Evaluating Uncertainty Quantification in Co-folding Models under Distribution Shift

Jaswin Hargun, Amogh Joshi, Manav Ramprasad, Walter Virany

Introduction and Background

Research Question & Significance

We aim to investigate how predicted Local Distance Difference Test (pLDDT), the standard confidence score reported by co-folding models, degrades as target complexes deviate from the training distribution. This question has direct practical relevance to drug discovery. In early-stage screening, practitioners rely on structure prediction models to rapidly screen candidate molecules *in silico* before committing to expensive wet lab experiments. However, these models have several failure modes, such as out-of-distribution (OOD) generalization. We want to assess whether pLDDT is a well-calibrated confidence score that allows practitioners to identify unreliable predictions under distribution shift.

Context & Datasets

Foundation models for modeling biomolecular interactions—including AlphaFold3, Boltz-2, Chai-1, and Protenix—have achieved remarkable progress in predicting the structure of biomolecular complexes [1-4]. These models have been rapidly adopted in therapeutic development, with Boltz-2 reportedly being used at all 20 of the largest pharmaceutical companies worldwide [5]. Despite their success, two primary failure modes have been identified for these models: poor predictions of intrinsically disordered regions (IDRs), and degraded performance on OOD targets. For IDRs, both prediction error and pLDDT calibration have been studied, showing that low pLDDT correlates with structural error [6, 7, 8]. For OOD generalization, Masters et al. (2025) demonstrate that structural *accuracy* degrades substantially for systems dissimilar to training examples [9], but do not assess pLDDT calibration in this regime. Shihab et al. (2026) show that pLDDT calibration degrades under distribution shift in terms of experimental method, but do not quantify the distribution shift in terms of sequence similarity to the training set [10]. Rayka & Naghavi evaluate classical uncertainty quantification (UQ) methods for binding affinity prediction and find that calibration degrades substantially under distribution shift, with out-of-distribution test sets showing degraded performance across all methods [11]. However, their analysis is restricted to small models for binding affinity prediction and does not examine pLDDT, nor do they quantify the distribution shift of the training set similarity for each target. To our knowledge, no study has characterized how pLDDT calibration degrades with relationship to training set similarity—this is the problem we aim to address.

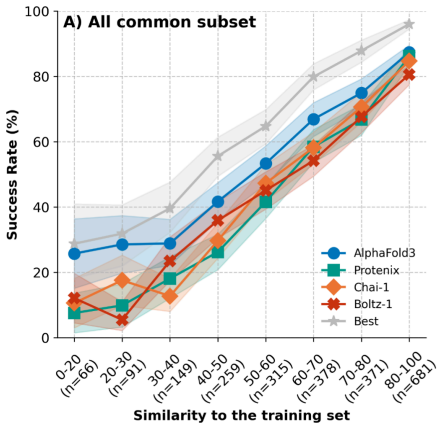

Figure 1. From the Runs-N-Poses paper, 2025. Describes success rate based on passing error metric thresholds under distribution shift.

We use Protenix as our primary model due to its fully open-sourced weights and training data [4], enabling explicit measurement of sequence and ligand similarity. Evaluation is performed on the Runs-N-Poses benchmark [12], an established protein-ligand co-folding evaluation set comprising more than 2,000 complexes deposited after September 30, 2021—the same training cutoff as Protenix.

Methods and Setup

Methods & Preprocessing

Our project investigates the reliability of structural UQ in generative protein-ligand folding models, systematically evaluating how calibration degrades under distribution shift. Foundation models for structure prediction currently rely on self-reported confidence metrics, such as the pLDDT. However, it remains unclear how well these internal confidence estimates correlate with actual physical prediction error when these models

are forced to extrapolate outside their training manifold. To address this, we treat structural UQ evaluation as a calibration problem and we have executed our primary inference pipeline using the Protenix model. For each inference run, we parse the generated artifacts to extract localized confidence scores, primarily focusing on metrics like pLDDT, which will serve as our primary scalar uncertainty metric. Concurrently, we compute the true structural prediction error by calculating the true Local Distance Difference Test (LDDT), which is what pLDDT is trying to predict, between the predicted complex coordinates and the experimental ground-truth structures. To formally evaluate calibration, we compute Expected Calibration Error (ECE) between the true LDDT and pLDDT. This metric is computed by partitioning predictions into bins based on pLDDT value and finding the average difference between mean pLDDT and observed LDDT for each bin. Our data processing pipeline is designed to quantify exact distribution shifts and rigorously prevent data leakage. Because Protenix was trained on PDB structures with a cutoff of September 30, 2021, we isolate a temporal holdout set of complexes deposited strictly after this date. Since official PDBBind general releases largely predate this cutoff, we curate our post-2021 temporal split by cross-referencing recent PDB releases with known binding data or utilizing updated, chronologically separated benchmarks like the PoseBusters or Runs-N-Poses test sets.

