

Co-folding confidence signals fail conditional coverage under structural distribution shift

fons Research
founders@fonsresearch.com

September 11, 2026

Abstract

Protein–ligand co-folding models ship a confidence signal with every predicted pose, and practitioners use it as an accept/reject criterion. We ask whether a threshold calibrated on that signal delivers the error rate it promises when the target is unlike anything in the model’s training window. Using 568 220 predictions from seven co-folding methods over 2448 protein–ligand systems released after each model’s own training cutoff, we show that split conformal risk control calibrated marginally meets a 10 % target error rate in aggregate (10.1 % realised) while delivering 35.8 % in the quartile of systems least similar to any pre-cutoff structure, violating its guarantee in 86 % of calibration splits. Group-conditional (Mondrian) calibration does not repair this at a modest cost in coverage: in that bin it accepts no predictions at all in 95 % of splits at $\alpha = 0.1$, and only begins accepting a meaningful share once the target error rate is loosened past 30 %. We further show that headline accuracy is substantially carried by memorisation: 27.5 % of post-cutoff systems have a ligand identical to one released before the cutoff, those cases are solved at 0.841 accuracy against 0.615 elsewhere, and they supply 34.3 % of all correct predictions. We release the per-prediction table, the calibration outputs, and the code that applies the thresholds to new predictions without a GPU.

1 Confidence signals are used as accept/reject criteria

Co-folding models (AlphaFold 3 [13], Boltz [4, 6], Chai-1 [7], Protenix [8]) predict how a small molecule binds a protein and report confidence alongside: predicted TM-score, interface predicted TM-score, predicted aligned error, per-atom pLDDT, and a composite ranking score. In practice these signals are used to triage: accept the confident predictions, discard or re-examine the rest.

That use is a selective-prediction problem with an implicit guarantee attached. Choosing a threshold τ on a confidence signal s and accepting when $s \geq \tau$ is a claim that the error rate among accepted predictions is acceptable. Whether that claim holds, and under what conditions, has not been systematically checked. It is checkable, and the conditions under which it fails are exactly the conditions of interest in drug discovery: new scaffolds, new pockets, target families with no close relative in the PDB.

2 Under structural distribution shift, they fail conditional coverage

2.1 Setup

We use the Runs N’ Poses release [1], which provides predictions from AlphaFold 3 (with and without templates) [13], Boltz-1 [4], Boltz-1x [5], Boltz-2 [6], Chai-1 [7], Protenix [8] and RoseTTAFold

All-Atom [9] for 2600 high-resolution protein–ligand systems released after 30 September 2021, together with symmetry-corrected ligand RMSD computed by OpenStructure and per-system similarity annotations against the full PDB.

Three choices define the analysis.

The label is uniform. A prediction is correct when its binding-site-superposed, symmetry-corrected ligand RMSD is below 2 Å. PoseBusters physical-validity checks [2] are available for only one pose per ligand and not at all for Boltz-2; folding validity into the label for some models and not others would make `correct` denote different things per model and confound the model comparison. Validity is retained as a secondary label for sensitivity analysis.

The grain is the pose a practitioner receives. Each model produces 25 poses per ligand (5 seeds $\times$ 5 diffusion samples). We analyse the model’s own top-ranked pose, one per ligand, because that is what an accept/reject decision is actually applied to.

Similarity is recomputed per model. Each method has its own training cutoff (Table 1). We recompute each system’s similarity to the nearest structure released before *that model’s* cutoff rather than applying one global date. This matters: Chai-1’s cutoff is more than eight months earlier than the date the benchmark was built around, and Boltz-2’s is twenty months later, so a single global cutoff would misstate the shift for both. RoseTTAFold All-Atom’s training set is described without a release-date cutoff, so it is marked unknown and excluded from all cutoff-dependent analysis rather than assigned a guessed date.

Table 1: Training-data cutoffs, each taken from the model’s own paper or repository. These dates determine what counts as “seen” for each model and therefore every similarity value in this paper; the machine-readable version with quoted source text is `config/models.yaml`.

Method	Variant analysed	PDB cutoff	Source
AlphaFold 3	with / without templates	2021-09-30	[13]
Boltz-1	v0.4	2021-09-30	[4]
Boltz-1x	v1.0.0, inference potentials	2021-09-30 [†]	[5]
Boltz-2	v2.0.3	2023-06-01	[6]
Chai-1	–	2021-01-12	[7]
Protenix	v0.5.0 weights	2021-09-30	[8]
RoseTTAFold All-Atom	–	<i>unknown</i>	[9]

[†] Boltz-1x applies inference-time steering potentials to Boltz-1 without retraining, so it inherits Boltz-1’s cutoff.

