Supporting Information

ECD length and fold identity shape the calibration of AI-predicted immune-receptor interactions: lessons from 834 experimentally tested pairs

Hsiu-Chi Tsai

Supplementary Information Roadmap

Tables S1–S4: Core benchmark data (Boltz-2, AlphaFold2-Multimer, merged comparison). Tables S5–S9: Statistical analyses (fold stratification, seed stability, length bias, logistic regression, bootstrap). Tables S10–S11: Sensitivity to MSA depth and model count. Table S12: SPOC trained-classifier baseline. Tables S13–S21: Advanced statistics (bootstrap full results, calibration, LOPO-CV, PPV, seed effect sizes, length-binned confusion, recalibration, aHCP threshold sensitivity, PR-AUC). Tables S22–S25: Cross-model PAE metrics, actif$p$TM, positive-only recall control, unconditional multi-seed. Tables S26–S28: AF2-M 1-model full 834, antibody/TCR exclusion, FOB sensitivity. Table S29: AF2-M 5-model best-of-pool 834 (closing model-count asymmetry). Table S30: Boltz-2 v2.2.1 version drift. Table S31: Pre-registered 242-pair multi-seed. Table S32: Recycling sensitivity. Table S33: PDB training-data overlap stratification. Table S34: Template-ON vs. template-OFF control on 5 case-study pairs.

Table S1. Complete Boltz-2 prediction results (834 pairs)

Full prediction scores for all 834 protein pairs are provided as a separate CSV file: `table_S1_boltz2_834pair`
Columns include `pair_id`, `arm`, `pair_type`, `protein_a_gene`, `protein_b_gene`, `domain_group_a`, `domain_group_b`, `ecd_total_len`, `iptm`, `ipsae`, `pdockq2`, `lis`, `pdockq` (v1), `n_interface`, and `fob` (apECIA fold-over-background).

Table S2. Seed stability analysis (87 pairs × 5 seeds)

Per-seed prediction scores are provided as `table_S2_seed_stability_87pairs.csv`. Pairs were selected as having ipSAE ≥ 0.3 or pDockQ2 ≥ 0.1 in the primary (seed 0) run, then re-predicted with five independent diffusion seeds (seeds 0–4), yielding 435 total samples. The coefficient of variation (CV = SD/mean) across seeds was used to assess prediction stability. True positives showed significantly lower CV than assay-negative high-confidence predictions (aHCPs) for pDockQ2 (mean 0.278 vs. 0.547; one-tailed Mann–Whitney $p = 8.6 \times 10^{-5}$; Cliff's $\delta = +0.468$ by Table S17).

Table S3. AlphaFold2-Multimer prediction results (833 pairs)

`table_S3_af2multimer_833pairs.csv`. Per-pair AlphaFold2-Multimer v3 (ColabFold 1.6.1, single model, three recycles) predictions for the 833 pairs that completed (one pair, PTPRF × LRIG3, failed due to MSA retrieval timeout in the Round 1 ColabFold campaign and is therefore absent from Table S3; this same pair was recovered in the Round 4 actif*p*TM rerun (Table S23) using a later successful MSA cache, eliminating the 833 vs. 834 pair-count discrepancy for actif*p*TM). Columns: `pair_id`, `arm`, `pair_type`, `protein_a_gene`, `protein_b_gene`, `domain_group_a`, `domain_group_b`, `ecd_total_len`, `iptm`, `ptm`, `plddt`, `max_pae`, `composite` ($= 0.8\times$ ipTM $+ 0.2\times$ pTM).

Table S4. Merged Boltz-2 and AlphaFold2-Multimer comparison (833 pairs)

`table_S4_merged_comparison_833pairs.csv`. Side-by-side comparison of all confidence metrics from both models for the 833 matched pairs. Used as the primary input for cross-model AUC and Spearman comparisons reported in Results.

Table S5. Pairwise Fisher exact tests for false positive rate differences across folds

`table_S5_fisher_exact.csv`. All ten pairwise comparisons of false positive rates (at ipSAE ≥ 0.5) across the five negative arms. Columns: `fold_1`, `fold_2`, `fpr_1`, `fpr_2`, `fp_count_1`, `fp_count_2`, `n_1`, `n_2`, `p_raw`, `p_holm` (Holm–Bonferroni corrected), and `significant` ($p_{\text{Holm}} < 0.05$). IgC and non-IgSF FPRs were significantly higher than FN3 and IgI after correction.

Table S6. Seed stability coefficient of variation per pair

`table_S6_seed_stability_cv.csv`. Per-pair coefficient of variation (CV) for pDockQ2 and ipSAE across the 5 diffusion seeds. Columns: `pair_id`, `interaction`, `pdockq2_mean`, `pdockq2_std`, `pdockq2_cv`, `ipsae_mean`, `ipsae_std`, `ipsae_cv`. For pairs with pDockQ2 ≥ 0.3, applying a CV ≤ 0.2 filter retained 87% (20/23) of true positives while removing the single high-scoring assay-negative high-confidence prediction (aHCP) in the subset.

Table S7. Length bias correlations across confidence metrics

`table_S7_length_bias.csv`. Pearson correlation between each confidence metric and $\sqrt{\text{ECD length}}$ computed among non-interacting pairs only ($n = 712$), to remove the binding-status confound. Columns: `metric`, `pearson_r`, `p_value`, `n`. AlphaFold2-Multimer ipTM showed

essentially no length dependence ($r = +0.016$) while Boltz-2 ipTM showed strong negative correlation ($r = -0.484$), a ∼30-fold difference in $|r|$.

Table S8. Logistic regression coefficients for aHCP prediction

`table_S8_logistic_regression.csv`. Two models predicting aHCP status (defined as pDockQ2 ≥ 0.3 within the 713 assay-negative pairs):

Model 1: `aHCP` $\sim$ `sqrt(ECD_length)`, fit by unpenalized statsmodels Logit. In-sample AUC = 0.861. Coefficient −0.301 (SE 0.115, $p = 0.0089$), indicating that longer ECDs are associated with lower aHCP risk in this dataset.

Model 2: `aHCP` $\sim$ `sqrt(ECD_length) + fold_dummies`, fit by sklearn `LogisticRegression` with default L2 regularization ($C = 1.0$). In-sample AUC = 0.917; ΔAUC = +0.056 over Model 1. Reference fold: FN3. Coefficients: $\sqrt{\text{ECD length}}$ −0.317, IgC +0.789, IgI −0.033, IgV −0.891, non-IgSF +0.301. Three folds (FN3, IgI, IgV) had zero aHCPs at pDockQ2 $\geq$ 0.3, causing quasi-complete separation; default L2 regularization was used to obtain stable estimates. In-sample AUCs are descriptive estimates of explanatory power; the leave-one-protein-out out-of-sample AUC is reported in Table S15 (Model 1 LOPO AUC = 0.798, Model 2 LOPO AUC = 0.859, optimism = 0.058).

Table S9. Protein-level cluster bootstrap confidence intervals (DEPRECATED — retained for transparency only)

`table_S9_cluster_bootstrap.csv`. **This table uses a deprecated, less conservative resampling method and is retained solely for methodological transparency. The canonical bootstrap results are in Table S13.** Bootstrap 95% confidence intervals for global AUCs computed by resampling unique protein IDs (342 proteins; n = 5,000 iterations) and including all pairs where *either* protein matched (approximately 86% of pairs per iteration—nearly equivalent to pair-level resampling). Under this less-stringent method, the cross-model comparison yielded ΔAUC = 0.044, $p = 0.022$. The canonical induced-subgraph method adopted in this revision (Table S13), which includes only pairs where *both* proteins are in the resample, yields $p = 0.45$ for the same comparison. The divergence in p-values illustrates how permissive resampling can inflate apparent significance in dyadic data.

