SAIR: Enabling Deep Learning for Protein-Ligand Interactions with a Synthetic Structural Dataset

Pablo Lemos[1] * **Zane Beckwith**[1] * **Sasaank Bandi**[1] **Jordan Crivelli-Decker**[1]

Benjamin J. Shields[1] **Thomas Merth**[1] **Punit K Jha**[1] **Nicola De Mitri**[1]

Tiffany J. Callahan[1] **Romelia Salomon-Ferrer**[1] **Martin Ganahl**[1]

SandboxAQ[1]

Abstract

Predicting protein-ligand binding affinities is crucial for drug discovery but is limited by the scarcity of high-quality 3D structures with measured activity. We present the **Structurally Augmented IC50 Repository (SAIR)**, the largest public dataset of protein-ligand 3D structures with activity data, containing $5,244,285$ **million structures across** $1,048,857$ **protein-ligand systems** from ChEMBL and BindingDB, computationally folded using Boltz-1x. The PoseBusters algorithm shows that approximately 97% of structures are physically valid. Benchmarking traditional scoring functions (Vina, Vinardo) and machine learning models (OnionNet-2, AEV-PLIG) indicates that ML models outperform classical methods but still correlate poorly with true affinities, emphasizing the need for models adapted to synthetic structures. SAIR provides a foundation for developing next-generation binding-affinity prediction methods.

1 Introduction

Understanding protein-ligand interactions is critical for drug discovery, as binding affinity determines both on-target efficacy and off-target effects. While binding affinity is fundamentally encoded in the 3D structure of the protein-ligand complex, experimental determination of these structures is limited in coverage and throughput, and computational methods face trade-offs between accuracy and cost (Wang et al., 2019; Cournia et al., 2017; Crivelli-Decker et al., 2024; York, 2023). Deep learning approaches using 3D structural inputs are generally more accurate than sequence-based models (Öztürk et al., 2019; Jiang et al., 2022; Limbu & Dakshanamurthy, 2022; Jiménez et al., 2018; Zheng et al., 2019; Wang et al., 2021; Son & Kim, 2021) but are constrained by the limited availability of high-quality structures with measured affinities (Askr et al., 2023; Libouban et al., 2023; Wang, 2024; Zeng et al., 2024). Existing datasets, such as PLINDER (Durairaj et al., 2024) and CrossDocked (Francoeur et al., 2020), either lack complete affinity annotations or sufficient coverage across protein and ligand space.

To address this, we introduce the **Structurally Augmented IC50 Repository (SAIR)**, the largest public dataset of protein-ligand 3D structures with annotated potencies, containing **1,048,857 complexes**

