and have learned scale and bias. All listed affine layers have a bias except V.

Initial values and random-number order. The ordinary head parameters are constructed on CPU in FP32. Except for the final output affine, Linear and Conv1d weights and biases use PyTorch's

default uniform initialization on $[-1/\sqrt{n_{\text{in}}}, 1/\sqrt{n_{\text{in}}}]$, where n_{in} is fan-in (including kernel width for convolutions). LayerNorm scales start at one and biases at zero. After constructing the head, both weight and bias of its final output affine are set to zero; the encoder and earlier affine layers retain their default initialization. This gives zero factor increments or zero G+ residual without zeroing the sequence representation.

Exact seeded reconstruction also requires the original constructor order. `FeatureResiduePredictor` first constructs a 21-token embedding, the three convolutions, V, and the Factor output head, then replaces the embedding with the feature projection P. G+ subsequently replaces the Factor output head with its node projection and pair head. The discarded modules still consume random numbers; omitting them changes seeded initial tensors despite preserving the final architecture. The accompanying source under `reproducibility/src/engramfold/models/` retains these constructors. For ESMC, the original 480-input writer is built first and only P, p are replaced using an isolated CPU random stream seeded by the training seed plus 10^6, with the caller's CPU random state restored. Thus the wider input does not reinitialize the remaining head parameters. This pairing is within the same backbone and head, not a claim of identical initial tensors across different backbone-specific head dimensions.

A.2 Ordinary dense orthogonal sampling

For each ordinary rotation seed, a dedicated CPU PyTorch generator samples $M \in \mathbb{R}^{128\times128}$ with independent standard-normal entries in float64. NumPy QR gives $M = QT$. We set

$$R = Q\,\mathrm{diag}(s), \qquad s_k = \begin{cases} -1, & T_{kk} < 0, \\ +1, & T_{kk} \geq 0, \end{cases} \tag{11}$$

and store R in float32. There is no determinant correction: $R \in O(128)$ in ideal arithmetic, so reflections are allowed. The ordinary rotation seeds are 20261001/02/03, separate from the training seeds 20260923/24/25. Sampling a rotation neither reads nor advances the global training random stream. One sampled matrix is a fixed buffer, shared over pairs, targets, and all calls to the adapted interface during training and prediction. Column-vector residuals receive $R\delta$ (row-vector storage uses $\delta R^\top$), while the baseline is unchanged. Factor fuses this operation into the frozen output weight as RW; G+ rotates its emitted residual. The source implementation is `orthogonal_rotation` in `native_geometry.py`. These ordinary controls are distinct from the mean-preserving family in Appendix O, the signed-permutation controls in Appendix P, and the trainable compensator in Appendix R.

A.3 Scope of the local identities

For identical inputs, anchors and writer parameters, changing only the output rotation gives $(RJ)^\top(RJ) = J^\top J$ for the local writer Jacobian. A single-pair Jacobian with respect to free factor coordinates is distinct from the Jacobian with respect to shared predictor parameters, which couples residues and targets. Full single-pair rank would not establish rotation closure of the shared update family; a smaller local rank would not establish a native task advantage. This local identity neither compares independently trained writers nor asserts spectral equivalence of the full recycled model. Likewise, a nonzero injected residual, a change between predicted structures, and improved accuracy against a reference structure are distinct observations. For G+, the free final layer proves Eq. 6. Ordinary Euclidean updates admit an orthogonal parameter correspondence under matched conditions; AdamW's coordinate-wise moment accumulation generally does not commute with a general rotation. Equality of G+'s representable classes therefore supplies a control, not a prediction of identical training.

B Per-predictor implementation and training recipes

The Protenix task-only direction recipe uses AdamW with learning rate 5×10^{-5}, weight decay 10^{-4}, betas $(0.9, 0.999)$, $\epsilon = 10^{-8}$ and gradient clipping at 1. The original Mini direction matrix includes Full and tangent at Train24/384 updates. Larger-set direction comparisons use Full at 1536

Table 3: Core recipes; folding and PLM weights stay frozen. M, B, S, D: MSE, bond, smooth-lDDT and distogram losses; M_w: weighted MSE; χ_s: supervised torsion loss. Anchor definitions and full recipes are in Appendices A–B.

Predictor	Interface / anchor	Task supervision	Structure inference
Protenix	MSA-OPM; cached	$4M + 4B + 4S + 0.03D$	4 trunk; 5 diffusion
OpenFold	First of 48 OPMs; recycled	$\mathrm{FAPE} + 0.3D + \chi_s$	4 trunk; structure module
AtlasFold	First LM diff/product; MLM	$0.4D + 2(M_w + S)$	5 trunk; 20/30 diffusion*

*20 steps for $L \leq 512$; 30 for $513 \leq L \leq 1024$. Protenix's bond mask is empty here.

updates. Frozen ESM2-35M representations, the folding checkpoint, native decoder and query-feature contracts are preserved. The loss is $4\mathcal{L}_{\mathrm{MSE}} + 4\mathcal{L}_{\mathrm{bond}} + 4\mathcal{L}_{\mathrm{smooth-lDDT}} + 0.03\mathcal{L}_{\mathrm{distogram}}$; the existing protein-only bond mask is empty and confidence is excluded. Only the last of four recycle passes retains the task gradient.

OpenFold loads official AF2 `model_3_ptm` weights and preserves all 48 main Evoformer blocks. The first block's OPM is adapted on each pass. The selected common learning rate is 10^{-4}, with AdamW weight decay 0.01 and global clipping at 1. The loss is FAPE plus 0.3 distogram plus supervised-chi; confidence and masked-MSA auxiliary losses are excluded. Three recycling iterations plus the initial pass give four trunk passes, with task backpropagation through the last. The added rotated-G+ study reuses this Train96/ESM2 recipe, three training seeds, data order and final step without new learning-rate selection. That follow-up's primary endpoint is the Confirm96-B interaction.

AtlasFold uses the 260703 structure checkpoint and native AtlasLM-3B, adding ESM2-35M only as adapter input. A bounded development calibration selects one common 10^{-4} learning rate using mean Dev8 performance pooled equally across Native, a fixed rotation, and G+; formal fits use an independent set of training seeds. Training uses four trunk calls, BF16 autocast with FP32 native diffusion/loss, activation checkpointing, and no stochastic dropout. The loss weights are 0.4 distogram and 2 diffusion (weighted MSE plus smooth lDDT), with confidence excluded. Inference's recycle argument of four produces five trunk calls; one sample uses the pinned native length rule of 20 diffusion steps through 512 residues and 30 through 1024.

The common formal training seeds are 20260923/24/25 and ordinary rotation seeds 20261001/02/03. The preserved-mean Protenix controls use their separately locked rotation family. Protenix prediction uses seed 101, four recycles, five steps, FP32 torch kernels, TF32 and MC dropout off. OpenFold uses seed 20260921 and four trunk passes without early stopping; AtlasFold uses seed 1. Different backbones use native losses and inference budgets: equal update counts do not mean equal computation.

C Data, study chronology, and cumulative exposure

Experimental structures come from an existing local PDB/mmCIF collection, with candidate identities fixed by the archived `monomer_clean` catalog hash. The original download date of the underlying collection was not retained; the 2021-09-30 criterion is an entry-release eligibility cutoff, not a database snapshot date.

The inherited Train24 and Dev8 are the original interface-cache queries, including later releases; the 2021-09-30 cutoff applies to newly selected candidates, not retrospectively to these sets. Train96 retains Train24 and adds 18 chains per short stratum; Train384 retains Train96 and adds another 72 per stratum. Realized lengths and target identities are provided with the data.

Confirm96-A uses the existing `monomer_clean` candidate catalog and experimental CIFs. Within each stratum, candidates are ordered by SHA256 of `engramfold-independent96-v1|target_id`, without model scores. At most one chain is retained per PDB entry in each panel; old Train24/Dev8 PDBs and qualifying sequence matches are excluded. Confirm96-B uses the same candidate quality criteria, with ordering `native-direction96-v1|confirmation|target_id`; it also excludes the previously observed Confirm96-A. Confirmation B is selected before the Train96 additions, which are then

Table 4: Identity of the four-cell studies. A study-specific primary endpoint on observed targets is not new-target confirmation. C96-B denotes Confirm96-B. All intervals condition on fitted models.

Study	Panel status	Declared primary	Role of Ψ
OpenFold Train96	Observed	C96-B Ψ	Primary; L48 secondary
OpenFold signed permutations	Observed	C96-B Ψ	Primary; L48 secondary
Protenix Train96	Observed	C96-B Ψ	Primary; no L48 study
OpenFold Fresh96	New targets	Fresh96 Ψ	Sole primary; fixed models
Protenix Fresh192	New targets	Fresh192 Ψ	Sole primary; fixed models
Protenix budget	Observed Fresh192	Ψ_{3072}	Primary; paired budget change secondary
OpenFold schedules	Observed C96-B	Three-rotation $\Psi_{\mathrm{All}} - \Psi_{\mathrm{First}}$ (not completed)	All–First Ψ difference using R1 only (amended scope)
OpenFold features	Observed	C96-B ESMC Δ_F	C96-B ESMC Ψ: secondary, Holm-adjusted p; confidence intervals unadjusted
Protenix/Atlas features	Observed	C96-B ESMC Δ_F, per backbone	Follow-up; no joint correction

Table 5: Data composition, separating inherited sets from additions. Quotas use four short-chain strata (128–191, 192–255, 256–319, 320–384) or three long-chain strata (385–512, 513–640, 641–768); the range column gives realized lengths. Q: existing monomer-clean catalog, release by 2021-09-30, single-model X-ray, resolution $\leq$ 2.5 Å, standard amino acids and mapped Cα coverage $\geq 90\%$. Q is not imposed on inherited Train24/Dev8. "New" refers to first scoring in the original study; all confirmation panels are later reused.

Set / origin	N	Source / conditions	Quota; actual range	Nesting / exclusions	Use / first scoring
Inherited Train24	24	Original cache; no new cutoff	4 x 6; 152–370	Subset of Train96	Adaptation training
Inherited Dev8	8	Original cache; no new cutoff	4 x 2; 166–342	Disjoint from training	Development; already observed
Train96 additions	72	Short-chain catalog (Q)	4 x 18; 163–375	18/stratum added to Train24; excludes B	Adaptation training; B selected first
Train384 additions	288	Short-chain catalog (Q)	4 x 72; 154–383	72/stratum added to Train96; excludes Dev8/A/B	Adaptation training
Confirm96-A	96	Short-chain catalog (Q)	4 x 24; 166–375	Hash order; excludes old queries/MSA and accepted neighbors	New for Factor–Generic
Confirm96-B	96	Short-chain catalog (Q)	4 x 24; 155–380	Hash order; excludes A, old queries/MSA and accepted neighbors	New for direction study
Length48	48	Long-chain catalog; v2 rules below	3 x 16; 392–750	Excludes recorded training/development and A/B	New for Protenix length transfer
Fresh96	96	Short-chain eligibility (Q); BLAST v2	4 x 24; 129–384	Excludes recorded exposure and accepted neighbors	New for fixed OpenFold Train96 interaction
Fresh192	192	Short-chain eligibility (Q); BLAST v2	4 x 48; 128–383	Refreshed exposure exclusions and accepted neighbors	New for fixed Protenix Train384/ESMC interaction

excluded against it. Train384 additions exclude the recorded Train96/Dev8 and both confirmation panels, plus previously selected additions and documented old MSA rows.

For these short-chain protocols, global affine alignment uses match 2, mismatch -1, gap-open -8 and gap-extend -1, taking the first optimal alignment. A sequence is excluded when it has at least 50 paired positions, identity at least 0.30 among paired positions, and coverage at least 0.70 of the shorter sequence for query/within-panel comparisons. For old MSA rows, coverage instead uses the candidate length; query duplicates retain the stricter rule. Confirm96-A's initial shorter-sequence rule for all MSA rows failed to fill its shortest stratum. Its v2 changed only that MSA denominator before any new prediction or score, retaining the ranking and quotas. These are sequence exclusions, not family labels.

The original Confirm96-A study compared Factor with Generic, which did not explicitly receive query factors. The later Confirm96-B direction study fixed its primary tangent comparison and prespecified Full as secondary. Subsequent Tiny, mean-preserving, and training-scale studies reuse that panel. The separate Train24 G+ study used its own seeds and head-specific development-selected rates; its results cannot fill matched Train384 cells.

Length48 was selected before its original new structure scores, with 16 complete proteins in each range 385–512, 513–640, and 641–768. Candidates were single-model X-ray structures, at most 2.5 Å resolution, with standard amino acids, initial release no later than 2021-09-30, verified entity mapping, and at least 90% reference Cα coverage. The realized lengths were 392–750. Deterministic hash ordering, exclusion rules and quotas were locked; targets were not cropped or replaced based on scores. All 672 original model–target attempts preceded opening the structural scores.

The length screening corpus contained 15,605 reference sequences including 584 known queries (Train384, Dev8, and both confirmation panels) and historical teacher rows. The filter-development examples were excluded from the panel. BLASTP 2.17.0+ from BLAST+ (Camacho et al., 2009) used a fixed search configuration and retained HSPs without a small hit-count cap. A candidate was rejected for a single HSP meeting all of: $E \leq 10^{-5}$, identity at least 0.30 with internal gaps included in the alignment-column denominator, at least 50 residue–residue columns, and paired coverage at least 0.70. Coverage is relative to the shorter sequence for known queries and to the candidate length for other teacher rows. HSPs are not unioned. Within-panel screening adds previously accepted targets to the reference corpus. CATH annotations (Orengo et al., 1997) were used to audit this filter. Four CATH-linked examples were missed at retrieval, so this is sequence screening, not homologous-superfamily isolation. The annotation-feasibility audit used the archived daily-release file `cath-b-newest-all.gz`, retrieved on 2026-09-19 with HTTP Last-Modified 2025-01-15. Its entries include `v4_3_0` and `putative` labels. Separate v4.4.0 chain and unclassified lists, whose headers are dated 2024-12-16, were retrieved on 2026-09-20 for the missing-annotation audit; these are distinct sources, not one CATH release.

The later OpenFold, AtlasFold and matched-G+ studies use these panels as observed follow-ups. Documented cumulative query exposure reaches 632 records, but does not certify complete historical teacher or development exposure. Unresolved CATH annotations and the added Train384 chains preclude the planned strict family-held-out claim. Foundation-model pretraining exposure is unknown.

D Scoring, failure handling, and statistical analysis

Reference Cα positions and mappings are fixed across systems. Define $\Omega_i = \{(a,b) : a < b,\ 0 < d_{ab}^{\mathrm{ref}} < 15\,\text{Å}\}$ and $T = \{0.5, 1, 2, 4\}$ Å. The primary per-target score is

$$S_i = \frac{1}{|\Omega_i|} \sum_{(a,b)\in\Omega_i} \frac{1}{4} \sum_{\tau\in T} \mathbf{1}\{a,b \text{ predicted and finite}\}\mathbf{1}\{|d_{ab}^{\mathrm{pred}} - d_{ab}^{\mathrm{ref}}| < \tau\}. \tag{12}$$

Prediction failure or missing coordinates do not remove reference pairs from the denominator. Global pair averaging and residue averaging can differ because residues have different numbers of reference neighbors. We report both, and use the true fixed-correspondence TM-score tool for supplementary structure scoring; the earlier Kabsch-based proxy is not substituted for that metric.

