Proteins: Structure, Function, and Bioinformatics, 2026, pp. 1–12

Research Article

Near-cognate decoys hide partner identity in structural benchmarks of intrinsically disordered regions and their partners

Ivan Meić,[1] Buğra Alptekin Sarı,[1] Steve Muller,[1] Fedor Chishchin[1] and Gabriel Richman[1,*]

[1]Omic Inc., Seattle, Washington, USA

[*]Corresponding author. Gabriel Richman, gabe@omic.ai

Abstract

Benchmarks for partner-conditioned prediction of intrinsically disordered region (IDR) structure share a recipe: short chains from PDB complexes, depositor labels, splits on identity or date, and a model credited with partner use when it beats a partner-free ablation. We built one, 5,683 complexes, and audited it. Polyproline II has no three-state code: 17.9% of coil residues in proline-rich regions are PPII against 5.9% elsewhere. 60% of post-2021 depositions share a super-cluster with an earlier entry, so a date split leaks. Across a partner change a region moves 2.3% against a 5.8% replicate-noise floor, dominated by structure-to-coil transitions. The cross-model ablation gap of +0.0035 is 0.74 times the SD of retraining that configuration over 45 runs. A within-model swap avoids that variance but reports its decoys: composition-chosen substitutes are near-cognate, with 2,726 of 4,768 complexes taking one from the true partner's own accession, and cost 0.0024. Homology screening gives 0.0066; matching those decoys partway back toward the composition-selected embedding distance gives 0.0073, negative in all five folds and every label-clean subset, though its interval includes zero on the cleanest 76: dissimilarity explains only part of the rise. A wrong partner costs 0.0076 in 15 of 15 paired runs. Where an ablation gap is offered as evidence of partner use, report that swap with its decoys stated, and a calibration task with a known partner-dependent rule. Our model recovers 84-90% of a coarse form of that rule, the axis these effects are measured on; on a finer form the recovery is seed-dependent.

Keywords benchmarking, intrinsically disordered regions, negative controls, data quality, protein language models, evaluation

Introduction

Intrinsically disordered regions make up a large fraction of eukaryotic proteomes and frequently adopt defined conformations upon binding a partner [1, 2, 3]. The natural computational framing is conditional: given an IDR and a partner, predict the bound conformation. At least 38 predictors of binding disordered regions have been published [4], and the most recent adopt exactly this framing, encoding both sequences with a protein language model and coupling them [5]. Structure-based predictors co-fold the complex instead [6, 7, 8]; disorder predictors such as flDPnn [9] address the prior question of whether a region is disordered at all [17, 18].

Every such method needs a benchmark, and benchmarks here follow a common recipe. Short chains are taken from PDB complexes on the assumption that a short chain in a complex is a disordered region; secondary structure is read from depositor `HELIX_P` and `SHEET` records and mapped to three states; positions without coordinates are labelled coil; independence is established by sequence identity, or by deposition date when an independent test set is wanted; and a model is credited with using the partner when it outperforms an ablated model trained without one. Each step is a shortcut with an obvious justification, and no audit of the whole recipe against independent evidence on the same data appears to exist. What follows is that audit.

We built it to current best practice – chain identity verified from the polymer entity rather than the entry, per-residue observation masks, unambiguous partner assignment, homology-aware splits – then audited each step against curated resources [11, 12, 13, 14]. None of those provides per-residue bound secondary structure with the partner attached, which is why the benchmark had to be built rather than downloaded. For five of the six findings below we can show the step is wrong without showing how far any downstream conclusion moves; only for the temporal split do we measure the cost of correcting it, and that cost is 60% of the candidate holdout. The sixth is a limit of the data rather than a step that can be corrected.

The reasoning is established elsewhere – networks trained on DUD-E learn entirely from ligand properties despite being given protein structure [15], and sequence-based interaction predictors collapse to chance once train-test similarity is removed [16] – but has not reached partner-conditioned prediction of disordered-region structure. Disorder prediction is assessed communally and blind by CAID [17, 18], but CAID supplies no partner, so this task has neither an equivalent assessment nor a community benchmark to inherit.

Two checks carry the argument. A **within-model partner swap** holds one trained model fixed and replaces the partner with a length-

Table 1 Construction flow. The two partitions of the final 5,683 answer different questions and are not alternative accounts of the same split: the upper block is pipeline stage, the lower is provenance of the surviving samples.

Stage	Count
PDB entries returned by the search query	8,030
Two-entity samples after entity, length, resolution and mask filters	5,461
Multi-entity samples added, partner by heavy-atom contact	397
Sum before deduplication	5,858
After deduplication on (IDR sequence, partner sequence, PDB ID)	**5,683**
of which mined in the current pass (`PDB_mined_v4`)	5,445
of which carried from the predecessor mining pass (`PDB_mined`)	156
of which DisProt-derived and label-verified (`DisProt`)	82

and composition-matched alternative; unlike an ablation it holds model, training run and parameter count identical, isolating partner identity from input configuration. Substituting the partner is not new – AlphaFold-Multimer's scores separate cognate IDR-receptor pairs from decoys built by swapping in non-binding fragments [19] – but that control belongs to the co-folding literature and is absent from sequence-based benchmarks. A **calibration task** with a known partner-dependent rule establishes whether an architecture can use partner information at all, which is what makes a small effect interpretable. A model result here is evidence about that model on this benchmark, not about proteins; that AlphaFold2 recovers conditionally folded states from monomer sequence alone [20] is a stronger, prior result about the biology.

Results

A benchmark with per-residue observation masks and unambiguous partner assignment

Built as Methods describes: two-entity PDB complexes with a chain of 10–100 residues at $\leq$2.5 Å, each IDR's accession resolved from its own polymer entity rather than the entry, depositor HELIX/SHEET records for secondary structure, and a per-residue observation mask. The final set contains 5,683 complexes with 193,028 observed residues, composed of 42.8% helix, 17.4% strand and 39.8% coil, spanning 1,921 distinct IDR accessions and 1,943 distinct partner accessions (Table 2). Homology-aware clustering at 30% identity with containment linking yields 1,817 super-clusters. Audited independently of the clustering that produced them – by Smith-Waterman nearest-neighbour identity against a composition-matched shuffled null, exact substring containment, and shared UniProt accession – every fold passes with zero containment and zero shared accessions between test fold and training set.

Label provenance decided what could be kept. In the dataset this one replaced, stored labels matched the deposited annotation for the same chain at a mean of 1.0 across 335 PDB-mined samples but 0.597 across 203 DisProt-derived ones, and not because of NMR: DisProt X-ray samples disagreed at 68%. The benchmark is accordingly PDB-mined: 5,445 samples from the current pass, 156 from its predecessor, and 82 DisProt samples retained where their labels passed the same verification.

Four construction steps, audited

Each of the first four steps substitutes an available proxy for the property wanted, and each is wrong in a direction that flatters the benchmark. Table 3 gives the rate for each on this benchmark and whether the three curated resources we could check share the defect; the paragraphs below carry the part a rate cannot.

