When Frozen Features Help or Harm: Correspondence and Auxiliary Supervision in Antibody–Antigen Prediction

Anonymous authors
Paper under double-blind review

Abstract

Frozen protein representations and auxiliary losses can augment an interaction predictor, but different comparisons answer different questions about their value. We separate the benefit of an added representation from sensitivity to its residue correspondence, and distinguish contact supervision from partner-ranking supervision. On a SAbDab-derived development set, aligned features outperform features permuted throughout downstream training and evaluation by 0.00820 macro pair AUPRC for ESM-IF1 and 0.00319 for ProteinMPNN, with positive differences in all ten paired seeds. A prespecified evaluation of the frozen checkpoints on a 217-complex test partition yields a mixed confirmation: ProteinMPNN retains a 0.00359 effect (95% seed-bootstrap interval [0.00156, 0.00588]; nine positive seeds), whereas ESM-IF1 falls to 0.00226 (interval $[-0.000007, 0.004901]$; six positive seeds). A prespecified 185-complex sensitivity subset excludes prior exposure through an unrelated baseline evaluation without replacing the primary partition. Thus, strong development-set correspondence sensitivity need not transfer with the same magnitude. Separately, a four-arm AsEP development study shows that removing partner ranking improves pair AUPRC by 0.0214 and epitope MCC by 0.0369, both in all ten seeds. The node advantage appears at matched epochs before learning rates diverge; gradient diagnostics do not identify a unique mechanism. Together, the studies distinguish correspondence sensitivity, baseline-relative benefit, and auxiliary-objective utility, while separating development findings from held-out confirmation.

1 Introduction

Protein language models and structure-conditioned inverse-folding models make large-scale pretraining available to small downstream predictors (Lin et al., 2023; Hsu et al., 2022; Dauparas et al., 2022). For residue-level tasks, transferring a representation entails more than supplying a collection of vectors: each vector must be assigned to the residue it describes. This creates two distinct questions. Does an additional representation improve prediction relative to the existing inputs? And how much does prediction depend on preserving the representation's correspondence to those inputs? A model can be highly sensitive to a feature assignment without gaining much from the correctly assigned feature.

Auxiliary supervision raises a related experimental question: which component of a composite objective produces an observed change? Antibody–antigen prediction can involve residue-level interface labels, residue-pair contacts, and preferences between matched and alternative partners. These objectives concern related biological objects but differ in target, granularity, and optimization pathway. Turning a complete relational branch on or off cannot determine which objective contributes to an observed change. Nor does a favorable local gradient statistic establish that an auxiliary objective improves the eventual prediction task.

We investigate both questions in a common partner-conditioned predictor. The representation study compares frozen aligned features, fixed-permuted features, sequence-only inputs, and random-projection controls. The objective study independently enables contact-edge and partner-ranking supervision while retaining node supervision. The aim is to isolate transfer effects within this

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.

controlled system, not to introduce a new pretrained model, a universal theory of alignment, or a benchmark-leading antibody predictor.

Crucially, we distinguish discovery from confirmation. Development experiments show large correspondence sensitivity for ESM-IF1 and a smaller effect for ProteinMPNN. After freezing the checkpoints and recording a confirmation protocol, we evaluate these two contrasts on a separate test partition. ProteinMPNN retains a clear positive effect, but the ESM-IF1 estimate contracts and its interval includes zero. The larger development effect is therefore not the one most clearly supported on test. This is a descriptive contrast, not a statistically established difference in encoder generalization. A documented overlap with an earlier, unrelated baseline evaluation is disclosed and addressed through a prespecified exclusion analysis.

The supervision study provides a second, development-bounded result. Direct contact supervision substantially improves pair prediction without a detected node penalty. In contrast, removing a complex-level ranking term improves both measured prediction endpoints. Training-history checks establish that the node advantage appears before learning-rate schedules diverge, narrowing a plausible procedural explanation without identifying a unique gradient mechanism.

The paper makes two empirical contributions. **First**, it separates baseline-relative representation benefit from correspondence sensitivity, and subjects the latter to a prespecified test whose mixed outcome limits cross-encoder claims. **Second**, it separates contact and ranking objectives in a factorial intervention and identifies where the resulting performance difference does—and does not—support a mechanistic interpretation. Controls, uncertainty estimates, and reproduction records make the evidence for each conclusion inspectable; numerical reproducibility is not treated as evidence of universal validity.

2 Setting, interventions, and inferential scope

2.1 Prediction tasks and data

For complex c, let $Y_c \in \{0,1\}^{L_c^{\mathrm{ag}} \times L_c^{\mathrm{ab}}}$ denote the annotated contact matrix over the cohort's modeled residues. Epitope and paratope labels indicate whether a row or column contains a contact. Our model produces pair scores and separate antigen- and antibody-node probabilities. The independent node heads are not constrained to equal marginals of the predicted contact matrix. Correct node labels alone consequently do not determine the pair assignment.

The representation study uses a project-specific SAbDab-derived cohort, denoted SAbDab2, with 3,229 training, 217 validation, and 217 test complexes (Dunbar et al., 2014). Membership is fixed and disjoint within this split. Its arrays retain observed antibody variable-domain and antigen-chain residues. The objective study uses the installed AsEP epitope-group partition, with 1,383 training and 169 validation complexes, modeling antibody CDR and antigen-surface residues (Liu et al., 2024). Its 169-complex test partition and the separate epitope-ratio test were used only for earlier frozen benchmark measurements. These are different evaluation populations; all models are trained within their own cohort. Appendix B.2 distinguishes the installed release from the publication and describes input boundaries.

What was held out? The SAbDab2 confirmation evaluates existing checkpoints on a partition not used to select those checkpoints. It is *not* described as wholly unexposed throughout the broader project. An access audit found that 32 of its 217 complexes had per-complex test outputs from an unrelated EpiFormer baseline, using a different 455-complex split. Before evaluating the focal models, the protocol specified both the full 217-complex primary population and a 185-complex exclusion sensitivity population. The latter has no overlap with that documented prior evaluation; it is not a guarantee against every possible pretraining or homology exposure. Neither population is selected according to the new outcomes.

2.2 Predictor and feature interventions

The common predictor projects 1,280-dimensional frozen ESM-2 features into separate antigen and antibody streams. One gated bidirectional cross-attention block, with width 128 and eight heads, supplies partner context. A low-rank bilinear decoder with pair width 64 produces contact scores;

independent linear heads produce node probabilities. Dropout is 0.2. Auxiliary pretrained features are concatenated before the projections with source-presence indicators. Appendix B gives the loss and execution details. The predictor is an experimental substrate, not a claimed architectural contribution.

For encoder q, write B for sequence-only inputs, A_q for aligned auxiliary features, and S_q for the same auxiliary features with a fixed residue-row permutation. Each downstream model is trained under its assigned input condition. The permutation is deterministic for each stored sequence and retained during training and evaluation; the base ESM-2 rows, labels, and residue masks are not permuted. Aligned and permuted conditions therefore preserve auxiliary vector multisets, input dimensions, source coverage, and trainable parameter count. They change the joint assignment of those vectors to the base inputs and targets. The contrast isolates this intervention, not a particular geometric component within the representation.

ESM-IF1 provides 512-dimensional encoder states from N/Cα/C backbone coordinates (Hsu et al., 2022). ProteinMPNN provides 128-dimensional encoder states before sequence decoding, using the same backbone store and modeled rows (Dauparas et al., 2022). Its selected full-backbone checkpoint requires oxygen positions, reconstructed here by fixed ideal geometry. The models share an inverse-folding task but differ in architecture, width, and preprocessing. AntiBERTy supplies a sequence-conditioned comparator on antibody VH/VL, not an antigen-side structural encoder (Ruffolo et al., 2021).

Random-projection controls require a separate qualification. ProteinMPNN's control mirrors its source width and coverage. The historical ESM-IF1 random arm instead mirrors a combined AntiBERTy-plus-ESM-IF1 configuration and is wider on the antibody side. It is not an exactly matched control for ESM-IF1 alone. Only the aligned–permuted comparison within that encoder has the exact capacity match (896,781 trainable parameters in each arm). More capacity is not assumed to make a control conservative under a fixed training procedure.