The shift coordinate is SuCOS-pocket similarity to the closest pre-cutoff system, binned at 0/30/60/90/100. Ligand ECFP4 Tanimoto is a secondary axis used as a robustness check; the joint grid is too sparse for conditional calibration and is reported only as a population table.

2.2 A kill condition, fixed in advance

Before computing any curve we committed a falsification criterion in writing: if, within every shift bin, the best native signal achieved normalised excess AURC ≤ 0.25 with a monotone risk–coverage curve and degradation across bins ≤ 0.10 , then confidence signals already support selective prediction

under shift and the central claim is dead. Normalised excess AURC is

$$nE = \frac{\text{AURC}_{\text{signal}} - \text{AURC}_{\text{oracle}}}{\text{AURC}_{\text{random}} - \text{AURC}_{\text{oracle}}},$$

zero for perfect ranking and one for an uninformative signal.

It did not fire. No model satisfied criterion 1 in all four bins, as the condition required. Five of twenty-eight model-bin cells reach $nE \leq 0.25$ (Boltz-1x at 0.195 in the 0–30 bin; Boltz-1x, Boltz-2 and Boltz at 0.141 to 0.209 in the 60–90 and 90–100 bins), and no model clears more than two of four. The monotonicity tolerance failed in every bin. The admissible signal set is closed and excludes anything computable only with the ground-truth structure; the analysis code asserts this by column name.

2.3 Result

The best signal is a chain-pair ipTM variant for every model, with normalised excess AURC from 0.204 (Boltz-2) to 0.558 (Protenix). Signals are informative but far from sufficient. Dispersion of ipTM across the 25 samples performs essentially identically to the point signal (mean 0.367 against 0.361), so sample disagreement adds little beyond the confidence value itself. Figure 1 shows the shape of the failure: in the least-similar bin risk stays high across the whole coverage range for every model.

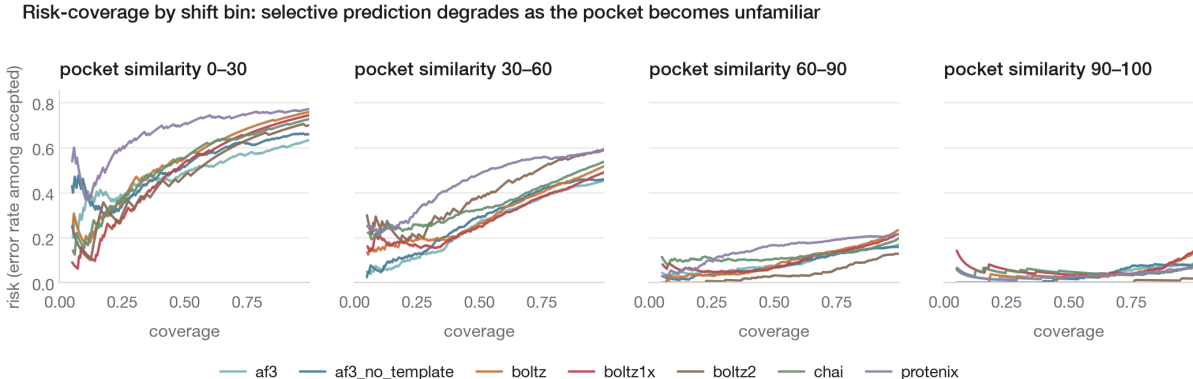


Figure 1: Risk against coverage, per model, by pocket similarity. Each curve is the model’s own best confidence signal. A signal that supported selective prediction would bend toward zero risk as coverage falls; in the least-similar bin none of them do.

Base error rate among top-ranked poses rises steeply as pocket similarity falls. For AlphaFold 3 it runs 0.635, 0.456, 0.170, 0.085 across the four bins; for Protenix 0.772, 0.588, 0.215, 0.072. A threshold calibrated on the pooled distribution inherits an error rate near 0.3 while the least-similar bin sits between 0.64 and 0.77.

Split conformal risk control at $\alpha = 0.1$, calibrated on a random half and evaluated on the other half over 50 splits and averaged across models, yields Table 2.