Finding 1: the selection rule

Tested against MobiDB, which aggregates DisProt, IDEAL, DIBS, MFIB and ELM and separates curated annotation from prediction. Absence of curation is not evidence of order, so 22.4% is a lower bound, but one measured over most of the benchmark. The audit method matters as much as the answer: exact substring matching against the canonical sequence checks only 50.0% of samples and reports 19.7% supported, resolving preferentially the well-studied proteins whose construct matches the canonical sequence – which are the curated ones. The folded-domain rate is over the 4,077 entities whose chain resolved, 749 of them; counting the 1,606 unresolved and 425 unmapped entities as clean gives 13.2% over all 5,683, a floor. Coverage is a residue overlap, not a ratio of lengths: chains carry several domains, one mapped elsewhere sharing no residue. Ubiquitin (Pfam PF00240) is the largest single family at 80 entities and the most partner-promiscuous entity here, its two accessions taking 54 distinct partners across 95 samples. Partner annotation is absent besides: `partner_family` is null for 5,601 of 5,683 samples.

Finding 2: the label alphabet

Polyproline II is not a marginal state: it is the conformation an IDR adopts when binding SH3, WW, EVH1, GYF and profilin domains. It is almost never deposited as a `HELIX_P` record and three-state DSSP has no code for it, so it is labelled coil by default. Recomputing from coordinates and assigning PPII from backbone dihedrals (Methods) gives a **3.1-fold** enrichment inside PxxP regions (95% CI 1.74-5.50, bootstrapped over 95 and 90 UniProt accessions, since residues are not independent), over 534 samples and 2,367 observed residues; the comparator is a background rate, not a matched control. 17.9% is a minority of coil, so the claim is that a defined modality is unrepresented, not that the label is mostly wrong.

Finding 3: coil, and where labels fray

Coil mixes measured irregularity with unmeasured absence, inseparably after the fact; on the dataset this benchmark replaced the figure was 45%. CAID's Disorder-NOX benchmark already excludes X-ray missing residues rather than scoring them as negatives [18], but bound-structure benchmarks have not inherited the fix, which is why the observation mask here is per residue and unresolved positions are

Table 2 Benchmark composition. Every sample carries a verified chain identity, a per-residue mask of resolved positions, secondary structure from the deposited annotation, and a partner that is unambiguous by construction (two-entity complexes, the majority) or contact-verified by heavy-atom distance (multi-entity complexes).

Property	Value
Complexes	5,683
Observed residues	193,028
Composition (H / E / C)	42.8% / 17.4% / 39.8%
Distinct IDR accessions	1,921
Distinct partner accessions	1,943
Homology super-clusters (30% identity)	1,817
Largest super-cluster	268 samples
Samples with a partner sequence	5,683 (100%)
Cross-validation folds	5, composition matched to within 0.0023 pp

Table 3 The four construction steps, their measured failure rates here, and whether DisProt, MFIB and DIBS share each defect. A vocabulary either contains a code or it does not, so the external columns are categorical for the first three rows and a rate for the fourth; the external rates are not directly comparable to ours because they are measured on each resource's own date field.

Step	What it assumes	Measured here	DisProt / MFIB / DIBS
Selection: a short chain in a complex is an IDR	disorder	**22.4%** have curated support, 26.2% of the 85.3% checkable; **41.2%** have a record with no curated disorder or MoRF at all, 21.8% fall below the half-of-region threshold, 14.6% have no SIFTS range; **18.4%** are covered over at least 80% of their residues by a Pfam or CATH family	not applicable: these are the curated resources
Label alphabet: H, E and C suffice	no fourth state	PxxP residues are **90.2%** C against 3.8% H and 6.0% E; **17.9%** of coil in PxxP regions is PPII against 5.9% elsewhere	no PPII term in any of the three
Coil: an unresolved position is an irregular one	coordinates	**27.4%** of C-labelled residues have no coordinates, against 0 for H and E; deposited-versus-recomputed agreement 0.8497	no per-residue coordinate status in any of the three
Split: deposition date gives independence	novelty	**766 (60%)** of 1,275 post-2021 samples share a super-cluster with an earlier entry	a date split leaks 66% / 96% / 27%

excluded from training and from every metric. The disagreement with an independent assignment – 272 structures, 11,871 residues, 60% of structures fall below 0.90 and 22% below 0.80 – is positional rather than substantive: terminal residues agree **38.3%** of the time against **95.5%** for residues four or more positions in, over 681 elements of five residues or more from the 307 of 400 sampled structures containing one. Depositors extend records roughly one residue past where the hydrogen-bond criterion stops. Two readings fit, terminal fraying or a 3-10 turn the recomputation suppresses where it abuts an alpha-helix, and the consequence here is specific either way: bound IDR elements are short, so 38% of residues even in elements of five or more sit within one position of an end.

Finding 4: the split

Recent depositions here mostly re-solve known families, so testing on them measures memorisation. Removing them leaves 509 samples in 380 super-clusters, separated in both time and homology from the 3,746 training samples, on which the model reaches 0.9142 binary AUC-ROC and 0.8204 Q3 – slightly above its cross-validated figures, so performance does not decay out of time, at the cost of 60% of the candidate holdout. This is the one step whose correction has a measured price.

A model trained on this benchmark

The remaining findings need a model: a conventional ESM-2 encoder with cross-attention to the partner. Under five-fold cross-validation clustered on IDR identity it reaches 0.8986 binary AUC-ROC (95% CI 0.8819-0.9118) and 0.8128 Q3, every sample held out once, intervals bootstrapped over super-clusters. A logistic regression on the same frozen embeddings reaches 0.8778, so the architecture adds 0.021 AUC-ROC over a linear readout of its own input; classical baselines reach 0.6404 for a fitted propensity table and 0.7002 for Chou-Fasman [30]. Baselines and per-class figures are in Supplementary Section S1.

Finding 5: an ablation gap does not measure how much a model uses the partner

Ablation. Supplying the partner sequence changes pooled binary AUC-ROC by +0.0035 against a model given neither partner sequence nor class, and +0.0045 against one given class only; pooled the same way, the LoRA gap is larger at +0.0073, and averaging the five per-fold gaps rather than pooling their residues gives +0.0081. Whether the partner was seen in training moves performance more: 0.9084 against 0.8940 for unseen (Supplementary Section S6).

That gap does not survive inference at the unit where samples are independent. Resampling the 1,817 homology super-clusters rather than the five folds gives a 95% interval of -0.0018 to +0.0088, including zero.

Gaps of this size need a noise floor, and it has to be a floor for retraining rather than for resampling. Retraining the same configuration over five folds and three seeds, 45 runs, the ablation gap has a total standard deviation of 0.00476 – fold 0.00347, seed

0.00144, residual 0.00292 – so the pooled +0.0035 is 0.74 times its own noise, and the two arms of that comparison are two different training runs (Supplementary Section S10). Seed-paired within one fold the gap is positive in all five seeds on fold 0 (+0.0051, SD 0.0025; Supplementary Section S2), and over the full 45-run design it is +0.0065 with the same sign in 14 of 15 (fold, seed) cells (Supplementary Section S8). The LoRA gap seed-paired on fold 0 is +0.0008 (SD 0.0015), positive in three seeds of five, so no LoRA benefit is claimed at that unit – which is a claim about five runs on one fold, not about the +0.0073 pooled figure above. Seven estimates of one quantity is what the ablation design produces rather than a disagreement to resolve; Table 6 reconciles them.

Controls. The ablation compares two separately trained models, so we varied the partner within each fixed model instead, on every complex in all five folds (Table 4, Figure 1). Every intervention costs accuracy, and the ordering is informative: rotating the partner's keys and values costs most (-0.0062 frozen), scrambling its residues about two thirds of that (-0.0045), a different real partner about half again (-0.0023). Rotating keys alone costs nothing resolvable (-0.0008, SD 0.0015).