2.3 Three representation contrasts, two evaluation stages

For a population D, let $M_D(F)$ be a model condition's mean per-complex pair average precision, averaged over the paired seeds. We distinguish

$$U_{q,D} = M_D(A_q) - M_D(B), \qquad H_{q,D} = M_D(S_q) - M_D(B), \qquad C_{q,D} = M_D(A_q) - M_D(S_q). \tag{1}$$

Here U is baseline-relative benefit, H is the baseline-relative effect of permuted features, and C is correspondence sensitivity. For common seeds and complexes, $C = U - H$. Thus a positive C does not imply a positive U: it can arise because H is negative. Conversely, a smaller C need not mean a poorer aligned predictor.

The retrained comparison also differs from an evaluation-only perturbation. Writing $f_{q,A}$ and $f_{q,S}$ for models trained under aligned and permuted inputs, respectively, the two interventions are

$$C_{q,D} = M_D(f_{q,A}, Z_q) - M_D(f_{q,S}, PZ_q), \tag{2}$$

$$T_{q,D}(k) = M_D(f_{q,A}, P_k Z_q) - M_D(f_{q,A}, Z_q). \tag{3}$$

Base inputs and masks are suppressed in the notation but fixed within each comparison. P is a fixed global row permutation; P_k is the specified evaluation-time offset or permutation. The first contrast permits adaptation to the corrupted assignment during training. The second measures reliance of an aligned-trained predictor on its expected assignment. Neither is a test of whether information is destroyed in an information-theoretic sense: an invertible permutation preserves the vector collection while changing the correspondence available to the learner. Their effect sizes should therefore not be pooled as repeated measurements of one quantity.

The development study estimates all three quantities. The later confirmation estimates *only* C for ESM-IF1 and ProteinMPNN. It uses the ten existing paired best checkpoints, without retraining, new checkpoint selection, thresholds, ensembling, or added test contrasts. In particular, it does not test whether either aligned encoder beats ESM-2 on test, or whether either permuted test model is below that baseline. The protocol was recorded after development evidence was frozen and before focal test evaluation; the 217-complex and exclusion readouts come from one evaluation. Appendix A records the chronology and full statistics.

2.4 METRICS, UNCERTAINTY, AND DECISION RULES

The primary pair endpoint is macro AUPRC: non-interpolated, tie-aware average precision is computed per complex and then averaged, rather than pooling residue pairs. The factorial study's primary node endpoint is per-complex epitope MCC at threshold 0.5. Earlier benchmark MCC uses validation-selected thresholds and is labeled separately. For ten paired seeds, we report

$$\widehat{\Delta}_D = \frac{1}{10}\sum_{s=1}^{10} d_{s,D}, \qquad d_{s,D} = \frac{1}{|D|}\sum_{c\in D}\left[\mathrm{AP}_{A,s,c} - \mathrm{AP}_{S,s,c}\right]. \tag{4}$$

The reported percentile-bootstrap intervals resample seed differences and condition on the evaluated complexes. They quantify training-randomness uncertainty, not uncertainty for a newly sampled antigen family. A separate development analysis averages over seeds within each complex and resamples complex rows to describe heterogeneity. We do not treat seed–complex cells as independent or combine these two intervals into a joint guarantee.

The holdout protocol explicitly uses exclusion of zero by a 95% seed-bootstrap interval as its decision criterion and reports exact seed-sign bounds alongside it. The two primary intervals are nominal, contrast-wise intervals, not simultaneous family-wise guarantees. Earlier protocols differ, and the ProteinMPNN development protocol left its decision rule ambiguous; both relevant readouts are retained rather than selecting the favorable rule. An unresolved comparison is neither equivalence nor evidence of absence. Development resampling does not undo model-selection reuse (Cawley and Talbot, 2010). Common-epoch checks are sensitivity analyses over correlated observations, not additional independent replications.

3 CORRESPONDENCE SENSITIVITY: DISCOVERY AND HELD-OUT CONFIRMATION

3.1 DEVELOPMENT SEPARATES A USEFUL FEATURE FROM A HARMFUL ASSIGNMENT

Table 1: SAbDab2 development pair AUPRC; ten common seeds. The baseline is the ten-seed mean used in the paired contrasts, not the separate twenty-seed mean. $R^{\dagger}$ is wider on the antibody side; ProteinMPNN's random control is source/width-matched.

Auxiliary encoder	Baseline B	Aligned A	Permuted S	Random R
ESM-IF1	0.022460	0.027571	0.019367	0.021865†
ProteinMPNN	0.022460	0.022825	0.019632	0.021387

ESM-IF1 improves over ESM-2 by 0.005111 AUPRC (95% seed-bootstrap CI $[0.003780, 0.006439]$). Training with the same vectors assigned to different residues instead lowers performance below ESM-2 by 0.003093 (CI $[-0.004391, -0.001807]$). The resulting aligned–permuted difference is 0.008204 (CI $[0.006928, 0.009452]$), positive in all ten seeds. This is a 42.4% ratio-of-means difference, but the absolute effect is our primary report. Both arms use input widths 1,793/1,793 and identical downstream capacity.

The corresponding ProteinMPNN gap is 0.003193 (CI $[0.00220, 0.00421]$; ten positive seeds). Its permuted condition is also below ESM-2: $H = -0.002828$ (CI $[-0.00406, -0.00173]$). Yet its aligned benefit is unresolved, $U = 0.000365$ (CI $[-0.00134, 0.00200]$; seven positive seeds). These measurements exemplify the distinction in Equation 1: correspondence sensitivity can coexist with an unresolved incremental benefit. They do not establish that benefit is absent or that correct correspondence is a universal necessary condition.

The matched ProteinMPNN aligned–random difference is 0.001438, with a positive mean-effect interval but only eight of ten positive seeds. Its sign-probability lower bound is 0.4931, below the alternative 0.5 criterion; because its development protocol did not specify which criterion governed, this contrast remains rule-dependent. The historical random expansion also fails to show a resolved gain over ESM-2. That outcome concerns one parameterization and does not rule out capacity or optimization explanations generally. Appendix C reports every development contrast and the actual matching contract.

The mean does not describe every complex. For ESM-IF1, 162/217 development complexes (74.7%) favor alignment after averaging across seeds. The complex-resampled interval is $[0.006229, 0.010306]$. Effects are right-skewed: the median is 0.004340, versus fifth and ninety-fifth percentiles of -0.010448 and 0.032129. The improvement is not confined to a handful of favorable cases, but approximately one quarter favor the permuted condition. These are development-complex statistics, not test-population or new-antigen-family guarantees.

3.2 The larger development effect has less conclusive test evidence

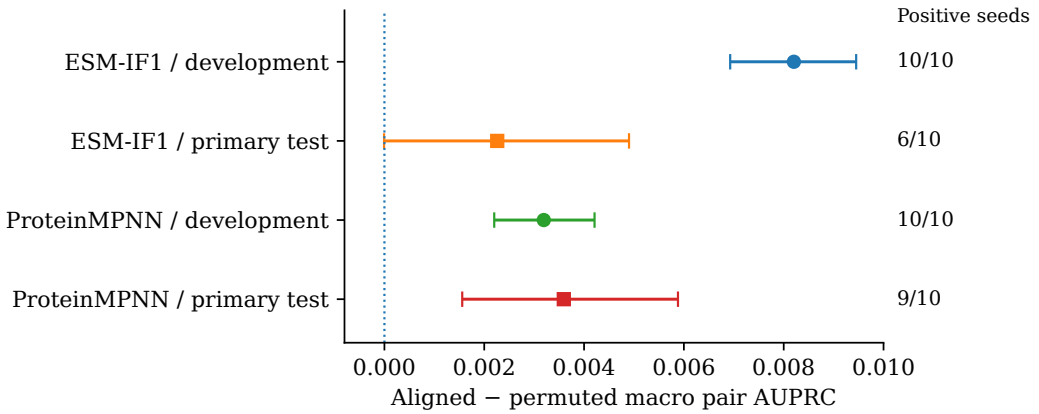

Figure 1: Correspondence sensitivity before and after confirmation. Each point is aligned minus permuted macro pair AUPRC; bars are nominal 95% percentile intervals across ten paired seeds. Development and test contain different complexes. ProteinMPNN retains a positive interval on the primary test partition; ESM-IF1's smaller test interval includes zero. The 217-complex test has the documented prior-exposure qualification in Section 2. This is not an encoder-by-split significance test.