*{pablo.lemos,zane.beckwith}@sandboxaq.com

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

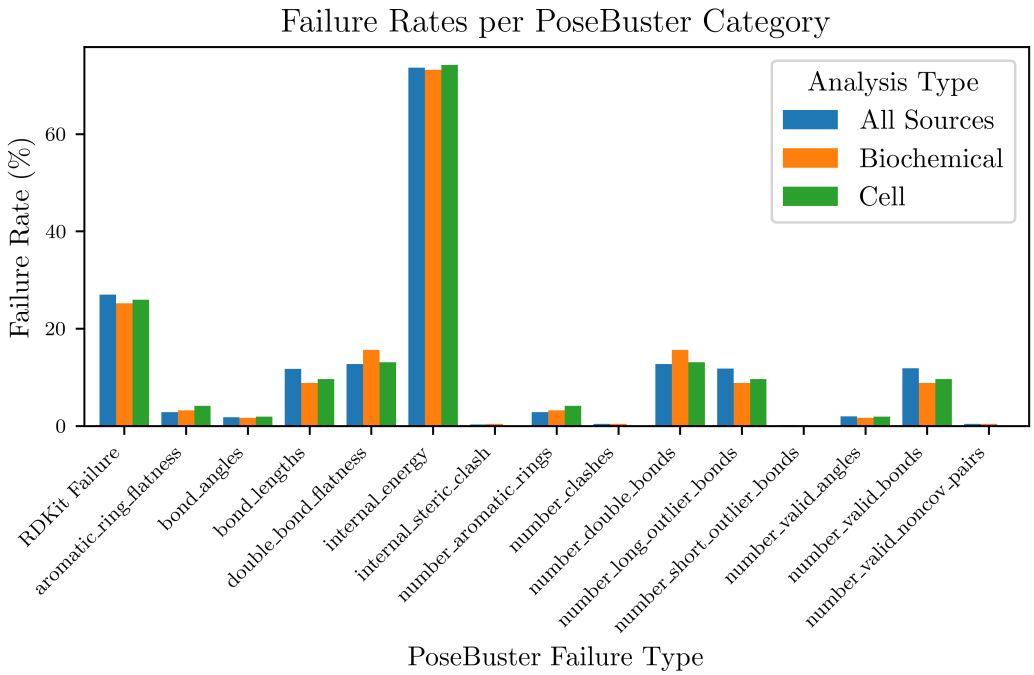

Figure 1: Failure rate for each PoseBusters check, defined as the number of structures failing a given check divided by the total number of failed structures. Bar colors indicate assay type. The first entry in the x-axis corresponds to cases where RDKit failed to load the ligand.

derived from ChEMBL (Gaulton et al., 2012; Zdrazil et al., 2024) and BindingDB (Liu et al., 2007, 2025), and folded using the Boltz-1x model (Wohlwend et al., 2024). SAIR provides a foundation for training and benchmarking large-scale machine learning models for binding affinity prediction. The dataset is publicly available at `https://www.sandboxaq.com/sair`.

2 Dataset Construction

2.1 Dataset Curation

Bioactivity data were obtained from the ChEMBL35 release (Gaulton et al., 2012) and BindingDB (1Q2025) (Liu et al., 2007), and subsequently curated using a minimal set of filters designed to retain a large volume of high-quality data. The specific filters are described in Appendix §A.1.

This results in $1,048,857$ complexes, with $936,702$ from ChEMBL and $613,597$ from BindingDB (see Table 1). Note the number of complexes from each source adds up to more than the total number, because of complexes that appear in both sources. For structure generation, duplicated complexes were only folded once. The distribution of pIC50 values is shown in Fig. 3. Further details about the data curation can be found in Appendix §A.1.

We used the Boltz-1x folding model to generate 3D structures for all protein-ligand complexes. Details about this can be find in Appendix §A.2.

3 Results

3.1 PoseBusters

We evaluated all generated protein-ligand structures using PoseBusters (Buttenschoen et al., 2024). Results are summarized in Appendix §B.4. Overall, Boltz-1x performs well in generating physically valid structures, with only approximately 3% of structures failing any PoseBusters check. This number is consistent with the performance reported in Wohlwend et al. (2024). Notably, only 0.53%