Binary AUC-ROC would mislead alone. Table 7 adds metrics that see a helix-strand reassignment or a calibration shift, and the ordering holds on all of them: rotating keys and values costs 0.017 Q3 and moves 23.7% of residues by more than 0.1 in probability, scrambling 0.014 Q3 and 18.3%, a different real partner 0.004 Q3 and 8.0%. LoRA gives the same ordering on binary AUC-ROC with smaller effects, but not on Q3, where scrambling costs more than the rotation (-0.0146 against -0.0084).

The swap gets the ablation gap's treatment: resampling the homology super-clusters that contain both label classes. The matched swap costs 0.0024 (95% CI 0.0014-0.0036 over 1,454 super-clusters and 4,768 complexes) and 0.0014 (95% CI 0.0004-0.0025) with LoRA, both excluding zero. That number depends on how the substitutes are chosen. Choosing by composition alone selects near-cognate partners: 2,726 of 4,768 complexes receive at least one substitute from the true partner's own UniProt accession, and the per-complex effect falls to zero in the closest quartile (Supplementary Section S6). Screening to under 30% identity, a different accession and a different entry within 5% of the true length, the cost rises to 0.0066 (95% CI 0.0044-0.0094 over 1,443 super-clusters), negative in all five folds; 5,633 of 5,683 complexes admit ten such decoys. But the screen moved the substitutes farther in composition and in embedding space, and the per-complex effect rises with both (Supplementary Section S6). A third set separates distance from identity: keeping the screen, but drawing each complex's decoys as near as that pool allows to its composition-selected decoys' embedding distance, the cost is 0.0073 (95% CI 0.0042-0.0112), negative in all five folds. The match is one-sided – median 0.0776 against a 0.0400 target – so these decoys still sit farther out than the published ones. Paired on the same complexes and resampled clusters, it exceeds the composition-selected cost by 0.0050 (95% CI 0.0019-0.0091) and is not distinguishable from the screened cost, whose difference is 0.0008 (95% CI -0.0021 to +0.0042); the ordering among the screened sets is not claimed. At the matched set's median distance the published gradient predicts 0.0029, 39% of it. Dissimilarity accounts for part of the rise, not most of it; what the composition-selected decoys also hide is cognate identity, and a swap control's headline number measures its decoys as much as the model.

Label quality is where this argument is weakest, so the weakness belongs here. Restricted to the 76 complexes whose regions are fully observed and whose labels pass verification, the ordering inverts: the composition-selected swap costs 0.0086 and is the only one of the three whose interval excludes zero, against 0.0066 matched and 0.0045 screened (Supplementary Section S11). Taken at face value that reverses the mechanism. Paired on those same 76 complexes, though, the contrast is unresolved rather than reversed: matched minus composition-selected is +0.0020, 95% interval -0.0058 to +0.0079, which admits a reversal as readily as no difference, against -0.0050 excluding zero on the full set, and unresolved in either direction on every label-verified subset (Supplementary Section S11). The mechanism rests on the full set, where it is resolved; these 76 complexes neither confirm nor refute it.

On the independent temporal test set the ablation gap is +0.0078 (95% CI 0.0044-0.0117 over 2,000 cluster-level resamples), excluding zero where the cross-validation gap did not, while the matched swap on that same fixed model costs 0.0027 – the same relation between the two measurements, on data no fold shares.

Capacity, or partner information? The cross-model ablation confounds the two: `full` and `no_partner` differ in whether a sequence flows through cross-attention at all, so a positive gap could reflect the extra pathway rather than knowledge of the partner. Five conditions separate them (Table 5). Each holds architecture, parameter count and training budget identical to `full` and feeds a real partner sequence through cross-attention; only its information content changes. The clean comparison is `wrong_partner`, which removes the true sequence and the true partner class while matching `full` in every other respect: on one run per fold it sits 0.0065 below `full` in five folds of five. A fold is not a training replicate, so the three conditions carrying the argument were retrained across folds and seeds together: 5 folds x 3 seeds x 3 conditions, 45 runs, every contrast differenced within a seed (Supplementary Sections S8 and S10). Two claims separate there. `wrong_partner` below `full` is robust: -0.0076 with the same sign in all 15 (fold, seed) cells and 1.73 times its own total retraining SD of 0.0044. `wrong_partner` below `no_partner` is not: -0.0010 at 0.40 times its total SD, with the same sign in only 10 cells of 15. A misleading partner is therefore reliably worse than the true one, so the ablation effect cannot be attributed solely to spare capacity – which is not to say it carries no architectural or training-path contribution at all. Whether a wrong partner is worse than no partner is not resolved by these replicates. The ablation gap across the same runs is +0.0065, reversing in one cell of fifteen.

Taken with Table 7: partner information reaches the model during training and is worth 0.004-0.007 AUC-ROC, which is not what a substitute measures at inference.

Every estimate whose design retrained more than once exceeds the retraining SD, and both estimates that fall below it retrained each arm once. Two single-run estimates exceed it as well, which is what the last column is for: exceeding a noise floor and having measured one are different things. The gap is real; its size is not established to better than a factor of two, which is the most an ablation of this kind supports.

The same control detects partner conditioning in a published method

To establish that the controls detect conditioning rather than merely fail to find it, we applied the matched swap to Disobind, a published sequence-based predictor of inter-protein contact maps and interface

Table 4 Within-model control effects, as change in binary AUC-ROC relative to the true partner. Mean across five folds, standard deviation in parentheses; rotations are means over ten Haar-uniform draws per fold. The cross-model ablation gap is not a within-model control and is given for scale: its two model columns are means over five retraining seeds on fold 0, the unit at which a gap can be seed-paired, and not over folds like every other row; the pooled five-fold gaps are +0.0035 frozen and +0.0073 with LoRA, and Table 6 reconciles all seven estimates. Every complex carries a partner sequence and all are scored (1,209, 1,085, 1,024, 1,069, 1,296). The final column restricts to IDRs the benchmark contains with more than one distinct partner – the subset where partner-dependence is learnable at all. Its intervention rows are over three folds; its ablation gap is over five, matching the folds available to each measurement. Finding 6 takes up whether being seen with several partners means being partner-dependent.

Intervention	Frozen ESM-2	LoRA ESM-2	Frozen, multi-partner IDRs
Rotation (keys)	-0.0008 (0.0015)	-0.0003 (0.0010)	-
Rotation (keys+values)	-0.0062 (0.0026)	-0.0050 (0.0043)	-
Scrambled partner	-0.0045 (0.0026)	-0.0030 (0.0023)	-0.0101 (0.0087)
Matched partner swap	-0.0023 (0.0008)	-0.0015 (0.0013)	-0.0012 (0.0012)
Cross-model ablation gap	+0.0051 (0.0025)	+0.0008 (0.0015)	+0.0072 (0.0084)

Table 5 Fixed-architecture partner corruptions, five folds, paired against baselines retrained under matched conditions (`full` 0.89745, `no_partner` 0.89338, gap +0.0041, positive in four folds of five). Differences are computed from these unrounded values, so three of the `vs full` entries do not reproduce from a four-decimal display. Every `vs no_partner` entry is smaller than the 0.00476 total standard deviation of retraining this configuration (Supplementary Section S10), and none of them is resolved by this design; the `vs full` entries exceed it. Standard deviations across folds in parentheses; folds share training data, so the sign count is given rather than a p-value. `wrong_partner` also collapses the partner class to unknown, matching `no_partner` in information while matching `full` in architecture.

