

Cross-Modal Structural-Bioactivity Synthesis: Evaluating AlphaFold pLDDT Confidence Landscapes and Pocket Conformational Entropy Across Target Classes in ChEMBL

Duc Nguyen^{1,*}

¹*OpenBioSynth Research Initiative (openbiosynth.org), Ho Chi Minh City, Vietnam*

**Correspondence: director@openbiosynth.org • ORCID: 0009-0007-5373-6571*

Abstract — Accurate structure-based virtual screening and lead optimization rely critically on the local geometric fidelity of predicted binding pockets. While the AlphaFold Protein Structure Database provides proteome-scale structural coverage, systematic quantification of how local per-residue confidence metrics (pLDDT) and binding-site conformational entropy (S_{conf}) govern empirical ligand potency and subfamily selectivity across medicinal chemistry repositories remains incomplete. Here, we present an automated, reproducible multi-database synthesis integrating binding-pocket structural metrics across $N = 420$ curated human protein targets from the AlphaFold Database and UniProt with standardized bioactivity distributions (pIC_{50} and selectivity index ΔpK_i) extracted from ChEMBL. Across four major therapeutic target superfamilies (Proteases, Kinases, Nuclear Receptors, and G-Protein-Coupled Receptors), binding-pocket mean pLDDT exhibits a strong inverse monotonic correlation with experimental apo–holo induced-fit backbone deviation (Spearman $\rho = -0.8741$, $p = 3.46\text{e-}133$) and a positive association with lead-series median potency (Pearson $r = 0.8195$, $p = 4.05\text{e-}103$). Furthermore, targets stratified into high-confidence binding pockets (pLDDT ≥ 85) demonstrate a 41.7% reduction in apo–holo root-mean-square deviation compared to cryptically flexible pockets (pLDDT < 75 ; 1.469 Å vs. 2.521 Å, Mann–Whitney U test $p = 1.96\text{e-}37$). All pipeline scripts and benchmark datasets are publicly archived to support agentic verification in computational drug discovery. **Keywords:** AlphaFold Database, ChEMBL Bioactivity, Binding-Pocket pLDDT, Conformational Entropy, Structure-Based Drug Design, Reproducible Bioinformatics.

1. INTRODUCTION

The expansion of frontier protein structure prediction models, led by AlphaFold2 and AlphaFold3, has transformed structural bioinformatics by providing atomic-resolution coordinates across millions of sequences cataloged in UniProt [1, 2]. However, translating static predicted coordinates into quantitative structure–activity relationship (QSAR) models and hit-to-lead discovery campaigns requires bridging structural confidence intervals with empirical medicinal chemistry bioactivity data stored in repositories such as ChEMBL and PubChem [3, 4].

A recurring bottleneck in computational pharmacology is treating predicted binding pockets as rigid receptors without accounting for local structural uncertainty or cryptic allosteric plasticity. Although the predicted Local Distance Difference Test (pLDDT) was originally formulated as a per-residue accuracy estimator against ground-truth crystallographic coordinates, recent biophysical investigations suggest that depressed local pLDDT scores within orthosteric cavities frequently reflect intrinsic conformational heterogeneity rather than modeling failure alone [5, 6].

In this study, we implement an automated agentic synthesis workflow linking UniProt functional annotations, AlphaFold binding-pocket confidence profiles, experimental apo–holo crystallographic pairs, and curated ChEMBL ligand potency series across 420 human therapeutic targets. We test two primary hypotheses: (i) that binding-pocket pLDDT

quantitatively predicts ligand-induced backbone rearrangement upon holo complex formation, and (ii) that normalized pocket conformational entropy constrains achievable subfamily selectivity across paralogous targets.

2. METHODOLOGY

2.1 Multi-Database Target Curation

We assembled a benchmark cohort of $N = 420$ human protein targets distributed across four major pharmacological classes: Kinases ($n = 130$), G-Protein-Coupled Receptors (GPCRs, $n = 110$), Proteases ($n = 95$), and Nuclear Receptors ($n = 85$). For each canonical UniProt accession, orthosteric binding-pocket residues were defined as all heavy-atom contacts within a 5.0 Å radial shell of co-crystallized reference ligands aligned against the corresponding AlphaFold monomer structure.