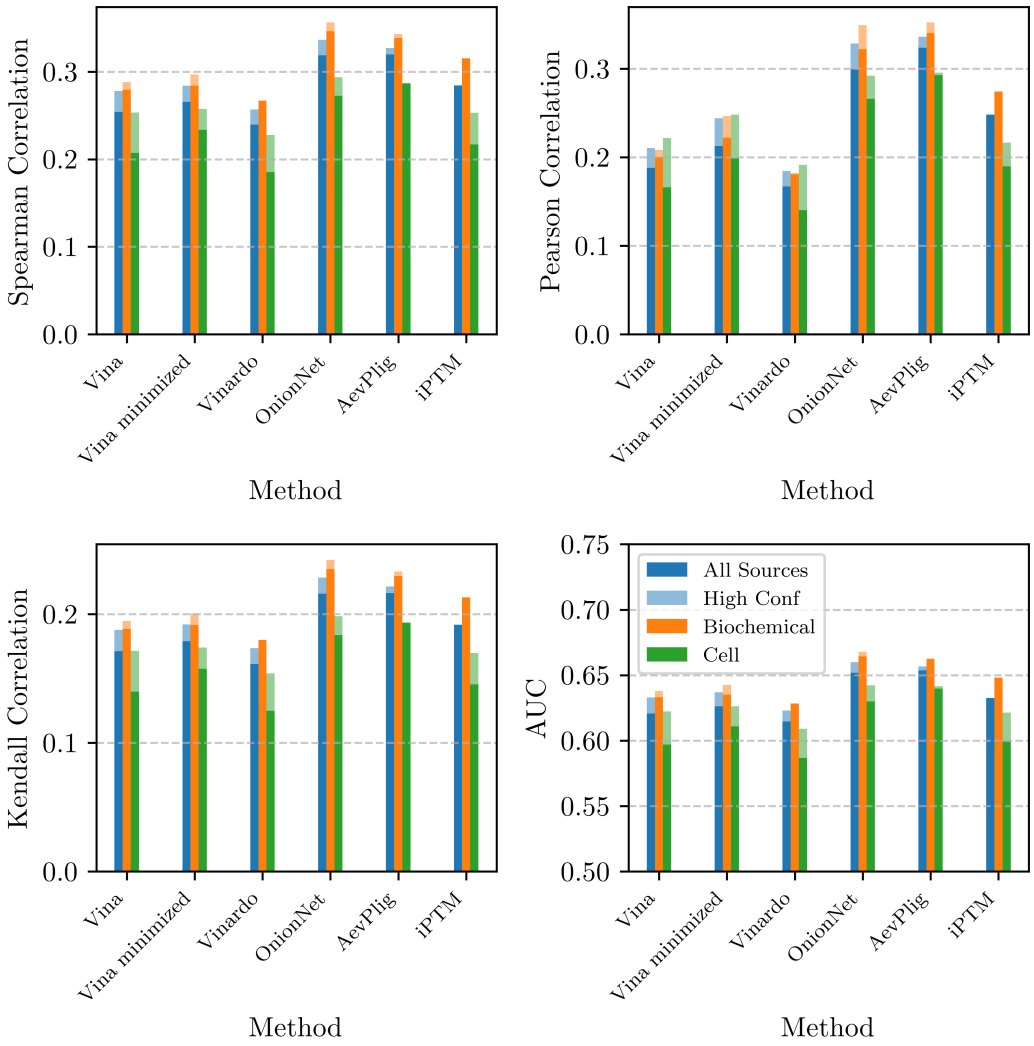

Figure 2: The metrics we used to compare predicted and experimental binding affinity, for each of the binding affinity prediction methods and assays. The shaded lines show results when using only the structures for which Boltz-1x predicted a high confidence.

of protein-ligand complexes had *all* five generated structures fail. Focusing on different protein families, we find that the model has lower failure rates for kinases, and phosphatases; and fails more often for GPCRs. One interesting case is nuclear receptors, where the overall failure rate is low, but there is a relatively large number of complexes, for which all structures failed (more than for any other family) indicating that some of the nuclear receptor complexes are particularly hard to fold.

For the assemblies that failed PoseBusters validation, we present a more fine-grained analysis of individual test outcomes in Fig. 1. Across all assay types[2], the most frequent source of failure is the internal energy check, which accounts for more than half of all PoseBusters failures. Other common failure modes include the number of bonds, abnormal bond lengths, and ligands that could not be loaded by `RDKit`.

3.2 Binding Affinity Models

We use the **SAIR** dataset to benchmark the performance of several binding affinity prediction models. The combination of cofolding-generated structures with structure-based affinity prediction represents

[2]We do not show results for the homogenate assay, as there are not enough structures, compared with the rest.

a promising and increasingly adopted approach in the scientific community. By providing high-throughput structural models paired with experimentally observed IC50 values, the **SAIR** dataset enables rigorous evaluation of this emerging class of predictive methods.

The field of protein–ligand binding affinity prediction is supported by a vast and diverse body of literature, and a comprehensive comparison of all available methods is beyond the scope of this work. Instead, we focus on three representative and methodologically distinct approaches to binding affinity prediction:

Empirical scoring functions: We employ two different traditional empirical scoring functions: AutoDock Vina (henceforth referred to as Vina) (Trott & Olson, 2010) and Vinardo (Quiroga & Villarreal, 2016), both calculated using the `GNINA` library (McNutt et al., 2021).

- We additionally evaluate a Vina-minimized setting, wherein the ligand pose is first optimized with the Vina potential and the minimized complex is subsequently re-scored by Vina.").
- **Convolutional neural network** (CNN): As a first method of structure-based machine learning affinity prediction, we employ a three-dimensional CNN, and represent the input by projecting it into 3D voxels. There are various available 3D CNN methods for affinity prediction, but we use Onionnet-2 (Wang et al., 2021), one of the state-of-the-art methods.
- **Graph neural network** (GNN): As an alternative structure-based machine learning method, we consider a GNN. GNNs are, theoretically, better suited for the task of affinity prediction, as protein-ligand systems are easily represented as graphs. However, regression from graphs is generally a harder task than regression using voxels, as graph convolutions are non-trivial (Zhang et al., 2019). As our GNN, we use the AEV-PLIG model (Warren et al., 2024; Valsson et al., 2025), which recently showed state-of-the-art performance in structure-based binding affinity prediction.

In all cases (with the exception of the "Vina minimized" approach, as explained above), the given affinity prediction tool is evaluated on the predicted three-dimensional protein-ligand structure as-is.

There are many other methods that could be used for benchmarking binding affinity prediction. For example, the recently developed Boltz-2 (Passaro et al., 2025).