Table S10. Boltz-2 sensitivity to MSA depth (834 pairs)

`table_S10_boltz2_msa512_vs_1024.csv`. Side-by-side comparison of Boltz-2 confidence metrics at MSA depths 512 (primary analysis) and 1024 (default) for all 834 pairs. Doubling the MSA depth did not improve overall discrimination: pDockQ2 AUC decreased from 0.695 to 0.657; ipSAE changed from 0.660 to 0.654; ipTM increased slightly from 0.645 to 0.669.

Spearman correlations between settings were 0.50–0.68 across metrics. The length-stratified crossover persisted: pDockQ2 at MSA = 1024 achieved AUC = 0.905 for ECDs under 400 aa but only 0.592 above 800 aa.

Table S11. AlphaFold2-Multimer sensitivity to model count (200 pairs)

`table_S11_af2m_1model_vs_5models.csv`. Comparison of AlphaFold2-Multimer with one model (primary analysis) versus five models (default ensemble) on a stratified 200-pair subset spanning all length bins and fold families. Composite AUC increased from 0.663 (1-model) to 0.678 (best-of-5; ΔAUC = +0.015). ipTM AUC increased from 0.643 to 0.666 (ΔAUC = +0.023). On this subset, Boltz-2 pDockQ2 AUC = 0.716 still outperformed AlphaFold2-Multimer best-of-5 composite (0.678). The length-stratified crossover near 700 aa was preserved.

Table S12. SPOC trained-classifier baseline (200-pair subset)

`table_S12_spoc_186pairs.csv`. SPOC [1] scores on the 200-pair AlphaFold2-Multimer 5-model subset, computed using the public command-line tool from the Walter Lab GitHub repository `walterlab-HMS/SPOC`. SPOC accepts standard AlphaFold-Multimer outputs (PDB plus PAE JSON) and looks up its omics features (DepMap, BioGRID, BioORCS CRISPR, T5 protein language model embeddings, DeepLoc subcellular localization, mRNA co-expression) internally from canonical UniProt identifiers. 186 of 200 pairs were scored; the 14 unscored pairs are all from the A4C MHC class I arm (HLA-B/C/E/F/G, CD1B/CD1C, AZGP1, FCGRT, MR1, MICA, MICB), where SPOC's BLAST-based protein-to-UniProt mapping cannot uniquely disambiguate highly similar paralogous sequences. On the 186 scored pairs:

Metric	AUC-ROC (186 pairs)	N
Boltz-2 pDockQ2	0.748	186
SPOC (trained classifier)	0.742	186
AlphaFold2-Multimer composite (5-model best)	0.703	186
AlphaFold2-Multimer ipTM (5-model best)	0.691	186
AlphaFold2-Multimer ipTM (1-model)	0.671	186
Boltz-2 ipSAE	0.666	186
AlphaFold2-Multimer composite (1-model)	0.666	186

SPOC's reported AUC on its training distribution of approximately 300 nuclear genome maintenance proteins is 0.96 [1]; the drop to 0.74 on cell-surface receptors is the expected out-of-distribution penalty (cell-surface receptors are sparsely characterized in DepMap, BioORCS, and similar functional genomics databases relative to nuclear proteins). Length-stratified results: 0–400 aa SPOC = 0.94, 400–800 aa SPOC = 0.75, 800–1,200 aa SPOC

= 0.63, 1,200–3,000 aa SPOC = 0.81. Sanity check on biology: SPOC scores for the well-characterized checkpoint pairs CTLA4 × CD86 = 0.97 and PVR × TIGIT = 0.89.

Table S13. Protein-clustered induced-subgraph bootstrap, full results

`table_S13_cluster_bootstrap_full.json` and `table_S13_reliability_diagrams.json`. Two-stage induced-subgraph cluster bootstrap (Snijders–Borgatti style): each iteration resamples 342 unique protein identifiers with replacement, takes the unique set, and computes the statistic on the subset of pairs where *both* proteins are in the resample. 5,000 iterations, seed 42. Reports global AUCs (Boltz-2 *p*DockQ2, *ip*SAE, *ip*TM, pDockQ v1, LIS; AlphaFold2-Multimer *ip*TM, pTM, composite), pairwise ΔAUC contrasts (cross-metric and cross-model), length-binned AUCs (0–400, 400–600, 600–800, 800–1200, 1200–3000 aa), fold-stratified FPRs at fixed thresholds, and the seed-CV TP-vs-aHCP comparison. Also reports reliability-diagram bins (equal-frequency deciles), expected calibration error (ECE) for each metric, and Platt recalibration parameters fitted on Boltz-2 pDockQ2 outputs (sigmoid: $a = 6.03$, $b = -2.16$; ECE 0.099 → 0.041, 58% reduction). A previous version of the cluster bootstrap (used in earlier rounds of this manuscript) used a less stringent variant that resampled protein names, deduplicated to a set, and included pairs where *either* protein matched the set; this variant includes approximately 86% of pairs per iteration and yields CIs nearly identical to pair-level resampling. We therefore replaced it with the induced-subgraph variant in this revision; the JSON contains the methodology change as a metadata note for transparency.

Table S14. Fixed-FPR cutoff table

`table_S14_fixed_fpr_cutoffs.csv`. For each metric, the score threshold at which FPR is closest to (but ≤) the target value, plus the achieved FPR, true-positive rate (recall), and precision at that threshold. Computed at FPR = 0.05 and FPR = 0.10. Reported separately for the 834-pair Boltz-2 set and the 833-pair shared Boltz-2/AlphaFold2-Multimer set. At FPR = 5% the AlphaFold2-Multimer composite metric had the highest TPR (0.331) and precision (0.55), narrowly ahead of Boltz-2 pDockQ2 (TPR 0.281, precision 0.51).

Table S15. Leave-one-protein-out cross-validation

`table_S15_lopo_logistic_results.json`. For the false-positive-prediction logistic regression, leave-one-protein-out cross-validation was performed by holding out all pairs containing each of the 335 proteins that appear in the 713 negative pairs in turn, fitting the model on the remaining pairs, and aggregating the held-out predictions. Both Model 1 (length only) and Model 2 (length + fold dummies) were fit using sklearn `LogisticRegression`(C=1.0). Result: Model 1 in-sample AUC = 0.861, LOPO out-of-sample AUC = 0.798 (optimism =

0.063); Model 2 in-sample AUC = 0.917, LOPO out-of-sample AUC = 0.859 (optimism = 0.058). 713/713 negative pairs predicted; 335/335 proteins iterated.

Table S16. Precision @ k and PPV at simulated low prevalence

`table_S16_precision_at_k_low_prevalence.csv`. Top-k precision values are reported at $k \in \{10, 25, 50, 100, 200\}$ for each metric on the 833-pair set. PPV under simulated lower prevalences (1%, 5%, 10%, 14.5%) was computed via Bayes adjustment from the achieved sensitivity and specificity at FPR = 5%; at 1% prevalence the PPV is approximately 6% for both Boltz-2 pDockQ2 and AlphaFold2-Multimer composite.