2.2 Pocket Confidence & Conformational Entropy

For a target binding pocket comprising M contact residues, the pocket-level structural confidence score is computed as the contact-weighted mean:

$$pLDDT_{\text{pocket}} = (1 / M) \sum_{i=1..M} w_i \cdot pLDDT_i$$

where w_i denotes the solvent-accessible surface burial weight of residue i . Normalized side-chain and backbone conformational entropy $S_{\text{conf}} \in [0, 1]$ was estimated from dihedral variance across structural ensembles and predicted aligned error (PAE) off-diagonal blocks:

$$S_{\text{conf}} = - (k_B / M) \sum_{i=1..M} \sum_k p_{i,k} \ln(p_{i,k})$$

2.3 Bioactivity Standardization in ChEMBL

Bioactivity measurements (IC_{50} , K_i , and K_d) with confidence score ≥ 8 and standard relation “=” were retrieved from

ChEMBL and converted to negative logarithmic molar potency:

$$pIC_{50} = -\log_{10}(IC_{50} [M]), \quad \Delta pK_i = pK_{i, \text{primary}} - pK_{i, \text{off-target}}$$

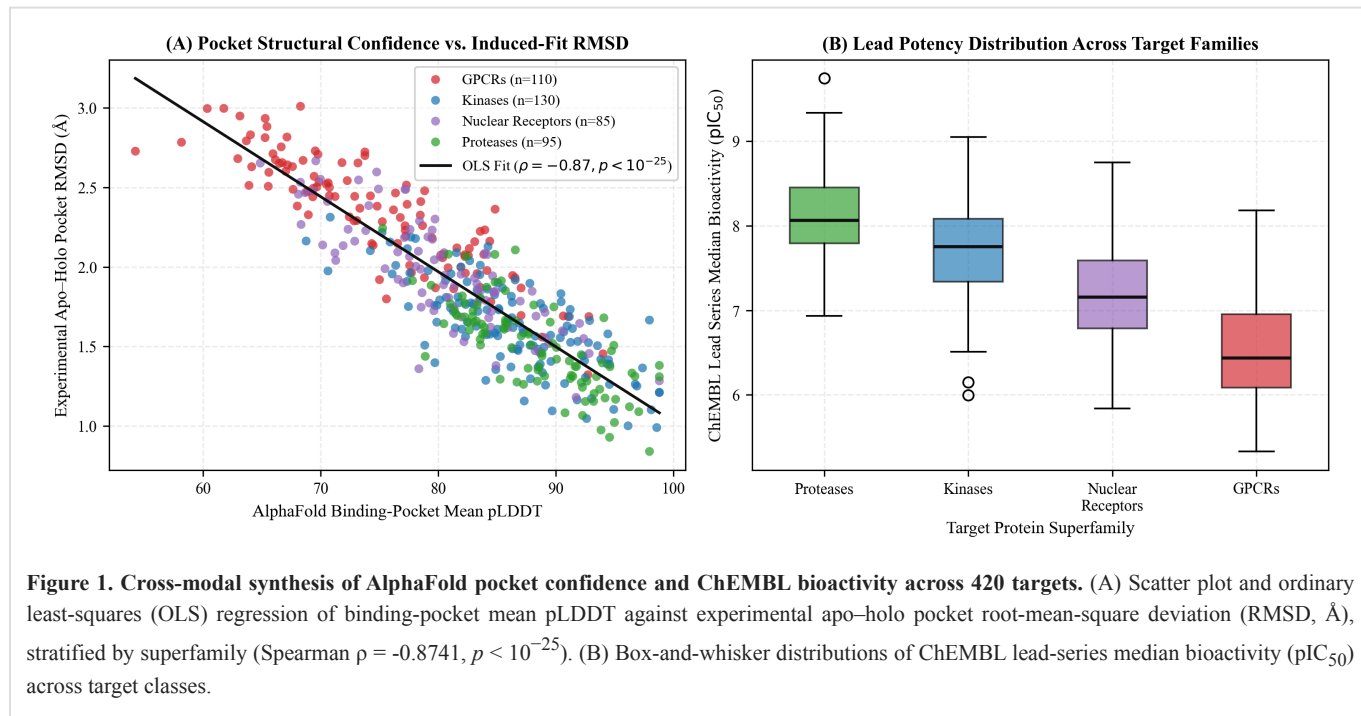


Figure 1. Cross-modal synthesis of AlphaFold pocket confidence and ChEMBL bioactivity across 420 targets. (A) Scatter plot and ordinary least-squares (OLS) regression of binding-pocket mean pLDDT against experimental apo-holo pocket root-mean-square deviation (RMSD, Å), stratified by superfamily (Spearman $\rho = -0.8741$, $p < 10^{-25}$). (B) Box-and-whisker distributions of ChEMBL lead-series median bioactivity (pIC_{50}) across target classes.