We use four metrics to compare the performance of the various binding, described in Appendix §B.3. We present the results of the model comparison in Fig. 2. Across all assay types, the GNN method achieved the highest performance, followed by the CNN method, with the empirical scoring functions performing the worst. However, none of the methods achieved a very high correlation, with Spearman correlations comparable to the ones achieved by some of the interface confidence metrics (even though those were not specifically tuned for binding affinity prediction), such as iPTM, also shown the figure. The shaded bars in the figure show the results when we only keep structures for which Boltz-1x predicts a high confidence (> 0.8). We find that keeping only these structures improves performance of almost all models, as the structures are more likely to be correct.

It is important to note that both the GNN and CNN models were originally trained on experimental structures, and our evaluation is conducted on synthetic structures generated via cofolding. Fine-tuning these models on a subset of the synthetic dataset would likely improve their performance and better align them with the structural distribution seen at inference time.

4 Conclusions

In this work, we introduce the Structurally Augmented IC50 Repository (**SAIR**), a large-scale dataset of protein–ligand 3D structures paired with annotated binding affinities. Comprising $5,244,285$ synthetically-generated structures representing $1,048,857$ protein-ligand complexes, each annotated with experimentally-determined potency, the dataset is designed to significantly expand the volume of data available for training and evaluating structure-based deep learning models in drug discovery.

We rigorously evaluated the quality of the generated structures using PoseBusters and observed a low overall failure rate of approximately 3%. To assess the utility of the dataset for predictive modeling, we benchmarked several structure-based binding affinity prediction methods. GNNs performed best, followed by CNNs and empirical scoring function methods. However, all models achieved only modest correlations, comparable to those obtained from the folding model's interface confidence

metrics. This suggests that models trained on experimental structures may not generalize well to synthetic data, highlighting the potential need for fine-tuning on generated complexes.

SAIR represents a valuable resource for inverse design tasks, enabling generative approaches to create new ligands conditioned on a target protein—further expanding the possibilities for structure-based drug discovery.

References

Embl-ebi: Rest calls based on sifts mappings. https://www.ebi.ac.uk/pdbe/api/doc/sifts.html. Accessed: 2025-06-13.

Rcsb pdb data api. https://data.rcsb.org/#data-api. Accessed: 2025-06-13.

Heba Askr, Enas Elgeldawi, Heba Aboul Ella, Yaseen AMM Elshaier, Mamdouh M Gomaa, and Aboul Ella Hassanien. Deep learning in drug discovery: an integrative review and future challenges. *Artificial Intelligence Review*, 56(7):5975–6037, 2023.

Martin Buttenschoen, Garrett M Morris, and Charlotte M Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024.

Chai Discovery. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, 2024. doi: 10.1101/2024.10.10.615955. URL https://www.biorxiv.org/content/early/2024/10/11/2024.10.10.615955.

Zoe Cournia, Bryce Allen, and Woody Sherman. Relative binding free energy calculations in drug discovery: recent advances and practical considerations. *Journal of chemical information and modeling*, 57(12):2911–2937, 2017.

Jordan E Crivelli-Decker, Zane Beckwith, Gary Tom, Ly Le, Sheenam Khuttan, Romelia Salomon-Ferrer, Jackson Beall, Rafael Gómez-Bombarelli, and Andrea Bortolato. Machine learning guided aqfep: a fast and efficient absolute free energy perturbation solution for virtual screening. *Journal of Chemical Theory and Computation*, 20(16):7188–7198, 2024.

Janani Durairaj, Yusuf Adeshina, Zhonglin Cao, Xuejin Zhang, Vladas Oleinikovas, Thomas Duignan, Zachary McClure, Xavier Robin, Daniel Kovtun, Emanuele Rossi, et al. Plinder: The protein-ligand interactions dataset and evaluation resource. *bioRxiv*, pp. 2024–07, 2024.

Paul G. Francoeur, Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B. Iovanisci, Ian Snyder, and David R. Koes. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of Chemical Information and Modeling*, 60(9):4200–4215, 2020. doi: 10.1021/acs.jcim.0c00411. URL https://doi.org/10.1021/acs.jcim.0c00411.

Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

Stephen R Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. Inchi, the iupac international chemical identifier. *Journal of cheminformatics*, 7:1–34, 2015.

Mingjian Jiang, Shuang Wang, Shugang Zhang, Wei Zhou, Yuanyuan Zhang, and Zhen Li. Sequence-based drug-target affinity prediction using weighted graph neural networks. *BMC genomics*, 23(1):449, 2022.