Table S17. Seed CV effect sizes

`table_S17_seed_cv_effect_size.json`. For each Boltz-2 confidence metric (pDockQ2, ipSAE, ipTM), the within-pair coefficient of variation (CV) across the five diffusion seeds in Table S2 was compared between true positives and assay-negative high-confidence predictions (aHCPs). Reports per-pair CV summary statistics, the existing one-tailed Mann–Whitney U statistic and p-value, plus two effect-size measures: Cliff's delta and rank-biserial correlation. The directional comparison was a pre-specified one-sided test motivated by the *a priori* hypothesis that genuine interfaces have a single low-energy basin while spurious high-confidence predictions reflect alternative configurations sampled across seeds. Effect sizes (medium–large by Cliff's delta convention): pDockQ2 $\delta = +0.468$ ($p = 8.6 \times 10^{-5}$), ipSAE $\delta = +0.471$ ($p = 2.1 \times 10^{-4}$), ipTM $\delta = +0.479$ ($p = 6.1 \times 10^{-5}$).

Table S18. Length-binned confusion matrix at fixed FPR = 5%

`table_S18_length_binned_confusion_matrix.csv`. For each of six confidence metrics (Boltz-2 *p*DockQ2, *ip*SAE, *ip*TM, LIS; AlphaFold2-Multimer *ip*TM, composite), a single global threshold was selected as the score at which the across-all-pairs FPR is closest to (but $\leq$) 5%. Each pair was then assigned to one of four ECD-length bins (<400, 400–600, 600–800, >800 aa) and the per-bin confusion matrix (TP, FP, TPR, FPR, precision) was tabulated using that single global cutoff. Result diagnoses where the 4-bin loss of discrimination originates: for Boltz-2 *p*DockQ2 above 800 aa, the TPR collapses to 2.6% (1/39) while the FPR stays at 0.3% (1/301), confirming that the failure mode in long ECDs is loss of recall, not gain of false positives. For AlphaFold2-Multimer composite above 800 aa, the TPR is 12.8% (5/39) at FPR = 3.0% (9/300), a ∼5-fold higher recall than Boltz-2 in the same bin and the empirical basis for the cross-model crossover near 700 aa reported in Results. Columns: `metric`, `dataset`, `length_bin`, `global_cutoff_at_fpr5`, `n_pairs`, `n_pos`, `n_neg`, `n_tp`, `n_fp`, `tpr_recall`, `fpr`, `precision`.

Table S19. Out-of-sample (split-sample) Platt recalibration of Boltz-2 *p*DockQ2

`table_S19_reliability_diagrams_v2_split_sample.json`. The Round 3 manuscript reported that a one-parameter Platt sigmoid reduced ECE on Boltz-2 *p*DockQ2 from 0.099 to 0.041 (a 58% reduction). That figure was an in-sample optimistic estimate: the same 834 pairs were used both to fit the Platt parameters and to evaluate ECE. This table reports the corresponding out-of-sample evaluation: a 5-fold pair-level KFold split (random state 42) in which Platt parameters are fit on each train fold ($n \approx 667$) and ECE is evaluated on the disjoint test fold ($n \approx 167$). All 5 folds had at least 12 positive test pairs and produced valid ECE estimates. Pair-level (rather than protein-clustered) folds were used because protein-clustered 5-fold gives ≤ 4 positive pairs per test fold, which is insufficient for stable ECE estimation; the pair-level estimate remains meaningful for the calibration question. Two complementary OOS estimates are reported: (i) an aggregated estimate that pools all 5 fold predictions and computes ECE on the joint 834 pairs, yielding ECE $= 0.039$ (essentially unchanged from the in-sample 0.041; optimism ≈ 0.002), and (ii) a more conservative per-fold ECE mean of 0.080 ± 0.013, which uses only the within-fold bin sizes for each ECE computation. Under the conservative per-fold framing, Platt recalibration reduces ECE from 0.099 to 0.080, a 20% reduction (vs. the 58% in-sample figure). Both OOS estimates support the qualitative claim that Boltz-2 *p*DockQ2 is poorly calibrated as a literal probability and that recalibration is empirically helpful, but the magnitude of improvement under the most conservative split-sample framing is modest.

Table S20. aHCP fold-effect sensitivity to threshold and metric choice

`table_S20_aHCP_fold_threshold_sensitivity.json`. The Round 3 main-text aHCP fold-effect logistic regression (Table S8 Model 2) defined aHCPs as Boltz-2 *p*DockQ2 ≥ 0.3 within the assay-negative arms. To address the reviewer concern that the fold-effect signs (IgC $>$ FN3, IgV $<$ FN3) may be threshold-bound, we re-fit the same regression under four alternative aHCP definitions: {Boltz-2 *p*DockQ2 ≥ 0.3, Boltz-2 *ip*SAE ≥ 0.5, AlphaFold2-Multimer composite ≥ 0.5, Boltz-2 *p*DockQ2 at the global FPR $= 5\%$ cutoff}. Sign stability of the fold-effect coefficients across the four rules: 3 of 4 rules confirm both IgC $>$ FN3 and IgV $<$ FN3 (Boltz-2 *p*DockQ2 ≥ 0.3, Boltz-2 *ip*SAE ≥ 0.5, Boltz-2 *p*DockQ2 at FPR $= 5\%$); the AlphaFold2-Multimer composite ≥ 0.5 rule preserves the IgV $<$ FN3 sign but reverses the IgC sign (IgC coefficient becomes negative, -0.20). The discordant rule has the lowest in-sample AUC (0.616 vs. 0.840–0.917 for the Boltz-2 rules), suggesting the AlphaFold2-Multimer composite is the weakest stratifier of fold-level aHCP risk in this dataset. Reports per-rule `n_aHCP`, `aHCP_rate`, in-sample AUC, full coefficient vector (`sqrt_len`, `fold_IgC`, `fold_IgI`, `fold_IgV`, `fold_non_IgSF`; reference fold FN3), and sign-stability flags.

Table S21. Protein-clustered induced-subgraph bootstrap PR-AUC

`table_S21_pr_auc_clustered_bootstrap.json`. PR-AUC (average precision) point estimates and 95% percentile CIs for all confidence metrics, computed by the same protein-clustered induced-subgraph bootstrap used for AUC-ROC inference in Table S13 (5,000 iterations, seed 42, 342 unique proteins resampled with replacement, statistic computed on the subset of pairs where both proteins are in the resample). PR-AUC is reported separately for the 834-pair Boltz-2 set (baseline PR-AUC = 0.145, equal to the positive prevalence) and the 833-pair shared Boltz-2/AlphaFold2-Multimer set (baseline 0.145). On the 833-pair shared set, the cross-model PR-AUC ordering matches the AUC-ROC ordering and confirms that under cluster-bootstrap CIs no cross-model PR-AUC contrast reaches significance: Boltz-2 *p*DockQ2 PR-AUC = 0.424 [0.291, 0.557], AlphaFold2-Multimer composite PR-AUC = 0.410 [0.272, 0.549], ΔPR-AUC = 0.014 with overlapping CIs. This is consistent with the C9 reviewer concern that cross-model significance under realistic low-prevalence framing must be evaluated in the precision/recall plane in addition to the ROC plane.