A target's complete seed/rotation record is the paired resampling unit. The original Protenix direction bootstrap uses seed 20260926, the original Length48 bootstrap uses 20260928, and the OpenFold four-cell study uses 20260921, each with 20,000 draws. The matched Protenix G+ study uses seed 20260926 and 20,000 draws on each panel. The original prespecified endpoints retain their roles; later comparisons do not replace them. Unadjusted secondary intervals and limited training-seed coverage should not be interpreted as a universal simultaneous confidence guarantee.

E Complete direction and comparator results

Table 6 provides the exact direction contrasts underlying Figure 3. Per-target scores and analysis code accompany the paper.

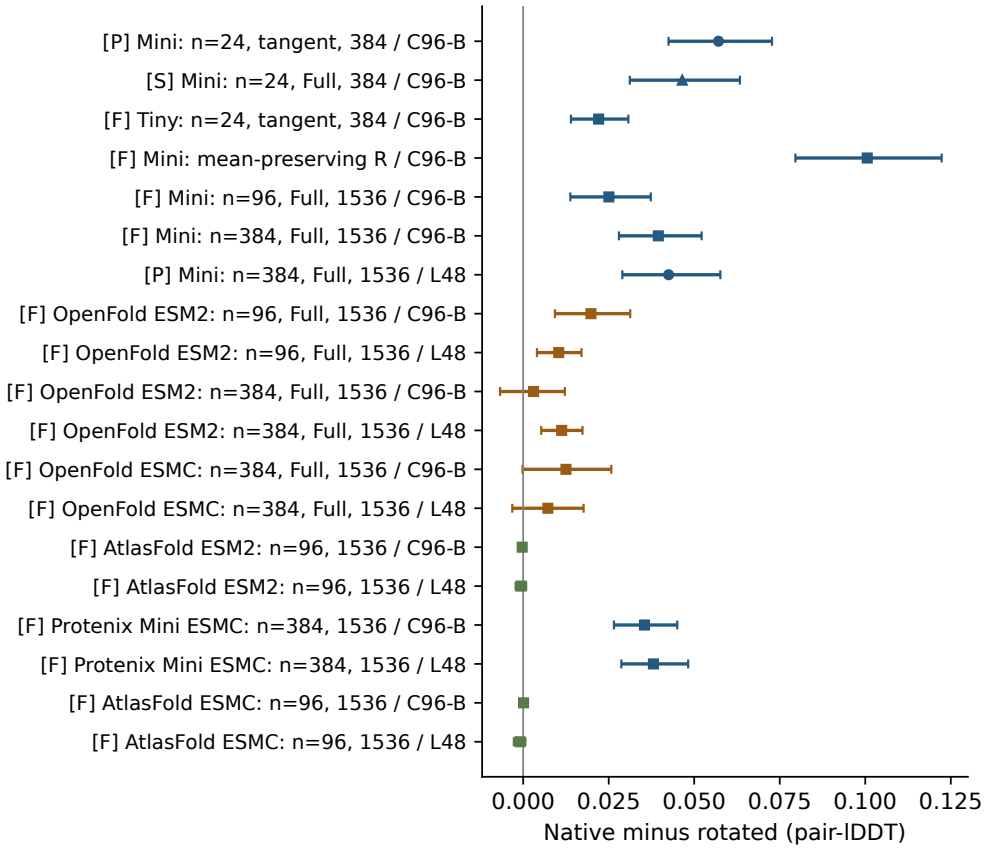

Figure 3: Scope of native-minus-rotation effects. P: primary comparison on a newly locked panel at the time of that study; S: prespecified secondary comparison in the original Protenix direction study; F: subsequent observed-panel study or extension. The mean-preserving row uses Train24/tangent/384. All intervals are conditional target-bootstrap intervals; rows share targets and are not independent replications. Null and negative conditions are retained. OpenFold labels distinguish ESM2 from ESMC features, including subsequent Protenix/AtlasFold comparisons; this figure shows Δ_F, not Ψ.

The original Protenix-Mini tangent comparison gives a Confirm96-B Native–Rotated advantage of +0.05710 [+0.04249, +0.07274], with 78/96 positive targets. Separate comparisons using the Full residual, the Tiny checkpoint's own interface, and mean-preserving rotations also retain a Native–Rotated advantage (Figure 3). These controls neither establish tangent as a better adapter nor isolate all downstream normalization.

Without further training, ESM2 Train384 Full adapters retain a Native–Rotated advantage on Length48 of +0.04252 [+0.02899, +0.05765], with 41/48 positive targets. Native and rotated means are 0.5292 and 0.4867, both above the unadapted mean of 0.3212 (Table 2). Supplementary metrics are concordant. This tests complete-sequence length transfer, not family/pretraining novelty. Official Mini-ESM scores 0.9214, leaving a substantial system-quality gap.

The two Protenix scaling terms are Full Train96/1536 (0.6165 versus 0.5915) and Full Train384/1536 (0.6257 versus 0.5862). Their paired interaction is +0.01446 [+0.00350, +0.02586]; the native training-set gain is only +0.00919 and is not positive for all training seeds. The tangent Train24 result does not enter this interaction.

The original OpenFold Train96/ESM2 interaction has positive seed and rotation marginals on both panels, with 8/9 and 7/9 positive crossed cells. AtlasFold's original ESM2 native–rotation crossed cells are positive in only 3/9 and 1/9 cases. These summaries and target distributions test depen-

Table 6: Direction contrasts by tested setting. P/S/F study identities are as in Figure 3; secondary intervals are unadjusted.

Setting	Study role	Difference	95% interval
Protenix Mini: Train24, tangent, 384 / C96-B	P	+0.05710	$[+0.04249, +0.07274]$
Protenix Mini: Train24, Full, 384 / C96-B	S	+0.04650	$[+0.03118, +0.06336]$
Protenix Tiny: Train24, tangent, 384 / C96-B	F	+0.02206	$[+0.01397, +0.03075]$
Protenix Mini: mean-preserving R / C96-B	F	+0.10059	$[+0.07958, +0.12233]$
Protenix Mini: Train96, Full, 1536 / C96-B	F	+0.02505	$[+0.01378, +0.03730]$
Protenix Mini: Train384, Full, 1536 / C96-B	F	+0.03951	$[+0.02800, +0.05216]$
Protenix Mini: Train384, Full, 1536 / L48	P	+0.04252	$[+0.02899, +0.05765]$
OpenFold ESM2: Train96, Full, 1536 / C96-B	F	+0.01981	$[+0.00928, +0.03129]$
OpenFold ESM2: Train96, Full, 1536 / L48	F	+0.01038	$[+0.00403, +0.01705]$
OpenFold ESM2: Train384, Full, 1536 / C96-B	F	+0.00303	$[-0.00675, +0.01218]$
OpenFold ESM2: Train384, Full, 1536 / L48	F	+0.01126	$[+0.00527, +0.01734]$
OpenFold ESMC: Train384, Full, 1536 / C96-B	F	+0.01251	$[-0.00019, +0.02578]$
OpenFold ESMC: Train384, Full, 1536 / L48	F	+0.00722	$[-0.00318, +0.01766]$
AtlasFold ESM2: Train96, 1536 / C96-B	F	-0.00029	$[-0.00132, +0.00069]$
AtlasFold ESM2: Train96, 1536 / L48	F	-0.00063	$[-0.00221, +0.00087]$
Protenix Mini ESMC: Train384, 1536 / C96-B	F	+0.03547	$[+0.02654, +0.04505]$
Protenix Mini ESMC: Train384, 1536 / L48	F	+0.03808	$[+0.02869, +0.04823]$
AtlasFold ESMC: Train96, 1536 / C96-B	F	+0.00012	$[-0.00103, +0.00128]$
AtlasFold ESMC: Train96, 1536 / L48	F	-0.00094	$[-0.00273, +0.00063]$

dence on individual instances; they neither remove selection bias nor establish family independence. Structural-separation analysis retains its original bins 1–11, 12–23 and at least 24 residues. It describes which reference relationships improve, not recovery of evolutionary covariance or a causal mechanism for the scores.

On Confirm96-B, Factor's difference is +0.01981, while G+'s is +0.00099. The Confirm96-B residue-average Cα lDDT and TM-score interactions are +0.01737 [+0.00571, +0.02996] and +0.01847 [+0.00250, +0.03528]. They score the same structures, not independent samples. Intervals for G+'s Native–Rotated score difference include zero on both panels without establishing training equivalence.

E.1 Matched Protenix Train384 generic-head comparison

The three G+ fits use the same 1536-update order, paired noise levels, frozen representations, learning rate and optimizer as Native. There is no additional G+ learning-rate search. Confirm96-B Native–G+ is the fixed primary endpoint of this observed-panel follow-up; Length48 is secondary. Table 7 retains all three scoring variants. Each metric has positive paired-seed means on both panels. Primary-score target medians are +0.04631 and +0.05966, with 78/96 and 43/48 positive targets, respectively. These counts and medians complement the means without treating metrics or training instances as independent proteins.

G+ itself improves over query-only by +0.17011/+0.12659; it is a functioning explicit-factor comparator under the shared recipe, not an optimized performance bound.

Table 7: Matched Protenix Train384 Native minus G+. Supplementary metric intervals are unadjusted and all intervals condition on the fitted models.

Panel	Metric	Difference	95% interval
C96-B	Pair-lDDT	+0.07630	$[+0.05677, +0.09795]$
C96-B	Residue-lDDT	+0.07184	$[+0.05321, +0.09251]$
C96-B	TM-score	+0.08256	$[+0.05889, +0.10803]$
L48	Pair-lDDT	+0.08138	$[+0.05968, +0.10571]$
L48	Residue-lDDT	+0.07830	$[+0.05731, +0.10164]$
L48	TM-score	+0.09515	$[+0.06814, +0.12610]$

These results establish competitive performance under the shared recipe; they do not estimate a separately tuned generic head's performance limit. This matched-G+ follow-up did not include rotated-

G+ fits, so these comparisons alone do not estimate the Protenix parameterization-by-rotation interaction Ψ. The subsequent complete controls are reported in Appendix K.

F LOCAL COMPATIBILITY AND SHARED-TRANSFER DIAGNOSTICS

At a shared query-only state, let A_x map free factor increments to the first-order update and let g_x be the native task-loss gradient with respect to that update. The diagnostic uses $P_\lambda = A_x(A_x^\top A_x + \lambda I)^{-1}A_x^\top$ and separately records $g_x^\top P_\lambda g_x/\|g_x\|^2$ and $\|P_\lambda g_x\|^2/\|g_x\|^2$. Ridge shrinkage is not an exact orthogonal projection. Equal-norm negative projected-gradient interventions verify local task response with the preceding recycle states fixed. Their first-order response is related by construction to the alignment quantities, so they are not independent mechanisms.

The observed-panel local advantage did not establish a positive association with the previously measured per-target folding gains. This oracle chooses increments using each target's structure supervision, bypassing the shared sequence predictor's learning problem. A later parameter-space differential study encountered responses near downstream numerical resolution, preventing a resolved differential comparison. A separate finite-step Train8-to-Dev8 study produced measurable responses but did not establish mean positive transfer or a native advantage. These outcomes do not establish a mechanism for the trained folding gains.

F.1 REACHABLE VERSUS LEARNED UPDATES

This diagnostic study uses nine paired noise conditions on Observed96. The oracle diagnostic constructs directions at the common query-only state. The learned-update diagnostic evaluates 12 archived Train24/tangent/384 adapters, keeping the first three recycles query-only and changing only the fourth update. For learned residual δ, it compares $\ell(U_q) - \ell(U_q + \eta\delta/\|\delta\|)$, normalized by $\eta\|g\|$, at $\eta = 10^{-3}\|U_q\|$, with the unnormalized actual-scale change $\ell(U_q) - \ell(U_q + \delta)$. These quantities have different units and scales; their difference is not an amplitude attribution.

Table 8: Task-loss diagnostics: native-minus-rotation effects. Small matched-norm learned response is the primary comparison; other rows are secondary. Intervals condition on fitted models. Final-residual retention uses different models and states.

Diagnostic	Difference	95% interval
Oracle direction (normalized)	+0.00134	$[+0.00114, +0.00154]$
Learned, small equal norm (normalized)	+2.17711e-05	$[-3.49056e-05, +7.77411e-05]$
Learned, actual norm (raw loss)	+0.04153	$[+0.03466, +0.04858]$
Final-residual retention (raw loss)	-0.00325	$[-0.00690, +0.00049]$

The mean norm ratios $\|\delta\|/\|U_q\|$ are 0.48283 for Native and 0.68642 for Rotated. Actual-scale response favors Native although its residuals are smaller; this is not evidence that the smaller norm causes the improvement. Relative to the small common amplitude, these average ratios correspond to roughly 483- and 686-fold larger interventions. The retention diagnostic instead uses Train96/Full/1536 on 24 fixed targets with three noises, retaining or removing only the final residual after each model's own preceding adapted states. It finds no established native advantage in that marginal task response, and does not generate free-sampling structures. Neither learned-update nor retention diagnostics identify an explanation of the Train384 structural gains.

At the primary oracle step, 58 of 3456 dynamic finite differences differed by more than 5% from the predicted response; these cases remain included. This numerical comparison concerns oracle directions, not every learned direction.

G FULL-INFERENCE SOURCE–NORM INTERVENTION

On 24 observed Confirm96-B targets, we cross learned residual source and global Frobenius norm for fixed Protenix Train384/Full/1536 models. Write $\delta_a = m_a d_a$, where $m_a = \|\delta_a\|_F$, $\|d_a\|_F = 1$, and $a \in \{N, R\}$. Each trained rotation is paired with the same-seed native model. We inject $m_N d_N$,

Table 9: Complete-inference source–norm intervention on 24 observed targets. Top: absolute Cα pair-lDDT. Bottom: native-minus-rotated source effects at each actual learned norm, and the primary symmetric average, with 95% stratified target-bootstrap intervals. The norm labels identify donors, not universally smaller/larger amplitudes.

Direction source	Native norm m_N	Rotated norm m_R
Native source	0.6373	0.6321
Rotated source	0.5906	0.5924

Source contrast	Difference	95% interval
At native norm	+0.04669	$[+0.02678, +0.06748]$
At rotated norm	+0.03967	$[+0.01948, +0.05978]$
Average source effect	+0.04318	$[+0.02378, +0.06320]$

$m_R d_N$, $m_N d_R$, or $m_R d_R$ throughout c4/s5 inference, scaling completed pair residuals rather than factor increments. Norms use no reference-structure mask. All 720 pure and mixed predictions use the same backend. The mixed residuals are inference-time interventions and need not correspond to outputs of the trained factor head or states encountered during training.

Here, *direction source* includes learned content and relative pair/channel magnitudes after global normalization, unlike the original channel-rotation control. For score Y_{ab} with direction source a and norm source b, the primary effect is

$$E_{\text{dir}} = \tfrac{1}{2}[(Y_{NN} - Y_{RN}) + (Y_{NR} - Y_{RR})]. \tag{13}$$

Whole targets are resampled within four fixed length strata, after averaging paired seeds and rotations; intervals condition on these fitted models.

Native-source updates remain better at both matched norms (Table 9). The average effect is positive for 18/24 targets and all three seed and rotation marginals; residue-average lDDT and TM-score agree in direction. Global norm alone is therefore insufficient to explain this structural gap. The norm-source effect is +0.00173 [-0.00299, +0.00733], and the interaction is +0.00702 [-0.00294, +0.01721]; neither interval excludes zero. These intervals do not establish norm irrelevance or additivity; conditional contrasts are reported below.