José Jiménez, Miha Skalic, Gerard Martinez-Rosell, and Gianni De Fabritiis. K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of chemical information and modeling*, 58(2):287–296, 2018.

Greg Landrum, Paolo Tosco, Brian Kelley, Ricardo Rodriguez, David Cosgrove, Riccardo Vianello, sriniker, Peter Gedeck, Gareth Jones, Eisuke Kawashima, NadineSchneider, Dan Nealschneider, Andrew Dalke, Matt Swain, Brian Cole, tadhurst cdd, Samo Turk, Aleksandr Savelev, Alain Vaucher, Maciej Wójcikowski, Ichiru Take, Rachel Walker, Vincent F. Scalfani, Hussein Faara,

Kazuya Ujihara, Daniel Probst, Niels Maeder, Jeremy Monat, Juuso Lehtivarjo, and guillaume godin. rdkit/rdkit: 2025_03_3 (q1 2025) release, June 2025. URL https://doi.org/10.5281/zenodo.15605628.

Gregory A. Landrum and Sereina Riniker. Combining ic50 or ki values from different sources is a source of significant noise. *Journal of Chemical Information and Modeling*, 64(5):1560–1567, 2024. doi: 10.1021/acs.jcim.4c00049. URL https://doi.org/10.1021/acs.jcim.4c00049.

Pierre-Yves Libouban, Samia Aci-Sèche, Jose Carlos Gómez-Tamayo, Gary Tresadern, and Pascal Bonnet. The impact of data on structure-based binding affinity predictions using deep neural networks. *International Journal of Molecular Sciences*, 24(22):16120, 2023.

Sarita Limbu and Sivanesan Dakshanamurthy. A new hybrid neural network deep learning method for protein–ligand binding affinity prediction and de novo drug design. *International Journal of Molecular Sciences*, 23(22):13912, 2022.

Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N Jorissen, and Michael K Gilson. Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, 35(suppl_1):D198–D201, 2007.

Tiqing Liu, Linda Hwang, Stephen K Burley, Carmen I Nitsche, Christopher Southan, W Patrick Walters, and Michael K Gilson. Bindingdb in 2024: a fair knowledgebase of protein-small molecule binding data. *Nucleic acids research*, 53(D1):D1633–D1644, 2025.

Andrew T McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. Gnina 1.0: molecular docking with deep learning. *Journal of cheminformatics*, 13(1):43, 2021.

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature methods*, 19(6):679–682, 2022.

Hakime Öztürk, Elif Ozkirimli, and Arzucan Özgür. Widedta: prediction of drug-target binding affinity. *arXiv preprint arXiv:1902.04166*, 2019.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707. URL https://www.biorxiv.org/content/early/2025/06/18/2025.06.14.659707.

Rodrigo Quiroga and Marcos A Villarreal. Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. *PloS one*, 11(5):e0155183, 2016.

Jeongtae Son and Dongsup Kim. Development of a graph convolutional neural network model for efficient prediction of protein-ligand binding affinities. *PloS one*, 16(4):e0249404, 2021.

Söding J. Steinegger, M. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35:1026–1028, 2017.

Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

Ísak Valsson, Matthew T Warren, Charlotte M Deane, Aniket Magarkar, Garrett M Morris, and Philip C Biggin. Narrowing the gap between machine learning scoring functions and free energy perturbation using augmented data. *Communications Chemistry*, 8(1):41, 2025.

Ercheng Wang, Huiyong Sun, Junmei Wang, Zhe Wang, Hui Liu, John ZH Zhang, and Tingjun Hou. End-point binding free energy calculation with mm/pbsa and mm/gbsa: strategies and applications in drug design. *Chemical reviews*, 119(16):9478–9508, 2019.

Huiwen Wang. Prediction of protein–ligand binding affinity via deep learning models. *Briefings in Bioinformatics*, 25(2):bbae081, 2024.

Zechen Wang, Liangzhen Zheng, Yang Liu, Yuanyuan Qu, Yong-Qiang Li, Mingwen Zhao, Yuguang Mu, and Weifeng Li. Onionnet-2: a convolutional neural network model for predicting protein-ligand binding affinity based on residue-atom contacting shells. *Frontiers in chemistry*, 9:753002, 2021.

Matthew Warren, Charlotte Deane, Aniket Magarkar, Garrett Morris, Philip Biggin, et al. How to make machine learning scoring functions competitive with fep. 2024.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, et al. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, pp. 2024–11, 2024.