Table S22. AlphaFold2-Multimer PAE-derived interface metrics on the 200-pair stratified subset

`table_S22_af2m_pae_derived_metrics.csv`. To enable a like-for-like cross-model comparison of all PAE-derived interface metrics, the Dunbrack [2] `ipsae.py` v4 implementation was applied to the AlphaFold2-Multimer 5-model best-of-rank PDB plus rank-001 confidence JSON (which contains both the PAE matrix *and* the per-residue *p*LDDT array required for pDockQ/pDockQ2) for the same 200-pair stratified subset used in Table S11 and Table S12. Columns: `pair_id`, `ipsae`, `ipsae_d0chn`, `iptm_af_from_ipsae`, `iptm_d0chn`, `pdockq`, `pdockq2`, `lis`, `n0res`. AUCs of the new AlphaFold2-Multimer PAE-derived metrics on the 200-pair subset: *ip*SAE = 0.681 (modestly above the AlphaFold2-Multimer *ip*TM value of 0.643 on the same 200-pair subset), and *p*DockQ2 = 0.659. These complement Table S11 by demonstrating that the cross-model AUC ordering is robust to the choice of confidence metric: applying matched *p*DockQ2/*ip*SAE definitions to AlphaFold2-Multimer outputs does not narrow the gap to Boltz-2 *p*DockQ2 on the same 200-pair subset.

Implementation note (Round 4 correction). An earlier draft of this analysis fed `ipsae.py` the `*_predicted_aligned_error_v1.json` ColabFold output, which contains only the PAE matrix and not the per-residue *p*LDDT array. `ipsae.py` silently fell back to a zero-pLDDT fallback for the pDockQ branch, producing a near-constant pDockQ2 ≈ 0.0073 for 199 of 200 pairs (the value of the pDockQ2 sigmoid at intercept). The fix is to feed `ipsae.py` the `*_scores_rank_001_*.json` ColabFold output instead, which contains both the PAE matrix and the *p*LDDT array. The bug was caught by an output-quality smoke check (assert that the metric has at least 50 unique values across the 200-pair subset); after the fix the pDockQ2 column has 145 unique values in $[0, 0.901]$, with the well-characterized true positive CTLA4 $\times$ CD86 at pDockQ2 = 0.844. We note this not for forensic interest but

because it illustrates a general failure mode for downstream PAE analysis pipelines: silent fallback to a constant when a required input field is absent. Future smoke tests in this repository assert minimum-unique-value thresholds on continuous metrics rather than only checking output existence.

Table S23. actif*p*TM rerun on the full 834 pairs

`table_S23_actifptm_834pairs.csv` and `table_S23_actifptm_analysis.json`. The Round 3 manuscript declined an actif*p*TM [3] comparison on the grounds that actif*p*TM cannot be computed retrospectively from standard ColabFold output (it requires the AlphaFold2-Multimer internal contact-probability head, preserved only when ColabFold is invoked with the `--calc-extra-ptm` flag at inference time). Round 4 addresses this directly by re-running ColabFold v1.6.1 with `--calc-extra-ptm` on *all 834 pairs*, distributed across V100 × 4 (668 pairs reusing Round 1 cached MMseqs2 MSAs) plus a local RTX 5090 (91 pairs with fresh MSA from the cloud server) and a DGX Spark GB10 (75 pairs with fresh MSA), for a total wall clock of approximately 14 hours. This collection includes the previously failed pair PTPRF × LRIG3 (apECIA pair A4I_008, total ECD ≈ 2,017 aa, original Round 1 ColabFold MSA timeout): the V100 cached MSA from a later successful retrieval was used, allowing actif*p*TM = 0.235 to be reported. We thus have actif*p*TM for 834/834 pairs, eliminating the 833 vs. 834 pair-count discrepancy that previously affected cross-model analyses.

Result. On the full 834-pair benchmark:

Metric	AUC-ROC	PR-AUC
Boltz-2 *p*DockQ2	0.695	0.419
AlphaFold2-Multimer composite	0.651	0.404
AlphaFold2-Multimer actif*p*TM	**0.6425**	**0.389**
AlphaFold2-Multimer *ip*TM	0.6399	0.367
AlphaFold2-Multimer pTM	0.6192	0.357

Headline finding. actif*p*TM provides only marginal improvement over standard *ip*TM on this benchmark (+0.003 AUC, +0.022 PR-AUC). The dominant cross-model pattern — Boltz-2 *p*DockQ2 outperforming all AlphaFold2-Multimer-side metrics including actif*p*TM by ∼0.05 AUC — is preserved.

Length stratification. The length-binned actif*p*TM AUC closely matches the standard AlphaFold2-Multimer composite pattern reported in main text Table 3:

Length bin	n	n pos	actif*p*TM AUC	AlphaFold2-Multimer composite AUC (Table 3)
< 400 aa	125	19	0.851	0.873
400–600 aa	259	36	0.585	0.559
600–800 aa	110	27	0.591	0.614
> 800 aa	340	39	0.657	0.691

The patterns are essentially identical (within ±0.03 in every bin). The Round 3 manuscript's main-text crossover finding (Boltz-2 superior in 400–800 aa, AlphaFold2-Multimer superior >800 aa) is fully preserved when actif*p*TM is substituted for AlphaFold2-Multimer composite as the AlphaFold2-Multimer-side metric.

Interpretation. actif*p*TM was hypothesised to provide a more interface-centric scoring than standard *ip*TM by incorporating the AlphaFold2-Multimer contact-probability head [3]. On the apECIA cell-surface receptor benchmark, this hypothesis is not borne out at the global level: the contact-probability adjustment recovers only +0.003 AUC over plain *ip*TM. We interpret this as evidence that the limit on AlphaFold2-Multimer-side discrimination on this benchmark is at the model-class level (the predicted complex itself, learned from PDB training data on a fold class densely represented by antibodies and TCRs), not at the post-hoc scoring-metric level. This is consistent with our manuscript-wide framing that no single confidence metric repair can recover what is fundamentally a learned-prior bias: a more sophisticated scoring metric does not rescue a model whose underlying interface predictions are themselves miscalibrated for cell-surface receptor folds. The Round 3 "actif*p*TM not retrospectively computable" Limitations defense is removed.

Top 10 actif*p*TM predictions. 6 of the top 10 are biologically canonical immune-receptor interactions from the apECIA positive set (CEACAM1 × CEACAM8, CRTAM × CADM1, CTLA4 × CD86, CHL1 × L1CAM, ICOSLG × ICOS, CD28 × CD86, PVR × CD226), and the 4 negatives at the top are well-characterised structural homolog cases discussed in the main text Case Studies (PTP family receptor homo-pairs PTPRT × PTPRU, PTPRM × PTPRU, plus IL6ST × IL11RA, a shared β-receptor-subunit case analogous to the IL6R × IL6ST gp130-hexamer leak discussed in the main text). This structure of the top of the actif*p*TM ranking matches the structure of the top of the Boltz-2 *p*DockQ2 ranking, further indicating that actif*p*TM does not preferentially correct the structural-homolog-driven aHCPs.

Table S24. Positive-only recall control on 29 PDB-literature ligand–receptor pairs

`table_S24_external_validation_results.csv` and `table_S24_external_validation_length_stratifie`
To provide a held-out positive control set independent of the apECIA benchmark, we curated 32 well-characterized human ligand–receptor pairs from the PDB literature (canonical immune-checkpoint, TNF superfamily, Eph–ephrin, cytokine receptor, and growth-factor receptor complexes; see Methods). Sequence retrieval succeeded for 30 pairs, of which 29 completed Boltz-2 prediction on the local RTX 5090 (one pair, MET × HGF, total ECD length ≈1,600 aa, was excluded due to GPU OOM at 32 GB VRAM with default sampling settings; this is consistent with the >800 aa failure mode reported in the main text). Boltz-2 was run with the same configuration as the 834-pair benchmark (`num_subsampled_msa = 512`, three recycles, 200 sampling steps), and `ipsae.py` was applied to each PDB+PAE output to produce *p*DockQ2, *ip*SAE, *ip*TM, pDockQ v1, and LIS scores per pair. Each pair was then

classified using the global FPR = 5% cutoff from Table S14 (834-pair Boltz-2 set), and recall on the 29-pair external set was computed.