Earlier Train24/tangent diagnostics found an oracle-space advantage but no established learned-direction advantage at $10^{-3}\|U_q\|_F$; actual-scale task response favored Native despite smaller mean norms (Appendix F). The model recipe, preceding state, amplitude and endpoint differ here. These results therefore do not form a same-model mechanism chain or identify how training produced the residuals.

The 24 observed Confirm96-B targets are chosen by the fixed hash order, six in each historical length stratum 128–191, 192–255, 256–319 and 320–384. All residuals used for normalization are finite and nonzero. The model is fixed Protenix-Mini Train384/Full/1536; no target supervision enters residual formation, norms or input masks. The completed pair residual is globally scaled before adding it to the unchanged query baseline. Scaling factor increments would change Full's linear and quadratic terms differently and is not used. The static update is replayed on each of four cycles; the subsequent pair states still evolve. Each rotation is paired separately with the same-seed native model.

Besides Eq. 13, define

$$E_{\text{amp}} = \tfrac{1}{2}[(Y_{NN} - Y_{NR}) + (Y_{RN} - Y_{RR})], \tag{14}$$

$$I = Y_{NN} - Y_{RN} - Y_{NR} + Y_{RR}. \tag{15}$$

The identity $E_{\text{dir}} + E_{\text{amp}} = Y_{NN} - Y_{RR}$ is a symmetric allocation under this intervention, not a natural causal percentage. Conditional norm contrasts are $A_{d_N} = Y_{NN} - Y_{NR}$ and $A_{d_R} = Y_{RN} - Y_{RR}$. Table 10 retains both and the supplementary score variants.

All 720 predictions (288 pure, 432 mixed) are computed on MI250 with c4/s5, seed101 and a single sample; pure baselines are not reused from another backend. The primary bootstrap has 20,000

Table 10: Full-inference intervention: primary direction-source effect and secondary norm/interaction contrasts. Supplementary intervals are unadjusted.

Metric	Contrast	Difference	95% interval
Pair-lDDT	Edir	+0.04318	$[+0.02378, +0.06320]$
Pair-lDDT	Eamp	+0.00173	$[-0.00299, +0.00733]$
Pair-lDDT	I	+0.00702	$[-0.00294, +0.01721]$
Pair-lDDT	A_dN	+0.00525	$[-0.00316, +0.01470]$
Pair-lDDT	A_dR	-0.00178	$[-0.00673, +0.00312]$
Residue-lDDT	Edir	+0.04083	$[+0.02221, +0.05991]$
Residue-lDDT	Eamp	+0.00134	$[-0.00320, +0.00682]$
Residue-lDDT	I	+0.00710	$[-0.00257, +0.01705]$
Residue-lDDT	A_dN	+0.00489	$[-0.00325, +0.01423]$
Residue-lDDT	A_dR	-0.00221	$[-0.00691, +0.00241]$
TM-score	Edir	+0.05246	$[+0.02838, +0.07774]$
TM-score	Eamp	+0.00228	$[-0.00349, +0.00863]$
TM-score	I	+0.01040	$[-0.00170, +0.02230]$
TM-score	A_dN	+0.00748	$[-0.00190, +0.01758]$
TM-score	A_dR	-0.00292	$[-0.00997, +0.00447]$

draws, seed20260926, with six targets resampled in each stratum. Inference noise, seed and rotation pairing are retained within a target; the repeated native baseline in three rotation pairs is not three independent observations. The intervals remain conditional on these models and observed proteins.

H NUMERICAL IMPLEMENTATION LIMITS

Backend reruns provide numerical comparisons, not new biological samples. AtlasFold evaluation required recovery of outputs after infrastructure failures; the same final checkpoints and target set were retained, without quality-based selection. Detailed recovery records accompany the artifact.

For matched Protenix G+, Confirm96-B uses MI250 for both sides. Length48 reuses historical H100 Native/Rotated predictions and new MI250 G+ predictions. Short-chain replay was exact, while a 750-residue replay had relative coordinate L2 error 0.00032018. This limited replay does not establish a numerical error bound for the entire long-chain panel.

I HISTORICAL EXPLORATORY BRANCHES

Earlier conditional-memory and feature-supervision explorations motivated the interface audit but are not the source of the present direction claim. The Factor–Generic study predates explicit-factor G+, and local oracle work is separate from trained folding outcomes. These historical studies are retained in the evidence archive with their original roles; they are not additional successful replications of the head-by-rotation interaction.

J OPENFOLD: FROZEN ESMC FEATURE SUBSTITUTION

This study fixes Train384, Full, 1536 updates, three paired training seeds and three rotations. It reuses the 15 historical ESM2 Factor/Rotated Factor/G+ fits and adds nine ESM2 rotated G+ fits plus the complete 24-fit ESMC matrix. The new fits are evaluated on the same 96 Confirm96-B and 48 Length48 targets. This is distinct from the original Train96 four-cell study; its primary endpoint is ESMC Native–Rotated on Confirm96-B, not a newly selected interaction endpoint.

ESMC-600M uses final post-layer-normalization representations from layer 36, width 1152, cached in BF16. The old ESM2 layer-12 cache has width 480 in FP32. The input projection changes from $480 \to 128$ to $1152 \to 128$; all other adapter initial weights are preserved using a separate initialization stream. The added projection weights number $(1152 - 480)128 = 86{,}016$ per head. Final output weights and biases remain zero-initialized. The PLM and AF2 weights stay frozen;

the first live OPM interface, four trunk passes, native loss, AdamW learning rate 10^{-4}, decay 0.01, clipping at 1 and final-step selection remain fixed. No new quality-based learning-rate search is performed. Consequently, this is a feature-source comparison under a common recipe, not an equal-parameter or pure PLM-size intervention. It does not resolve which ESM2 layer is preferable or whether multilayer concatenation helps.

Intervals use 20,000 target-paired bootstrap draws with seed 20260926. Recomputed historical intervals may differ slightly from earlier tables that used another bootstrap stream; point estimates and target scores are unchanged. The three prespecified key secondary tests on Confirm96-B are ESMC's Ψ, $K = \Delta_{F,C} - \Delta_{F,E}$, and $J = \Psi_C - \Psi_E$. Their Holm-adjusted centered-bootstrap p values are 0.73786, 0.62547, and 0.73786, respectively. None establishes the corresponding effect. These are conditional bootstrap tests, not randomization inference; supplementary intervals below are unadjusted.

For pair-lDDT, K is +0.00948 [-0.00499, +0.02480] on Confirm96-B and -0.00404 [-0.01700, +0.00880] on Length48. The corresponding J values are +0.00781 [-0.01050, +0.02654] and -0.00379 [-0.01936, +0.01225]. The supplementary Confirm96-B residue-average lDDT Native–Rotated interval for ESMC is [+0.00004, +0.02521], narrowly above zero without multiplicity correction. It does not replace the primary pair-lDDT endpoint. These direct estimates do not support claiming that feature substitution strengthens or eliminates the rotation effect. ESMC Factor's seed-marginal Native–Rotated differences are negative for one of three seeds on each panel; positive targets number 50/96 and 30/48. The full target arrays and all seed and rotation marginals are retained in the accompanying evidence.

Table 11: Feature-substitution pair-lDDT contrasts. $\Delta_F = F - RF$, $\Delta_{G+} = G + -RG+$, and $\Psi = \Delta_F - \Delta_{G+}$. Negative Native–G+ favors the generic head. Historical point estimates are unchanged; these intervals use the feature study's bootstrap stream.

Panel	Features	Contrast	Difference	95% interval
C96-B	ESM2-35M	Δ_F	+0.00303	$[-0.00656, +0.01218]$
C96-B	ESM2-35M	Δ_{G+}	+0.00335	$[-0.00480, +0.01112]$
C96-B	ESM2-35M	Ψ	-0.00032	$[-0.01194, +0.01157]$
C96-B	ESM2-35M	Native–G+	-0.00648	$[-0.01772, +0.00548]$
C96-B	ESMC-600M	Δ_F	+0.01251	$[-0.00019, +0.02578]$
C96-B	ESMC-600M	Δ_{G+}	+0.00502	$[-0.00616, +0.01555]$
C96-B	ESMC-600M	Ψ	+0.00749	$[-0.00849, +0.02388]$
C96-B	ESMC-600M	Native–G+	-0.00588	$[-0.01997, +0.00905]$
L48	ESM2-35M	Δ_F	+0.01126	$[+0.00531, +0.01734]$
L48	ESM2-35M	Δ_{G+}	+0.00424	$[-0.00069, +0.00929]$
L48	ESM2-35M	Ψ	+0.00702	$[-0.00021, +0.01451]$
L48	ESM2-35M	Native–G+	-0.00703	$[-0.01357, -0.00052]$
L48	ESMC-600M	Δ_F	+0.00722	$[-0.00318, +0.01766]$
L48	ESMC-600M	Δ_{G+}	+0.00399	$[-0.00353, +0.01226]$
L48	ESMC-600M	Ψ	+0.00323	$[-0.00965, +0.01618]$
L48	ESMC-600M	Native–G+	-0.01647	$[-0.02919, -0.00451]$

K PROTENIX AND ATLASFOLD FOUR-CELL FOLLOW-UPS

The Protenix Train96 follow-up adds three Native G+ and nine rotated G+ fits, reusing the 12 fixed Full Factor models. All 12 new fits use 1536 updates, learning rate 5×10^{-5}, weight decay 10^{-4}, global clipping at 1, the original seeds and rotations, and frozen ESM2 layer-12 features. Train96 and Dev8 membership matches the OpenFold experiment. This matches training membership and update count, not backbone loss or computation cost. The study evaluates the already observed Confirm96-B panel, with the direct pair-lDDT interaction as its primary endpoint; no Train96 Length48 factorial is added.

The separate Protenix/AtlasFold feature study adds 33 fits per backbone: nine ESM2 rotated-G+ fits and the complete 24-fit ESMC factorial. Protenix uses Train384 and AtlasFold Train96, both at 1536 updates under the retained backbone-specific recipe. Atlas uses learning rate 10^{-4} and

Table 12: Feature-substitution supplementary residue-average Cα-lDDT contrasts.

Panel	Features	Contrast	Difference	95% interval
C96-B	ESM2-35M	Δ_F	+0.00342	$[-0.00575, +0.01215]$
C96-B	ESM2-35M	Δ_{G+}	+0.00317	$[-0.00453, +0.01058]$
C96-B	ESM2-35M	Ψ	+0.00025	$[-0.01085, +0.01171]$
C96-B	ESM2-35M	Native–G+	-0.00587	$[-0.01664, +0.00565]$
C96-B	ESMC-600M	Δ_F	+0.01235	$[+0.00004, +0.02521]$
C96-B	ESMC-600M	Δ_{G+}	+0.00487	$[-0.00581, +0.01496]$
C96-B	ESMC-600M	Ψ	+0.00748	$[-0.00797, +0.02330]$
C96-B	ESMC-600M	Native–G+	-0.00577	$[-0.01925, +0.00858]$
L48	ESM2-35M	Δ_F	+0.01090	$[+0.00507, +0.01681]$
L48	ESM2-35M	Δ_{G+}	+0.00418	$[-0.00060, +0.00906]$
L48	ESM2-35M	Ψ	+0.00672	$[-0.00035, +0.01405]$
L48	ESM2-35M	Native–G+	-0.00757	$[-0.01408, -0.00112]$
L48	ESMC-600M	Δ_F	+0.00742	$[-0.00220, +0.01704]$
L48	ESMC-600M	Δ_{G+}	+0.00406	$[-0.00317, +0.01195]$
L48	ESMC-600M	Ψ	+0.00336	$[-0.00873, +0.01538]$
L48	ESMC-600M	Native–G+	-0.01556	$[-0.02753, -0.00416]$

Table 13: Feature-substitution supplementary fixed-correspondence TM-score contrasts.

Panel	Features	Contrast	Difference	95% interval
C96-B	ESM2-35M	Δ_F	+0.00056	$[-0.01218, +0.01299]$
C96-B	ESM2-35M	Δ_{G+}	+0.00774	$[-0.00372, +0.01874]$
C96-B	ESM2-35M	Ψ	-0.00718	$[-0.02408, +0.01031]$
C96-B	ESM2-35M	Native–G+	-0.01155	$[-0.02718, +0.00513]$
C96-B	ESMC-600M	Δ_F	+0.01457	$[-0.00226, +0.03179]$
C96-B	ESMC-600M	Δ_{G+}	+0.00526	$[-0.00744, +0.01742]$
C96-B	ESMC-600M	Ψ	+0.00931	$[-0.01078, +0.02957]$
C96-B	ESMC-600M	Native–G+	-0.00786	$[-0.02538, +0.01036]$
L48	ESM2-35M	Δ_F	+0.01012	$[+0.00068, +0.01997]$
L48	ESM2-35M	Δ_{G+}	+0.00009	$[-0.00695, +0.00745]$
L48	ESM2-35M	Ψ	+0.01003	$[-0.00029, +0.02090]$
L48	ESM2-35M	Native–G+	-0.00135	$[-0.01050, +0.00753]$
L48	ESMC-600M	Δ_F	+0.00628	$[-0.01138, +0.02306]$
L48	ESMC-600M	Δ_{G+}	+0.00235	$[-0.01134, +0.01685]$
L48	ESMC-600M	Ψ	+0.00393	$[-0.01909, +0.02599]$
L48	ESMC-600M	Native–G+	-0.01701	$[-0.03915, +0.00353]$

Table 14: Complete head-by-rotation interactions Ψ on observed panels; 95% target-bootstrap intervals condition on fitted models. Confirm96-B Ψ is primary only in the original OpenFold Train96 and Protenix Train96 four-cell follow-ups. The Protenix/AtlasFold feature study's primary is instead ESMC Δ_F on Confirm96-B per backbone; its Ψ entries are follow-up intervals without joint multiplicity correction. Length48 is not new confirmation here. —: not run. These are not direct tests of differences between rows.

Predictor / training set / features	Confirm96-B	Length48
OpenFold 96 / ESM2	+0.01882 $[+0.00662, +0.03209]$	+0.01400 $[+0.00604, +0.02203]$
OpenFold 384 / ESM2	-0.00032 $[-0.01194, +0.01157]$	+0.00702 $[-0.00021, +0.01451]$
OpenFold 384 / ESMC	+0.00749 $[-0.00849, +0.02388]$	+0.00323 $[-0.00965, +0.01618]$
Protenix 96 / ESM2	+0.01833 $[+0.00686, +0.03039]$	—
Protenix 384 / ESM2	+0.03666 $[+0.02502, +0.04931]$	+0.04296 $[+0.02945, +0.05783]$
Protenix 384 / ESMC	+0.02700 $[+0.01549, +0.03883]$	+0.02809 $[+0.01753, +0.03860]$
AtlasFold 96 / ESM2	-0.00052 $[-0.00193, +0.00080]$	-0.00006 $[-0.00204, +0.00178]$
AtlasFold 96 / ESMC	-0.00040 $[-0.00166, +0.00071]$	-0.00045 $[-0.00297, +0.00175]$

decay 0.01; Protenix retains 5×10^{-5} and 10^{-4}. No new quality-based hyperparameter search is introduced. Atlas's native AtlasLM-3B is retained; ESMC replaces only the added adapter features. As in OpenFold, this changes 480-dimensional FP32 ESM2 features to 1152-dimensional BF16 ESMC features and adds 86,016 input-projection weights per head. These are not pure PLM-size or equal-parameter comparisons.