Darrin M. York. Modern alchemical free energy methods for drug discoveryexplained. *ACS Physical Chemistry Au*, 3(6):478–491, 2023. doi: 10.1021/acsphyschemau.3c00033.

Vinicius Zambaldi, David La, Alexander E Chu, Harshnira Patani, Amy E Danson, Tristan OC Kwan, Thomas Frerix, Rosalia G Schneider, David Saxton, Ashok Thillaisundaram, et al. De novo design of high-affinity protein binders with alphaproteo. *arXiv preprint arXiv:2409.08022*, 2024.

Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J Manners, James Blackshaw, Sybilla Corbett, Marleen de Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, et al. The chembl database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic acids research*, 52(D1):D1180–D1192, 2024.

Xin Zeng, Shu-Juan Li, Shuang-Qing Lv, Meng-Liang Wen, and Yi Li. A comprehensive review of the recent advances on predicting drug-target affinity based on deep learning. *Frontiers in Pharmacology*, 15:1375522, 2024.

Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6(1):1–23, 2019.

Liangzhen Zheng, Jingrong Fan, and Yuguang Mu. Onionnet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. *ACS omega*, 4 (14):15956–15965, 2019.

A Dataset Construction

A.1 Data Curation

Initial curation proceeded as follows:

ChEMBL35: 1. Removed entries missing ligand SMILES or pchembl values. **2.** Removed entries which: did not have a UniProt ID for the protein target, referenced multiple protein targets, or referenced a protein variant. **3.** Removed entries with a data validity comment[3]. **4.** Removed entries where *standard relation* was $<$ or $>$. This step ensures that any measured values obtained are within the limit of detection for the assay. **5.** Only included assays that were flagged by ChEMBL as measuring binding (e.g., K_i, IC50, K_d). **6.** Removed measurements outside of a reasonable biochemical assay dynamic range (1 pM $< x <$ 100 μM).

BindingDB: 1. Removed entries missing molecule SMILES or IC50 values. **2.** Removed entries without a UniProt ID for the protein target or referenced multiple protein targets. **3.** Removed entries where reported IC50 values contained inequalities (i.e. $<$ or $>$). This step ensures that any measured values obtained are within the limit of detection for the assay. **4.** Remove measurements outside of a reasonable biochemical assay dynamic range (1 pM $< x <$ 100 μM).

After initial curation, data from both sources were merged into a single table. While this curation strategy can introduce variability in IC50 values by combining data from different assays Landrum & Riniker (2024), this dataset is still fully compatible with the maximal curation strategy outlined in

[3]The data validity comment was introduced in ChEMBL15 and includes information about the quality of the entry and to allow users to make an informed decision on whether to include that value in their analyses (https://chembl.blogspot.com/2020/10/data-checks.html).

Landrum & Riniker (2024) for data points from ChEMBL. Because BindingDB does not perform curation at the level of specific assays, the maximal curation strategy is not compatible with that source. For protein-ligand complexes that appear in both ChEMBL and BindingDB, we keep the information from both datasets.

All bio-activity values were converted to pIC50 units ($-\log_{10}$). SMILES strings for the ligand molecular structures were standardized by the removal of salts, protonation at neutral pH (where possible), and canonicalization using RDkit. Note that the choice of neutral pH for the ligand protonation is immaterial for the subsequent computational prediction of the protein-ligand structures, as current cofolding models do not predict the positions of hydrogen atoms.

A coarse ligand library filter was applied to exclude likely false positives/false negatives by removing PAINS and molecules with molecular weights exceeding 1250 Da. Protein-ligand complexes containing proteins with more than 2000 amino acid residues were excluded, in order to increase the probability of successful prediction by the cofolding model on current GPU hardware. Next, duplicate entries were removed based on UniProt accession and canonical SMILES.

The amino acid sequence for each protein was obtained from its UniProt entry using the accession number provided in the ChEMBL or BindingDB dataset. Note that this canonical sequence from UniProt may differ from the one used in the original bioactivity assay. For instance, the experimental protein may have been a truncated construct, a mutant, or a specific quaternary structure (e.g., a homodimer), whereas our analysis used the monomeric sequence from the database.

Finally, to avoid data leakage when using this dataset to train or evaluate models that use structural data from the PDB for training, protein-ligand systems that already have experimentally-solved structures in the PDB were removed. The existence of a corresponding structure in the PDB was determined by finding the Chemical Component Dictionary (CCD) identifier of the ligand (by first computing its InChIKeyHeller et al. (2015) using RDKit) and looking for matches to this (Uniprot ID, CCD ID) pair in the PDB. This search utilized the RCSB GraphQL search APIrcs, and the PDBe REST API provided by EMBL-EBISIF.