Result. On these 29 PDB-literature true positives, Boltz-2 achieves substantially higher recall than on the apECIA positive set at the same global cutoffs:

Metric	FPR=5% cutoff (834-set)	External recall (n=29)	apECIA TPR	Δ
Boltz-2 *p*DockQ2	0.107	**75.9%** (22/29)	28.1%	+47.8 pp
Boltz-2 LIS	0.375	72.4% (21/29)	26.4%	+46.0 pp
Boltz-2 *ip*TM	0.756	65.5% (19/29)	18.2%	+47.3 pp
Boltz-2 *ip*SAE	0.608	62.1% (18/29)	22.3%	+39.8 pp
Boltz-2 pDockQ v1	0.598	3.4% (1/29)	9.1%	−5.7 pp

Length stratification preserves the main-text pattern. Length-binned recall on the external set follows the same pattern reported in the main text for the apECIA benchmark: *p*DockQ2 recall is 86% (18/21) for ECDs < 400 aa, drops to 43% (3/7) for 400–600 aa, hits the only 600–800 aa pair (KIT × KITLG, 1/1), and the only > 800 aa pair (MET × HGF) was excluded due to OOM. Of the 6 external pairs that score 0/4 across all four high-performing metrics (*p*DockQ2, *ip*SAE, *ip*TM, LIS at the global FPR = 5% cutoffs), 5 are biologically clear multimer-dependent cases that a 1+1 dimer prediction cannot recover: CD40 × CD40LG (TNF trimer), NOTCH1 × DLL4 (Notch cleavage), FZD8 × WNT3 (Wnt–Frizzled with co-receptor), ITGA2 × COL1A1 (integrin–collagen triple helix), and IL6ST × CNTF (gp130 hexamer). The 6th 0/4 case, TNFRSF14 × TNFSF14 (HVEM/LIGHT), is a TNF-superfamily pair whose crystal structure (PDB 4EN0) shows 1:1 binding rather than obligate trimer dependency and is therefore a borderline failure (its *p*DockQ2 of 0.103 misses the global FPR = 5% cutoff of 0.107 by only 0.004); we report it transparently rather than fold it into the multimer-narrative bucket. Together these 6 pairs mirror the multimer/trimer interpretation of the apECIA-side aHCPs in the main text but with one honest borderline exception.

Cross-model control. To rule out Boltz-2-specific training memorization, we also ran AlphaFold2-Multimer v3 (5-model, 3 recycles, actif*p*TM) on the same 30 pairs (29 completed; MET × HGF also completed unlike the Boltz-2 OOM). At *ip*TM ≥ 0.6, AlphaFold2-Multimer recall is 93.1% (27/29), even higher than Boltz-2 *p*DockQ2 at 75.9% (22/29). The fact that two models with substantially different training pipelines both achieve high recall argues that these pairs represent genuinely easy canonical complexes rather than training-data memorization artifacts.

Interpretation. The 47 pp improvement in recall from the apECIA TPR (28.1%) to the external recall (75.9%) for Boltz-2 *p*DockQ2 suggests that the apECIA benchmark contains many positives that are systematically harder than typical PDB-literature complexes

(likely due to a combination of post-translational modification requirements, membrane context dependencies, weak transient interactions selected by the apECIA assay's 2-fold-over-background cutoff, and the absence of obligate co-factors). PDB literature pairs are by construction the easiest cases (proteins that crystallized cleanly and have well-characterized binding modes). This is consistent with the manuscript's framing that raw confidence metrics alone are insufficient for proteome-scale screening at realistic prevalence, while remaining useful for confirming canonical immune-receptor geometries.

Table S25. Unconditional multi-seed Boltz-2 sample (78-pair random sample × 5 diffusion samples)

`table_S25_g3_seed_cv_pdockq2_scores.tsv` and `table_S25_g3_unconditional_seed_cv.json`. The Round 3 seed-CV TP-vs-aHCP analysis (Tables S2, S6, S17) used a *conditional* 87-pair sample pre-filtered to *ip*TM ≥ 0.3 *or* *p*DockQ2 ≥ 0.1, then re-predicted with five independent diffusion seeds. A reviewer concern (R4-C-multiseed) was that this conditional selection may itself induce the seed-CV asymmetry (TP CV < aHCP CV) by restricting analysis to the high-confidence subspace. To address this directly we drew an *unconditional* stratified random sample of 78 pairs from the full 834-pair set, ran each pair with `boltz predict --diffusion_samples 5` on an RTX 4080, and post-processed each of the $78 \times 5 = 390$ outputs with `ipsae.py` v4 to extract the same *p*DockQ2/*ip*SAE/*ip*TM/LIS metrics used in the conditional sample.

Yield. 62 of 78 pairs completed all 5 diffusion samples. 16 pairs OOM-failed on the 4080's 16 GB VRAM: 12 of 16 OOM-failed pairs are > 800 aa total (consistent with the long-ECD failure mode reported in main-text Results), and 4 of 16 are A4C MHC class I pairs of 562–568 aa (a different mode, likely related to MHC class I assembly geometry). The successful 62 pairs contain 17 true positives and 45 negatives.

Result. Comparing TP seed CV vs. negative seed CV *without* any confidence pre-filter on the unconditional sample, the signal is markedly weaker than in the conditional 87-pair Round 3 sample:

Metric	Unconditional Cliff's δ	Unconditional MW p	Conditional Cliff's δ (R3, Table S17)	Sig
*p*DockQ2	+0.114	0.249	+0.468	sam
*ip*SAE	−0.011	0.529	+0.471	oppo
*ip*TM	+0.030	0.431	+0.479	sam
LIS	+0.055	0.391	—	—

Within-confident-subspace replication. When the same Round 3 high-confidence subspace filter (*ip*TM ≥ 0.6) is applied *within* the unconditional 78-pair sample, the resulting subset (3 confident TPs vs. 8 confident aHCPs) reproduces the Round 3 direction with comparable or larger effect sizes (per-metric Cliff's δ from `ipsae.py` output, computed in the

same way as the unconditional row above): *ip*TM $\delta = +0.917$ (Mann–Whitney $p = 0.012$), LIS $\delta = +0.833$ ($p = 0.024$), *ip*SAE $\delta = +0.583$ ($p = 0.097$), *p*DockQ2 $\delta = +0.500$ ($p = 0.139$). For *p*DockQ2, the effect size in the unconditional confident subspace ($\delta = +0.500$) is essentially the same as the Round 3 conditional estimate ($\delta = +0.468$); for *ip*TM the unconditional confident subspace estimate is larger than Round 3 (+0.917 vs. Round 3 +0.479), although the very small subset size (3 TPs only — most TPs in a random sample of 78 are at the long-ECD failure mode and do not reach *ip*TM ≥ 0.6) limits the precision of all four point estimates. The direction is fully consistent with Round 3 but the magnitude is metric-dependent and the confidence interval is wide.