Condition	Binary AUC-ROC	vs `no_partner`	vs `full`	below `full`
Constant partner	0.8890	-0.0044 (0.0014)	-0.0084 (0.0046)	5/5
Shuffled partners	0.8911	-0.0023 (0.0029)	-0.0063 (0.0010)	5/5
Resampled each draw	0.8917	-0.0017 (0.0024)	-0.0058 (0.0037)	4/5
Length-matched partner	0.8920	-0.0014 (0.0020)	-0.0055 (0.0029)	5/5
Wrong partner, class removed	0.8910	-0.0024 (0.0031)	-0.0065 (0.0017)	5/5

residues for IDRs, built on ProtT5 [21, 5]. It takes protein pairs as UniProt ranges, so the substitution is entirely at the input level and the model untouched; 59 of the 61 pairs in its v_19 test set were scored (Methods). None appears among the 5,342 pairs of its v_19 training set, so the baseline is not inflated by leakage, though for 21 of the 59 both partners occur in training otherwise.

Scored against its own target contact maps, the swap reduces Disobind's AUC-ROC from 0.7031 to 0.5220, near chance: a mean paired drop of 0.1811 degrading 47 of 57 complexes (paired Wilcoxon, $p = 8.9 \times 10^{-7}$). The complex is the inferential unit, since residues within one are strongly correlated; the per-residue shift in interface probability, 0.176 on average (Figure 2; median 0.139 over 3,455 residues), is a magnitude rather than a test. The control therefore spans 0.18 AUC-ROC on a partner-conditioned model against 0.0024 to 0.0073 on ours.

Calibrating whether the architecture can use a partner at all

The Disobind contrast shows the controls have power, but across a different task, encoder and dataset. A null from a model that cannot read its partner channel would be uninformative.

Holding benchmark, architecture, hyperparameters and training budget fixed, we relabelled fold 0 so the correct answer depends on the partner: a bit from the partner's first residue determines whether helix and strand are transposed (Methods). A model that reads the partner can learn the rule; one that ignores it is at chance.

Transposing H and E leaves the structured-versus-coil partition untouched, so binary AUC-ROC is invariant under that rule by construction and it can only be read on Q3, where the gap is **+0.0496** against **+0.0007** on the real task. Against a partner-blind floor of 0.5961 Q3 and fold 0's own real-task ceiling of 0.8425 the model reached 0.6457, recovering **20.1%**.

A companion rule keyed the same way but exchanging structured and coil classes makes binary AUC-ROC informative, and comparable with every effect above. Here the partner-reading model reaches 0.8795 against 0.5234 for the partner-blind one – a gap of **+0.356**, 75 times the 0.00476 retraining SD that governs every cross-model gap in this paper – and against the same fold's ceiling of 0.9249 it recovers **88.7%**.

Those two figures are single runs, so we retrained both rules under further seeds and only one replicates (Supplementary Section S12). The coarse rule does: three seeds give a gap of +0.3442, SD 0.0139, and a captured fraction of 84.4% to 90.3%. The fine rule does not: two further seeds of the same configuration recover **73.8%** and **79.3%** against the first's 20.1%. The partner-blind floor barely moves across the three, 0.5961 to 0.6075, while the partner-reading arm goes from 0.6457 to 0.7938 – a rule this budget finds in two runs of three, rather than one the architecture cannot represent. The 20.1% is one run and the worst of three, and any reading of it as a limit is withdrawn.

What survives is the coarse figure, and it is the one the paper needs. Exchanging structured and coil inverts the whole output from one bit of the partner, and that is the axis every effect here is measured on. Recovering 84% to 90% of its own real-task ceiling, the architecture can read a partner and act on that axis, so a small measured response is not the metric failing. It does not follow that a *localised* effect would be found: this rule moves every position at once, where partner dependence moves a few per region. The calibration bounds the global case and bounds nothing on finer distinctions. No ablation or swap establishes either point, and neither does a single run.

Finding 6: what partner-dependence there is, is a difference in order, not in fold

The controls put the partner's contribution at a few thousandths of AUC-ROC; they do not say why it is so small. For each IDR

Table 6 All 7 estimates of the cross-model ablation gap in this paper, reconciled. They differ in what was retrained, in the unit inference is drawn at, and in how uncertainty was quantified, and the spread between them – +0.0035 to +0.0078 – is the paper's point rather than an inconsistency in it: an ablation gap is not one quantity. The fifth column divides each estimate by the 0.00476 total retraining standard deviation (Supplementary Section S10); the sixth asks whether that estimate's own design measured retraining variance at all, since a cluster bootstrap on one pair of runs does not, however tight its interval. Generated by `scripts/build_gap_reconciliation.py`. Within-model swap effects are absent by construction: both arms come from one trained model, training variance cancels in the difference, and the cluster bootstrap governs them instead.

Estimate	Value	What was retrained	Unit of inference	/ SD	Resolved
Pooled cross-validation, frozen	+0.0035	one run per arm per fold, the published CV	homology super-cluster (1,817), 2,000 resamples	0.74	no
Matched-condition retrain baselines	+0.0041	both arms under Table 5's matched budget, one run per fold	fold, as a sign count	0.86	no
Seed-paired, fold 0, frozen	+0.0051	five seeds of both arms, fold 0 only	seed	1.07	yes
Paired-seed hierarchy, folds 0-1	+0.0060	3 seeds x 2 folds, differenced within a seed	seed within fold	1.26	yes
Full retraining design, 5 folds x 3 seeds	+0.0065	45 runs: 3 conditions x 3 seeds x 5 folds	(fold, seed) cell	1.37	yes
Multi-partner IDR subset, frozen	+0.0072	nothing further; the published runs scored on a subset	fold	1.51	no
Independent temporal test set	+0.0078	one run per arm on the pre-2021 training split	super-cluster, 2,000 resamples	1.64	no

Table 7 The same interventions on metrics that binary AUC-ROC cannot see. Mean across five folds. Delta-Q3 is per-residue three-state accuracy; mean |dP| is the mean absolute change in per-residue P(structured); the last column is the fraction of residues whose probability moves by more than 0.1. A control that moves binary AUC-ROC by six thousandths can still move a quarter of the residues by more than 0.1. Standard deviations across the five folds are in parentheses. They describe fold-to-fold spread, not sampling uncertainty: no column carries a confidence interval, and none should be read as an independent confirmatory test.

Intervention	Model	Delta binary	Delta Q3	mean |dP|	moved > 0.1
Rotation (keys)	Frozen	-0.0008 (0.0016)	-0.0004 (0.0079)	0.022 (0.005)	4.8% (2.0)
Rotation (keys+values)	Frozen	-0.0062 (0.0029)	-0.0168 (0.0052)	0.070 (0.006)	23.7% (2.2)
Scrambled partner	Frozen	-0.0045 (0.0029)	-0.0141 (0.0035)	0.056 (0.012)	18.3% (5.2)
Matched partner swap	Frozen	-0.0023 (0.0009)	-0.0042 (0.0008)	0.028 (0.004)	8.0% (1.4)
Rotation (keys)	LoRA	-0.0003 (0.0011)	-0.0027 (0.0058)	0.019 (0.009)	4.4% (4.3)
Rotation (keys+values)	LoRA	-0.0050 (0.0048)	-0.0084 (0.0083)	0.059 (0.014)	18.9% (4.4)
Scrambled partner	LoRA	-0.0030 (0.0025)	-0.0146 (0.0167)	0.048 (0.012)	15.2% (4.9)
Matched partner swap	LoRA	-0.0015 (0.0014)	-0.0012 (0.0020)	0.022 (0.002)	6.2% (0.9)

region observed with two or more distinct partners we compared its conformation across a partner change against its own replicate noise (Methods; three requirements each changed the answer, and each is stated there).