A.2 Folding Model

We used the Boltz-1x folding model to generate 3D structures for all protein-ligand complexes described in §2. Boltz-1 is a publicly available implementation of AlphaFold 3[4]. Additionally, Boltz-1x extends this model by introducing a guiding potential to the diffusion process to prevent clashes, resulting in more physically realistic binding poses. Due to the lack of quaternary structure information, all complexes were treated as monomeric assemblies. We generated five structure samples per complex, as this represents the maximum number we can compute on a single GPU for the longest protein sequence in the dataset. While it is common practice to increase sample diversity by varying random seeds across multiple runs, we did not apply this technique due to resource constraints[5].

The Boltz-1x model was run using three recycling steps and 200 sampling steps (any other settings are the defaults as of the boltz-1x release). Note that, as all these systems are monomeric, the pairing strategy is irrelevant.

Multiple sequence alignments (MSAs) for input to the model were generated using the MMseqs2 tool (Steinegger, 2017) (via the ColabFold Mirdita et al. (2022) project). This used the UniRef30 sequence database version 2302 and the ColabFoldDB metagenomic sequence database version 202108.

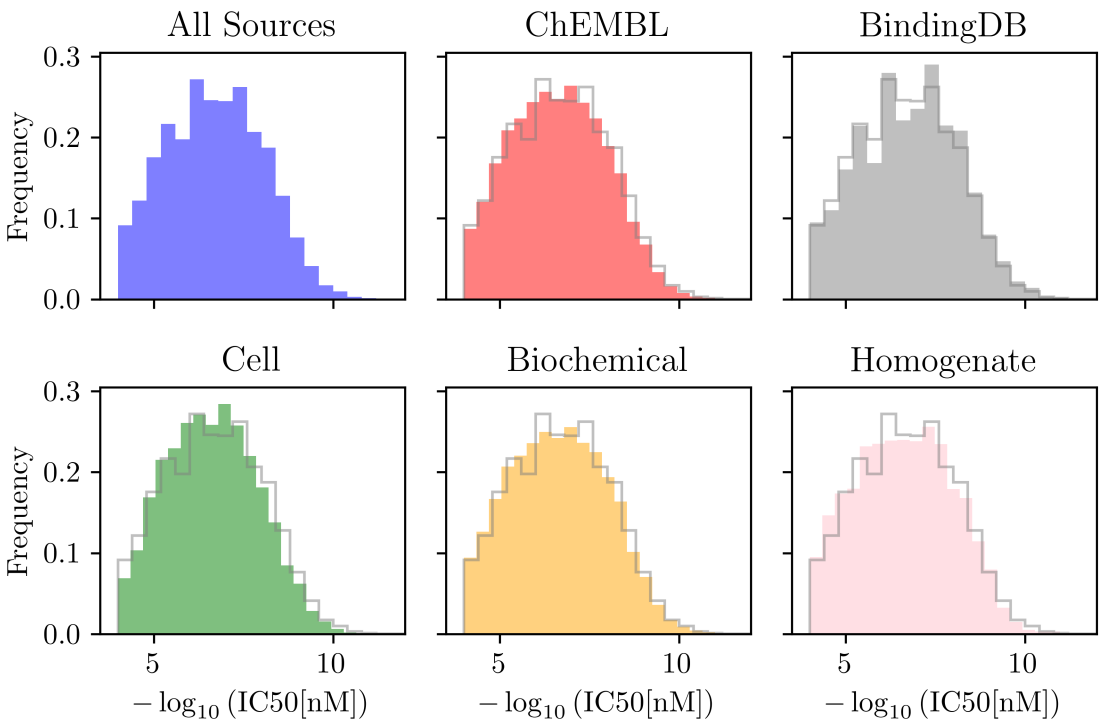

Figure 3: Histogram showing the distribution of pIC50 values stratified by data source (ChEMBL and BindingDB) and assay type (biochemical, cell-based, and unknown). Note that, as shown in Table 1, the assay type could not be inferred for the majority of complexes based on the curated assay descriptions. As a result, the bottom panels include only those complexes with known assays. For all histograms except the first one, we also plot the overall distribution, in grey.