The prespecified 217-complex evaluation gives different conclusions for the two encoders (Table 2). ProteinMPNN's aligned–permuted difference is 0.003594 (CI $[0.001559, 0.005881]$), positive in nine of ten seeds. Its one-sided positive-seed lower bound is 0.6058. It therefore meets the prespecified mean-interval criterion and also the supplementary directional criterion. The point estimate is close to, and slightly above, its development value of 0.003193; we do not infer that the two effects are equal.

ESM-IF1's test difference is 0.002260 (CI $[-0.0000072, 0.004901]$), positive in six of ten seeds. It does not meet the prespecified criterion. The tiny negative lower endpoint is retained rather than rounded to a positive result. The point estimate is approximately 72% below its development effect, but the reduction is not attributed uniquely to model-selection adaptation: the two populations also contain different complexes. Unresolved evidence does not establish that the test effect is zero.

Table 2: Prespecified primary test results, all 217 complexes. Scores and intervals are reported with enough precision to retain the negative ESM-IF1 lower endpoint. "Positive seeds" counts paired differences. The 185-complex exclusion analysis is secondary (Appendix A).

Encoder	Aligned	Permuted	Difference [95% CI]	Positive seeds
ESM-IF1	.027919	.025659	.002260 [−.0000072, .004901]	6/10
ProteinMPNN	.028490	.024896	.003594 [.001559, .005881]	9/10

The 185-complex sensitivity readout increases the estimates to 0.003022 for ESM-IF1 and 0.004909 for ProteinMPNN. Both mean-effect intervals are positive, although ESM-IF1's sign bound remains below 0.5. We retain the primary result rather than promoting the subset that gives ESM-IF1 a positive interval. Excluding the documented exposure does not reverse either point estimate; this narrower observation is not proof that all exposure or dependence has been removed.

The confirmation changes the representation conclusion. Development supports sensitivity for both encoders, whereas the prespecified test criterion resolves only ProteinMPNN. The mean effects remain positive for both: there is no sign reversal. The ordering of their effect magnitudes changes descriptively, but a significant result for one encoder and an unresolved result for another does not establish a significant difference between encoders. Because the confirmation contains no test baseline or random arm, it also cannot establish below-baseline harm or positive baseline-relative transfer on unseen data.

3.3 Perturbation sensitivity is not exclusive to structural features

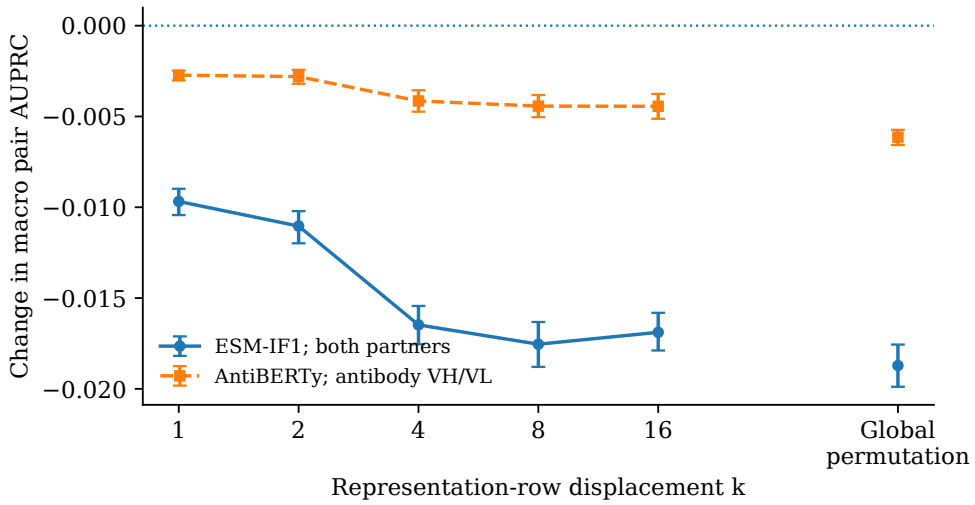

Figure 2: Development-only, evaluation-time perturbations of aligned-trained models. Curves connect finite offsets; global permutation is a separate condition. Bars are 95% seed-bootstrap intervals. Coverage differs: ESM-IF1 augments both partners, AntiBERTy antibody VH/VL. These are losses relative to each unperturbed model, not to ESM-2, and are not the retrained contrasts in Figure 1.

A separate intervention holds an aligned-trained checkpoint fixed and cyclically shifts its auxiliary rows at evaluation. For ESM-IF1, a one-row offset costs 0.009680 AUPRC; an eight-row offset costs 0.017541, approximately 94% of the 0.018717 global-permutation loss. For independently extracted AntiBERTy features, the corresponding losses are 0.002728, 0.004431, and 0.006141. Every tested perturbation lowers performance in every seed for both representations. Thus the sensitivity is not structure-exclusive.

The curve shapes differ, but their causes are not isolated. AntiBERTy augments the antibody side whereas ESM-IF1 augments both partners; encoder training, feature width, and feature pathways also differ. We therefore do not conclude that structure is intrinsically local and sequence intrinsically global. The response is not strictly monotone: ESM-IF1's offset-16 mean is slightly less damaging than offset 8. "Approximately 94%" describes two measured effects; it is not a demonstrated saturation threshold.

We retain row displacement as the operational variable. SAbDab2 arrays predominantly follow observed chain order, while AsEP's CDR/surface selection introduces segment jumps. Chain junctions, unresolved residues, and cyclic wrapping remain exceptions even on SAbDab2. The intervention is a computational reassignment of features, not a physical displacement of atoms or an inferred biological interaction radius. Its evaluation-time loss cannot be equated with the smaller gap obtained after allowing a downstream model to train on permuted features.

4 CONTACT AND RANKING SUPERVISION HAVE DIFFERENT EFFECTS

4.1 A FOUR-ARM INTERVENTION ON THE TRAINING PROCEDURE

All four AsEP development arms retain node supervision and vary only the configured contact and ranking weights,

$$\mathcal{L} = \mathcal{L}_{\mathrm{N}} + \lambda_E \mathcal{L}_{\mathrm{E}} + \lambda_R \mathcal{L}_{\mathrm{R}}, \qquad (\lambda_E, \lambda_R) \in \{0, 1\} \times \{0, 0.5\}. \tag{5}$$

Here $\mathcal{L}_{\mathrm{N}}$ combines epitope/paratope BCE and Dice losses, $\mathcal{L}_{\mathrm{E}}$ supervises matched residue contacts, and $\mathcal{L}_{\mathrm{R}}$ ranks a matched complex against sampled partner substitutions. The substitutions are not experimentally established non-binders and are not assigned all-zero contact maps. They supply a different learning signal from labeled contact edges.

Table 3: AsEP group-validation factorial study. Ten paired seeds per condition; means over complexes and then seeds. Node MCC uses fixed threshold 0.5. Partner retrieval, the ranking objective's own target, is not an endpoint in this table.

Objective	λ_E	λ_R	Pair AUPRC	Node MCC	Node AUPRC
Node	0	0	.00724	.2798	.4299
Node + Edge	1	0	.09055	.2773	.4244
Node + Rank	0	.5	.00670	.2050	.3558
Node + Edge + Rank	1	.5	.06917	.2404	.3884

Adding contact supervision to node-only training raises pair AUPRC from 0.00724 to 0.09055, with positive paired differences in all ten seeds. The node MCC change is -0.0026, positive in three seeds, and unresolved under the recorded rule; node AUPRC is likewise unresolved. This is an absence of detected node harm, not an equivalence or non-inferiority result. Pair improvement from direct pair labels is expected; the informative comparison is what happens when ranking is added to that useful configuration.