Replicates of the same complex disagree on **5.8%** of co-resolved positions. Changing the partner costs a further **2.3%** for the median region (95% CI 0.99-5.76% over 46 regions, Wilcoxon p = 0.00017), with 34 of 46 regions moving in the expected direction. The effect is robust to the replicate-independence threshold: +0.0204 with no threshold (71 regions) and +0.0164 at one year (43 regions). It is also robust to the two ways these regions are not independent. Collapsing to one observation per accession leaves 42 accessions and raises the median to +0.029; excluding those an established Pfam or CATH domain covers – the contamination Finding 1 measures – 30 remain at +0.049, still excluding zero.

The kind of change is the informative part. Only **0.5%** of disagreeing positions convert one secondary structure into another; the rest exchange with coil – 62.5% coil-helix, 20.7% coil-PPII and 16.2% coil-strand, so a fifth of the partner-dependent change is the very state Finding 2 shows the alphabet cannot express – and replicates of the same complex show the same profile. The PPII category is not in the deposited labels and could not be: those are three-state with no PPII code, which is Finding 2. This comparison is made in the dihedral-derived four-state space of the Methods, recomputed from coordinates for both members of every pair, so what a coil-PPII transition names is a position the benchmark would label C on both sides. Partner-associated differences here are changes in how much order is induced, not in fold, on regions heterogeneous in construct boundaries, resolution and origin (Supplementary Section S7). The signal is real and small against the noise available to measure it.

A construct defect surfaced alongside: **163 samples across 86 accessions** carry a six-histidine tag inside the putative IDR.

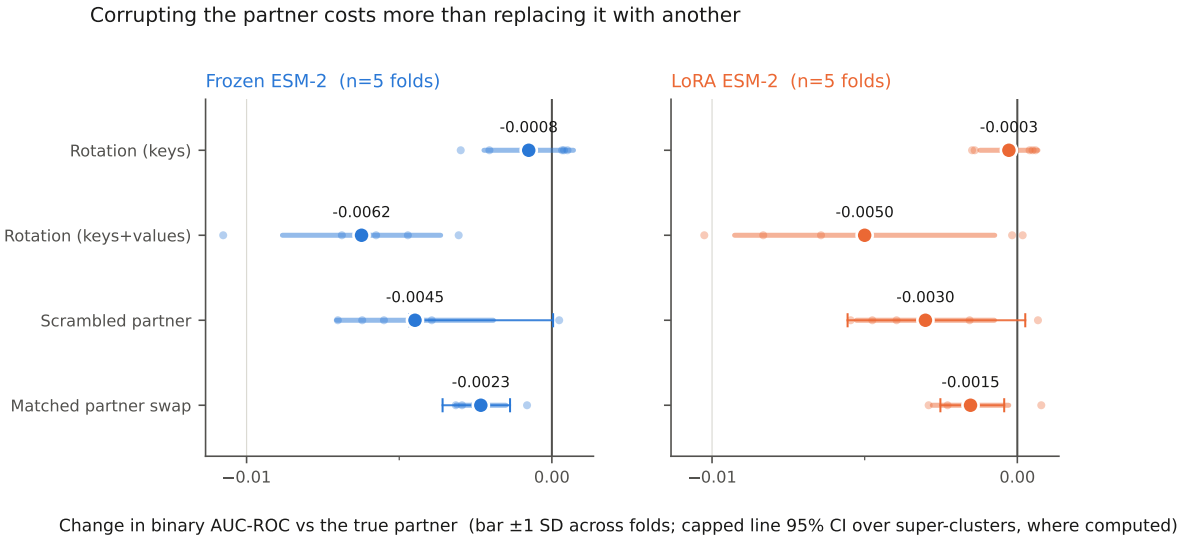

Figure 1 Within-model controls. Change in binary AUC-ROC when the partner is altered but the trained model is held fixed, across all five folds and every complex in them. Rotations and swaps are means over ten draws per fold, scrambles over five. Faint points are individual folds and the thick bar is one standard deviation about their mean, which describes fold-to-fold spread and supports no inference: folds share training data. The capped line is a 95% interval from resampling homology super-clusters, drawn for the two interventions whose per-residue predictions were saved – the matched swap, 0.0024 (95% CI 0.0014-0.0036) frozen, excluding zero, and the scramble, whose interval includes zero despite a larger mean because it rests on five draws rather than ten. The interval's own point estimate, 0.0024, is the pooled one; the dot it surrounds is the mean of the five fold effects, 0.0023. They are two estimators of one quantity and differ in the fourth decimal. The rotations carry no interval and are descriptive. No noise band is drawn: these are within-model effects, both arms from one trained model, so the retraining standard deviation that governs the cross-model gaps does not apply to them and the cluster interval is the whole of their uncertainty. Every intervention costs accuracy, and the ordering is the finding: rotating keys and values costs most, scrambling the partner's residues about two thirds of that, and substituting a different real partner about half again.

Discussion and conclusions

Each shortcut substitutes an available proxy for the property wanted: a short chain for a disordered region, a depositor annotation for a measurement, an unresolved position for an irregular one, deposition date for independence, an ablation for evidence of use – each wrong in a direction that flatters the benchmark.

The ablation gap is positive and, by convention, evidence of conditioning. It is also 0.74 times the standard deviation of retraining the same configuration, so by itself it is not measurable. A within-model swap escapes that variance, both arms coming from one trained model, but what it reports depends on how its decoys were chosen: composition alone hands back the true partner's own protein for more than half the benchmark, and screening those out triples the effect whether or not the distance is matched back afterwards. The response is to partner identity, and the low published figure is an artefact of the decoys. Nor is the gap only capacity: a wrong partner is worse than the true one in every paired run, though whether it is worse than no partner stays below the same noise floor.

The control is not blind to conditioning where it exists: the same substitution costs Disobind 0.18, an order of magnitude more than ours, and AlphaFold-Multimer separates cognate pairs from 73 decoys at 0.91 AUC-ROC [19]. Neither contrast is architecture-controlled, which is what the calibration supplies: without one a small effect cannot be told apart from a model barely able to condition. The calibration answers that on the coarse axis these effects are measured on: 84% to 90% capture across three seeds, so the channel is not the limit there. It answers nothing about finer distinctions, where three seeds of one configuration recover 20.1%, 73.8% and 79.3% and the figure is a property of the run rather than of the architecture – which is this paper's argument about single runs, arriving at the one experiment we ran to establish credibility.

Pretraining limits the positive claim only: our splits control train-test overlap but not overlap with ESM-2's corpus, so "predictable from sequence" means predictable given a pretrained language model. That could inflate absolute accuracy; it cannot manufacture the small size of the response to which partner is shown.

Limitations

Every model result describes one benchmark built by one group, though the label defects are not ours alone – Table 3's last column measures the same four steps on DisProt, MFIB and DIBS. Two further couplings reproduce both partner effects (Supplementary Section S4), but all three were trained by us. Coverage is uneven: rotation and swap cover five folds, the multi-partner stratification three, the agnostic panel one. The Disobind comparison rests on 59 pairs, 26 of them self-pairs, though its effect survives excluding them – the 33 hetero-pairs drop by 0.157 at the median over the possible assignments of the two rows whose identity its deposited outputs do not fix, and by between 0.119 and 0.187 under every one of them (Methods) – and its test set is held out. Two figures are floors: the curated-IDR fraction, since 14.6% of samples have no SIFTS range, and the 17.9% PPII fraction, since crystal structures miss conformations transient in solution. The calibration is idealised, so its figures bound capacity in one direction only. Finding 6 rests on 46 regions, 42 accessions, tabulated in Supplementary Section S7.