Endpoint identity and uncertainty. For each backbone in the feature study, the prespecified primary is ESMC Factor–Rotated on Confirm96-B. The new Ψ, G+ rotation effects, Native–G+, cross-feature contrasts, Length48 and supplementary metrics retain follow-up roles. Completing four cells does not give all contrasts a new confirmatory identity. The feature study did not prespecify OpenFold's three-test Holm family; that correction is not transferred here. All feature-study intervals are unadjusted for joint multiplicity. Both studies resample whole target records 20,000 times with seed 20260926. The feature study's Length48 analysis resamples 16 targets within each of its three fixed length strata. Historical point estimates are unchanged; re-estimated intervals can differ from older tables using another seed or resampling rule. Neither study replaces Fresh96's unestablished primary interaction.

Instance consistency. Protenix Train96 pair-lDDT Ψ is positive for all three seed and rotation marginals, all nine crossed cells and 61/96 targets. Its residue-lDDT and TM-score interaction intervals are also positive; the third TM-score seed marginal is negative, so main-metric consistency is not generalized to every metric. Protenix Train384 has positive pair-lDDT interaction marginals for both feature sources and panels. Atlas's complete matrices do not establish the interaction. A tiny unadjusted positive Atlas ESMC Confirm96-B G+ rotation interval is retained; this is not a Factor benefit or an interaction. Supplementary metrics reuse the same structures, not independent samples.

For Protenix Train384/ESMC on Confirm96-B, the Native head gap is $F - G^{+} = +0.11131$, while $RF - RG^{+} = +0.08431$ after rotation. Their difference is $\Psi = +0.02700$: a substantial head gap survives rotation. This identity is not a causal percentage decomposition.

Backend limitations. Protenix's ESMC four cells run on MI250. Historical ESM2 Length48 Factor/query predictions run on H100, while its G+ cells run on MI250. Atlas's ESMC seed 20260925 has all eight models on H100; the other two ESMC seeds and all ESM2 models run on MI250. Backend is therefore confounded with that seed, and its ESMC–ESM2 comparison changes backend too. This is not an independent hardware replication. Atlas training required checkpoint recovery and revisions to numerical diagnostics and software compatibility. These changes did not select targets, model arms, final checkpoints, activations or learning rates by their structure scores; the execution history is documented in the artifact.

Table 15: Pair-lDDT interaction marginals. Each entry averages whole targets and the other model index; marginal signs are descriptive, not independent tests. Seeds and rotations retain their fixed order.

Predictor/Train	Panel	Features	Three seed means	Three rotation means
Protenix/96	C96-B	ESM2	+0.02293, +0.02674, +0.00532	+0.01169, +0.02609, +0.01721
Protenix/384	C96-B	ESM2	+0.03451, +0.02648, +0.04899	+0.02420, +0.05634, +0.02943
Protenix/384	C96-B	ESMC	+0.01509, +0.02445, +0.04147	+0.02417, +0.03370, +0.02312
Protenix/384	L48	ESM2	+0.03492, +0.03525, +0.05870	+0.02357, +0.05768, +0.04762
Protenix/384	L48	ESMC	+0.02064, +0.02599, +0.03765	+0.02376, +0.03166, +0.02885
AtlasFold/96	C96-B	ESM2	-0.00073, -0.00115, +0.00030	-0.00056, -0.00047, -0.00054
AtlasFold/96	C96-B	ESMC	-0.00035, +0.00070, -0.00156	-0.00076, +0.00063, -0.00108
AtlasFold/96	L48	ESM2	-0.00024, -0.00119, +0.00124	-0.00020, +0.00009, -0.00009
AtlasFold/96	L48	ESMC	+0.00046, -0.00212, +0.00031	+0.00012, -0.00125, -0.00023

Table 16: Protenix four-cell comparisons: pair-lDDT contrasts. $\Delta_F = F - RF$, $\Delta_{G+} = G + -RG+$ and $\Psi = \Delta_F - \Delta_{G+}$. Endpoint roles and unadjusted intervals are as specified above.

Train/features	Panel	Contrast	Difference	95% interval
96/ESM2	C96-B	Δ_F	+0.02505	$[+0.01378, +0.03730]$
96/ESM2	C96-B	Δ_{G+}	+0.00672	$[+0.00029, +0.01342]$
96/ESM2	C96-B	Ψ	+0.01833	$[+0.00686, +0.03039]$
96/ESM2	C96-B	F–G+	+0.06086	$[+0.04455, +0.07855]$
384/ESM2	C96-B	Δ_F	+0.03951	$[+0.02800, +0.05216]$
384/ESM2	C96-B	Δ_{G+}	+0.00285	$[-0.00168, +0.00726]$
384/ESM2	C96-B	Ψ	+0.03666	$[+0.02502, +0.04931]$
384/ESM2	C96-B	F–G+	+0.07630	$[+0.05677, +0.09795]$
384/ESMC	C96-B	Δ_F	+0.03547	$[+0.02654, +0.04505]$
384/ESMC	C96-B	Δ_{G+}	+0.00847	$[+0.00093, +0.01616]$
384/ESMC	C96-B	Ψ	+0.02700	$[+0.01549, +0.03883]$
384/ESMC	C96-B	F–G+	+0.11131	$[+0.09418, +0.12871]$
384/ESM2	L48	Δ_F	+0.04252	$[+0.02906, +0.05696]$
384/ESM2	L48	Δ_{G+}	-0.00043	$[-0.00564, +0.00454]$
384/ESM2	L48	Ψ	+0.04296	$[+0.02945, +0.05783]$
384/ESM2	L48	F–G+	+0.08138	$[+0.06032, +0.10449]$
384/ESMC	L48	Δ_F	+0.03808	$[+0.02869, +0.04823]$
384/ESMC	L48	Δ_{G+}	+0.00999	$[+0.00274, +0.01786]$
384/ESMC	L48	Ψ	+0.02809	$[+0.01753, +0.03860]$
384/ESMC	L48	F–G+	+0.16397	$[+0.13799, +0.19110]$

Table 17: Protenix four-cell comparisons: residue-average Cα-lDDT contrasts. $\Delta_F = F - RF$, $\Delta_{G+} = G + -RG+$ and $\Psi = \Delta_F - \Delta_{G+}$. Endpoint roles and unadjusted intervals are as specified above.

Train/features	Panel	Contrast	Difference	95% interval
96/ESM2	C96-B	Δ_F	+0.02381	$[+0.01286, +0.03570]$
96/ESM2	C96-B	Δ_{G+}	+0.00609	$[-0.00002, +0.01239]$
96/ESM2	C96-B	Ψ	+0.01772	$[+0.00655, +0.02943]$
96/ESM2	C96-B	F–G+	+0.05842	$[+0.04269, +0.07558]$
384/ESM2	C96-B	Δ_F	+0.03711	$[+0.02609, +0.04919]$
384/ESM2	C96-B	Δ_{G+}	+0.00288	$[-0.00159, +0.00717]$
384/ESM2	C96-B	Ψ	+0.03423	$[+0.02322, +0.04623]$
384/ESM2	C96-B	F–G+	+0.07184	$[+0.05321, +0.09251]$
384/ESMC	C96-B	Δ_F	+0.03413	$[+0.02559, +0.04326]$
384/ESMC	C96-B	Δ_{G+}	+0.00827	$[+0.00094, +0.01579]$
384/ESMC	C96-B	Ψ	+0.02586	$[+0.01484, +0.03718]$
384/ESMC	C96-B	F–G+	+0.10736	$[+0.09125, +0.12375]$
384/ESM2	L48	Δ_F	+0.04029	$[+0.02735, +0.05417]$
384/ESM2	L48	Δ_{G+}	-0.00070	$[-0.00575, +0.00403]$
384/ESM2	L48	Ψ	+0.04100	$[+0.02788, +0.05550]$
384/ESM2	L48	F–G+	+0.07830	$[+0.05783, +0.10059]$
384/ESMC	L48	Δ_F	+0.03557	$[+0.02665, +0.04520]$
384/ESMC	L48	Δ_{G+}	+0.00956	$[+0.00262, +0.01711]$
384/ESMC	L48	Ψ	+0.02601	$[+0.01597, +0.03598]$
384/ESMC	L48	F–G+	+0.15631	$[+0.13146, +0.18216]$

Table 18: Protenix four-cell comparisons: fixed-correspondence TM-score contrasts. $\Delta_F = F - RF$, $\Delta_{G+} = G + -RG+$ and $\Psi = \Delta_F - \Delta_{G+}$. Endpoint roles and unadjusted intervals are as specified above.

Train/features	Panel	Contrast	Difference	95% interval
96/ESM2	C96-B	Δ_F	+0.02512	$[+0.01177, +0.03899]$
96/ESM2	C96-B	Δ_{G+}	+0.01007	$[+0.00096, +0.01966]$
96/ESM2	C96-B	Ψ	+0.01506	$[+0.00043, +0.02978]$
96/ESM2	C96-B	F–G+	+0.06117	$[+0.04171, +0.08197]$
384/ESM2	C96-B	Δ_F	+0.04051	$[+0.02660, +0.05528]$
384/ESM2	C96-B	Δ_{G+}	+0.00234	$[-0.00447, +0.00902]$
384/ESM2	C96-B	Ψ	+0.03817	$[+0.02393, +0.05321]$
384/ESM2	C96-B	F–G+	+0.08256	$[+0.05889, +0.10803]$
384/ESMC	C96-B	Δ_F	+0.03400	$[+0.02292, +0.04608]$
384/ESMC	C96-B	Δ_{G+}	+0.00802	$[-0.00242, +0.01840]$
384/ESMC	C96-B	Ψ	+0.02598	$[+0.01044, +0.04257]$
384/ESMC	C96-B	F–G+	+0.10958	$[+0.08681, +0.13321]$
384/ESM2	L48	Δ_F	+0.04782	$[+0.03056, +0.06662]$
384/ESM2	L48	Δ_{G+}	+0.00015	$[-0.00811, +0.00901]$
384/ESM2	L48	Ψ	+0.04766	$[+0.03070, +0.06721]$
384/ESM2	L48	F–G+	+0.09515	$[+0.06901, +0.12459]$
384/ESMC	L48	Δ_F	+0.03853	$[+0.02743, +0.05030]$
384/ESMC	L48	Δ_{G+}	+0.01167	$[-0.00016, +0.02466]$
384/ESMC	L48	Ψ	+0.02686	$[+0.01046, +0.04271]$
384/ESMC	L48	F–G+	+0.19642	$[+0.16152, +0.23287]$

Table 19: AtlasFold four-cell comparisons: pair-lDDT contrasts. $\Delta_F = F - RF$, $\Delta_{G+} = G + -RG+$ and $\Psi = \Delta_F - \Delta_{G+}$. Endpoint roles and unadjusted intervals are as specified above.

Train/features	Panel	Contrast	Difference	95% interval
96/ESM2	C96-B	Δ_F	-0.00029	$[-0.00132, +0.00067]$
96/ESM2	C96-B	Δ_{G+}	+0.00024	$[-0.00045, +0.00106]$
96/ESM2	C96-B	Ψ	-0.00052	$[-0.00193, +0.00080]$
96/ESM2	C96-B	F–G+	-0.00080	$[-0.00213, +0.00045]$
96/ESMC	C96-B	Δ_F	+0.00012	$[-0.00103, +0.00128]$
96/ESMC	C96-B	Δ_{G+}	+0.00052	$[+0.00003, +0.00106]$
96/ESMC	C96-B	Ψ	-0.00040	$[-0.00166, +0.00071]$
96/ESMC	C96-B	F–G+	-0.00040	$[-0.00163, +0.00085]$
96/ESM2	L48	Δ_F	-0.00063	$[-0.00220, +0.00089]$
96/ESM2	L48	Δ_{G+}	-0.00057	$[-0.00169, +0.00037]$
96/ESM2	L48	Ψ	-0.00006	$[-0.00204, +0.00178]$
96/ESM2	L48	F–G+	-0.00048	$[-0.00297, +0.00187]$
96/ESMC	L48	Δ_F	-0.00094	$[-0.00273, +0.00063]$
96/ESMC	L48	Δ_{G+}	-0.00049	$[-0.00191, +0.00085]$
96/ESMC	L48	Ψ	-0.00045	$[-0.00297, +0.00175]$
96/ESMC	L48	F–G+	-0.00157	$[-0.00423, +0.00055]$

Table 20: AtlasFold four-cell comparisons: residue-average Cα-lDDT contrasts. $\Delta_F = F - RF$, $\Delta_{G+} = G + -RG+$ and $\Psi = \Delta_F - \Delta_{G+}$. Endpoint roles and unadjusted intervals are as specified above.

Train/features	Panel	Contrast	Difference	95% interval
96/ESM2	C96-B	Δ_F	-0.00027	$[-0.00144, +0.00085]$
96/ESM2	C96-B	Δ_{G+}	+0.00023	$[-0.00051, +0.00110]$
96/ESM2	C96-B	Ψ	-0.00050	$[-0.00203, +0.00099]$
96/ESM2	C96-B	F–G+	-0.00093	$[-0.00234, +0.00041]$
96/ESMC	C96-B	Δ_F	+0.00009	$[-0.00111, +0.00136]$
96/ESMC	C96-B	Δ_{G+}	+0.00054	$[-0.00001, +0.00112]$
96/ESMC	C96-B	Ψ	-0.00045	$[-0.00172, +0.00072]$
96/ESMC	C96-B	F–G+	-0.00054	$[-0.00180, +0.00076]$
96/ESM2	L48	Δ_F	-0.00064	$[-0.00213, +0.00080]$
96/ESM2	L48	Δ_{G+}	-0.00049	$[-0.00154, +0.00042]$
96/ESM2	L48	Ψ	-0.00015	$[-0.00202, +0.00161]$
96/ESM2	L48	F–G+	-0.00046	$[-0.00288, +0.00181]$
96/ESMC	L48	Δ_F	-0.00110	$[-0.00280, +0.00034]$
96/ESMC	L48	Δ_{G+}	-0.00049	$[-0.00183, +0.00078]$
96/ESMC	L48	Ψ	-0.00061	$[-0.00300, +0.00143]$
96/ESMC	L48	F–G+	-0.00150	$[-0.00407, +0.00055]$

Table 21: AtlasFold four-cell comparisons: fixed-correspondence TM-score contrasts. $\Delta_F = F - RF$, $\Delta_{G+} = G + -RG+$ and $\Psi = \Delta_F - \Delta_{G+}$. Endpoint roles and unadjusted intervals are as specified above.

Train/features	Panel	Contrast	Difference	95% interval
96/ESM2	C96-B	Δ_F	-0.00131	$[-0.00318, +0.00040]$
96/ESM2	C96-B	Δ_{G+}	+0.00017	$[-0.00049, +0.00090]$
96/ESM2	C96-B	Ψ	-0.00148	$[-0.00363, +0.00044]$
96/ESM2	C96-B	F–G+	-0.00093	$[-0.00320, +0.00124]$
96/ESMC	C96-B	Δ_F	-0.00048	$[-0.00287, +0.00163]$
96/ESMC	C96-B	Δ_{G+}	+0.00106	$[+0.00022, +0.00211]$
96/ESMC	C96-B	Ψ	-0.00154	$[-0.00420, +0.00068]$
96/ESMC	C96-B	F–G+	-0.00128	$[-0.00379, +0.00068]$
96/ESM2	L48	Δ_F	-0.00152	$[-0.00384, +0.00058]$
96/ESM2	L48	Δ_{G+}	-0.00253	$[-0.00812, +0.00082]$
96/ESM2	L48	Ψ	+0.00101	$[-0.00342, +0.00635]$
96/ESM2	L48	F–G+	-0.00186	$[-0.00497, +0.00083]$
96/ESMC	L48	Δ_F	-0.00135	$[-0.00413, +0.00096]$
96/ESMC	L48	Δ_{G+}	-0.00388	$[-0.01114, +0.00066]$
96/ESMC	L48	Ψ	+0.00253	$[-0.00334, +0.01051]$
96/ESMC	L48	F–G+	-0.00177	$[-0.00616, +0.00194]$

Table 22: Direct ESMC-minus-ESM2 changes, pair-lDDT. These paired follow-up intervals are unadjusted; backend qualifications above also apply.