Removing ranking from the full objective improves pair AUPRC by 0.0214 and node MCC by 0.0369, both in every paired seed. Node AUPRC also increases by 0.0361. Conversely, adding ranking to node-only training lowers mean node MCC by approximately 0.0748. The two relational signals therefore cannot be treated interchangeably in this system. These observations concern the two prediction endpoints, not a Pareto improvement over every task: the ranking objective's own retrieval performance is outside Table 3.

4.2 THE NODE DIFFERENCE PRECEDES LEARNING-RATE DIVERGENCE

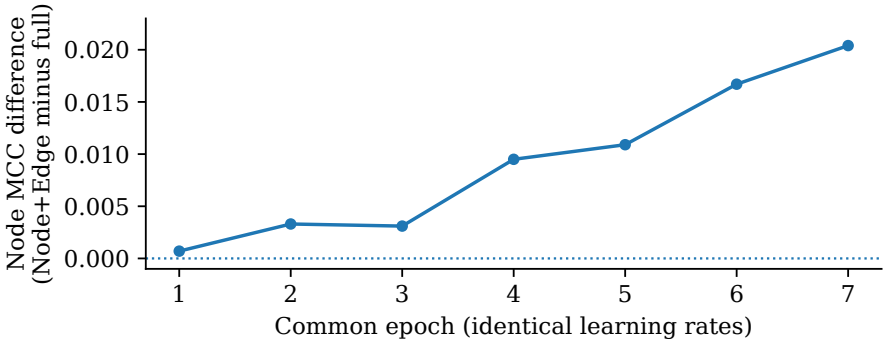

Figure 3: Common-epoch node MCC difference between Node+Edge and the full objective. Both conditions use their initial learning rate in every seed throughout epochs 1–7. Points are the recorded ten-seed means, not independent experiments or reconstructed seed trajectories. The dip at epoch 3 precludes a strict monotonicity claim.

All 40 runs complete twelve epochs. However, `selection_metric=total` denotes a different weighted sum in each arm. This quantity drives ReduceLROnPlateau, best-checkpoint selection,

and the stopping counter. The fixed-weight intervention therefore includes its downstream effects on scheduling and selection; identical configuration strings do not hold the monitored numerical objective fixed.

We examine common epochs to determine whether different selected checkpoints alone explain the node advantage. The Node+Edge minus full-objective estimate is positive at every recorded common epoch. More specifically, the learning rate remains at its initial value for both arms and every seed through epoch 7, while the difference increases overall from 0.0007 at epoch 1 to 0.0204 at epoch 7 (Figure 3). The advantage thus appears before either schedule reduces its rate, not merely after divergent checkpoint choices. Later scheduling can still affect the final magnitude.

These checks strengthen an intervention-level claim: under this training procedure, disabling ranking improves the measured prediction endpoints. They do not isolate a unique gradient-level mechanism. The objective still changes parameter trajectories, adaptive optimizer state, and potentially paths outside the shared trunk. The common-epoch and shared-rate analyses were sensitivity checks on existing histories, not independent holdout confirmations of the factorial result.

4.3 Gradient diagnostics inform interpretation but do not establish a mechanism

Training traces show a negative opening node–edge cosine followed by a rapid durable transition to positive agreement. Recorded primary-stack crossing updates are 104/110/77 on AsEP and 50/39/38 on SAbDab2; a second toolchain gives 104/110/77 and 48/40/35. These are three seeds per cohort with paired technical replays, not twelve independent seeds. The convergence-only AsEP cosine near 0.70 would miss the early regime. Relative gradient magnitude is less stable across cohorts: edge/node norms are approximately 1.1–1.3 on AsEP and 0.3 on SAbDab2.

This node–edge geometry is not the node–ranking comparison in the removal intervention. A short-window ranking gradient was small and nearly orthogonal on the shared trunk, but that observation cannot bound its long-run effect through AdamW, other parameters, or later training. We therefore neither rule out ranking because its local gradient is small nor attribute its performance effect to the node–edge sign transition. Appendix E specifies the trace summaries and their limitations.

5 Related work

Frozen sequence and inverse-folding representations provide useful protein features (Lin et al., 2023; Hsu et al., 2022; Dauparas et al., 2022). AntiBERTy specializes sequence pretraining to antibodies (Ruffolo et al., 2021), and DiscoTope-3.0 already uses inverse-folding latent representations for B-cell epitope prediction (Høie et al., 2024). Residue-level alignment also appears as a representation-learning objective: Birnbaum et al. (2024) jointly embed protein sequences and structures. We do not claim to introduce structural transfer or the importance of residue identity. Our distinction is experimental: baseline-relative benefit, fixed-correspondence corruption, and evaluation-time sensitivity are measured separately, then the central correspondence contrast is checked on a prespecified test population.

AsEP formalizes antibody-specific epitope prediction and includes residue-level structural context (Liu et al., 2024). PEPNet and LF3DRoPE develop stronger task-specific geometric architectures (Chen et al., 2026; Fan et al., 2026). Our system is not a replacement for those models. Its frozen AsEP group/ratio MCCs are 0.1572/0.2584; the formal WALLE table reports 0.152/0.305, and PEPNet reports 0.139/0.401. These are contextual published comparisons, not matched reimplementations or evidence that our representations cause greater out-of-distribution robustness (Appendix F).

Auxiliary objectives and multitask interference are established topics. Gradient similarity has been used to adapt auxiliary losses (Du et al., 2018), GradNorm adjusts task scaling (Chen et al., 2018), and PCGrad projects conflicting gradients (Yu et al., 2020). Our contribution is not that negative transfer or changing gradient geometry exists. It is a factorial separation of two biologically related objectives in a particular predictor, with common-epoch and shared-learning-rate checks, while keeping local gradient agreement distinct from demonstrated endpoint utility.

6 Discussion and limitations

What the held-out result changes. The correspondence effect that was larger and more uniform in development is not the one most clearly resolved on the prespecified test. This limits a practical inference from ablation strength: the size of a development perturbation gap is not, in this study, a reliable guide to how strongly that gap will persist on another set of complexes. It does not establish ProteinMPNN's superiority as a representation, because the confirmation estimates sensitivity rather than baseline-relative benefit, and no encoder-by-split interaction was tested. The ESM-IF1 estimate remains compatible with a positive effect. Reporting the full interval avoids turning a threshold decision into either a claim of zero or an almost-positive success.

Two distinct uses of controls. Within an encoder, aligned–permuted comparisons hold the vector multiset and model capacity fixed. Against ESM-2, feature injection changes both information and parameterization. Random projections probe a particular expansion, not all capacity effects. On the supervision side, a composite-loss ablation would conceal the different contributions of edge and ranking terms. Together these distinctions support a concrete evaluation practice: report the baseline, the targeted intervention, and the intended endpoint separately rather than treating any one favorable contrast as evidence for all three.

Limits that remain after confirmation. The primary test has documented prior baseline exposure for 32 complexes; the prespecified exclusion analysis addresses that record, not every possible indirect influence. Its intervals condition on fixed complexes and do not quantify sampling of new biological families. Development and factorial analyses remain subject to selection reuse, and the latter has no new held-out contrast. Homologous complexes and pretrained-model overlap can limit both cohorts. Two inverse-folding encoders, one downstream architecture, and one main biological task do not establish a universal transfer law. ProteinMPNN's reconstructed oxygen, unequal representation widths, AntiBERTy's narrower coverage, and one fixed permutation realization per sequence constrain cross-model interpretation. We do not equate geometric row continuity with an exact biological length scale.