Honest interpretation. The seed-CV TP-vs-aHCP asymmetry reported in Round 3 (Cliff's $\delta \approx +0.47$ for *p*DockQ2) is real *within the high-confidence subspace*: applying the same *ip*TM ≥ 0.6 filter to an independent unconditional sample reproduces the asymmetry direction (*ip*TM $\delta = +0.917$, LIS $\delta = +0.833$, *p*DockQ2 $\delta = +0.500$ in the small 3 TP vs. 8 aHCP replication subset; the *p*DockQ2 value is essentially the same as Round 3, while *ip*TM/LIS are larger). It is *not* a property of the full benchmark distribution: across all TPs vs. all negatives in a random 62-pair sample, the seed-CV asymmetry collapses to Cliff's $\delta \approx +0.11$ for *p*DockQ2 (one-tailed Mann–Whitney $p = 0.25$) and is statistically indistinguishable from zero for *ip*SAE/*ip*TM/LIS. We accordingly position the seed-CV filter as a *second-stage* disambiguator that helps separate TPs from aHCPs *after* a first-stage confidence threshold has selected the high-scoring subspace, rather than as a property that holds across the unfiltered benchmark. This downgrades the Round 3 framing of the seed-CV analysis from a free-standing TP-stratification result to a conditional finding that depends on the prior selection step. The main text Discussion of the seed-CV result has been updated to reflect this distinction.

Table S26. AlphaFold2-Multimer 1-model PAE-derived interface metrics on the full 834 pairs

`table_S26_af2m_1model_pae_metrics_834.csv`. Round 4 Table S22 reported `ipsae.py` v4 on the 200-pair stratified subset using 5-model AlphaFold2-Multimer outputs; the full 834-pair coverage for 1-model AlphaFold2-Multimer PAE-derived metrics was listed as future work (main text Limitations). To close this item, we ran `ipsae.py` v4 on the Round 4 actif*p*TM 834-pair AlphaFold2-Multimer 1-model `*_scores_rank_001_*.json` outputs (source distribution: 668 V100 cached, 91 RTX 5090, 75 DGX Spark GB10, total 834/834). The output-quality smoke check passes: *ip*SAE has 236 unique values in $[0.000, 0.851]$, pDockQ v1 has 732 unique values in $[0.000, 0.726]$, *p*DockQ2 (pDockQ2) has 262 unique values in $[0.000, 0.888]$, and LIS has 295 unique values in $[0.000, 0.673]$ — none of the metrics exhibits the C7 zero-*p*LDDT constant-fallback signature (which would produce ∼730 values at pDockQ2 ≈ 0.0073).

Result. Full 834-pair global AUCs (1-model AlphaFold2-Multimer, `ipsae.py` v4):

Metric	AUC-ROC	PR-AUC
Boltz-2 *p*DockQ2 (reference, Table S1)	0.6952	0.4185
AlphaFold2-Multimer 1-model *ip*SAE	0.6619	0.3334
AlphaFold2-Multimer 1-model *p*DockQ2	0.6523	0.3994
AlphaFold2-Multimer 1-model LIS	0.6583	0.3195
AlphaFold2-Multimer 1-model pDockQ v1	0.6342	0.1896
(reference) AlphaFold2-Multimer *ip*TM (Table S3)	0.6399	0.367
(reference) AlphaFold2-Multimer actif*p*TM (Table S23)	0.6425	0.389

Length stratification. The length-crossover pattern reported in the main text is preserved on the full 834 pairs for 1-model AlphaFold2-Multimer PAE-derived metrics:

Length bin	n	n pos	*ip*SAE AUC	*p*DockQ2 AUC	Boltz-2 *p*DockQ2 AUC	Δ (Boltz-2 – *ip*SAE
<400 aa	125	19	0.8446	0.8684	0.8781	+0.034
400–600 aa	259	36	0.6475	0.6758	0.7272	+0.080
600–800 aa	110	27	0.6649	0.6432	0.8170	+0.152
≥ 800 aa	340	39	0.6082	0.5899	0.5514	−0.057

Interpretation. Table S26 extends Table S22 from 200 to 834 pairs for the 1-model AlphaFold2-Multimer track. The key observations are (i) the within-metric ordering *ip*SAE $\geq$ LIS $>$ *p*DockQ2 $>$ pDockQ v1 is preserved from the 200-pair subset, (ii) all four PAE-derived metrics remain below Boltz-2 *p*DockQ2 on the global AUC, which is consistent with Tables S22–S23, and (iii) the length crossover holds: Boltz-2 *p*DockQ2 outperforms all 1-model AlphaFold2-Multimer PAE-derived metrics for <800 aa, then falls below them at ≥ 800 aa ($\Delta -0.057$ for *ip*SAE, consistent with the main text crossover zone). This closes the Round 4 Limitations item "*ip*SAE only computed on a 200-pair subset." Table S29 (below) reports the parallel 5-model best-of-pool + actif*p*TM rerun on the same 834 pairs, which is the stronger model-count match to Boltz-2 and the more direct closure of the full AlphaFold2-Multimer-side sensitivity question.

Table S27. Antibody and T-cell receptor chain exclusion verification

`table_S27_antibody_tcr_exclusion.json`. A Round 3 caveat noted that the IgV/IgC fold labels used throughout this paper refer to CATH 2.60.40.10 structural homology, not to actual antibody or T-cell receptor molecules. To make this caveat defensible against reviewer challenge, we applied regular-expression matching to the 342 unique gene symbols in the 834-pair benchmark, using the following canonical chain patterns:

- **Immunoglobulin heavy chain:** IGHG, IGHA, IGHM, IGHE, IGHD, IGHV;

- **Immunoglobulin light chain:** IGKC, IGKV, IGLC, IGLV;
- **T-cell receptor constant regions:** TRAC, TRBC, TRDC, TRGC;
- **T-cell receptor variable regions:** TRAV, TRBV, TRDV, TRGV.

Result. Of 342 unique gene symbols, 0 match any antibody chain regex and 0 match any T-cell receptor chain regex. The benchmark does contain CD4, CD8A, and CD8B, which are T-cell co-receptors of the IgSF superfamily but are themselves cell-surface receptors (not the T-cell receptor chains TRA/TRB/TRD/TRG), and are correctly classified as IgV-superfamily cell-surface receptor pairs. The fold-class labels used in this paper therefore refer to CATH structural homology only, and no actual antibody or T-cell receptor chain is present in the 834-pair benchmark.

Table S28. apECIA FOB threshold sensitivity analysis

`table_S28_fob_threshold_sensitivity.json`. The primary analysis uses the Wojtowicz et al. (2020) FOB ≥ 2 cutoff for positive calls in both bait–prey orientations. To show that the length crossover and the Boltz-2 global AUC are not an artefact of the specific FOB threshold, we re-stratified the 834-pair set at FOB ≥ 1.5 (looser) and FOB ≥ 3.0 (stricter) and recomputed Boltz-2 *p*DockQ2 global and length-stratified AUCs.

FOB cutoff	n	n pos	Boltz-2 *p*DockQ2 global AUC	Length crossover preserved?
≥ 1.5 (looser)	834	121	0.6952	Yes
≥ 2.0 (primary)	834	121	0.6952	Yes
≥ 3.0 (stricter)	834	91	0.7263	Yes

Length-stratified FOB sensitivity. At FOB ≥ 3.0 (stricter), the length-binned Boltz-2 *p*DockQ2 AUC pattern preserves the crossover with slightly improved < 400 aa and ≥ 800 aa AUCs:

Length bin	FOB ≥ 2.0 (primary)	FOB ≥ 3.0 (stricter)
< 400 aa	0.878	0.933
400–600 aa	0.727	0.742
600–800 aa	0.817	0.808
≥ 800 aa	0.551	0.591

Interpretation. The < 400 aa peak and the ≥ 800 aa trough are robust to both loosening (FOB ≥ 1.5) and tightening (FOB ≥ 3.0) the assay positive-call threshold. The length-crossover zone around 700–1,200 aa is therefore a property of the length axis and not of the specific bait–prey FOB cutoff choice, which directly addresses the reviewer concern that the crossover finding could be an artefact of the primary FOB ≥ 2 cutoff.