The pooled row is the guarantee working exactly as the theory promises: 10.1% against a 10% target, straddling α about half the time, which is what marginal validity in expectation looks like at finite sample. Every other row shows that the pooled number is an average over a population whose difficulty spans an order of magnitude, and that the easy majority is paying for the hard minority.

Table 2: Conditional coverage of marginally calibrated split conformal risk control at $\alpha = 0.1$. The pooled row meets the target; every stratified row shows who pays for it.

Pocket similarity	Realised error	Abstention	Splits violating α
0–30	0.358	0.855	86 %
30–60	0.213	0.730	95 %
60–90	0.076	0.543	15 %
90–100	0.029	0.442	0 %
pooled	0.101	0.613	53 %

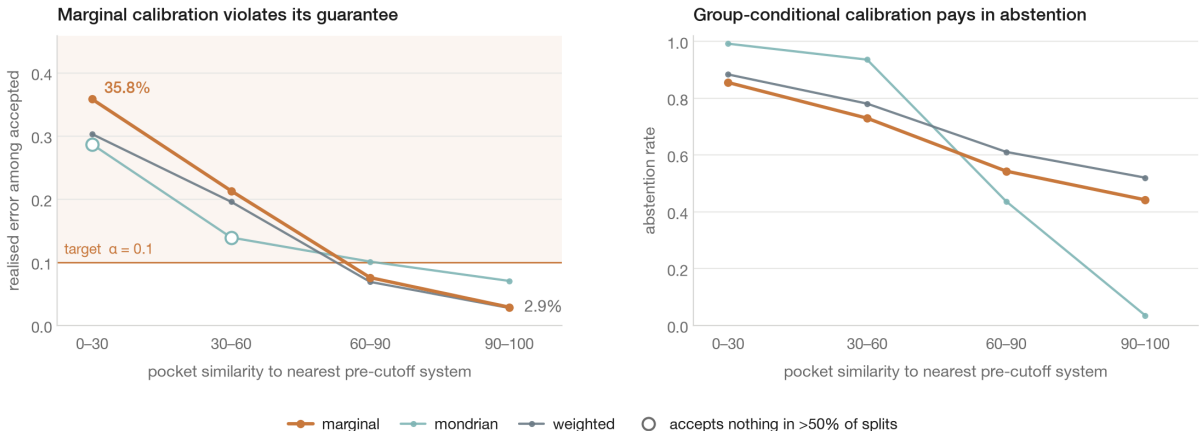


Figure 2: Realised error and abstention by pocket similarity, averaged over seven models and 50 random calibration/test splits. Hollow markers mark bins where the method accepts nothing in more than half of splits, so its error rate there is undefined rather than good.

2.4 The failure is driven by pocket novelty, not ligand novelty

Repeating the identical protocol grouping on ligand ECFP4 Tanimoto instead of pocket similarity reproduces the direction but at roughly a third of the magnitude: marginal calibration realises 0.151 error in the least-similar ligand bin against 0.061 in the most-similar, versus 0.358 against 0.029 on the pocket axis. Both axes violate the guarantee, and violation frequency in the hardest bin is comparable (84 % of splits on the ligand axis, 86 % on the pocket axis), but the severity is not. An unfamiliar *pocket* degrades selective prediction far more than an unfamiliar *ligand*. This is consistent with the mechanism being pose memorisation rather than chemistry generalisation, and it is why the pocket axis is used for conditional calibration.

3 Restoring the guarantee costs the entire acceptance budget

We implement three calibrations with one interface: marginal split conformal [12, 10], group-conditional (Mondrian) [11] with a separate threshold per bin, and weighted conformal under covariate shift [3] with likelihood ratios estimated by a density-ratio classifier over the shift features.

Mondrian restores conditional validity, in the only way available to it. In the 0–30 bin at $\alpha = 0.1$ it accepts nothing in 93 % of splits; its abstention rate is 0.992. This is correct behaviour, not a failure: no threshold on this signal certifies this bin at this error rate with this much calibration

data, and the method declines rather than issue one it cannot back. The alternative, borrowing a threshold from a better-populated bin, is precisely the marginal failure above.

Sweeping the target error rate shows the shape of the trade in the least-similar bin (Table 3).

Table 3: Least-similar bin (pocket similarity 0–30) across target error rates. Marginal calibration misses its target at every level, by a factor of five at $\alpha = 0.05$ and three at $\alpha = 0.10$, narrowing to roughly twice the target by $\alpha = 0.4$; conditional calibration holds the target only by declining almost everything. Twenty splits per level, against fifty for Table 2.