We ask two things of the field: report the within-model swap alongside the ablation, with its decoy construction stated, and

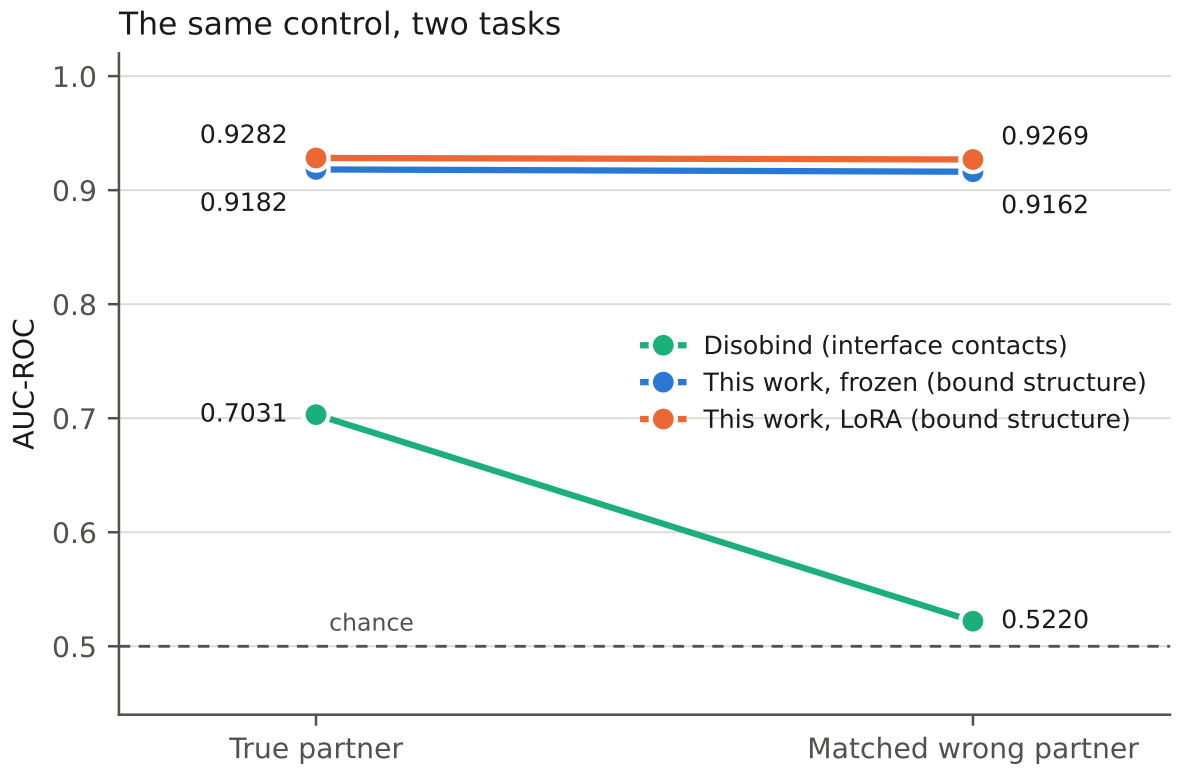

Figure 2 The same control applied to two tasks. Each line joins a model's accuracy with the true partner to its accuracy with a length- and composition-matched partner from a different complex, chosen by composition alone as in the published control rather than screened for homology (Supplementary Section S9). Disobind, predicting interface contacts, falls from 0.7031 to 0.5220 — approximately chance. Both of our models, predicting bound secondary structure, fall by about a hundredth as much: on fold 0 the swap costs 0.0020 frozen and 0.0013 with LoRA, against 0.1811 for Disobind. Our two series are the 964 of fold 0's 1,209 complexes that carry both label classes among their scored residues and were scored by all ten draws, which is the basis every pooled swap estimate in this paper uses; scoring all 1,209 instead moves the frozen cost to 0.0025 and the LoRA cost to 0.0015 without changing the comparison. The comparison is across tasks and not architecture-controlled: Disobind uses ProtT5 embeddings and a different architecture. Table 4 reports means over all five folds. Plotted and quoted from one source.

report a calibration establishing that the architecture can use its conditioning channel – under replicate seeds, because a calibration read from one run is the error this paper is about, and we made it. The two asks differ in cost, and the difference should be stated rather than glossed: the swap needs no new data and no new training, since it rescores a model you already have, while the calibration is a retraining and a single run of it can be off by a factor of three (Supplementary Section S12). Five of the six findings are correctable by construction; the sixth is not, because a benchmark assembled correctly at every step can still lack the resolving power for the phenomenon it tests. None of this implies partner-conditioned prediction is a bad idea – only that an ablation gap does not measure how much a model uses its partner, and that measuring that takes a within-model substitution whose decoys are screened and whose inference is at the unit where samples are independent.

Methods

Dataset construction

Candidate complexes were retrieved from the RCSB PDB search API [22] (accessed 2026-08-10) with exactly two protein polymer entities, at least one of 10-100 residues, and resolution <= 2.5 A, returning 8,030 entries. Polymer entities, chains, sequences and UniProt accessions came from the RCSB GraphQL API; the short entity was taken as the IDR and the other as the partner, with partners under 30 residues rejected, though 9 survive the multi-entity addition below, the shortest 7. With exactly two entities the partner is unambiguous and no contact calculation is required.

Secondary structure was assigned from `HELIX_P` and `SHEET` instance features – depositor records rather than a DSSP computation [23, 24] – mapped to H and E with all other positions C. Per-residue observation came from the PDBe residue-listing API [25]; samples with fewer than 10 observed residues were rejected. IUPred3 scores [26] were computed on the parent sequence and retained as metadata, not used as a filter. This yielded 5,461 samples.

A further 397 samples derived from three-to-five-entity complexes were added, for which the partner was determined by heavy-atom contact (5.0 Å cutoff, minimum three contacting residues) computed from mmCIF coordinates with Biopython [27], and the observation mask taken from `_atom_site.label_seq_id`. Merging and deduplication on (IDR sequence, partner sequence, PDB ID) gave 5,683 unique complexes. Of these, 82 come from DisProt rather than from the mining query, retained where their labels agreed with

the deposited annotation; they did not pass the numeric filters above, and 25 exceed the resolution limit (Supplementary Section S5).

Limitations of filtering: we did not require the IDR to be predicted disordered. Calibrating an IUPred3 threshold against 397 samples already verified as IDR–partner complexes gave a median disorder fraction of 0.11, and DisProt-curated regions — experimentally verified disorder — scored no higher than PDB-mined ones, so a threshold would have removed genuine IDRs. The score is retained as a stratification variable.

Splits

IDR and partner sequences were clustered with MMseqs2 [28] at 30% identity and 80% coverage. Super-clusters were formed by union-find over shared IDR clusters, shared grouping key (the IDR's own accession), shared PDB entry, and exact substring containment between IDR sequences. Partner-sequence clusters were deliberately not used as union edges: common partners chain the dataset into a single component under single linkage, reducing 1,817 clusters to 896 with the largest holding half the samples. Partner novelty is instead recorded per sample.

Super-clusters were assigned to five folds by a greedy pass on observed-residue count followed by local refinement minimising secondary-structure composition imbalance, with clusters kept atomic. For each fold, that fold is the test set, the next is validation and the remaining three are training.