Predictor	Panel	Contrast	Difference	95% interval
Protenix	C96-B	$\Delta_{F,C} - \Delta_{F,E}$	-0.00404	$[-0.01998, +0.01119]$
Protenix	C96-B	$\Delta_{G+,C} - \Delta_{G+,E}$	+0.00562	$[-0.00243, +0.01362]$
Protenix	C96-B	$\Psi_C - \Psi_E$	-0.00966	$[-0.02707, +0.00686]$
Protenix	C96-B	Native gain	+0.20424	$[+0.17425, +0.23545]$
Protenix	C96-B	G+ gain	+0.16924	$[+0.14552, +0.19417]$
Protenix	C96-B	Change in F–G+	+0.03501	$[+0.01103, +0.05769]$
Protenix	L48	$\Delta_{F,C} - \Delta_{F,E}$	-0.00445	$[-0.02185, +0.01242]$
Protenix	L48	$\Delta_{G+,C} - \Delta_{G+,E}$	+0.01042	$[+0.00239, +0.01885]$
Protenix	L48	$\Psi_C - \Psi_E$	-0.01486	$[-0.03566, +0.00464]$
Protenix	L48	Native gain	+0.22733	$[+0.19772, +0.25697]$
Protenix	L48	G+ gain	+0.14474	$[+0.12176, +0.16759]$
Protenix	L48	Change in F–G+	+0.08259	$[+0.05298, +0.11278]$
AtlasFold	C96-B	$\Delta_{F,C} - \Delta_{F,E}$	+0.00040	$[-0.00060, +0.00148]$
AtlasFold	C96-B	$\Delta_{G+,C} - \Delta_{G+,E}$	+0.00028	$[-0.00046, +0.00096]$
AtlasFold	C96-B	$\Psi_C - \Psi_E$	+0.00012	$[-0.00121, +0.00150]$
AtlasFold	C96-B	Native gain	+0.00040	$[-0.00079, +0.00160]$
AtlasFold	C96-B	G+ gain	-0.00001	$[-0.00100, +0.00079]$
AtlasFold	C96-B	Change in F–G+	+0.00040	$[-0.00124, +0.00223]$
AtlasFold	L48	$\Delta_{F,C} - \Delta_{F,E}$	-0.00030	$[-0.00164, +0.00113]$
AtlasFold	L48	$\Delta_{G+,C} - \Delta_{G+,E}$	+0.00009	$[-0.00149, +0.00186]$
AtlasFold	L48	$\Psi_C - \Psi_E$	-0.00039	$[-0.00296, +0.00218]$
AtlasFold	L48	Native gain	-0.00098	$[-0.00291, +0.00072]$
AtlasFold	L48	G+ gain	+0.00010	$[-0.00125, +0.00171]$
AtlasFold	L48	Change in F–G+	-0.00108	$[-0.00400, +0.00162]$

L PAIRED COMPETITIVENESS COMPARISONS

Table 23: Native minus G+: paired follow-up comparisons. Negative values favor G+. Intervals condition on fitted models. Protenix Confirm96-B is the primary endpoint of its matched-G+ follow-up; Length48 is secondary. Other rows are supplementary comparisons; intervals are not jointly multiplicity-adjusted. All panels were already observed for these studies. These estimates address method competitiveness, separately from Native–Rotated direction effects.

Backbone / recipe	Panel	Difference	95% interval
Protenix, Train384	C96-B	+0.07630	$[+0.05677, +0.09795]$
Protenix, Train384	L48	+0.08138	$[+0.05968, +0.10571]$
OpenFold, Train96	C96-B	+0.00656	$[-0.00515, +0.01845]$
OpenFold, Train96	L48	-0.00977	$[-0.01674, -0.00288]$
OpenFold, Train384	C96-B	-0.00648	$[-0.01777, +0.00537]$
OpenFold, Train384	L48	-0.00703	$[-0.01369, -0.00039]$
AtlasFold, Train96	C96-B	-0.00080	$[-0.00213, +0.00044]$
AtlasFold, Train96	L48	-0.00048	$[-0.00302, +0.00191]$

M FRESH96: FIXED-MODEL VALIDATION ON NEW TARGETS

This study reused the 24 OpenFold Train96/ESM2-35M adapters and query-only baseline without further training. Models were fixed before target selection; the execution recipe and analysis were specified before new predictions. The score grid contains all $25 \times 96 = 2,400$ model–target pairs.

Selection and its revision. Fresh96 contains four prespecified short-chain length strata with 24 targets each, with actual lengths 129–384. It uses the short-chain eligibility criteria (Q) in Table 5, one target per PDB, and deterministic hash ordering independent of model scores. Recorded training, development, evaluation and engineering exposures were excluded, including query/PDB identities and sequences. The historical teacher exclusion library was retained. An initial global-alignment rule failed to fill the first stratum and rejected all eight fixed shuffled negative controls, motivating the revision below.

Before any new folding prediction, selection v2 adopted the previously calibrated BLASTP 2.17.0+ rule used for Length48, without changing the models, 96-target quota, deterministic order or primary endpoint. Searches use BLOSUM62, gap-open 11/extend 1, word size 3, SEG and soft masking, composition adjustment 2, search E-value 10^{-3} and no HSP-count truncation. Exclusion requires one HSP with E-value $\leq 10^{-5}$, identity ≥ 0.30 over alignment columns including internal gaps, at least 50 paired residues and coverage ≥ 0.70. Coverage uses the shorter sequence for real queries/panel neighbors and the candidate length for historical teacher rows; HSPs are not merged. No known ID, PDB or exact-sequence overlap remains with the recorded exclusions. This is sequence-screened new-target validation relative to documented project exposure, not the identical Confirm96-B screening implementation, strict family isolation or exclusion from foundation-model pretraining.

Endpoint and complete reporting. The sole primary endpoint was pair-lDDT Ψ, using 20,000 whole-target paired bootstrap draws with seed 20260921. It is not established: the point estimate is positive but its interval includes zero. For the primary metric, 54/96 target interactions and 8/9 seed–rotation cells are positive; one seed marginal is negative (Table 24). The row-wise Factor contrast is positive, but cannot replace the primary endpoint. The other metrics, lengths and row-wise comparisons are secondary and unadjusted. The new panel is not pooled with the observed panels, and no targets are added in response to its result. Observed-panel target interactions remain positive for 61/96 and 34/48 targets; their narrower positive intervals do not establish a difference between panels.

Table 24: Fresh96 primary-interaction marginals in locked seed/rotation order. The three training seeds are 20260923/24/25; rotation seeds are 20261001/02/03.

Marginal	First	Second	Third
Seed	-0.00139	+0.01979	+0.01203
Rotation	+0.00333	+0.01416	+0.01293

Table 25: Fresh96 contrasts from complete saved scores. Only pair-lDDT interaction is primary. All other intervals are unadjusted secondary summaries, conditional on the fixed fitted models.

Metric	Contrast	Mean	95% interval
Pair-lDDT	factor_rotation	+0.01240	$[+0.00278, +0.02218]$
Pair-lDDT	gplus_rotation	+0.00226	$[-0.00606, +0.01018]$
Pair-lDDT	interaction	+0.01014	$[-0.00318, +0.02347]$
Pair-lDDT	native_minus_gplus	-0.00146	$[-0.01235, +0.00941]$
Pair-lDDT	factor_minus_query	+0.19817	$[+0.16864, +0.22779]$
Pair-lDDT	generic_plus_minus_query	+0.19963	$[+0.16964, +0.22962]$
Residue-lDDT	factor_rotation	+0.01148	$[+0.00237, +0.02075]$
Residue-lDDT	gplus_rotation	+0.00177	$[-0.00635, +0.00950]$
Residue-lDDT	interaction	+0.00971	$[-0.00312, +0.02252]$
Residue-lDDT	native_minus_gplus	-0.00154	$[-0.01208, +0.00895]$
Residue-lDDT	factor_minus_query	+0.18764	$[+0.15949, +0.21603]$
Residue-lDDT	generic_plus_minus_query	+0.18917	$[+0.16046, +0.21798]$
TM-score	factor_rotation	+0.01289	$[+0.00016, +0.02562]$
TM-score	gplus_rotation	+0.00412	$[-0.00562, +0.01361]$
TM-score	interaction	+0.00877	$[-0.00828, +0.02585]$
TM-score	native_minus_gplus	-0.00472	$[-0.01859, +0.00896]$
TM-score	factor_minus_query	+0.21904	$[+0.18312, +0.25465]$
TM-score	generic_plus_minus_query	+0.22376	$[+0.18782, +0.25968]$

N POST-HOC TRAINING-SETTING CHANGE OF THE COMPLETE INTERACTION

For the separate Factor-only contrast, the fixed-budget change $(N-R)_{384}-(N-R)_{96}$ is -0.01678 [-0.03214, -0.00261] on Confirm96-B and +0.00087 [-0.00728, +0.00863] on Length48. These quantities do not equal the change of the full interaction below.

After the ESM2 Train384 four-cell control was complete, a subsequent analysis computed $q_i = \Psi_{384,i} - \Psi_{96,i}$ for every shared target, pairing the same three seeds and three rotations. The analysis uses 20,000 paired whole-target bootstrap draws (seed 20260924), retaining all model cells for each sampled target. These are exploratory, unadjusted intervals, not differences between two marginal confidence intervals or new confirmatory endpoints.

Table 26: Direct ESM2 interaction change, Train384 minus Train96, under the same 1536-update budget. All six metric/panel combinations are retained.

Panel	Metric	$\Delta\Psi_{\mathrm{train}}$	95% interval
C96-B	Pair-lDDT	-0.01914	$[-0.03759, -0.00168]$
L48	Pair-lDDT	-0.00698	$[-0.01673, +0.00259]$
C96-B	Residue-lDDT	-0.01712	$[-0.03448, -0.00061]$
L48	Residue-lDDT	-0.00660	$[-0.01611, +0.00263]$
C96-B	TM-score	-0.02565	$[-0.05176, -0.00016]$
L48	TM-score	-0.00990	$[-0.02681, +0.00735]$

This estimand differs from the previously reported training change in the Factor-only contrast Δ_F. The pair-lDDT interaction decreases on Confirm96-B under this analysis; the Length48 interval includes zero. Neither result isolates dataset size: coverage, membership and repetition counts change together, and convergence at 1536 updates has not been established.

O A PROSPECTIVE RECONSTRUCTION-BASED PREDICTION TEST

This subsequent study tests a candidate predictor of the post-training Native–Rotated score difference. It fixes OpenFold/Train96/ESM2/Full, 1536 task updates and Confirm96-B, an already observed panel. Eight new rotations were fixed before diagnostic measurements; all preserve the all-ones channel direction and are conjugates of one fixed matrix in $SO(127)$ on its orthogonal complement. Thus their spectra and distance from identity are matched. They are not the three historical rotations used in the main four-cell comparison.

The primary predictor X_R is the held-out relative squared error when a shared rotated Factor student reconstructs a trained Native teacher's updates. Three historical Native teachers generate reconstruction targets on the same fixed query-only anchor trajectories. For each teacher and rotation, one student starts from that teacher's adapter weights with the added rotation and is fitted on Train24 for 768 Adam updates (Kingma & Ba, 2015) (learning rate 10^{-3}, no decay, global clip 1), with 2048 sampled pairs per target. The full sequence is encoded before pair sampling, and the same residue increments are used in all four rounds; only the supplied anchors vary. The locked final-node error averages equally over Holdout72 targets, rounds and teachers, using full-pair teacher energy and a training-side stabilizer. Holdout72 is the remainder of Train96, held out from reconstruction fitting, not from the original teacher's task training. Zero-output errors, student/teacher norm ratios, angles and absolute RMS were recorded; reduced error alone does not establish useful compensation.

The exploratory predictor D_R instead measures the Dev8 score loss when an extra R is applied to a fixed trained Native residual during full inference. All X_R, D_R, rankings and the low/high-X selections were locked before the new structure-training runs. Each of the eight rotated configurations was then trained from the paired historical encoder initialization and zero output layer for three seeds; reconstruction students were not used as initializers. The outcome Y_R is the mean Native-minus-retrained-Rotated pair-lDDT over the 96 targets and three seeds. The eight rotations, rather than proteins or fits, are the units for this prediction test.

Table 27: All eight locked rotations in the reconstruction prediction study. The sole primary association is between X and Y for pair-lDDT. The target intervals for Y condition on each fitted rotation and models.

Rotation	X	D	Rotated score	Y	95% target interval
r20270101	0.84137	+0.24557	0.4858	+0.01614	$[+0.00340, +0.02966]$
r20270102	0.82713	+0.20338	0.4916	+0.01026	$[-0.00356, +0.02414]$
r20270103	0.84164	+0.21774	0.4826	+0.01926	$[+0.00466, +0.03378]$
r20270104	0.84826	+0.18758	0.4938	+0.00811	$[-0.00487, +0.02115]$
r20270105	0.82295	+0.17191	0.5123	-0.01039	$[-0.02132, +0.00020]$
r20270106	0.81867	+0.16705	0.4871	+0.01483	$[+0.00199, +0.02823]$
r20270107	0.81058	+0.16573	0.4831	+0.01880	$[+0.00565, +0.03311]$
r20270108	0.82145	+0.19382	0.5026	-0.00074	$[-0.01425, +0.01259]$

The sole primary test is a positive Spearman association between X_R and Y_R, with eight rotations as units and all $8! = 40{,}320$ permutations. It gives $\rho = +0.02381$ and one-sided $p = 0.48839$. The exploratory D test gives $\rho = +0.19048$, $p = 0.33229$ and cannot replace the primary predictor. The X span is 0.03767, so the diagnostic varied across rotations. Its large absolute error reflects partial, finite-budget reconstruction, not a lower bound on representational error. The prediction hypothesis was not supported.

Six of the eight Y_R point estimates are positive and two are negative, ranging from -0.01039 to +0.01926. Neither negative estimate has an interval excluding zero. An average advantage over selected rotations does not imply superiority to every rotation instance, even within this controlled family and fixed training recipe.

For the primary outcome, the three seed-specific X correlations are -0.21429, +0.42857 and -0.04762; leaving one teacher out gives -0.04762, +0.04762 and +0.00000. Synchronous whole-target resampling (20,000 draws, seed 20260924) gives a conditional correlation interval of [-0.35714, +0.40476]. This sensitivity analysis does not turn the 96 proteins or 24 fits into independent prediction units. The eight rotations limit power, but neither the near-zero main estimate nor these

Table 28: Exact one-sided association tests on pair-lDDT. X is primary; D is exploratory, without joint multiplicity correction. Both use the same eight rotations. Supplementary structure-score arrays are retained in the artifact.

Outcome	Predictor	Spearman ρ	Exact p
Pair-lDDT	X (primary diagnostic)	+0.02381	0.48839
Pair-lDDT	D (exploratory)	+0.19048	0.33229

sensitivities provide a stable positive ranking. No new diagnostic, rotation or training budget was selected to rescue the prediction.