α	Marginal		Group-conditional	
	Realised error	Abstention	Abstention	Accepts nothing
0.05	0.264	0.969	0.999	99 %
0.10	0.351	0.858	0.994	95 %
0.20	0.465	0.634	0.954	59 %
0.30	0.664	0.202	0.880	31 %
0.40	0.711	0.001	0.760	9 %

Weighted conformal sits between the two: it tightens thresholds under shift and reduces violation rates relative to marginal, but it applies one global threshold and so cannot deliver per-bin validity.

An important asymmetry: Mondrian *relaxes* the threshold where the signal works well. In the 90–100 bin its abstention is 0.035 against marginal’s 0.442. The cost of conditional validity is per-bin, not aggregate, and pooling can even reduce total abstention while destroying the guarantee where it matters.

Cost of the guarantee: abstention required to hold α , by shift bin

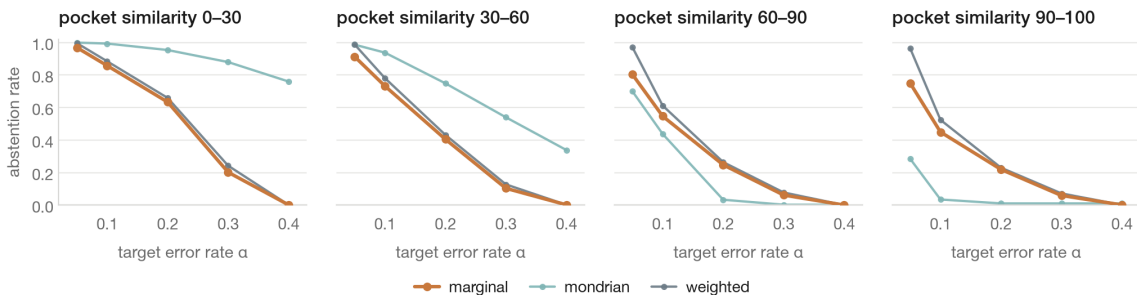


Figure 3: Abstention required to hold each target error rate, by bin and method. The familiar bins reach usable coverage quickly; the unfamiliar bin never does.

4 Headline accuracy is substantially memorisation

Benchmark accuracy computed on the full post-cutoff distribution is carried disproportionately by cases the model has effectively already seen. The ligand-similarity distribution has a hard mass point at Tanimoto 100: the identical ligand, in a different structure, released before the model’s cutoff.

- 27.5 % of post-cutoff systems fall in that exact-match group.

- Accuracy there averages 0.841, against 0.615 on everything else.
- They supply 34.3 % of all correct predictions.
- Excluding them and the next decile of ligand similarity, headline accuracy falls from 0.677 to 0.599.

Protenix depends on exact matches most (0.103 accuracy lost), Boltz-2 least (0.047), though Boltz-2 also has the smallest post-cutoff sample by a wide margin.

At matched abstention within the least-similar bin, AlphaFold 3 leads while little is discarded: at zero abstention accuracy spans 0.228 to 0.365, with AlphaFold 3 highest. The ordering does not survive higher abstention. Rejecting 75 % of cases puts Boltz-1x ahead at 0.864 against AlphaFold 3’s 0.614, but on 22 to 55 accepted poses per model, so the bin is too small to rank models there. The expensive models do not collapse on unfamiliar targets relative to each other. We cannot answer the more interesting question of whether they still beat cheap physics-based docking there: no docking baseline is included in this release, and RoseTTAFold All-Atom, the only non-diffusion method present, ships a single deterministic pose with no confidence score and therefore cannot be placed on an abstention curve at all.

5 Limitations

Signal coverage. The release ships scored CSVs, not predicted structures, so only ranking scores and chain-pair ipTM variants could be tested. pLDDT, PAE and pTM, which practitioners also cite, are absent. If per-atom confidences rank correctness materially better than ipTM, the quantitative picture would improve, though the conditional-coverage mechanism would not change.

RMSD verification. We re-implemented the release’s binding-site-superposed symmetry-corrected RMSD independently and verified it against the released values on 28 poses sampled across the accuracy range, over-weighting the 1 Å to 3 Å band where a disagreement could flip the correctness label. 27 of 28 agree to within 1×10^{-4} Å, all 17 poses in the 1 Å to 3 Å band agree exactly, and no label flips across the 2 Å threshold. The single disagreement is a system with four copies of the same ligand where our chain-mapping search selects a different copy than the release did; it is a copy-assignment difference rather than a geometry difference.