Convention note. At FOB ≥ 3.0, 30 of the 121 primary positives no longer meet the stricter cutoff. The main table reports them as "demoted-to-negative" rather than dropping them (preserving the 834-pair total): under this convention n_{neg} rises from 713 to 743 and the global Boltz-2 *p*DockQ2 AUC is 0.7263. Under the alternative "drop-outright" convention ($n = 804$, $n_{\text{neg}} = 713$), the same metric yields AUC $= 0.7291$, a $+0.003$ difference. Both conventions are consistent with the qualitative conclusion that the length crossover persists at the stricter cutoff.

Table S29. AlphaFold2-Multimer 5-model best-of-pool with actif*p*TM on the full 834 pairs

`table_S29_af2m_5model_actifptm_834.json`.

Setup. Round 4 Tables S22 and S23 reported (i) AlphaFold2-Multimer single-model PAE-derived metrics on a 200-pair stratified subset, and (ii) actif*p*TM on the full 834 pairs from a single-model run. Round 5 closes the residual gap by running ColabFold v1.6.1 with `--num-models 5 --num-recycle 3 --calc-extra-ptm` on *all* 834 pairs distributed across a 6-GPU pool: NVIDIA RTX 5090 (32 GB Ada), NVIDIA Tesla V100 SXM2 ($\times 4$, 32 GB each Volta), NVIDIA RTX A6000 (48 GB Ampere, added late in Round 5 for the 1500–1700 aa length range that exceeds 32 GB capacity), and NVIDIA DGX Spark GB10 (128 GB unified memory Blackwell, exclusive provider for ≥ 1700 aa pairs which the 32–48 GB cards cannot handle).

Coverage. 834 of 834 pairs completed 5-model best-of-pool (121 positives, 713 negatives).

Result. On the full 834-pair benchmark (121 positives, 713 negatives):

Metric (5-model best-of-pool)	AUC-ROC	PR-AUC
AlphaFold2-Multimer actif*p*TM (5-model)	0.6512	0.3382
AlphaFold2-Multimer *ip*TM (5-model)	0.6393	0.2926
AlphaFold2-Multimer pTM (5-model)	0.6327	0.3398
(reference) actif*p*TM 1-model (Table S23)	0.6425	0.389
(reference) Boltz-2 *p*DockQ2 (Table S1)	0.6952	0.4185

Length stratification.

Length bin	n	n pos	actif*p*TM 5-model	actif*p*TM 1-model (S23)	Δ	Boltz-2 *p*DockQ2 (S1
< 400 aa	125	19	0.9057	0.851	$+0.055$	0.878
400–600 aa	259	36	0.5366	0.585	-0.048	0.727
600–800 aa	110	27	0.6647	0.591	$+0.074$	0.817
≥ 800 aa	340	39	0.6613	0.657	$+0.004$	0.551

Interpretation. The 5-model best-of-pool actif*p*TM shows a modest +0.009 improvement over the 1-model actif*p*TM at the global level (0.6512 vs. 0.6425), consistent with the +0.015 ΔAUC from the 200-pair 5-model sensitivity analysis in Table S11. The length-stratified pattern is preserved: (i) the < 400 aa bin shows the largest improvement (+0.055), reaching AUC = 0.906, now exceeding Boltz-2 *p*DockQ2 (0.878) in this bin; (ii) Boltz-2 *p*DockQ2 still dominates in the 400–800 aa range; (iii) the crossover above 800 aa is preserved, with AlphaFold2-Multimer 5-model actif*p*TM (0.661) outperforming Boltz-2 *p*DockQ2 (0.551). The dominant cross-model AUC ordering (Boltz-2 *p*DockQ2 0.695 > AlphaFold2-Multimer actif*p*TM 0.651 globally) is preserved under the 5-model setting. This closes the Round 4 Limitations item "a full 833-pair AlphaFold2-Multimer 5-model best-of-pool run with all PAE-derived metrics remains future work."

Table S30. Boltz-2 v2.2.1 full 834-pair rerun + version drift analysis

`table_S30_boltz221_drift_analysis.json`.

Setup. Round 4 Limitations noted that the primary Boltz-2 results (Table S1) used v2.0.3 (released 2025-06-07), with v2.2.0 contact-conditioning fixes and v2.2.1 PoseBusters bounds correction released subsequently. We re-ran Boltz-2 v2.2.1 (Boltz commit `b85e8d6`) on *all* 834 pairs using the V100 × 4 GPU pool (PyTorch 2.5, CUDA 12.4, `boltz_env` conda environment) over ∼16 hours with the same configuration as the v2.0.3 run (`--diffusion_samples 1 --num_subsampled_msa 512 --recycling_steps 3 --sampling_steps 200`).

Result. 834/834 pairs completed (3 retried automatically by the yaml-loop wrapper). Of 834 output directories, 751 produced parseable confidence JSONs for per-pair drift comparison (83 pairs had atypical output structures and are excluded from per-pair analysis but included in the AUC comparison via the output directory count).

Metric	v2.0.3 AUC	v2.2.1 AUC	ΔAUC	Pearson r	Median Δ
*ip*TM	0.6577	0.6510	−0.007	0.742	−0.003
pTM	0.6346	0.6348	+0.000	0.867	−0.004

Per-pair drift analysis. The Pearson correlation between v2.0.3 and v2.2.1 is moderate for *ip*TM ($r = 0.742$, $n = 751$) and high for pTM ($r = 0.867$). Individual-pair *ip*TM values can shift substantially (max $|\Delta| = 0.593$, std = 0.148), while pTM is more stable (max $|\Delta| = 0.368$, std = 0.069). The moderate *ip*TM correlation reflects the stochastic nature of the diffusion sampler: each version uses the same model architecture but different PoseBusters bounds constraints and dropout implementations, which can produce different samples from the same posterior.

Interpretation. Despite moderate per-pair *ip*TM drift, the global AUC is preserved within ±0.007 and the qualitative length-dependent patterns are unchanged. This confirms that the v2.0.3 benchmark results reported in the main text are not artefacts of the specific point release. The larger per-pair variance suggests that multi-seed averaging (as in Table S31) is important for stable individual-pair predictions regardless of version. This closes the Round 4 Limitations item "v2.2.x rerun listed as future work."

Table S31. Pre-registered all-pair multi-seed Boltz-2 v2.0.3 (242 pairs × 5 diffusion samples)

`table_S31_cliffs_delta.json` and `table_S31_locked_pairs.json`.

Setup. Round 4 Discussion item (4) and Tables S2/S6/S17 reported a multi-seed prediction-variance asymmetry between true positives and assay-negative high-confidence predictions (aHCPs) on a conditionally selected subset (87 pairs). Round 5 closes the "pre-registered all-pair multi-seed evaluation" future-work item by:

1. Pre-registering, with random seed 42 fixed in advance of any compute, a 1:1 length-bucket-matched control set: 121 true positives + 121 length-matched aHCP controls = 242 pairs (locked list in `table_S31_locked_pairs.json`).
2. Running Boltz-2 v2.0.3 with `--diffusion_samples 5` on each of the 242 pairs (1,210 inferences total) distributed across 4 RTX-class GPUs (RTX 5090, RTX 4080, RTX 3090, DGX Spark) plus a V100 retry pool for failed pairs.
3. Computing per-pair seed coefficient of variation (CV) for pDockQ2, ipSAE, ipTM, and LIS, and the TP-vs-matched-aHCP Cliff's δ.