The ranking intervention concerns a single weight and sampling configuration; retrieval and experimental binding utility were not tested. The results are predictive evaluations, not wet-laboratory validation or a recommendation to alter therapeutic candidates. Numerical reconstruction of saved analyses improves traceability but cannot itself verify biological labels, remove dataset exposure, or establish novelty. These boundaries are part of the result, rather than exceptions to a broader claim.

7 Conclusion

Frozen-feature benefit and correspondence sensitivity are distinct empirical quantities. Development experiments identify below-baseline corruption effects for two inverse-folding encoders, but prespecified test evaluation gives a mixed confirmation: ProteinMPNN retains a resolved aligned–permuted difference, while ESM-IF1's larger development effect attenuates and remains unresolved. Separately, contact and partner-ranking supervision have different endpoint effects in a controlled development study, with a ranking-removal node advantage preceding learning-rate divergence. The evidence supports targeted, endpoint-specific interventions and explicit held-out checks, not a universal claim that added information helps, harms, or transfers for the same reason.

AI USE STATEMENT

Generative AI tools provided limited assistance with literature search, minor hypothesis refinement, code debugging, interpretation of results, and figure preparation. They also assisted with limited wording edits during manuscript revision. AI-assisted suggestions were not treated as empirical evidence. The authors are responsible for the final manuscript, references, implementation, analyses, and scientific claims, including any material prepared with AI assistance.

ETHICS STATEMENT

This work analyzes existing protein-structure datasets and prediction systems. It does not report new human-subject or animal experiments, clinical efficacy, or experimental validation of designed molecules. Model outputs should not be interpreted as validated binding or safety assessments. Redistribution of upstream structures, pretrained weights, and derived artifacts remains subject to their original licenses and access conditions. Unresolved effects, prior evaluation exposure, and negative confirmation outcomes are reported to limit misleading claims about downstream utility.

REPRODUCIBILITY STATEMENT

The development evidence snapshot fixes analysis inputs, code references, result artifacts, and feature-store digests. Eight designated analyses were numerically reconstructed in clean processes from saved inputs within a predefined 10^{-10} tolerance; Appendix G lists their coverage. This is reanalysis of saved experiments, not independent retraining. It does not by itself certify all ProteinMPNN, factorial, or later holdout results. The post-freeze confirmation has a separate protocol and result record, and is kept distinct from the original eight-step reproduction. Appendix A records its populations, fixed checkpoints, endpoints, and exposure sensitivity. The manuscript-source archive builds the document and regenerates figures from reported numerical summaries; full experimental reproduction additionally requires the original numerical inputs, checkpoints, and recorded environment.

REFERENCES

Foster Birnbaum, Saachi Jain, Aleksander Madry, and Amy E. Keating. Jointly embedding protein structures and sequences through residue level alignment. *PRX Life*, 2:043013, 2024. doi:10.1103/PRXLife.2.043013.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning*, volume 80 of PMLR, pp. 794–803, 2018.

Jiayi Chen, Guixu Zhang, Zhijian Xu, and Qian Zhang. PEPNet: A two-stage point cloud framework with hierarchical embedding and antigen–antibody interaction modeling for epitope prediction. *Briefings in Bioinformatics*, 27(1):bbag067, 2026. doi:10.1093/bib/bbag067.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J. Ragotte, Lukas F. Milles, et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022. doi:10.1126/science.add2187.

Yunshu Du, Wojciech M. Czarnecki, Siddhant M. Jayakumar, Mehrdad Farajtabar, Razvan Pascanu, and Balaji Lakshminarayanan. Adapting auxiliary losses using gradient similarity. *arXiv preprint arXiv:1812.02224*, 2018.

James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M. Deane. SAbDab: The structural antibody database. *Nucleic Acids Research*, 42(D1):D1140–D1146, 2014. doi:10.1093/nar/gkt1043.

Chuanliu Fan, Nan Yu, Junjie Wu, and Guohong Fu. Inter-residue geometry attention for antibody-specific epitope prediction. *arXiv preprint arXiv:2608.01092v1*, 2026.

Magnus Haraldson Høie, Frederik Steensgaard Gade, Julie Maria Johansen, Charlotte Würtzen, Ole Winther, Morten Nielsen, and Paolo Marcatili. DiscoTope-3.0: Improved B-cell epitope prediction using inverse folding latent representations. *Frontiers in Immunology*, 15:1322712, 2024. doi:10.3389/fimmu.2024.1322712.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *International Conference on Machine Learning*, volume 162 of PMLR, pp. 8946–8970, 2022.

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.

Chu'nan Liu, Lilian Denzler, Yihong Chen, Andrew Martin, and Brooks Paige. AsEP: Benchmarking deep learning methods for antibody-specific epitope prediction. In *Advances in Neural Information Processing Systems*, 37, Datasets and Benchmarks Track, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Jeffrey A. Ruffolo, Jeffrey J. Gray, and Jeremias Sulam. Deciphering antibody affinity maturation with language models and weakly supervised learning. *arXiv preprint arXiv:2112.07782*, 2021.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, 33, 2020.

A HELD-OUT CONFIRMATION: DESIGN, EXPOSURE, AND COMPLETE READOUT

A.1 CHRONOLOGY AND UNCHANGED MODELS

The development evidence was fixed before a new confirmation protocol was recorded. The confirmation was motivated by a specific limitation of the original evidence: representation comparisons had been evaluated on a validation population that participated in development. The additional study did not select a new architecture, retrain the downstream models, change the feature extractor, or select a different checkpoint. It evaluated the ten existing paired best checkpoints for each of two contrasts: aligned minus fixed-permuted ESM-IF1, and aligned minus fixed-permuted ProteinMPNN.

The protocol named macro pair AUPRC as the endpoint. It ruled out new threshold tuning, ensembles, additional feature arms, and a companion-metric search on the test partition. The prespecified primary statistic was the mean of paired seed differences with a percentile-bootstrap interval. Exclusion of zero by that interval was the prespecified decision criterion; positive-seed counts and exact sign bounds were reported alongside it. No outcome-dependent choice between the criteria is used in the held-out analysis. This distinguishes the confirmation from the earlier ProteinMPNN development protocol, whose criterion was ambiguous.

The confirmation is a post-development test, not a prospective study designed before all modeling work. The development snapshot is preserved, and the confirmation record is an amendment rather than a replacement for it. A date-stamped protocol in the experimental repository documents that sequence; it is not described as independently registered at an external registry. The outcome limits the claims: one encoder meets the primary criterion and the other does not.

A.2 WHAT THE ACCESS AUDIT DOES AND DOES NOT ESTABLISH

The SAbDab2 manifest has 3,229 training, 217 validation, and 217 test members, mutually disjoint under its instance identifiers. Before the new evaluation, the audit found no test outputs for the focal representation arms, 141 relevant evaluation receipts naming validation, and 120 inspected configurations without a test manifest used for selection. These checks support separation of the focal models' training/selection workflow from the confirmation partition; they do not establish that every test complex was unknown to every previous project activity.

In particular, three scripts for an unrelated EpiFormer baseline evaluated a different, 455-instance test population. That population overlaps the present training set by 365 instances, validation by 58, and test by 32. The 32 prior test outputs create a possible researcher-exposure pathway even though the split, model, and protocol differ. Whether that information indirectly influenced any earlier modeling judgment cannot be established mechanically from file absence. We therefore report this exposure rather than calling the full test partition completely untouched.

Before the focal test scores were read, the confirmation specified an exclusion population containing the other 185 test complexes. The term "exposure-excluded" means no overlap with that documented earlier baseline evaluation. It does not certify the absence of homologues, overlap with pretraining corpora, or every conceivable exposure. Both populations are evaluated in the same once-authorized read. The 217-complex population remains primary regardless of which interval looks more favorable.

A.3 STATISTICAL INTERPRETATION

ProteinMPNN meets the primary mean-interval criterion on the 217-complex partition and also has a positive-seed probability lower bound exceeding 0.5. ESM-IF1 meets neither criterion on that partition. Its exact lower interval endpoint is $-7.1989706811 \times 10^{-6}$, not zero and not positive. We retain the endpoint's sign rather than changing a decision through rounding. At the same time, failure to exclude zero is not evidence that the effect is absent: the estimate and much of the interval are positive.