Model and training

The model encodes the IDR and the partner with ESM-2 650M [10] (`esm2_t33_650M_UR50D`), projects both to 256 dimensions, and applies multi-head cross-attention (4 heads) with the IDR as query and the partner as keys and values, followed by a three-class per-residue classifier. Where no partner sequence is available a learned partner-type embedding is substituted. Trainable parameters: 1,104,067.

Three ablations were trained: `full` (partner sequence and partner class), `type_only` (partner class, sequence withheld) and `no_partner` (sequence withheld and partner class collapsed to a single reserved index). Collapsing the class is what distinguishes `no_partner` from `type_only`; without it the two configurations are identical.

Loss was cross-entropy with label smoothing 0.15 and `ignore_index` on masked positions, so unobserved residues contribute to neither the loss nor the metrics. Optimisation used AdamW at learning rate 5×10^{-5}, batch size 8, early stopping on validation AUC-ROC with patience 20 evaluations. Frozen-ESM arms were trained from precomputed ESM-2 embeddings cached in float16; cached and live embeddings agree to 4×10^{-4} relative error and produce identical argmax predictions. LoRA arms [29] ($r = 4$, $\alpha = 4$, dropout 0.2) were trained with live encoding on an NVIDIA RTX 3090.

Evaluation

Binary AUC-ROC is computed over helix-or-strand versus coil; three-state AUC-ROC is the macro one-versus-rest mean; Q3 is per-residue three-state accuracy. All metrics use observed residues only.

No target sample size was set in advance: the benchmark is the complete set of PDB entries meeting the stated criteria, the study is observational, and no hypothesis test was powered. Confidence intervals are 95% percentile bootstrap over super-clusters (2,000 resamples for pooled metrics, 1,000 for per-class). Resampling super-clusters rather than residues is necessary because residues within one IDR are highly correlated: on a 97-sample subset of the predecessor dataset a residue-level interval spanned 0.023 where a sample-level interval over the same predictions spanned 0.071.

Controls

All controls hold one trained model fixed and alter only the partner it receives.

Rotation null. A Haar-uniform orthogonal matrix $R \in O(256)$ was drawn by QR decomposition of a Gaussian matrix with sign correction, and applied to the partner representation entering cross-attention — to keys only (preserving the semantics of delivered values) and to keys and values. Ten seeds. R preserves norms and all partner–partner inner products, so only the learned query–key alignment is destroyed. Verified by an identity-rotation check reproducing the unmodified predictions exactly.

Matched partner swap. Each IDR was rescored against partners drawn from other complexes in the training pool, within ±20% of the true partner's length and ranked by cosine similarity of 20-dimensional amino-acid composition; ten draws per sample. A scrambled arm used the true partner's residues in shuffled order, preserving length and composition exactly, over five draws. Every control scores every complex in the fold. Inference is at the super-cluster: clusters are resampled with replacement 2,000 times, the pooled binary AUC-ROC is recomputed under the true partner and under each draw on the same residues, and the intervention's mean over draws is taken inside each resample. Complexes whose scored residues are all one class carry no AUC and are excluded, leaving 4,768 of 5,683 in 1,454 super-clusters. No equivalence test is reported: an earlier version of this work tested effects against a ±0.0023 margin taken from one fold's seed SD, and a margin set equal to a noise estimate is not a statement about practical importance. Effects are reported against the measured retraining SD instead, with which design resolves which effect stated (Supplementary Section S10).

Agnostic panel. Per-residue P(structured) was marginalised over a fixed 24-partner panel excluding each complex's own partner, and compared to the correct-partner prediction.

Fixed-architecture corruptions (Table 5). Five conditions retrain the published architecture, parameter count, optimisation schedule and token budget, changing only which partner sequence each sample is shown. **const_partner** shows every sample one sequence, the median element of the split's sorted pool. **shuffled_partner** applies a permutation of the split's partner list, with identity pairings pushed one place along so no sample receives its own partner. **wrong_partner** applies the same permutation and additionally collapses `partner_class` to the reserved unknown class, so it matches `no_partner` in information while matching `full` in architecture. **random_partner** draws, per sample, uniformly from the pool restricted to within ±20% of the true partner's length and excluding the true sequence. **resampled_partner** draws afresh at every access from one seeded stream, so a sample sees a different partner in each epoch; it is the only condition without a fixed assignment. The other four are built once per dataset from a seed fixed by condition name, and each condition is applied to the training, validation and test splits alike, so a corrupted model is trained, selected and scored under the same corruption. Exclusion is by sequence identity, not by accession or family: on fold 0's test split 14 of 1,209 samples

under the shuffle and 24 under the length-matched draw receive a substitute sharing the true partner's accession, and length ratios span 0.814-1.178 for the length-matched condition against 0.208-4.576 for the shuffle. One training run per condition per fold, five folds. The complete assignments are deposited.

Disobind comparison

Disobind was obtained from https://github.com/isblab/disobind (Epsilon_3, Version_16 parameters) and its benchmark from Zenodo record 10.5281/zenodo.15873859 (accessed 2026-08-11), and run on an RTX 3090 with `--cmaps --coarse 1`. Its input format is a CSV of UniProt ranges, so the swap was applied at the input level with the same length and composition matching used above and no modification to the model. Predictions were read from the continuous arrays in `Predictions.npy`; contact-map predictions were scored against `Target_bcmap_test_v_19.h5` from the same release. 26 of the 59 listed pairs pair a protein with another copy of itself, and the deposited per-complex AUCs carry no identifiers, so the two pairs that dropped out of scoring cannot be assigned to a type. The hetero-pair effect is therefore computed over all 1,711 possible assignments of those two rows: the mean drop on the 33 hetero-pairs is 0.157 at the median across those assignments, spans 0.119 to 0.187 across all of them, and is positive under every one (`scripts/build_disobind_breakdown.py`).

Auditing the construction steps

Curated-IDR cross-reference. Every IDR accession was queried against MobiDB, counting `curated-disorder-*` and `curated-lip-*` (MoRF) separately. `prediction-*` was excluded, since validating a selection rule with a predictor is circular, and `derived-missing_residues` likewise, being the signal the observation mask is built from. A region counts as supported when at least half its residues fall under a curated span, placed in UniProt numbering through SIFTS.

Independent secondary structure. Structures were re-assigned from coordinates with pydssp [23], which assigns three states directly from backbone hydrogen bonds – there is no eight-state intermediate, and no separate 3-10 or pi code, so a 3-10 turn adjoining a helix is scored as coil. Deposited and recomputed labels were compared only at residues the benchmark marks observed and that a local gapped alignment maps with matching residue identity.

PPII assignment. PPII is defined geometrically rather than by hydrogen bonding, and is assigned from backbone dihedrals: phi in [-104, -46] and psi in [116, 174] (the conventional -75 ± 29 / 145 ± 29 basin), in runs of at least three consecutive residues, and only where the hydrogen-bond assignment gives neither H nor E. Dihedrals are computed only between residues consecutive in `auth_seq_id`. The method recovers poly-L-proline bound to profilin (8 of 10 residues), the Crk and Sem-5 proline-rich peptides, and the HIV-1 Nef PxxPxR motif, and calls no PPII in the folded domains they bind.

Partner-dependence. Regions are (accession, overlapping SIFTS residue range); exact sequence grouping splits construct variants of one region apart and, because identical constructs come from one laboratory, compares only the most similar structures. Labels are indexed by UniProt residue number and a pair is used only where the two samples agree on the amino acid at 95% of shared positions; one pair in six fails. Same-partner pairs are co-deposited far more often than different-partner pairs (17.5% against 3.1% within thirty days), so they are required to be at least 180 days apart.