Reaching that agreement required searching over protein chain mappings rather than assigning chains by sequence identity. In homomeric complexes the copies are 100 % identical, so a greedy assignment superposes the binding site onto the wrong monomer, producing an excellent backbone fit and a ligand RMSD wrong by tens of Å. Any reimplementations of this metric needs the mapping search; we note it because the failure is silent and looks like a successful superposition.

Physical validity. PoseBusters checks exist for one pose per ligand and none for Boltz-2, so the secondary label is partial.

Bin sparsity. The joint ligand-by-pocket grid has 6 of 15 occupied cells below 100 systems for every model, which is why conditional calibration groups on one axis. The two axes are correlated but not redundant, and the ligand-axis result (Section 2.4) is reported as a robustness check rather than a second independent finding.

Calibration set size. Mondrian’s near-total abstention in the least-similar bin is a joint property of the signal and the amount of calibration data available there (110–141 poses per model per half-split, and only 40 for Boltz-2). A larger benchmark would tighten the finite-sample correction and could make some of those bins certifiable. What would not change is the marginal method’s violation, which is a property of pooling and not of sample size.

6 Code and data availability

We release the master table (568 220 rows, one per ligand–model–seed–sample), the calibration thresholds for seven models at five target error rates, the figures, and every script that regenerates them, at <https://github.com/fonsResearch/fons-bench>.

The calibrated thresholds are a small JSON artifact in the repository, so scoring a new set of predictions needs neither a GPU nor the benchmark data:

```
uv run fons-bench evaluate
  --predictions my_preds.csv --model af3 --alpha 0.1
```

Given a confidence signal and, optionally, the pocket similarity of each target to the nearest structure predating the model’s training cutoff, it applies both the marginal and the group-conditional rule and reports which bins it cannot certify rather than silently accepting in them. The master table is a build product rather than a stored file: one command reassembles it byte-identically from the source release, whose checksums are recorded in the repository.

The benchmark is designed to be re-run against future models. The shift coordinate is recomputed per model from its own training cutoff, so a model released with a later cutoff is scored on its own terms rather than against a fixed date that would understate its exposure.

7 What this means

The practical recommendation is not “use group-conditional calibration”. It is that a confidence threshold quoted without reference to the similarity between the target and the model’s training window is not a guarantee about anything. Where similarity is high, the signals work and calibration is easy. Where it is low, which is where a structure prediction is actually load-bearing, no threshold on the currently available signals certifies a useful error rate, and a calibration method that says so is behaving correctly. Reporting a single benchmark accuracy or a single calibrated threshold across a mixed-similarity population conceals this completely.

References

- [1] Runs N’ Poses: have protein–ligand co-folding methods moved beyond memorisation? *bioRxiv* 2025.02.03.636309. Data: Zenodo record 18366081, <https://doi.org/10.5281/zenodo.14794785>.
- [2] M. Buttenschön, G. M. Morris, C. M. Deane. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 2024.
- [3] R. J. Tibshirani, R. Foygel Barber, E. Candès, A. Ramdas. Conformal prediction under covariate shift. *NeurIPS*, 2019.

- [4] J. Wohlwend et al. Boltz-1: democratizing biomolecular interaction modeling. *bioRxiv* 2024.11.19.624167, 2024.
- [5] Boltz developers. Boltz repository and prediction documentation. <https://github.com/jwohlwend/boltz>. Boltz-1x applies inference-time steering potentials to the Boltz-1 weights.
- [6] S. Passaro et al. Boltz-2: towards accurate and efficient binding affinity prediction. *bioRxiv* 2025.06.14.659707, 2025.
- [7] Chai Discovery. Chai-1: decoding the molecular interactions of life. *bioRxiv* 2024.10.10.615955, 2024.
- [8] ByteDance Protenix Team. Protenix: toward high-accuracy open-source biomolecular structure prediction. <https://github.com/bytedance/Protenix>.
- [9] R. Krishna et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* 384:ead12528, 2024.
- [10] H. Papadopoulos, K. Proedrou, V. Vovk, A. Gammerman. Inductive confidence machines for regression. *ECML*, 2002.
- [11] V. Vovk. Conditional validity of inductive conformal predictors. *ACML*, 2012.
- [12] V. Vovk, A. Gammerman, G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.
- [13] J. Abramson et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 630:493–500, 2024.