The result constrains this specific reconstruction diagnostic, not the observed direction effects or all possible measures of compensation. Reproducing one teacher's residual is a different objective from finding any residual that produces a useful structure; the null association does not identify which mismatch or optimization limitation matters. It also does not validate D as an alternative predictor. A separately locked compensation intervention remains a distinct question; it cannot retroactively establish this failed predictive link.

P SIGNED-PERMUTATION CONTROL UNDER THE FIXED ADAMW RECIPE

This observed-panel OpenFold/Train96/ESM2/Full/1536 follow-up trains three signed transforms across three seeds in Factor and G+, giving 18 fits, and reuses the six original Native/G+ fits and query outputs. Confirm96-B pair-lDDT Ψ is primary; Length48 and other metrics are secondary and unadjusted, not new-target confirmation.

Each transform is $(Tx)_j = s_j x_{\pi(j)}$ with $s_j \in \{-1, +1\}$ and fixed permutation π. The three matrices have 63, 63 and 55 negative signs, respectively, so $T\mathbf{1} \neq \mathbf{1}$: they are not pure permutations or mean-preserving controls. The transform is applied to the complete residual by gather-and-sign, leaving the baseline unchanged. Historical dense Factor rotations were fused into the decoder; numerical implementations are not claimed identical across these transformation families.

For G+'s free affine output, $W_T = T^\top W$ and $b_T = T^\top b$ yield the same function. In ideal arithmetic, a signed coordinate permutation also maps AdamW's first moment with signs and permutation, and its second moment by the unsigned permutation. With matched random conditions, parameter groups, scalar ϵ, isotropic decay and global L2 clipping, this coordinate mapping is compatible with the update. It differs from arbitrary dense rotations. A strict correspondence of complete FP32 training trajectories was not established. Non-significant G+ differences are not equivalence tests, nor do they rule out optimization or numerical contributions to Factor's behavior.

Table 29: Signed-permutation four-cell pair-lDDT means. RF/RG+ here denote the signed-transform arms; they are distinct from the historical dense controls.

Panel	Query	F	RF	G+	RG+
C96-B	0.3019	0.5019	0.4879	0.4953	0.4957
L48	0.2608	0.3887	0.3806	0.3985	0.3933

Table 30: All signed-transform contrasts, with original 20,000-draw target intervals (seed 20260921). These condition on fixed models and transforms.

Panel	Metric	Contrast	Mean	95% interval
C96-B	Pair-lDDT	Δ_F	+0.01405	$[+0.00465, +0.02372]$
C96-B	Pair-lDDT	Δ_{G+}	-0.00035	$[-0.00698, +0.00654]$
C96-B	Pair-lDDT	Ψ	+0.01440	$[+0.00379, +0.02520]$
C96-B	Residue-lDDT	Δ_F	+0.01303	$[+0.00410, +0.02217]$
C96-B	Residue-lDDT	Δ_{G+}	+0.00004	$[-0.00626, +0.00657]$
C96-B	Residue-lDDT	Ψ	+0.01300	$[+0.00296, +0.02312]$
C96-B	TM-score	Δ_F	+0.01452	$[+0.00273, +0.02657]$
C96-B	TM-score	Δ_{G+}	-0.00034	$[-0.00840, +0.00823]$
C96-B	TM-score	Ψ	+0.01486	$[+0.00160, +0.02826]$
L48	Pair-lDDT	Δ_F	+0.00808	$[+0.00204, +0.01409]$
L48	Pair-lDDT	Δ_{G+}	+0.00516	$[-0.00211, +0.01268]$
L48	Pair-lDDT	Ψ	+0.00292	$[-0.00622, +0.01206]$
L48	Residue-lDDT	Δ_F	+0.00805	$[+0.00231, +0.01383]$
L48	Residue-lDDT	Δ_{G+}	+0.00436	$[-0.00255, +0.01148]$
L48	Residue-lDDT	Ψ	+0.00368	$[-0.00495, +0.01241]$
L48	TM-score	Δ_F	+0.01059	$[+0.00017, +0.02100]$
L48	TM-score	Δ_{G+}	+0.00716	$[-0.00256, +0.01689]$
L48	TM-score	Ψ	+0.00344	$[-0.00942, +0.01633]$

The Confirm96-B interaction is positive in all seed and transform marginals and eight of nine individual cells; Length48's interaction is not established. Sensitivity is therefore not restricted to dense mixing, without identifying expressivity as its sole cause. The artifact retains every target, system and marginal; supplementary metrics need not share all signs.

Q DEVELOPMENTAL STRUCTURE CURVES AT SAVED CHECKPOINTS

Without further training, the 48 OpenFold/ESM2 fits from Train96 and Train384 were evaluated at saved 384, 768 and 1536 updates on the same eight Dev targets. The query baseline is shared across nodes. These are observed-development diagnostics, not a new confirmation panel or a basis for selecting the final checkpoint. The final-1536 rule in the structural comparisons is unchanged.

Table 31: Dev8 pair-lDDT means at the fixed saved checkpoints. The same targets are used at every node; three seeds and three rotations are aggregated within targets.

Train size	Updates	F	RF	G+	RG+
96	384	0.4891	0.4803	0.4615	0.4417
96	768	0.5390	0.5101	0.5507	0.5434
96	1536	0.5413	0.5500	0.5527	0.5572
384	384	0.4915	0.4887	0.4865	0.4672
384	768	0.5445	0.5155	0.5282	0.5258
384	1536	0.5660	0.5577	0.5599	0.5572

Table 32: Saved-checkpoint contrasts. All three metrics and all nodes are shown. Ψ intervals use the original synchronous target bootstrap and are unadjusted, conditional on fitted models; Dev8 does not supply precise convergence evidence.

Metric	Train	Updates	Δ_F	Δ_{G+}	Ψ	95% interval for Ψ
Pair-lDDT	96	384	+0.00886	+0.01984	-0.01098	$[-0.03370, +0.01279]$
Pair-lDDT	96	768	+0.02881	+0.00735	+0.02146	$[-0.01018, +0.05223]$
Pair-lDDT	96	1536	-0.00874	-0.00449	-0.00424	$[-0.01728, +0.00952]$
Pair-lDDT	384	384	+0.00277	+0.01929	-0.01653	$[-0.03728, +0.00285]$
Pair-lDDT	384	768	+0.02900	+0.00238	+0.02662	$[-0.01894, +0.06325]$
Pair-lDDT	384	1536	+0.00834	+0.00264	+0.00570	$[-0.03259, +0.03384]$
Residue-lDDT	96	384	+0.00887	+0.01833	-0.00946	$[-0.03128, +0.01279]$
Residue-lDDT	96	768	+0.02642	+0.00624	+0.02018	$[-0.00879, +0.04767]$
Residue-lDDT	96	1536	-0.00839	-0.00359	-0.00480	$[-0.01765, +0.00816]$
Residue-lDDT	384	384	+0.00265	+0.01901	-0.01636	$[-0.03656, +0.00263]$
Residue-lDDT	384	768	+0.02812	+0.00241	+0.02571	$[-0.01750, +0.06061]$
Residue-lDDT	384	1536	+0.00776	+0.00280	+0.00496	$[-0.03009, +0.03064]$
TM-score	96	384	+0.01557	+0.03093	-0.01536	$[-0.04795, +0.01721]$
TM-score	96	768	+0.01868	-0.00015	+0.01883	$[-0.02243, +0.05619]$
TM-score	96	1536	-0.01465	-0.00285	-0.01180	$[-0.03084, +0.00537]$
TM-score	384	384	+0.00059	+0.02145	-0.02086	$[-0.04322, +0.00109]$
TM-score	384	768	+0.04484	+0.00355	+0.04129	$[-0.00822, +0.08517]$
TM-score	384	1536	+0.00758	-0.01330	+0.02088	$[-0.00499, +0.04784]$

All four mean scores improve from 384 to 1536 in both training sets. The pair-lDDT interaction is nonmonotone, with all six node intervals including zero. Neither uniform early rotational delay nor a steadily widening native advantage describes both trajectories. Tail-node comparisons (768 to 1536) are additional descriptions, not substitutes for the 384-to-1536 window. The archived per-node Δ_F and Δ_{G+} are retained separately: a change in Ψ alone does not show which arm caught up. No new direct between-training-size curve-interaction test is inferred by subtracting interval endpoints. Train/Dev curves neither establish convergence nor negate the final results on the different Confirm96-B population.

R SHARED ORTHOGONAL CHANNEL COMPENSATION WITHIN FACTOR

The compensation study uses the same OpenFold/Train96/ESM2/Full/1536 recipe as the prediction study in Appendix O. It adds a shared channel map to Native and to r20270107/r20270104, selected as low/high reconstruction error before the new rotations' task training. Selection was not based on their later structure scores. There are three paired seeds per condition, giving nine fits evaluated on Confirm96-B. The original Native and rotated models provide the paired no-map controls.

A targeted output freedom. For the all-ones unit vector $u \in \mathbb{R}^{128}$ and orthonormal complement $B \in \mathbb{R}^{128\times127}$, we use an exponential-map parameterization (Lezcano-Casado & Martínez-Rubio, 2019) on that complement:

$$C = uu^\top + B\exp(A)B^\top, \qquad A^\top = -A. \tag{16}$$

Its 8,001 free upper-triangular entries parameterize the full skew-symmetric matrix. One C is shared across targets, pairs and recycling calls within each trained model. Native injects $C\Delta U_\theta$; Rotated injects $CR\Delta U_\theta$. The baseline is unchanged. For a fixed residual, C preserves its pairwise Euclidean norm and channel mean in ideal arithmetic. Joint training can still change ΔU_θ and its amplitude; the norm of baseline plus residual is not controlled by this identity.

The rotation family belongs to $\mathcal{H} = \{C \in SO(128) : Cu = u\} \cong SO(127)$ in ideal arithmetic. The unrestricted skew exponential covers this group: real orthogonal block diagonalization gives planar rotations, each with a real skew logarithm. For any (θ, C_N), choose $C_R = C_N R^\top \in \mathcal{H}$. Then

$$C_R R\Delta U_\theta = C_N \Delta U_\theta, \qquad \mathcal{F}^R_{F+C} = \mathcal{F}^I_{F+C}. \tag{17}$$

This is a function-class equality for the stated subgroup, not for arbitrary $O(128)$ transforms. With paired initial states and random conditions, equal injections imply equal frozen downstream states and hence equal subsequent live anchors, extending the argument through recycling. No per-target or per-round freedom is added to C.

Actual code uses $C = I + B[\exp(A) - I]B^\top$ for an exact identity at zero A; finite-precision bases and rotations only approximate the ideal group. Every model starts at $A = 0$, $C = I$ and zero residual, rather than an inverse-R hot start. The channel parameters receive zero gradient at that zero-residual point, then learn jointly with the original head. They use AdamW at 10^{-3} with no decay; head parameters retain 10^{-4} and decay 0.01; all use one global clipping norm of 1. Function-class equality therefore does not imply mapped initial parameters, equivalent optimizer trajectories or equal final performance. Orthogonal adaptation itself is not new (Qiu et al., 2023; Liu et al., 2024); here it is a controlled intervention on the Native–Rotated score difference.

Paired endpoints and their identity. Define

$$B_N = S(N{+}C){-}S(N), \quad B^C_R = S(R{+}C){-}S(R), \quad T_C = (N{-}R){-}(N_C{-}R_C) = B^C_R{-}B_N, \tag{18}$$

where the last expression uses scores, with the same target and seed paired in each term. Here T_C is the additional compensation gain for rotated Factor, relative to Native Factor's gain; it compares gains, not the absolute scores of the compensated models. The prespecified contrasts average B^C_R and T_C equally over the two rotations, with the prediction that both are positive; each rotation and Native's own change are also reported. These are prespecified mechanism contrasts, not a new-target confirmation or a newly declared sole-primary endpoint. The 20,000 synchronous whole-target draws (seed 20260924) retain the shared Native control; duplicated uses of it are not independent observations. Intervals condition on these fitted models and selected rotations and are not jointly multiplicity-adjusted. Confirm96-B was already observed.

In the original execution, the three pair-lDDT seed marginals of average T_C are positive, as are the six seed-by-rotation marginals. Only r20270107 independently establishes a positive T_C interval; both rotated arms improve in absolute score. Average residue-lDDT supports a positive T_C, whereas the average TM-score interval for T_C includes zero. This supplies a useful average intervention in the tested setting, without establishing equivalence, elimination for every rotation, a learned R^{-1}, or a predictive rule for unseen settings. It does not validate the failed reconstruction predictor. Changes in output freedom, parameterization and joint optimization remain coupled.

Table 33: Pair-lDDT before and after jointly training the shared map. Native also improves. Intervals for changes are conditional on fitted models and unadjusted; contrasts use unrounded scores.

Condition	Without C	With C	Own change	95% CI for change
Native	0.5019	0.5148	+0.01290	$[+0.00223, +0.02317]$
r20270107	0.4831	0.5214	+0.03826	$[+0.02570, +0.05150]$
r20270104	0.4938	0.5150	+0.02121	$[+0.01166, +0.03105]$

Table 34: All original compensation contrasts. Native's change, the two-rotation average and both individual rotations are retained for every metric.

Metric	Condition	Contrast	Mean	95% interval
Pair-lDDT	Native	B_N	+0.01290	$[+0.00223, +0.02317]$
Pair-lDDT	Two-R mean	B_R^C	+0.02973	$[+0.02212, +0.03768]$
Pair-lDDT	Two-R mean	T_C	+0.01684	$[+0.00497, +0.02923]$
Pair-lDDT	r20270107	B_R^C	+0.03826	$[+0.02570, +0.05150]$
Pair-lDDT	r20270107	T_C	+0.02537	$[+0.01010, +0.04159]$
Pair-lDDT	r20270104	B_R^C	+0.02121	$[+0.01166, +0.03105]$
Pair-lDDT	r20270104	T_C	+0.00831	$[-0.00530, +0.02190]$
Residue-lDDT	Native	B_N	+0.01268	$[+0.00257, +0.02244]$
Residue-lDDT	Two-R mean	B_R^C	+0.02780	$[+0.02059, +0.03530]$
Residue-lDDT	Two-R mean	T_C	+0.01512	$[+0.00397, +0.02683]$
Residue-lDDT	r20270107	B_R^C	+0.03580	$[+0.02380, +0.04848]$
Residue-lDDT	r20270107	T_C	+0.02312	$[+0.00868, +0.03840]$
Residue-lDDT	r20270104	B_R^C	+0.01981	$[+0.01076, +0.02903]$
Residue-lDDT	r20270104	T_C	+0.00713	$[-0.00577, +0.01999]$
TM-score	Native	B_N	+0.01683	$[+0.00214, +0.03134]$
TM-score	Two-R mean	B_R^C	+0.02839	$[+0.01865, +0.03856]$
TM-score	Two-R mean	T_C	+0.01156	$[-0.00453, +0.02831]$
TM-score	r20270107	B_R^C	+0.04152	$[+0.02488, +0.05920]$
TM-score	r20270107	T_C	+0.02469	$[+0.00443, +0.04569]$
TM-score	r20270104	B_R^C	+0.01526	$[+0.00351, +0.02727]$
TM-score	r20270104	T_C	-0.00157	$[-0.02005, +0.01706]$

The target distribution also limits a uniform-benefit interpretation. Averaging paired seeds and rotations within each target, B_R^C is positive on 77/96 targets with median +0.02200; T_C is positive on 53/96 with median +0.00429, below its mean of +0.01684. These medians and sign counts are post-hoc descriptive summaries of the saved arrays, not additional prespecified tests. A positive mean T_C need not imply equal or positive additional gains for each target. A remaining score difference whose interval includes zero establishes neither equivalence nor complete recovery; no recovery percentage is reported.