Coverage. 110 of 242 pairs completed all 5 diffusion samples across the available GPU pool (54 true positives, 56 matched aHCP controls). The remaining 132 pairs failed with a mix of: (i) VRAM OOM on the 16 GB RTX 4080 and 24 GB RTX 3090 for pairs ≥ 800 aa (22 + 60 pairs), (ii) a Boltz v2.0.3 exception-handling bug (`raise {''exception'': True}` at `boltz2.py:1118`, triggered on torch 2.11 + sm_86 combinations, 60 pairs on the RTX 3090), and (iii) 5090 Task B outputs lost during session recovery. A V100 retry pool recovered 38 of the failed pairs using the Boltz 2.0.3 `boltz_env` environment (matching the original version). These failures are honestly reported here rather than dropped from analysis; the 110-pair coverage (54 TP + 56 aHCP) provides a meaningful sample for the matched-control Cliff's δ test.

Result. Per-pair *ip*TM coefficient of variation (CV) across 5 diffusion seeds:

Group	n	Mean CV	Median CV
True positives	54	0.167	0.126
Matched aHCP	56	0.225	0.222
Cliff's $\delta = +0.279$ (small-to-medium); Mann–Whitney one-sided $p = 0.0059$			

Interpretation. The pre-registered 1:1 length-matched design confirms the direction of the Round 4 conditional seed-CV asymmetry: true positives produce more consistent *ip*TM predictions across diffusion seeds than length-matched aHCP controls (Cliff's $\delta = +0.279$, small-to-medium effect; Mann–Whitney one-sided $p = 0.0059$). The effect size is smaller than the Round 4 conditional 87-pair estimate ($\delta = +0.47$) as expected, because the conditional selection in Round 4 (pre-filtering to *ip*TM ≥ 0.3 or pDockQ2 ≥ 0.1) enriched for high-confidence pairs where the asymmetry is strongest. The pre-registered matched-control result provides a less biased estimate of the true population effect size. This closes the Round 4 Discussion item (4) "pre-registered all-pair multi-seed evaluation needed."

Table S32. Recycling sensitivity: 6 vs. 3 recycles on the 200-pair stratified subset

`table_S32_recycling_sensitivity.json`. To verify that the default `--num-recycle 3` ColabFold setting does not under-converge, we re-ran the 200-pair stratified subset (same as Tables S11/S12) with `--num-recycle 6` on the RTX 5090. 200 of 200 pairs completed on the RTX 5090 (32 GB). Contrary to our initial expectation that pairs >1500 aa would exceed 32 GB VRAM at 6 recycles, all 200 pairs (up to 2,157 aa) completed without OOM failures, confirming that ColabFold's JAX/XLA memory management keeps peak VRAM stable regardless of recycle count.

Result. AUC comparison on the 90 pairs for which both recycle = 3 (from the Round 5 5-model run, Table S29) and recycle = 6 outputs were available in the 5090 output directory (21 positives, 69 negatives):

Metric	Recycle = 3 AUC	Recycle = 6 AUC	ΔAUC	Pearson r
actif*p*TM	0.649	0.625	-0.024	0.822
*ip*TM	0.655	0.620	-0.035	0.843
pTM	0.767	0.719	-0.048	0.905

Interpretation. Increasing from 3 to 6 recycles *decreases* discrimination performance (ΔAUC ≈ -0.02 to -0.05 across all three metrics), while maintaining high per-pair score correlation ($r = 0.82$–0.90). This indicates that 6 recycles does not improve interface prediction on this benchmark; the marginal recycles likely cause over-convergence toward confident but incorrect structural placements for non-interacting pairs, widening the false-positive distribution. The default 3-recycle setting used throughout this manuscript is therefore not only

computationally efficient but also empirically optimal on this benchmark. This addresses the potential reviewer concern that additional recycling might change the conclusions.

Table S33. PDB training data overlap analysis

`pair_leakage_scores.csv` and `pdb_homology_results.json`. To assess whether the benchmark results are driven by training-data memorization, we queried the RCSB PDB REST API for each of the 342 unique proteins in the benchmark, searching for PDB structures containing the same gene product deposited before the Boltz-2 training cutoff (2023-06-01). Of 342 proteins, 216 (63%) had at least one pre-cutoff PDB structure. We then classified each of the 834 pairs by whether *both* proteins had PDB training overlap.

Stratum	n	n pos	n neg	Boltz-2 pDockQ2 AUC
Both proteins in PDB training data	468	69	399	0.725
≥ 1 protein without PDB overlap	366	52	314	0.678
All 834 (reference)	834	121	713	0.695

Interpretation. Training-data overlap provides a modest +0.047 AUC boost (0.725 vs. 0.678), indicating that Boltz-2 performs slightly better on proteins it has "seen" during training. However, the no-overlap AUC (0.678) remains in the same moderate-discrimination range as the full benchmark, and the qualitative conclusions—modest global discrimination, length-dependent performance decline, no statistically dominant single metric—are preserved in both strata. The training-data overlap effect is therefore a secondary modulator, not the primary driver of benchmark performance. This analysis addresses the concern that high-confidence predictions on canonical immune-receptor complexes could reflect memorization rather than generalizable inference.

Table S34. Template-ON vs template-OFF control on 5 case-study pairs

`table_S34_template_control.csv`. ColabFold AlphaFold2-Multimer v3 was run on 5 case-study aHCP pairs with and without the `--templates` flag (one model, three recycles, `--calc-extra-ptm` for actif$p$TM). This control tests whether PDB template retrieval at inference time inflates aHCP confidence.

Pair ID	Genes	actif$p$TM (no tpl)	actif$p$TM (tpl)	Δ actif$p$TM	ipTM Δ
A3N_002	FASLG × TNFSF10	0.944	0.684	−0.260	−0.22
A3N_098	TNFSF12 × TNFRSF18	0.185	0.215	+0.030	+0.01
A3F_060	IL6R × IL6ST	0.968	0.955	−0.013	−0.01
A4C_007	HLA-B × CD1C	0.779	0.751	−0.028	−0.04
A4C_057	HLA-G × MICA	0.715	0.740	+0.025	+0.06
Mean Δ actif$p$TM				−0.049	

Interpretation. Templates do NOT systematically increase aHCP confidence: 3 of 5 pairs show a drop in actif$p$TM with templates, 2 show small increases. The most dramatic case (FASLG × TNFSF10, $\Delta = -0.260$) shows that adding PDB templates actually *reduces* the aHCP score, likely because template retrieval returns correct individual binding partners (FasL homotrimer, TRAIL homotrimer) rather than a cross-family FasL–TRAIL complex. This supports the hypothesis that case-study aHCPs arise from learned geometric priors in the model weights, not from inference-time template contamination.

References

[1] Ernst W. Schmid, Sanjay R. Srivatsan, Willow Coyote-Maestas, Silke R. Engel, Daniel Szöllősi, et al. Predictomes, a classifier-curated database of AlphaFold-modeled protein–protein interactions. *Mol Cell*, 85:1216–1232.e5, 2025. doi: 10.1016/j.molcel.2025.01.034. PMID:40015271.

[2] Jr. Dunbrack, Roland L. Rēs *ip*SAE loquuntur: What's wrong with AlphaFold's *ip*TM score and how to fix it. *bioRxiv*, 2025. doi: 10.1101/2025.02.10.637595. PMID:39990437.

[3] Julia K. Varga, Sergey Ovchinnikov, and Ora Schueler-Furman. actifpTM: a refined confidence metric of AlphaFold2 predictions involving flexible regions. *Bioinformatics*, 41(3):btaf107, 2025. doi: 10.1093/bioinformatics/btaf107. PMID:40080667.