The ESM-IF1 sensitivity interval on 185 complexes excludes zero, but its sign bound remains below 0.5. This does not replace or overturn the primary result. The inclusion/exclusion populations are nested and their intervals should not be treated as independent replication. Similarly, the two test contrasts share models' base inputs, seeds, and complexes, and their nominal intervals do not provide

Table 4: Complete prespecified confirmation. Intervals resample paired seed differences, not complexes. The full 217-complex rows are primary; the 185-complex rows are exposure-exclusion sensitivities. A positive point estimate is distinguished from an interval excluding zero.

Encoder	Population	Aligned	Permuted	Mean difference
ESM-IF1	Primary 217	.027919	.025659	$+.002260$
ESM-IF1	Exclusion 185	.029271	.026248	$+.003022$
ProteinMPNN	Primary 217	.028490	.024896	$+.003594$
ProteinMPNN	Exclusion 185	.030425	.025516	$+.004909$

Encoder	Population	Nominal 95% interval	Positive seeds	Sign bound
ESM-IF1	Primary 217	$[-.0000072, .004901]$	6/10	.3035
ESM-IF1	Exclusion 185	$[.000310, .006084]$	7/10	.3934
ProteinMPNN	Primary 217	$[.001559, .005881]$	9/10	.6058
ProteinMPNN	Exclusion 185	$[.002662, .007365]$	9/10	.6058

a family-wise confidence statement. We do not retroactively add an outcome-favorable multiplicity policy or equivalence margin.

The primary test intervals quantify variability across the ten trained seeds while conditioning on these complexes. The included confirmation report does not supply a new complex-resampling or biological-cluster interval, and the manuscript does not invent one. The separate development heterogeneity analysis is not transferred to the test partition. This leaves uncertainty about sampling new complexes or antigen families beyond the evaluated population.

Table 5: Development and primary-test correspondence effects. The columns describe different complex populations. Ratios and retained fractions are descriptive arithmetic, not formal interaction tests or calibrated predictions of generalization.

Encoder	Development C	Primary test C	Positive seeds (dev/test)	Test/dev
ESM-IF1	.008204	.002260	10/10; 6/10	.275
ProteinMPNN	.003193	.003594	10/10; 9/10	1.126

The difference in declared outcomes is not itself a declared difference between encoder effects. Nor is ESM-IF1's attenuation uniquely evidence of validation overfitting: composition of the two complex populations and ordinary sampling variation can contribute. No model-by-split interaction test was prespecified or added after reading these results. The narrower conclusion is that the development effect estimates did not extrapolate uniformly, and only ProteinMPNN's test correspondence effect met the primary criterion.

A.4 What this test cannot answer

Only aligned–permuted comparisons were authorized. Neither ESM-2 alone nor a random control was evaluated as a new test arm, so this confirmation cannot establish test-set baseline-relative benefit U or below-baseline harm H. A positive C could be driven by differing behavior of either arm relative to an unmeasured baseline. The node/ranking factorial intervention was not tested here; its evidence remains developmental. The test cannot explain ProteinMPNN's feature quality under reconstructed oxygen, prove a universal need for correspondence, or establish a physical interaction length scale.

The test partition is closed after this evaluation. This manuscript revision uses the existing reported means, intervals, and counts; it adds no model evaluation, alternate endpoint, checkpoint search, or new test-derived hypothesis test. Reporting a sensitivity result and exposing a negative or unresolved primary outcome are parts of the original confirmation, not a license to continue adapting to the test set.

B MODEL, LOSSES, AND EXECUTION DETAILS

B.1 PARTNER-CONDITIONED ARCHITECTURE

For antigen residue i and antibody residue j, the input vectors concatenate a 1,280-dimensional frozen ESM-2 representation and the enabled auxiliary sources. Each source includes a presence indicator. Separate LayerNorm–linear–SiLU–dropout projections map the two streams to width $d = 128$. A bidirectional cross-attention block uses eight attention heads. Each direction has its own parameters and a scalar sigmoid gate initialized with logit -2. Residue-wise feed-forward layers expand to $4d$ and return to d. Padding masks are enforced throughout. Explicit intra-chain geometric attention is not enabled in this configuration; geometry enters through the frozen structural representations.

With final contextual states $h_i^{\mathrm{ag}}, h_j^{\mathrm{ab}}$, the bilinear decoder uses two directional projected affinities:

$$z_{ij} = \frac{1}{\sqrt{d_p}} \left[\alpha_1 \langle Q_{\mathrm{ag}} h_i^{\mathrm{ag}}, K_{\mathrm{ab}} h_j^{\mathrm{ab}} \rangle + \alpha_2 \langle K_{\mathrm{ag}} h_i^{\mathrm{ag}}, Q_{\mathrm{ab}} h_j^{\mathrm{ab}} \rangle\right] + b_{\mathrm{ag}}(h_i^{\mathrm{ag}}) + b_{\mathrm{ab}}(h_j^{\mathrm{ab}}) + b_0, \quad (6)$$

where $d_p = 64$ and (α_1, α_2) is a learned softmax. Pair probabilities are $\sigma(z_{ij})$. Epitope and paratope probabilities come from independent linear heads on the contextual states. The implementation additionally computes pooled pair marginals and a normalized complex score; these auxiliary outputs should not be confused with the node heads used for node metrics. The default node–pair consistency weights are zero.

The three loss components are

$$\mathcal{L}_{\mathrm{N}} = \tfrac{1}{4}\big(\mathcal{L}_{\mathrm{BCE}}^{\mathrm{epi}} + \mathcal{L}_{\mathrm{BCE}}^{\mathrm{para}} + \mathcal{L}_{\mathrm{Dice}}^{\mathrm{epi}} + \mathcal{L}_{\mathrm{Dice}}^{\mathrm{para}}\big), \quad (7)$$

$$\mathcal{L}_{\mathrm{E}} = \mathrm{WBCE}(\{z_{ij}\}, \{Y_{ij}\}; \text{valid pair mask}), \quad (8)$$

$$\mathcal{L}_{\mathrm{R}} = -\log \frac{\exp(s^+/\tau)}{\exp(s^+/\tau) + \sum_{m=1}^{K} \exp(s_m^-/\tau)}. \quad (9)$$

The ranking score s is the model's pooled, normalized interaction score; the implementation supports centered scores and log-mean-exp pooling. The recorded default ranking temperature is $\tau = 0.1$. Positive BCE weights are estimated on training data and capped at 100. Dice loss uses additive smoothing of 1. Count and node–pair consistency losses are disabled in the principal configuration. Training samples four candidate negatives from a pool of 64, considering partner substitutions in both directions. Candidate negatives are not experimentally verified non-binders; this is a limitation of the ranking target, not proof of the ranking result's mechanism.

Table 6: Principal executable configuration. The representation launcher uses partner-ranking validation loss for model selection; the factorial launcher uses each arm's total loss. The width-confirmation configuration is a separate experiment.

Quantity	Setting
Downstream width / pair projection width	128 / 64
Cross-attention blocks / heads	1 / 8
Dropout / gate initial logit	0.2 / -2
Optimizer	AdamW (Loshchilov and Hutter, 2019)
Initial learning rate / weight decay	5×10^{-5} / 5×10^{-4}
Gradient clipping norm	1.0
Learning-rate scheduler	ReduceLROnPlateau; factor 0.5; patience 3
Minimum learning rate	10^{-7}
Principal epoch budget	12
Node BCE / Dice weights	0.25 for each of four terms
Contact / ranking weights (full model)	1.0 / 0.5
Training negative count / pool size	4 / 64
Factorial node threshold	0.5

B.2 DATA AND FEATURE BOUNDARIES

The experiments retain each cohort's fixed contact labels and candidate-residue masks rather than relabeling either population after observing model scores. For SAbDab2, contact preparation copies

the authoritative pair table's labels and antibody/antigen indices into sparse maps; only listed candidate rows enter the loss and evaluation. Unlisted residue pairs are not silently converted into labeled negatives. This inherited labeling contract is not an independently reconstructed distance-based annotation. The original AsEP annotation uses a 4.5 Å heavy-atom contact criterion (Liu et al., 2024); its antibody CDR and antigen-surface selection must not be described as selecting only positive interface residues. The installed group manifest has 169 validation and 169 test items, while the publication reports 170 in each. This manuscript uses the installed membership, not the published counts as a substitute. Published benchmark rows therefore provide context rather than an asserted byte-identical reimplementation. The SAbDab-derived cohort is a project-specific release, not a new claim about all of SAbDab. Split authority, candidate masks, contact stores, and instance identifiers belong to its frozen input contract.