The maximum orthogonality and mean-direction errors of the learned maps are approximately 1.67×10^{-5} and 2.18×10^{-6}, respectively. These quantify finite-precision constraint errors, not full-model error bounds. A diagnostic dtype failure required restarting from initialization; no structure score was used to change the chosen rotations, training budget or retry rule.

R.1 Same-seed repetition of all nine compensated fits

After the original results, all nine $+C$ models were trained again from the original seeds, zero-output heads and identity C, for 1536 updates each. The original Train96 order, loss, AdamW groups, two selected rotations and four-pass live-anchor configuration were retained. The repeat reused the same H100 software environment and cached ESM2 features; it did not test a new installation or fresh feature extraction. The no-C models are the same fixed original fits, not independently retrained baselines. The repeat therefore adds neither new training seeds nor new targets, and is not pooled with the original execution.

Table 35: Testing shared output freedom within Factor on observed Confirm96-B. Mean Cα pairlDDT: (a) absolute scores over the same three seeds; (b) paired gains, with rotated gains averaged over both selected rotations. Both executions jointly train C from identity and share the original fixed no-C baseline. Intervals are unadjusted 95% target intervals conditional on each execution's fitted models. The repeat adds no independent seeds. Contrasts use unrounded scores.

(a) Absolute scores

Condition	Without C	Original $+C$	Repeat $+C$
Native	0.5019	0.5148	0.5315
r20270107	0.4831	0.5214	0.5256
r20270104	0.4938	0.5150	0.5118

(b) Compensation gains and their difference

Contrast	Original	95% interval	Repeat	95% interval
B_R^C	+0.02973	$[+0.02212, +0.03768]$	+0.03025	$[+0.02129, +0.03934]$
B_N	+0.01290	$[+0.00223, +0.02317]$	+0.02960	$[+0.01644, +0.04376]$
T_C	+0.01684	$[+0.00497, +0.02923]$	+0.00065	$[-0.01223, +0.01338]$

Both executions used H100, Torch 2.7.1+cu128, TF32 disabled and eight CPU threads. Frozen AF2 weights, rotation arrays and cached ESM2 inputs were the same. A full comparison of driver and kernel runtime states is unavailable. Independent 32-update FP32 trajectories differed despite exact restoration of parameters, optimizer state, recorded random states and data position; saved-state forward agreement did not establish training-trajectory equality. The original formal runs did not retain the intermediate tensors needed to identify their first divergent operation.

Post-hoc numerical diagnosis. A subsequent comparison of the stored starting checkpoints found identical writer parameters and buffers, optimizer configuration and state, saved Torch CPU/CUDA random states, learning rates and data positions across all nine original/repeated pairs; the complete target schedules also matched. Python/NumPy global RNG states were not recorded in the historical checkpoints and differed between later diagnostic executions, while the recorded Torch/CUDA states and actual inputs matched. Matching early log scalars did not establish matching hidden tensors. Two three-update replays of one Native+C condition on the same physical H100, with the original loss and optimizer, first exposed differences in encoder parameter gradients during the second backward pass, while the captured forward activations, coordinates and loss still agreed. With cuDNN benchmarking and TF32 disabled in both settings, enabling deterministic cuDNN eliminated the observed between-replay differences in a separate pair extended to eight updates with checkpoint restoration after update three, including the captured gradients, parameters and optimizer tensors.

This is an instrumented, post-hoc numerical diagnosis, not a full-budget repeat or a new estimate of compensation benefit. Tensor copies introduce synchronization, and the trace does not cover every backbone intermediate or adjoint. It supports a controllable source of early numerical variability under the tested execution, without uniquely locating a kernel in the historical trajectories, explaining their final effect-size difference, or establishing deterministic full training. No original run is replaced.

The repeated rotated-arm gain is positive, but Native now improves by +0.02960 [+0.01644, +0.04376]. Consequently the repeated T_C interval includes zero (Table 35). Neither selected rotation independently establishes positive T_C in the repeat, although both improve in absolute score. The original positive average T_C remains an observed result, but the additional compensation gain for rotated Factor relative to Native Factor's gain has not been re-established by the complete repeat. This does not prove zero additional gain, statistically different population effects between executions, or a particular cause for the training divergence. No run is replaced and no favorable executions are selected or averaged together. The full-training outcome variation has not been causally attributed.

Table 36: All 21 compensation contrasts in the repeated execution, with the original unadjusted 20,000-draw target bootstrap (seed 20260924).

Metric	Condition	Contrast	Mean	95% interval
Pair-lDDT	Native	B_N	+0.02960	$[+0.01644, +0.04376]$
Pair-lDDT	Two-R mean	B_R^C	+0.03025	$[+0.02129, +0.03934]$
Pair-lDDT	Two-R mean	T_C	+0.00065	$[-0.01223, +0.01338]$
Pair-lDDT	r20270107	B_R^C	+0.04248	$[+0.02888, +0.05644]$
Pair-lDDT	r20270107	T_C	+0.01287	$[-0.00154, +0.02783]$
Pair-lDDT	r20270104	B_R^C	+0.01802	$[+0.00541, +0.03096]$
Pair-lDDT	r20270104	T_C	-0.01158	$[-0.02903, +0.00551]$
Residue-lDDT	Native	B_N	+0.02784	$[+0.01546, +0.04110]$
Residue-lDDT	Two-R mean	B_R^C	+0.02825	$[+0.01973, +0.03691]$
Residue-lDDT	Two-R mean	T_C	+0.00041	$[-0.01181, +0.01251]$
Residue-lDDT	r20270107	B_R^C	+0.03963	$[+0.02675, +0.05288]$
Residue-lDDT	r20270107	T_C	+0.01179	$[-0.00201, +0.02614]$
Residue-lDDT	r20270104	B_R^C	+0.01686	$[+0.00490, +0.02904]$
Residue-lDDT	r20270104	T_C	-0.01097	$[-0.02745, +0.00509]$
TM-score	Native	B_N	+0.03123	$[+0.01366, +0.05008]$
TM-score	Two-R mean	B_R^C	+0.03101	$[+0.02012, +0.04228]$
TM-score	Two-R mean	T_C	-0.00022	$[-0.01749, +0.01668]$
TM-score	r20270107	B_R^C	+0.04976	$[+0.03205, +0.06792]$
TM-score	r20270107	T_C	+0.01853	$[-0.00111, +0.03889]$
TM-score	r20270104	B_R^C	+0.01226	$[-0.00214, +0.02720]$
TM-score	r20270104	T_C	-0.01897	$[-0.04097, +0.00222]$

The repeat neither validates the failed reconstruction predictor nor tests compensation at reference anchors.

S RESIDUAL-ANCHOR INTERVENTION

This study tests a specific feedback hypothesis: whether constructing the residual from the adapted trajectory's live factors helps rotated Factor and reduces the Native–Rotated score difference. It contains no channel compensator C.

Intervention and paired training. Let L denote live residual anchors and Q the query-only references for the same sequence and recycle index t. At the first OPM of the intact 48-block Evoformer, the two paths are

$$U_{L,t} = U_{0,t}^{\text{live}} + \mathcal{R}\big[G(a_t + \delta a, b_t + \delta b) - G(a_t, b_t)\big], \tag{19}$$

$$U_{Q,t} = U_{0,t}^{\text{live}} + \mathcal{R}\big[G(a_t^Q + \delta a, b_t^Q + \delta b) - G(a_t^Q, b_t^Q)\big], \tag{20}$$

where $\mathcal{R}$ is identity for Native and the fixed channel rotation for Rotated. Each adapted path retains its own live OPM baseline, pair state, four-pass recycling and structure module. The four reference pairs (a_t^Q, b_t^Q) come from four separate query-only passes, not a repeated first-pass anchor. Mask, normalization and frozen decoder are unchanged. One sequence-conditioned writer supplies the same factor increments at each pass; only the anchors differ. This replaces neither the live baseline nor the full recycling process, and does not make the update a cached Protenix residual.

The recipe is OpenFold model_3_ptm, Train96, ESM2-35M layer 12 and Full, with 1,536 updates and the original AdamW recipe. Native and the two rotations r20270107/r20270104 are independently retrained under Q, each with the three paired seeds. These rotations were selected and locked before their task scores in the prediction study. All nine new fits start from the original zero-output initialization, without a warm start or C. The original L models and query baseline are reused from the prediction study. All comparisons use the same already observed Confirm96-B panel; no target or rotation is replaced.

Predeclared quantities and statistical identity. Using scores with matched target and training seed, define

$$T_A = (N_Q - R_Q) - (N_L - R_L), \qquad B_R^{\text{anchor}} = R_L - R_Q. \tag{21}$$

Both are averaged equally over the two selected rotations. The hypothesis requires positive T_A and positive B_R^{anchor}, supported by their intervals. The latter differs from compensation's $B_R^C = S(R + C) - S(R)$. The paired identity

$$T_A = B_R^{\text{anchor}} + (N_Q - N_L) \tag{22}$$

holds for each target and seed. Thus a reduced gap can reflect a changed Native reference as well as a changed rotated arm.

As locked, 20,000 synchronous target-bootstrap draws (seed 20260924) preserve the shared Native control. Intervals condition on the three fitted seeds and two selected rotations; they are not jointly multiplicity-adjusted. Proteins, not predictions, seeds or residue pairs, are the sampling units. Pair-lDDT is primary; residue-lDDT and TM-score are supplementary measurements from the same structures. This is an observed-panel, Factor-only intervention, not a new-target confirmation of the Factor/G+ interaction.

Table 37: Anchor intervention: absolute Cα pair-lDDT over three seeds. L uses live residual factors; Q uses round-matched query-only references. The unadapted baseline is common. Scores have four decimals; contrasts use unrounded values. Rotated means average the same two selected rotations.

Condition	Live L	Reference Q
Query-only	0.3019	0.3019
Native	0.5019	0.4925
Rotated (two-R mean)	0.4884	0.4911
r20270107	0.4831	0.4978
r20270104	0.4938	0.4844
Native–Rotated	+0.01346	+0.00146

Results and interpretation. The two predeclared mean effects have negative point estimates and intervals including zero (Table 38). The expected average feedback benefit is not established. Table 37 shows why the smaller Q gap alone is insufficient evidence of useful compensation: it combines a Native decrease with a slight mean increase of Rotated. Neither corresponding average change has an interval excluding zero. Adaptation under Q still works: Native exceeds query-only by +0.19064 [+0.16124, +0.22135], and both rotated Q arms retain positive adaptation intervals. Failure of the feedback prediction cannot therefore be explained by all reference-anchor models losing their adaptation benefit.

Table 38: Pooled anchor contrasts. Positive T_A means a larger Native–Rotated score difference under Q; positive B_R^{anchor} favors live anchors for the rotated arm. Intervals are conditional target intervals, unadjusted. Sign counts use effects averaged over paired seeds and rotations per target.

Metric	Contrast	Mean	95% interval	Positive
Pair-lDDT	T_A	-0.01200	$[-0.02665, +0.00187]$	42/96
Pair-lDDT	B_R^{anchor}	-0.00264	$[-0.00894, +0.00332]$	46/96
Pair-lDDT	$N_Q - N_L$	-0.00936	$[-0.02016, +0.00122]$	45/96
Residue-lDDT	T_A	-0.01065	$[-0.02435, +0.00233]$	42/96
Residue-lDDT	B_R^{anchor}	-0.00213	$[-0.00803, +0.00348]$	45/96
Residue-lDDT	$N_Q - N_L$	-0.00852	$[-0.01869, +0.00145]$	44/96
TM-score	T_A	-0.00966	$[-0.02923, +0.00884]$	44/96
TM-score	B_R^{anchor}	-0.00053	$[-0.00988, +0.00803]$	47/96
TM-score	$N_Q - N_L$	-0.00913	$[-0.02283, +0.00457]$	42/96

The individual rotations respond in opposite directions (Table 39): r20270107 favors reference anchors, whereas r20270104 has a positive live-anchor benefit and nearly zero T_A. The latter benefit's pair-lDDT interval starts at only +3.82140e-07. These unadjusted secondary comparisons do not establish the average hypothesis or a general rotation-by-anchor heterogeneity rule; no direct heterogeneity test was specified. Pooled supplementary-metric intervals likewise include zero. The results establish neither equivalence nor a universal harm from live anchors. These anchor results neither validate the reconstruction ranking nor resolve the compensation findings, whose relative benefit over Native was not re-established in the repeat. Because C is absent, they do not test whether its benefit depends on live anchors, nor isolate expressivity as the unique mechanism.

Table 39: Both selected rotations in the anchor intervention. These secondary intervals use the same target draws without joint correction. The very small positive endpoint is retained in scientific notation rather than rounded to zero.

Metric	Rotation	Contrast	Mean	95% interval
Pair-lDDT	r20270107	T_A	-0.02407	$[-0.04227, -0.00759]$
Pair-lDDT	r20270107	B_R^{anchor}	-0.01471	$[-0.02528, -0.00499]$
Pair-lDDT	r20270104	T_A	+0.00007	$[-0.01471, +0.01488]$
Pair-lDDT	r20270104	B_R^{anchor}	+0.00943	$[+3.82140e-07, +0.01863]$
Residue-lDDT	r20270107	T_A	-0.02196	$[-0.03899, -0.00635]$
Residue-lDDT	r20270107	B_R^{anchor}	-0.01344	$[-0.02341, -0.00423]$
Residue-lDDT	r20270104	T_A	+0.00066	$[-0.01327, +0.01448]$
Residue-lDDT	r20270104	B_R^{anchor}	+0.00918	$[+0.00013, +0.01805]$
TM-score	r20270107	T_A	-0.02534	$[-0.04911, -0.00393]$
TM-score	r20270107	B_R^{anchor}	-0.01621	$[-0.03052, -0.00316]$
TM-score	r20270104	T_A	+0.00601	$[-0.01399, +0.02642]$
TM-score	r20270104	B_R^{anchor}	+0.01515	$[+0.00229, +0.02807]$

Training and evaluation reference caches are separate. The reference for the final pass is held fixed during backward checkpoint recomputation. A preliminary next-update comparison did not establish bitwise training-trajectory equivalence; subsequent agreement from a restored state does not resolve that limitation. The scientific settings and source implementation were unchanged.

T FRESH192: FIXED PROTENIX MODELS ON NEW TARGETS

This study selects the previously observed Protenix-Mini default Train384/Full/1536, ESMC-600M configuration for a new-target test. It freezes all 24 adapters: Factor and G+ each have three Native fits and nine rotated fits, using seeds 20260923/24/25 and rotations 20261001/02/03. Query-only is the 25th system. No model is retrained, no checkpoint is selected on this panel, and the historical choice of recipe is disclosed. This is a test of a fixed configuration's transfer, not a random sample of predictor architectures.

Panel, selection and chronology. Fresh192 uses the short-chain catalog criteria in Appendix C: single-model X-ray monomers released by 2021-09-30, resolution at most 2.5 Å, standard amino acids and at least 90% reference Cα coverage. Four prespecified strata, 128–191, 192–255, 256–319 and 320–384, each contribute 48 complete sequences; actual lengths are 128–383. There is one target per PDB. This panel is not a long-chain extrapolation study.

The exposure inventory includes Fresh96 and the later prediction, compensation and anchor-intervention targets, while retaining historical conservative exclusions. Inventory corrections preceded new target predictions and did not change the search thresholds or deterministic ranking. Selection uses the same BLAST screening rule as Fresh96: BLASTP 2.17.0+, a single qualifying HSP with $E \leq 10^{-5}$, identity at least 0.30, at least 50 paired residues and coverage at least 0.70. Recorded query comparisons use the shorter sequence length as coverage denominator; other teacher rows use candidate length. HSPs are not merged or count-limited. Selected panel members are screened against one another as well.