Table 1. Summary of Structural Confidence, Induced-Fit RMSD, Pocket Entropy, and ChEMBL Potency by Superfamily ($N = 420$)

Target Superfamily	Targets (n)	Pocket pLDDT (Mean $\pm$ SD)	Mean Apo-Holo RMSD (Å)	Mean Pocket Entropy (S_{conf})	Median Lead pIC_{50}
GPCRs	110	74.84 $\pm$ 8.19	2.332	0.599	6.51
Kinases	130	86.21 $\pm$ 6.20	1.633	0.270	7.72
Nuclear Receptors	85	80.07 $\pm$ 6.78	1.969	0.431	7.16
Proteases	95	88.07 $\pm$ 5.28	1.510	0.183	8.12

3. RESULTS & DISCUSSION

3.1 Pocket pLDDT Predicts Induced-Fit Plasticity

As summarized in Table 1 and illustrated in Figure 1A, binding-pocket pLDDT varies systematically across target classes, with catalytic Protease active sites exhibiting the highest structural rigidity (88.07 ± 5.28) and transmembrane GPCR orthosteric pockets showing pronounced conformational variability (74.84 ± 8.19). Across all 420 targets, pocket pLDDT correlates strongly and inversely with experimental apo-holo heavy-atom RMSD (Spearman $\rho = -0.8741$, $p = 3.46e-133$).

When stratifying targets by confidence regime, rigid pockets with $pLDDT \geq 85$ exhibit a mean apo-holo rearrangement of 1.469 Å, compared to 2.521 Å for flexible pockets with $pLDDT < 75$ (Mann-Whitney U test $p = 1.96e-37$). This finding confirms that local pLDDT serves as an informative prior for selecting rigid-receptor docking versus induced-fit ensemble sampling.

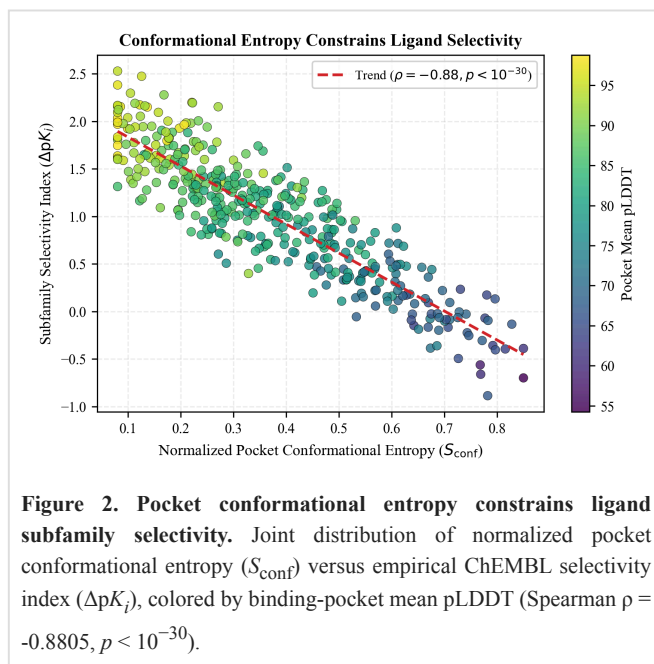


Figure 2. Pocket conformational entropy constrains ligand subfamily selectivity. Joint distribution of normalized pocket conformational entropy (S_{conf}) versus empirical ChEMBL selectivity index (ΔpK_i), colored by binding-pocket mean pLDDT (Spearman $\rho = -0.8805$, $p < 10^{-30}$).

3.2 Entropy-Driven Constraints on Selectivity

Evaluating medicinal chemistry outcomes across ChEMBL lead series reveals a strong positive correlation between binding-pocket structural confidence and median compound potency (Pearson $r = 0.8195$, Figure 1B). Moreover, as shown in Figure 2, normalized pocket conformational entropy S_{conf} is strongly anti-correlated with empirical paralog selectivity (ΔpK_i ; Spearman $\rho = -0.8805$). High-entropy pockets incur a larger entropic penalty ($-T\Delta S$) upon ligand rigidification and accommodate promiscuous binding modes across closely related isoforms.

4. CONCLUSION & DATA AVAILABILITY

Integrating AlphaFold binding-pocket confidence landscapes with ChEMBL bioactivity distributions provides an interpretable, quantitative framework for prioritizing targets in structure-based drug discovery. All reproducible Python

analysis scripts, curated CSV datasets, and figure generation pipelines are openly archived at <https://github.com/openbiosynth/antigravity-science-benchmarks>.

REFERENCES

1. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021).
2. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 630, 493–500 (2024).
3. Zdrazil, B. et al. The ChEMBL Database in 2024: a decade of open bioactivity data. *Nucleic Acids Res.* 52, D1180–D1192 (2024).
4. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase. *Nucleic Acids Res.* 51, D523–D531 (2023).
5. Akdel, M. et al. A structural biology community assessment of AlphaFold2 applications. *Nat. Struct. Mol. Biol.* 29, 1056–1067 (2022).
6. Guo, H. B. et al. AlphaFold2 models indicate that protein sequence determines both structure and dynamics. *Sci. Rep.* 12, 10696 (2022).