SAbDab2 contains 3,229 training, 217 validation, and 217 primary test members. Membership-disjoint partitions do not establish complete homology or pretraining separation. The structural encoders consume structure-derived inputs; these experiments are not sequence-only deployment evaluations. We do not establish that experimental-chain features are representative of arbitrary predicted unbound structures. Structural pretraining overlap, homologous complexes, and differences between full observed chains and selected residue arrays constrain generalization. Input/label contracts are held constant within each intervention, not asserted identical across cohorts.

C Representation controls and full reported contrasts

Table 7: The historical random control matches the combined two-source augmentation, not the ESM-IF1-only augmentation. More parameters are not assumed to give a conservative control, since parameter count need not improve a fixed-budget optimizer monotonically.

Condition	Antibody width	Antigen width	Trainable parameters
ESM-2	1,280	1,280	763,401
ESM-2 + aligned ESM-IF1	1,793	1,793	896,781
ESM-2 + permuted ESM-IF1	1,793	1,793	896,781
Historical random control	2,306	1,793	963,471
ESM-2 + AntiBERTy + ESM-IF1	2,306	1,793	963,471

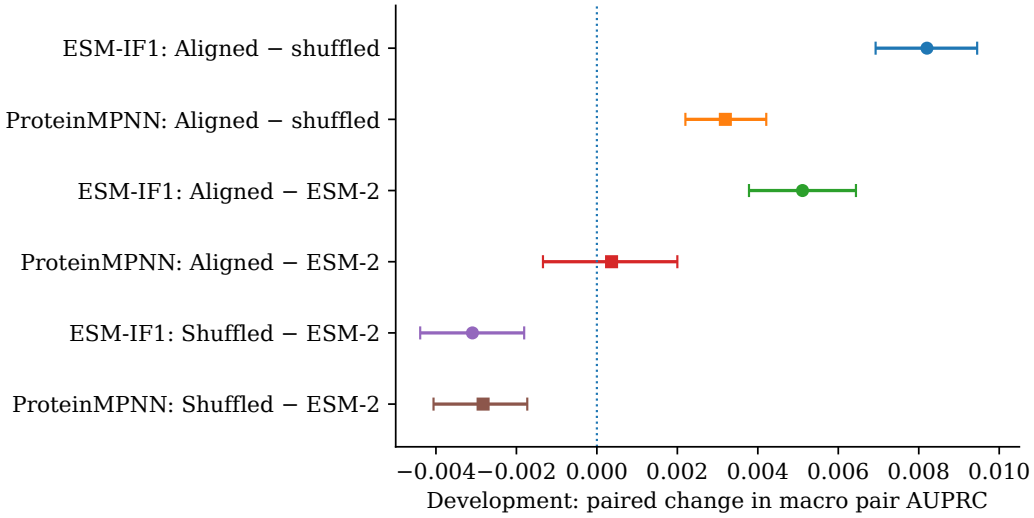

Figure 4: Full development comparison of the three estimands in Equation 1. The aligned–permuted gap is distinct from benefit relative to ESM-2. Intervals are the reported 95% seed-bootstrap intervals; none is inferred from a table of rounded arm means.

The random projection is a fixed function of base embeddings, not independent noise. It preserves no additional information conditional on the base input but can change parameterization, normalization, and optimization. ProteinMPNN uses a separately constructed control mirroring its 128-dimensional source and partner coverage. Aligned/permuted comparisons within each encoder have identical input dimensions. Cross-encoder magnitudes are not controlled estimates of an architecture or modality effect.

Table 8: ProteinMPNN paired contrasts on the ten shared seeds. Intervals are contrast-wise 95% percentile seed-bootstrap intervals. The aligned–random comparison passes the mean-interval criterion but not the alternative positive-seed criterion. The protocol's unspecified decision rule is retained as a limitation.

Contrast	Mean difference	95% interval	Positive seeds
Aligned − permuted	$+0.003193$	$[+0.00220, +0.00421]$	10/10
Aligned − random	$+0.001438$	$[+0.00030, +0.00260]$	8/10
Aligned − ESM-2	$+0.000365$	$[-0.00134, +0.00200]$	7/10
Random − ESM-2	-0.001073	$[-0.00255, +0.00031]$	2/10
Permuted − ESM-2	-0.002828	$[-0.00406, -0.00173]$	0/10
Permuted − random	-0.001755	$[-0.00253, -0.00096]$	1/10

The base mean in Table 1 is the ten-seed mean 0.022460, used in these paired contrasts. A separate existing twenty-seed ESM-2 report gives 0.021764. Substituting that latter mean into the ten-seed table would make the means and paired differences inconsistent; it is not used here.

The ProteinMPNN extraction reads encoder states after three message-passing layers, not amino-acid logits or autoregressive decoder states. The chosen checkpoint is `v_48_020`; the backbone store contains N/Cα/C, and oxygen positions are reconstructed. Shared preprocessing across aligned and permuted arms does not make reconstruction error cancel: it could interact with correspondence. Consequently, a detected corruption effect neither validates the oxygen approximation nor explains away an unresolved aligned benefit.

AntiBERTy is sequence-conditioned. The self-extracted SAbDab2 condition embeds VH and VL separately and concatenates them on the antibody side. The AsEP field named `igfold` contains an AntiBERTy representation accessed through that software pathway, not evidence that an IgFold structural encoder was tested. Numerical checks on two full-domain examples were consistent with AntiBERTy outputs. They do not reconstruct a missing original-cache hash, establish an exact checkpoint revision from two examples, or certify byte identity across the entire cache. We separate model identity, extraction evidence, row mapping, and derived-store digest.

D Statistical definitions and interpretation

AUPRC in the implementation is tie-aware, non-interpolated average precision. For a complex with $P > 0$ positive pairs, tied scores are processed as a group. Each group's positive-count increment divided by P is multiplied by precision after including the entire group. Complexes without a positive label have undefined AP and are not assigned an artificial AP of zero. The valid-complex population must be common across paired arms. MCC uses the usual confusion-matrix expression and is set to zero when its denominator vanishes; this convention is shared across arms.

For paired seed scores a_s, b_s, the primary estimand is

$$\widehat{\Delta} = \frac{1}{S}\sum_{s=1}^{S}(a_s - b_s). \tag{10}$$

Two possible relative summaries are

$$\widehat{r}_{\mathrm{RoM}} = \frac{\bar{a} - \bar{b}}{\bar{b}}, \qquad \widehat{r}_{\mathrm{MoR}} = \frac{1}{S}\sum_{s=1}^{S}\frac{a_s - b_s}{b_s}. \tag{11}$$

They are generally different. The main representation percentages use r_{RoM}; the separate pair-head baseline-margin report also retains its historical r_{MoR}. Neither quantity substitutes for the absolute difference, and seed pairing does not require choosing one over the other.

The representation comparison protocol uses exclusion of zero by a bootstrap interval for the mean difference. The corruption protocol uses a one-sided exact lower bound on the probability of a negative seed difference exceeding 0.5. The pair-head baseline-margin protocol adds an effect-size threshold. These decision policies are reported by experiment rather than harmonized after observing results. For ten of ten signs in a specified direction, the 95% one-sided exact lower bound is approximately 0.7411; for eight of ten it is approximately 0.4931. A test of mean improvement and a test of direction probability concern different estimands and need not agree.