Candidates are ordered within strata by SHA256 of the target ID prefixed by `protenix-fourcell-fresh192-v1|`, with at most 2048 candidates per stratum and a fixed screening time cap. Selection does not read model scores. No ID, PDB or exact-sequence overlap remains with the recorded exclusions. These claims concern the documented inventory, not certified complete project history, strict family separation or foundation-model pretraining exclusion. The panel, models, execution recipe and analysis were fixed before new folding predictions.

Inference and endpoint. The original complete-sequence c4/s5, one sample and inference seed101 are retained. $D = 508.5$ comes from the adapter checkpoint. ESMC uses final-layer 1152-dimensional features, the historical H100 FP32 extraction/BF16 cache path and FP32 loading; folding uses the fixed MI250 path. No target coordinates or observed-residue mask enter inference. The score grid contains all $25 \times 192 = 4,800$ model–target pairs. The predefined failure rule assigns zero credit to failed predictions, with no target replacement, crop or altered denominator; no formal prediction required this rule.

The sole primary endpoint is pair-lDDT Ψ in Equation 7. After averaging paired seeds and rotations inside each target, 20,000 paired bootstrap draws (seed20260925) resample 48 targets within each of the four locked strata. Strata retain equal weights. Intervals condition on the fixed 24 adapters and rotations and do not include retraining or historical recipe-selection uncertainty. All other contrasts, strata and structure metrics are secondary and unadjusted. New and observed panels are reported separately; neither targets nor rotations are added after the result.

Results and scope. The primary interaction is +0.02515 [+0.01832, +0.03213], with 140/192 positive targets. All three seed and rotation marginals, and all nine seed-by-rotation means, are positive (Table 40). Both supplementary interaction intervals are positive, but score the same structures. The full means are in Table 2; the secondary contrasts are retained below. Factor's positive direction effect and the positive direct interaction support greater Factor rotation sensitivity under this fixed recipe on new targets. G+'s direction interval includes zero without establishing equivalence. This confirmation does not replace OpenFold's Fresh96 result or identify a unique expressivity, optimization or downstream mechanism.

Table 40: Fresh192 primary-interaction marginals. The first two rows give seed and rotation marginals in ascending locked order. The final three rows give individual rotations across the three training seeds.

Margin / rotation	First / seed23	Second / seed24	Third / seed25
Seed	+0.02036	+0.02116	+0.03395
Rotation	+0.02126	+0.02916	+0.02504
r20261001	+0.02734	+0.01514	+0.02128
r20261002	+0.01602	+0.02046	+0.05101
r20261003	+0.01770	+0.02788	+0.02955

Table 41: Fresh192 contrasts. Only pair-lDDT Ψ is primary; the remaining intervals are unadjusted secondary summaries of the same fitted models.

Metric	Contrast	Mean	95% interval
Pair-lDDT	Δ_F	+0.02392	$[+0.01878, +0.02926]$
Pair-lDDT	Δ_{G+}	-0.00124	$[-0.00619, +0.00349]$
Pair-lDDT	Ψ	+0.02515	$[+0.01832, +0.03213]$
Pair-lDDT	F–G+	+0.10009	$[+0.08646, +0.11441]$
Pair-lDDT	F–Query	+0.43697	$[+0.41420, +0.45946]$
Pair-lDDT	G+–Query	+0.33688	$[+0.31468, +0.35868]$
Pair-lDDT	RF–Query	+0.41305	$[+0.39006, +0.43572]$
Pair-lDDT	RG+–Query	+0.33812	$[+0.31651, +0.35942]$
Residue-lDDT	Δ_F	+0.02341	$[+0.01844, +0.02867]$
Residue-lDDT	Δ_{G+}	-0.00111	$[-0.00592, +0.00347]$
Residue-lDDT	Ψ	+0.02452	$[+0.01792, +0.03128]$
Residue-lDDT	F–G+	+0.09695	$[+0.08379, +0.11074]$
Residue-lDDT	F–Query	+0.42057	$[+0.39841, +0.44241]$
Residue-lDDT	G+–Query	+0.32361	$[+0.30231, +0.34465]$
Residue-lDDT	RF–Query	+0.39716	$[+0.37494, +0.41917]$
Residue-lDDT	RG+–Query	+0.32472	$[+0.30395, +0.34531]$
TM-score	Δ_F	+0.02197	$[+0.01612, +0.02813]$
TM-score	Δ_{G+}	-0.00066	$[-0.00655, +0.00488]$
TM-score	Ψ	+0.02262	$[+0.01476, +0.03093]$
TM-score	F–G+	+0.09904	$[+0.08337, +0.11583]$
TM-score	F–Query	+0.47302	$[+0.44655, +0.49934]$
TM-score	G+–Query	+0.37398	$[+0.34757, +0.40035]$
TM-score	RF–Query	+0.45105	$[+0.42427, +0.47757]$
TM-score	RG+–Query	+0.37464	$[+0.34864, +0.40052]$

Table 42: Fresh192 secondary interaction estimates within the four prespecified length strata. These unadjusted intervals do not test a length trend.

Metric	Residues	Ψ	95% interval
Pair-lDDT	128–191	+0.01983	$[+0.00446, +0.03624]$
Pair-lDDT	192–255	+0.02552	$[+0.00880, +0.04311]$
Pair-lDDT	256–319	+0.02805	$[+0.01851, +0.03805]$
Pair-lDDT	320–384	+0.02722	$[+0.01629, +0.03939]$
Residue-lDDT	128–191	+0.01996	$[+0.00466, +0.03644]$
Residue-lDDT	192–255	+0.02526	$[+0.00951, +0.04178]$
Residue-lDDT	256–319	+0.02709	$[+0.01779, +0.03677]$
Residue-lDDT	320–384	+0.02578	$[+0.01525, +0.03752]$
TM-score	128–191	+0.01879	$[+0.00009, +0.03948]$
TM-score	192–255	+0.02516	$[+0.00608, +0.04492]$
TM-score	256–319	+0.02696	$[+0.01598, +0.03858]$
TM-score	320–384	+0.01958	$[+0.00774, +0.03251]$

U AtlasFold propagation and calibration

On two previously observed diagnostic chains, 7hln_A and 9qdy_A, we recorded five passes for Query and three Native parents with their C-only/head-only continuations (20 instances, no new training). Table 43 summarizes the parents: differences survive the LM stack, projection and fusion. Their smaller relative size at the trunk output does not imply deletion of the update; absolute difference RMS increases while the Query scale increases more. Final structures change little relative to Query; this is not a quality gain. These two-chain, Native-only observations identify neither the cause of absent adaptation gains nor a rotation interaction. Full per-pass measurements and coordinates accompany the artifact.

Table 43: Atlas parent-model pair-state differences. Each ratio is $\mathrm{RMS}(z_{\mathrm{adapted}} - z_Q)/\mathrm{RMS}(z_Q)$ at the same target, pass and position. Entries are separate medians over two chains, three fitted seeds and five passes, not independent-sample intervals or cross-layer signal-retention fractions.

Position	Relative difference	Difference RMS	Query RMS
After injection	0.3612	68.46	189.16
LM stack output	0.2794	102.44	370.06
After projection	0.2122	2.28	10.66
After fusion	0.1619	2.32	14.78
Trunk output	0.0311	3.16	101.29

A separate Native-only calibration continued each parent for 512 updates, either learning an identity-initialized shared orthogonal C with the original head frozen, or continuing the head with its AdamW state. On observed Confirm96-B, the primary C-only minus head-only pair-lDDT contrast is -0.000005 [-0.000697, +0.000749]; C-only minus Query is -0.001287 [-0.003015, +0.000408]. Neither establishes a benefit. Intervals use 20,000 target bootstrap draws after averaging three fitted seeds, with the Query comparison descriptive and unadjusted. These results do not establish information redundancy or motivate a specific alternative interface.

Native/ESMC readout. A fixed Train96/Dev8 64-bin distance probe compared native single/pair inputs plus a 128-dimensional auxiliary view of native single features, aligned ESMC, or residue-permuted ESMC. Each arm used 131,520 trainable parameters, three seeds and 1,536 updates; the folding model stayed frozen. Final Dev8 cross-entropies were 1.475322, 1.472815, and 1.477265. The sole primary native-minus-ESMC improvement was +0.002506 [-0.000186, +0.005430], with one seed mean negative. Intervals use 20,000 resamples of eight observed targets after seed aggregation. Aligned ESMC outperformed its permutation overall, but their ordering reversed in the prespecified close-pair subset; these unadjusted secondary comparisons do not replace the primary result. Fixed compression, readout capacity and optimization limit the inference: this does not establish ESMC redundancy or folding gains. Six additional ESMC Native propagation observations on the same two chains also retained trunk-state differences with small final structural changes. The supplement provides all fixed-node scores and secondary contrasts.

V Protenix four-cell budget extension

This observed-panel follow-up continues the 24 Protenix Train384/ESMC-600M Full models used for Fresh192. The training set, feature cache, three seeds, three rotations, loss, constant learning rate 5×10^{-5} and inference recipe remain fixed. We restore each model's step1536 weights and AdamW state and train to step3072. The original checkpoints lack complete RNG states, so this is not a claim of bitwise equivalence to uninterrupted historical training. No compensator is added, and the final checkpoint is fixed independently of Dev or evaluation scores.

The sole primary endpoint is Ψ_{3072}; the paired change $K = \Psi_{3072} - \Psi_{1536}$ is the key secondary endpoint. Fresh192 has already been observed, so this extends the budget range of the earlier recipe's evidence rather than providing another new-target confirmation. Four length strata of 48 targets each are resampled for 20,000 draws (seed20260926), with the same target indices for both budgets

and every model. Seeds and rotations are averaged within target. Intervals condition on the fitted instances and secondary comparisons have no joint multiplicity correction.

Table 44: Protenix/ESMC on observed Fresh192, before and after continuation. Absolute Cα pair-lDDT means use four decimals; paired gains and their 95% target intervals use unrounded scores. Query is fixed.

System	1536	3072	Gain	95% interval
Query	0.3912	0.3912	–	–
Factor Native	0.8281	0.8332	+0.00506	$[+0.00218, +0.00797]$
Factor Rotated	0.8042	0.8217	+0.01747	$[+0.01343, +0.02172]$
G+ Native	0.7280	0.7633	+0.03527	$[+0.02731, +0.04345]$
G+ Rotated	0.7293	0.7700	+0.04074	$[+0.03374, +0.04818]$

Table 45: Row effects and their direct budget changes on the same targets. The historical point estimates are unchanged. These intervals use the budget study's paired bootstrap stream; they do not overwrite the original Fresh192 intervals. A change interval containing zero establishes neither equivalence nor convergence.

Effect	1536	95% interval	3072	95% interval	Change	95% interval
Δ_F	+0.02392	$[+0.01886, +0.02927]$	+0.01150	$[+0.00743, +0.01604]$	-0.01241	$[-0.01736, -0.00754]$
Δ_{G+}	-0.00124	$[-0.00620, +0.00351]$	-0.00670	$[-0.01232, -0.00154]$	-0.00546	$[-0.01140, +0.00035]$
Ψ	+0.02515	$[+0.01835, +0.03228]$	+0.01820	$[+0.01176, +0.02504]$	-0.00695	$[-0.01423, +0.00046]$

All four cells improve (Table 44). Factor's signed gap decreases (paired change -0.01241 [-0.01736, -0.00754]), while G+'s contrast also shifts in point estimate. The difference between these paired changes is K, whose interval includes zero (Table 45). The final interaction is positive on 127/192 targets and in all three seed marginals, three rotation marginals and nine seed-by-rotation cells. Supplementary final interactions are +0.01776 [+0.01159, +0.02427] for residue-lDDT and +0.01529 [+0.00733, +0.02368] for TM-score. Their budget-change intervals include zero; one TM-score seed-by-rotation cell is negative. These metrics score the same structures.

Dev8 was evaluated at 1536, 2304 and 3072 updates. Increasing scores and three Dev nodes do not establish convergence.

W OpenFold injection-schedule retraining

We retrain First-only, All and Last-only injection with four trunk passes, holding OpenFold Train96/ESM2-35M/Full, final-pass task supervision and 1536 updates fixed. A single adapter remains shared across passes; the schedule gates its residual at the first, every, or last pass. Native and rotated Factor and G+ are separately trained for each schedule. Baseline OPMs and the other Evoformer blocks remain active. Unlike the historical final-pass-only gradient recipe, this study retains the differentiable MSA/pair paths across passes; native coordinate-distance binning remains nondifferentiable. Its All condition is therefore newly trained, not the old All model reused.

A bounded Train/Dev calibration compares learning rates 5×10^{-5} and 10^{-4} across all 12 schedule/arm cells, using a separate development seed and 192 updates per fit. The average absolute Dev8 pair-lDDT selects the common rate 5×10^{-5}; neither interaction size nor Confirm96-B selects it. The original design planned three rotations and 72 fits. After the first set of fits, computational-budget constraints limited the final scope to 36 fits: three schedules, four cells, three seeds and the first prespecified rotation R1=20261001. This reporting amendment preceded unified scoring of these predictions; the remaining two rotations were not trained. We report the single-R1 analogue of the planned All–First comparison, not completion of the original three-rotation primary matrix.

First has the largest interaction, yet both G+ arms are near Query and its absolute scores are lowest (Table 46). All's Factor contrast is +0.00206 [-0.00721, +0.01145]; its G+ contrast is -0.01279 [-

Table 46: Retrained OpenFold schedules on observed Confirm96-B. Each absolute pair-lDDT mean averages three seeds; RF/RG+ use only R1. Query is 0.3019. The final columns are Ψ and its conditional 95% target interval. Higher interaction need not mean higher absolute quality.

Schedule	F	RF	G+	RG+	Ψ	95% interval
First-only	0.3365	0.3063	0.3024	0.3026	+0.03031	$[+0.02103, +0.04055]$
All	0.4391	0.4371	0.3338	0.3466	+0.01485	$[+0.00505, +0.02504]$
Last-only	0.3907	0.3792	0.3898	0.3917	+0.01340	$[+0.00444, +0.02270]$

Table 47: Direct paired changes of Ψ, using shared target resamples. All–First is the single-R1 analogue of the original planned main comparison; the remaining comparisons are secondary. No familywise correction is applied.

Schedule contrast	Ψ difference	95% interval
All − First	-0.01546	$[-0.03047, -0.00089]$
First − Last	+0.01690	$[+0.00473, +0.02988]$
All − Last	+0.00145	$[-0.01387, +0.01623]$

0.01841, -0.00762]. Thus its positive Ψ does not establish a positive Factor Native–Rotated score difference. Last's Native Factor–G+ contrast is +0.00091 [-0.00644, +0.00852].

The All–First interaction difference is negative in all three seed means. Corresponding residue-lDDT and TM-score differences are -0.01430 [-0.02865, -0.00043] and -0.02006 [-0.03908, -0.00147]. We resample the 96 targets 20,000 times (seed20260926), averaging seeds within target and sharing draws across schedules and contrasts. These intervals are conditional on the fitted models and R1. The full row effects, query gains, target arrays and seed marginals are supplied, including mixed Factor seed responses under All. This schedule study changes training and inference together; it does not isolate repeated writing from learning dynamics or identify a universal feedback mechanism. It is distinct from changing the deployment schedule of fixed All-trained models and from changing residual anchors while retaining the injection schedule.