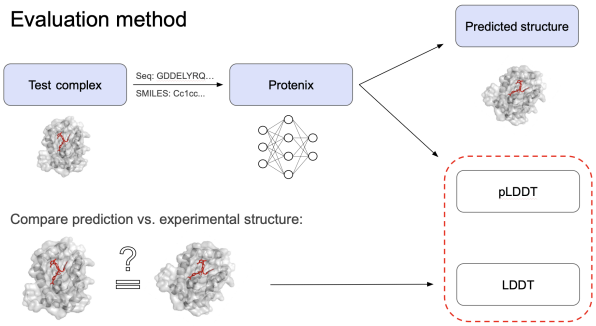

Figure 2. The proposed information flow. Each complex within the validation set is compared with the entire training set, resulting in similarity scores.

To mathematically formalize the "distance" from the training distribution, our preprocessing pipeline computes two specific metrics for every test complex: the maximum ligand SuCOS similarity for the test ligand against the pre-2021 training ligands [13], and the maximum global sequence identity for the target protein against the training set. Given the immense size of the pre-2021 training database, we utilize sequence alignment algorithms and standard cheminformatics libraries to manage the computational scale of determining these maximum similarities.

Preprocessing our primary evaluation complexes involves stripping all existing crystallographic ligands from the coordinate files and converting the inputs strictly into FASTA sequences and isomeric SMILES strings. Utilizing isomeric SMILES is critical to preserve stereochemistry, ensuring we evaluate the exact chiral constraints of the ligand *ab initio*. Furthermore, we establish an explicit parsing policy to handle non-target molecular entities present in the raw coordinate files, ensuring they are appropriately retained or masked during input generation to prevent unintended spatial leakage.

Evaluation Metrics

The primary metrics in our evaluation can be grouped into 2 categories: uncertainty metrics and error metrics. Within the uncertainty categorization, our selected metric is pLDDT; for the error categorization, our selected metric is per-residue LDDT for protein structures. This pairing is particularly appropriate because it allows us to ask whether the model's internal confidence estimate coincides with realized structural accuracy for the predicted ligand pose.

Each complex within the Runs-N-Poses evaluation dataset is evaluated using Protenix to generate a predicted protein-ligand structure, along with the associated uncertainty metrics. The predicted structure is then assessed against the experimentally validated structure to assess the structure prediction error in terms of LDDT. To quantify distribution shift, two similarity scores have been calculated for each complex: the maximum global sequence identity between the target protein and the Protenix training set, and the maximum ligand SuCOS similarity between the target ligand and ligands contained within the training set. This was repeated for all complexes within the evaluation dataset, creating a set of self-contained tuples with the associated uncertainty, error, and similarity metrics.

Our primary analysis aims to evaluate how calibration degrades under distribution shift. To do so, we have computed the ECE score for each predicted complex by relating per-residue pLDDT to corresponding LDDT. This calibration information is then visualized using multiple 2D graphs, each corresponding to a level of similarity. We have chosen buckets of low (30-60%), medium (60-90%) and high (90-100%) protein sequence similarity to the training set. Similarly, complexes were stratified into five bins based on maximum ligand SuCOS similarity to the training set (0-20%, 20-40%, 40-60%, 60-80%, and 80-100%).

Results, Discussion, and Future Directions

Results & Visualizations

We evaluated pLDDT calibration across 2,370 protein-ligand complexes from the Runs-N-Poses benchmark, stratifying by two OOD axes: protein sequence similarity and ligand SuCOS similarity to the Protenix training set. For each system, per-residue pLDDT (normalized to [0,1]) and LDDT were computed and used to calculate Expected Calibration Error (ECE) via 20 equal-width plDDT bins corresponding to intervals of length .05. 95% confidence intervals (CIs) were computed via bootstrapping by sampling 500 systems from the evaluation set with replacement and computing ECE over these subsets.

Figure 3 shows per-residue LDDT vs. pLDDT scatter plots stratified by protein similarity. Interestingly, the majority of residues cluster above a pLDDT score of 0.8, reflecting the model's uniform high confidence. Critically, the most OOD group (30–60% sequence similarity) exhibits a substantial fraction of residues with high pLDDT and low lDDT, indicating overconfidence in the model. This discrepancy narrows as similarity increases, with the other groups showing tighter clusters in regions of high LDDT and pLDDT. Notably, the vast majority of residues have a high LDDT. While this highlights the high accuracy of Protenix on the Runs-N-Poses test set, it also points to a gap in which we could not assess the calibration of pLDDT on more difficult targets.