For ESM-IF1 complex heterogeneity, let d_{sc} be the AP difference for seed s and complex c, and $\bar{d}_c = S^{-1} \sum_s d_{sc}$. Resampling $\bar{d}_c$ preserves each complex's seed bundle. The complex interval conditions on the trained seed set, whereas a seed interval conditions on the observed evaluation population. We do not treat all $S \times C$ entries as independent, claim a joint variance decomposition from the two marginal standard deviations, or interpret a complex interval as a new-antigen-family prediction interval.

All six ProteinMPNN contrasts are shown, including unresolved comparisons. No post-hoc equivalence or non-inferiority margin is imposed. The mean ordering alone does not establish the prespecified ordering in an inferential sense. The paired-epoch check is a post-hoc sensitivity analysis over correlated epochs; its multiple positive estimates do not represent additional independent seed replications. Seed-wise mean-effect intervals for the factorial contrasts should not be inferred from the rounded arm means. The factorial evidence reported here consists of the recorded paired mean differences and direction counts.

E TRAINING-HISTORY AND GRADIENT DIAGNOSTICS

Table 9: Objective-dependent training histories; ten runs per arm, all completing 12 epochs. Selected epochs are mean ± sample standard deviation. Reduction counts and final rates are means.

Objective	Selected epoch	LR reductions	Final learning rate
Node	7.8 ± 2.27	0.7	3.38×10^{-5}
Node + Edge	7.3 ± 1.62	0.8	3.00×10^{-5}
Node + Rank	4.0 ± 1.10	1.5	1.88×10^{-5}
Full	6.1 ± 1.45	1.1	2.38×10^{-5}

The fixed-epoch differences are positive at all 12 reported epochs. Their means at epochs 1, 2, 3, 4, 6, 8, 10, and 12 are respectively 0.0007, 0.0033, 0.0031, 0.0095, 0.0167, 0.0202, 0.0280, and 0.0479. These are reported means, not a reconstructed continuous trajectory. In particular, the decrease from epoch 2 to 3 prevents a claim of strict monotonicity.

The gradient analysis computes

$$\gamma_t = \frac{g_{\mathrm{N},t}^{\mathsf{T}} g_{\mathrm{E},t}}{\|g_{\mathrm{N},t}\| \, \|g_{\mathrm{E},t}\|} \tag{12}$$

on real training batches and shared parameters. The trace summarizer uses its recorded smoothing and durable-crossing logic; the opening summary averages the first 40 updates. A negative opening mean does not imply every initial update is negative. The plotting and aggregation code share the same smoothing implementation. Cosines averaged over steps are not equivalent to the cosine of a sum of step gradients, and we do not exchange these statistics.

Table 10: Recorded positive-crossing updates for three seeds per cohort. The second stack is a technical replay of the same seed/cohort combinations.

Cohort	Stack	Seed 1	Seed 2	Seed 3
AsEP	Primary	104	110	77
AsEP	Second	104	110	77
SAbDab2	Primary	50	39	38
SAbDab2	Second	48	40	35

The source traces contain 1,383 steps per AsEP epoch and 3,229 per SAbDab2 epoch. Crossing within a small fraction of the first epoch is a property of these runs, not a universal fraction of optimization. The three converged-checkpoint measurements of node–edge cosine, approximately 0.70, are on AsEP; an earlier cohort label in the experimental notes was corrected. Neither the crossing measurements nor converged cosines establish that the observed early conflict causes the ranking-removal benefit. The factorial intervention and node–edge geometry involve different task comparisons.

F Frozen benchmark context and a failed configuration confirmation

Table 11: Reported epitope MCC for context, not a controlled reimplementation. External numbers come from the named paper and configuration. No claim of statistical superiority is made across rows. Dashes indicate no number used for that configuration.

Method / configuration	Group	Ratio	Source
Frozen predictor (single-model mean)	0.1572	0.2584	This study
WALLE	0.152	0.305	Liu et al. (2024), Table 1
PEPNet	0.139	0.401	Chen et al. (2026), Table 2
PEPNet+LE	0.156	0.337	Chen et al. (2026), Table 2
LF3DRoPE, main configuration	0.171	0.410	Fan et al. (2026)
LF3DRoPE, CDR-type features	0.188	—	Fan et al. (2026)

WALLE's formal table reports standard errors of 0.019 and 0.023 for group and ratio MCC. These are not seed-bootstrap intervals and must not be compared as though they were. The lower WALLE values in an earlier repository summary are not substituted for the formal table. The LF3DRoPE CDR-type group result is reported as 0.188 ± 0.009 and is kept separate from its main-table configuration. The source is a preprint, not labeled a peer-reviewed accepted result.

The original frozen predictor's group/ratio AUROC values are approximately 0.6970/0.7644, and F1 values 0.2445/0.3280. These test measurements used validation-selected thresholds. They must not be compared directly with the fixed-0.5 factorial node MCCs on validation. Ensembles are secondary and do not replace single-model rows when an ensemble number is more favorable.

In a separate development comparison, both width-128 and width-256 models used total-loss selection. Width 256 improved same-validation, threshold-selected MCC by approximately 0.0278. This is a width effect conditional on that selection procedure, not an isolated benefit of switching away from partner-ranking selection. The selected wider configuration subsequently reached ratio-test MCC 0.2551, compared with 0.2584 for the original configuration. This does not support an improvement. The newer model's own interval containing the older mean does not by itself test their paired difference or establish equivalence. The negative confirmation is retained without reopening either test split for further model development.

G Evidence provenance and numerical reproduction

The original development evidence package distinguishes experimental inputs, analysis outputs, claims, and narrative explanations. The later held-out confirmation is an explicitly recorded extension and is not retroactively included in the original eight-step reproduction. Input identity requires the run root and artifact digest as well as a filename; files with the same name in different runs need not denote the same experiment. The reproduction harness compares explicitly mapped numerical fields rather than guessing that differently named fields are equivalent.

For the central ESM-IF1 difference, a saved-output reconstruction yields 0.008203640685771799 compared with the earlier report's 0.008203640685771797. Its final compare-only pass records the field mapping and links to the earlier reconstruction. A metadata refresh is not labeled a new scientific verification. Failed execution attempts and schema differences remain in the audit history. The eight-step list does not separately name the complete ProteinMPNN development analysis, the complete four-arm objective analysis, or the later held-out confirmation; their existence in the frozen

Table 12: Scope of the final eight-step numerical reproduction record. "Match" means no numerical discrepancy under the declared tolerance, not a fresh training run.

Analysis	Final numerical comparison
ESM-IF1 aligned–permuted pair contrast	Match
Pair-head margin over predicted-marginal outer product	Match
ESM-IF1 complex heterogeneity	Match
ESM-IF1 inference corruption ladder	Match
AntiBERTy inference corruption ladder	Match
Gradient crossing summaries	Match
Width comparison under total-loss selection	Match
Ratio-test configuration confirmation	Match

snapshot is not presented as equivalent to an independently rerun cold-reproduction step. The scope of a reproduction claim is the named analysis, not every number that appears in a manuscript.

The feature-provenance record distinguishes ESM-IF1 encoder output, AntiBERTy sequence output, and ProteinMPNN encoder output, with cohort-specific residue mappings and derived-store digests. For the copied AsEP antibody cache, the original cache is no longer available in the experimental environment. Tensor-level identity checks support its AntiBERTy attribution, but are not a substitute for the missing source digest. Some upstream weight identifiers are recorded as model revisions rather than full checkpoint hashes. These limitations are retained rather than replaced with invented provenance.

Saved prediction tensors, model weights, and upstream licensed data are not all embedded in the lightweight repository. Reconstructing a particular analysis requires its original numerical inputs, source manifest, recorded environment, and command. A source-only manuscript archive is therefore not the complete experimental reproduction package. Numerical agreement verifies the execution of the specified analysis on saved inputs; it does not prove the absence of implementation errors, data leakage, biological confounding, or selection bias.