Temporal holdout. Deposition dates were retrieved for all 5,396 distinct entries from `rcsb_accession_info.deposit_date`. Training used complexes deposited before 2021-01-01; every later sample sharing a super-cluster with a pre-cutoff sample was discarded rather than tested on. Validation was split off the pre-cutoff side by whole clusters, selected smallest-first to a target sample fraction, because the cluster-size distribution is skewed enough that taking a fraction of cluster identifiers places 48% of samples in validation.

Calibration task. Fold 0 was relabelled so the correct answer depends on the partner: `mode = AA_ORDER.index(partner_sequence[0]) % 2`, and where `mode == 1` the H and E labels are transposed. Complexes, IDR sequences, class balance, architecture and training budget are unchanged. The rule is keyed on first-residue identity rather than composition because the matched swap matches 20-dimensional composition and the scramble preserves it exactly, so a compositional rule would survive both controls by construction. Scrambling flips the mode 49% of the time against a maximum of 50%. Binary AUC-ROC is invariant to this relabelling and must not be used to score it.

Software

MMseqs2 18-8cc5c [28], Biopython 1.85, PyTorch 2.10.0 (local, MPS) and 2.0.1+cu117 (GPU server), scikit-learn 1.6.1, NumPy 2.2.6, SciPy 1.15.3, fair-esm 2.0.0, pydssp 0.9.1, IUPred3, h5py, peft 0.7.1 (GPU server). Statistical tests were two-sided Wilcoxon signed-rank tests as implemented in SciPy; no correction for multiple comparisons was applied, and comparisons are reported individually rather than as a family.

Acknowledgments

We thank the RCSB PDB, PDBe and UniProt teams, whose openly published structures and residue-level mappings this audit is built on.

Funding

This work was funded by Omic Inc. The funder had no role in the design of the audit, the analysis, or the decision to publish.

Author contributions

All authors contributed to the design of the audit, the analysis and the manuscript, and approved the submitted version.

Competing interests

G.R. is the founder of Omic Inc. All other authors are employees of Omic Inc. and declare no other competing interests.

Ethics

Not applicable. This is a computational study using publicly available macromolecular structure data (RCSB PDB, PDBe, UniProt). No

human participants, no animal subjects, no clinical data, and therefore no trial registration or reporting summary applies.

Data availability

The benchmark, cross-validation splits, temporal holdout, control outputs, calibration tasks, the per-residue predictions of every training run reported here, and a provenance table mapping each verified quantity to its source file are deposited at [CITATION NEEDED: deposit DOI]. The deposit is assembled from this repository by `scripts/build_deposit.py`, which checksums every file and runs the verification script against the deposited artifacts from inside the deposit; `deposit/DEPOSIT_INFO.json` records that result. No data are available only on request. Disobind and its benchmark are the work of their authors and are available at the URLs given in Methods (github.com/isblab/disobind; Zenodo 10.5281/zenodo.15873859).

Code availability

All analysis code, including the within-model control suite, the calibration task and the verification script that checks each reported result in this manuscript against its source file, is at [CITATION NEEDED: repository URL].

References

1. Oldfield CJ, Dunker AK. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annual Review of Biochemistry* (2014). doi:10.1146/annurev-biochem-072711-164947
2. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology* (2005). doi:10.1038/nrm1589
3. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology* (2015). doi:10.1038/nrm3920
4. Basu S, Kihara D, Kurgan L. Computational prediction of disordered binding regions. *Computational and Structural Biotechnology Journal* (2023). doi:10.1016/j.csbj.2023.02.018
5. Majila K, Ullanat V, Viswanath S. Disobind: a sequence-based, partner-dependent contact map and interface residue predictor for intrinsically disordered regions. *Cell Systems* (2026). doi:10.1016/j.cels.2025.101486
6. Jumper J, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* (2021). doi:10.1038/s41586-021-03819-2
7. Evans R, et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv* (2021). doi:10.1101/2021.10.04.463034
8. Abramson J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* (2024). doi:10.1038/s41586-024-07487-w
9. Hu G, et al. flDPnn: accurate intrinsic disorder prediction with putative propensities of disorder functions. *Nature Communications* (2021). doi:10.1038/s41467-021-24773-7
10. Lin Z, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* (2023). doi:10.1126/science.ade2574
11. Quaglia F, et al. DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. *Nucleic Acids Research* (2021). doi:10.1093/nar/gkab1082
12. Piovesan D, et al. MobiDB: 10 years of intrinsically disordered proteins. *Nucleic Acids Research* (2022). doi:10.1093/nar/gkac1065
13. Schad E, et al. DIBS: a repository of disordered binding sites mediating interactions with ordered proteins. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx640
14. Fichó E, Reményi I, Simon I, Mészáros B. MFIB: a repository of protein complexes with mutual folding induced by binding. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx486
15. Chen L, Cruz A, Ramsey S, Dickson CJ, et al. Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLOS ONE* (2019). doi:10.1371/journal.pone.0220113
16. Bernett J, Blumenthal DB, List M. Cracking the black box of deep sequence-based protein-protein interaction prediction. *Briefings in Bioinformatics* (2024). doi:10.1093/bib/bbae076
17. Del Conte A, Mehdiabadi M, Bouhraoua A, Monzon AM, Tosatto SCE, Piovesan D. Critical assessment of protein intrinsic disorder prediction (CAID) - results of round 2. *Proteins* (2023). doi:10.1002/prot.26582
18. Mehdiabadi M, Del Conte A, Nugnes MV, Aspromonte MC, Tosatto SCE, Piovesan D. Critical assessment of protein intrinsic disorder round 3 - predicting disorder in the era of protein language models. *Proteins* (2025). doi:10.1002/prot.70045
19. Omidi A, Moller MH, Malhis N, Bui JM, Gsponer J. AlphaFold-Multimer accurately captures interactions and dynamics of intrinsically disordered protein regions. *PNAS* (2024). doi:10.1073/pnas.2406407121
20. Alderson TR, Pritišanac I, Kolarić Đ, Moses AM, Forman-Kay JD. Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. *PNAS* (2023). doi:10.1073/pnas.2304302120
21. Elnaggar A, et al. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE TPAMI* (2022). doi:10.1109/TPAMI.2021.3095381
22. Burley SK, et al. Updated resources for exploring experimentally-determined PDB structures. *Nucleic Acids Research* (2025). doi:10.1093/nar/gkae1091
23. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* (1983). doi:10.1002/bip.360221211
24. Hekkelman ML, et al. DSSP 4: FAIR annotation of protein secondary structure. *Protein Science* (2025). doi:10.1002/pro.70208
25. Armstrong DR, et al. PDBe: improved findability of macromolecular structure data in the PDB. *Nucleic Acids Research* (2020). doi:10.1093/nar/gkz990
26. Erdős G, Pajkos M, Dosztányi Z. IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation. *Nucleic Acids Research* (2021). doi:10.1093/nar/gkab408
27. Cock PJA, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* (2009). doi:10.1093/bioinformatics/btp163
28. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* (2017). doi:10.1038/nbt.3988

29. Hu EJ, et al. LoRA: low-rank adaptation of large language models. *arXiv* 2106.09685 (2021).
30. Chou PY, Fasman GD. Prediction of protein conformation. *Biochemistry* (1974). doi:10.1021/bi00699a002