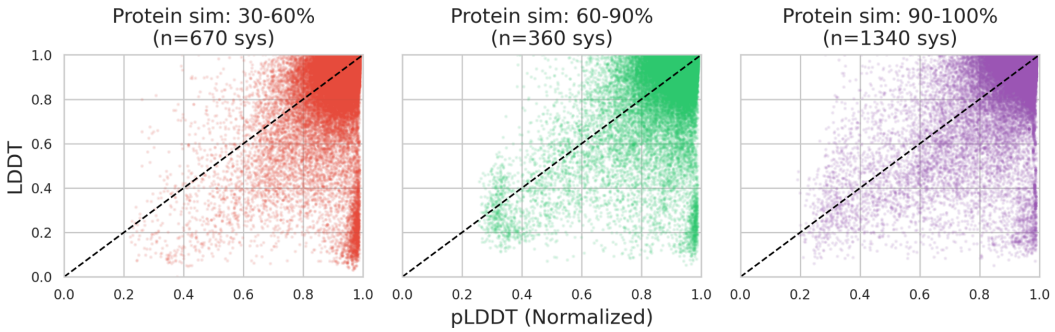

Figure 3. Per-residue LDDT vs. normalized pLDDT for 2,370 protein-ligand complexes stratified by protein sequence similarity to the Protenix training set: 30–60% (left), 60–90% (middle), 90–100% (right). Each panel displays a random sample of 155,955 residues corresponding to the minimum across similarity bins. Dashed diagonal indicates perfect calibration.

Figure 4 shows ECE stratified by protein similarity (left) and ligand similarity (right). Both exhibit a monotonic trend: ECE is highest in the most OOD bin and decreases inversely with similarity to the training set. For protein sequence similarity, ECE drops from 0.015 in the 30–60% bin to 0.003 in the 90–100% bin. A similar trend can be seen for ligand SuCOS similarity, with ECE decreasing from 0.016 in the 0–20% bin to 0.004 in the 80–100% bin. These trends are consistent across bootstrap resamples, though CIs are wide for small n (Table 1).

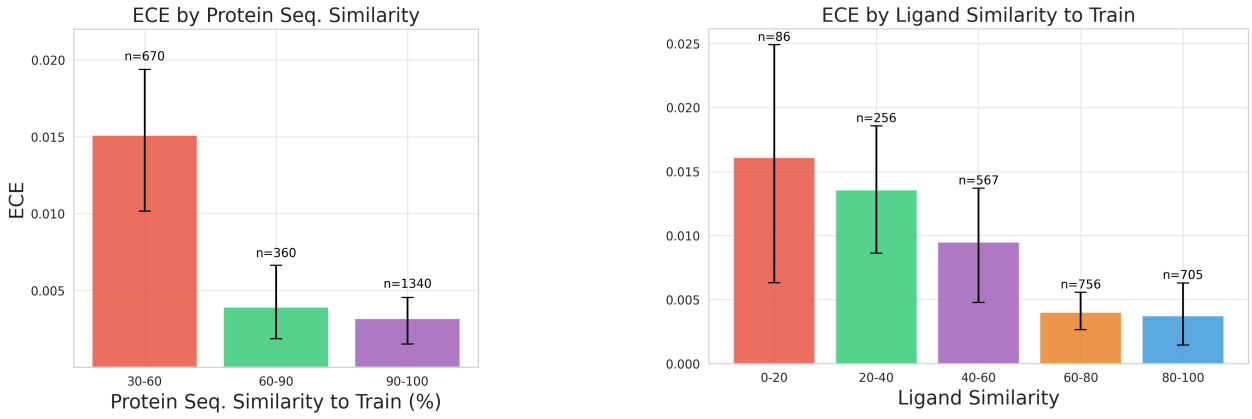

Figure 4. ECE stratified by protein sequence similarity (left) and ligand SuCOS similarity (right) across 2,370 protein-ligand complexes. Error bars denote 95% bootstrap CI obtained by resampling 500 systems.

Protein Seq. Similarity	ECE	Number of Systems	Number of residues
30-60%	0.0151 [0.0105, 0.0189]	670	303, 407
60-90%	0.0039 [0.0022, 0.0067]	360	155,955
90-100%	0.0032 [0.0019, 0.0047]	1340	573,798
Ligand SuCOS Similarity			
0-20%	0.0161 [0.0063, 0.0249]	86	45,843
20-40%	0.0136 [0.0086, 0.0186]	256	113,435
40-60%	0.0095 [0.0048, 0.0137]	567	221,210
60-80%	0.0040 [0.0026, 0.0056]	756	319,278
80-100%	0.0037 [0.0015, 0.0063]	705	333,394

Table 1. ECE with 95% bootstrap CIs stratified by protein sequence similarity and ligand SuCOS similarity to Protenix training set.

Discussion

Our results support the hypothesis that pLDDT calibration in Protenix degrades under distribution shift. ECE exhibits a consistent monotonic increase across both similarity axes as complexes become more dissimilar to the training set. This suggests that practitioners should treat pLDDT scores with skepticism when applying co-folding models to OOD protein targets or ligands precisely the regime most relevant to early-stage drug discovery.

However, the preliminary analysis has a key limitation: we compute LDDT over protein residues only, hence calibration is not evaluated for ligand pose. This is consequential because protein backbone prediction is generally reliable and high-confidence across all OOD bins - as visible in the domination of high-LDDT residues in Figure 3. The OOD signal we observe in ECE is therefore indirect: novel ligands likely induce conformational changes in the binding pocket that the model fails to anticipate, while still assigning high pLDDT to the surrounding protein backbone.

Next steps

We highlight three directions that would meaningfully extend this work. First, our preliminary analysis identified systems with high ECE but 100% protein sequence similarity. These outliers suggest that sequence similarity alone is insufficient as an OOD proxy, and that ligand-induced conformational changes may be the dominant source of miscalibration. Structural comparison of predicted vs. experimental binding pockets in these cases could yield actionable mechanistic insight.

Second, we aim to extend the per-residue pipeline to per-atom ligand analysis. Since our current ECE is computed over protein backbone residues, the calibration of pLDDT on ligand atoms remains uncharacterized. Using RDKit graph isomorphism for atom correspondence and biotite's LDDT implementation, we can construct a direct measure of ligand pose calibration stratified by the same OOD axes.

Finally, this work focuses on confidence calibration for structural predictions. Recent co-folding models also predict binding affinity, which depends critically on the accuracy of the predicted structure. It would be natural to ask whether structural confidence - pLDDT - can serve as a proxy for uncertainty in binding affinity predictions.

References

[1] Josh Abramson et al. "Accurate structure prediction of biomolecular interactions with AlphaFold 3". In: *Nature* (2024). DOI: 10.1038/s41586-024-07487-w.

[2] Saro Passaro et al. *Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction*. 2025. DOI: 10.1101/2025.06.14.659707.

[3] Jacques Boitreaud et al. *Chai-1: Decoding the molecular interactions of life*. en. 2024. DOI: 10.1101/2024.10.10.615955. (Visited on 02/23/2026).

[4] Yuxuan Zhang et al. *Protenix-v1: Toward High-Accuracy Open-Source Biomolecular Structure Prediction*. 2026. DOI: 10.64898/2026.02.05.703733.

[5] Saro Passaro, Gabriele Corso, and Jeremy Wohlwend. *Boltz-2 — Towards Accurate and Efficient Binding Affinity Prediction; MIT Jameel Clinic — jclinic.mit.edu*. 2025.

[6] Mehmet Akdel et al. "A structural biology community assessment of AlphaFold2 applications". In: *Nature Structural & Molecular Biology* (2022). DOI: 10.1038/s41594-022-00849-w.

[7] Kathryn Tunyasuvunakool et al. "Highly accurate protein structure prediction for the human proteome". In: *Nature* (2021). DOI: 10.1038/s41586-021-03828-1.

[8] Jessica L. Binder et al. "AlphaFold Illuminates Half of the Dark Human Proteins". In: *Current opinion in structural biology* (2022). DOI: 10.1016/j.sbi.2022.102372. (Visited on 03/10/2026).

[9] Matthew R. Masters, Amr H. Mahmoud, and Markus A. Lill. "Investigating whether deep learning models for co-folding learn the physics of protein-ligand interactions". In: *Nature Communications* (2025). DOI: 10.1038/s41467-025-63947-5.

[10] Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. *CalPro: Prior-Aware Evidential–Conformal Prediction with Structure-Aware Guarantees for Protein Structures*. 2026. DOI: 10.48550/arXiv.2601.07201.

[11] Milad Rayka and S. Shahab Naghavi. "Uncertainty quantification enables reliable deep learning for protein–ligand binding affinity prediction". In: *Scientific Reports* (2025). DOI: 10.1038/s41598-025-27167-7.

[12] Peter Škrinjar et al. *Have protein-ligand co-folding methods moved beyond memorisation?* 2025. DOI: 10.1101/2025.02.03.636309.

[13] Susan Leung et al. *SuCOS is Better than RMSD for Evaluating Fragment Elaboration and Docking Poses*. May 2019. DOI: 10.26434/chemrxiv.8100203.