Table 1: Data distribution by source and assay. Note that the total number between both assays is larger than the number of generated structures. That is because of protein-ligand pairs that appear in both datasets

Assay	**Source**	
	BindingDB	**ChEMBL**
biochem	2,293	416,331
cell	0	183,286
homogenate	0	13,980
na	934,409	0
Total	**936,702**	**613,597**

A.3 Data Statistics

A.3.1 Proteins

The $1,048,857$ protein-ligand systems in the dataset comprise $5,149$ unique protein sequences. Of these $5,149$ proteins present in our dataset, $2,150$ are believed to have no structures deposited in the PDB. The distribution of sequence lengths is shown in Fig. 4. Most sequences fall within the

[4]There are some minor changes between AlphaFold 3 and Boltz-1, such as the strategy used for Multiple Sequence Alignment (MSA) subsampling.

[5]Further, Boltz-1's MSA subsampling is deterministic with respect to the random seed, unlike other cofolding models such as AlphaFold3 and Chai-1 (Chai Discovery, 2024)), where seed variation is a primary source of stochasticity. As a result, we do not expect significant diversity gains from seed variation in Boltz-1x.

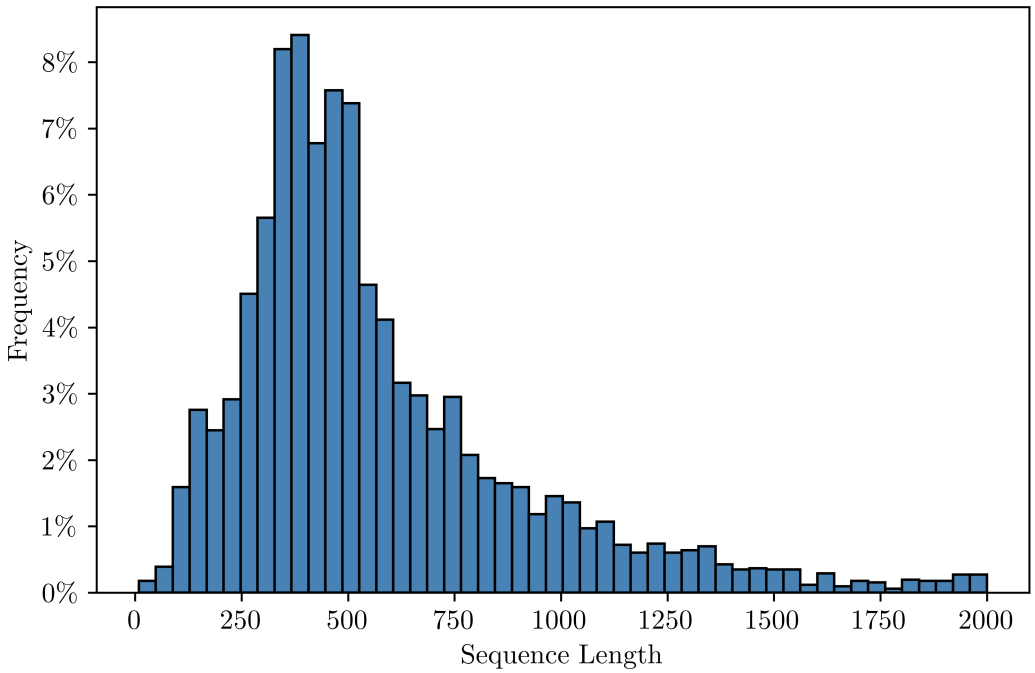

Figure 4: Distribution of protein sequence lengths across all unique entries in the dataset.

300-500 amino acid range. Beyond 500 residues, the frequency decreases steadily, with very long sequences (e.g., > 1500 amino acids) appearing only rarely.

Sequence clustering with MMseqs2 (Steinegger, 2017) using reasonable values for the minimum sequence identity and minimum coverage (MMseqs2 flags `-min-seq-id` and `-c`, respectively) revealed the presence of a large number of singleton clusters. For example, setting `-min-seq-id` and `-c` to $[0.8, 0.8]$, $[0.5, 0.7]$, and $[0.3, 0.2]$ resulted in 3793, 2818, and 1862 clusters, respectively.

Proteins were assigned to a family by using metadata provided by the UniProt database. We first classified them into enzymes or non-enzymes by looking at the presence of an enzymatic activity number (EC number). Enzymes were further subdivided into kinases (EC $= 2.7.x$), phosphatases (EC $= 3.1.x$) and other enzymes. Non-enzymes were subdivided by looking at their gene ontology codes (GO code). For example, the presence of GO $= 0004879$ implies that the protein is a nuclear receptor.

The distribution of protein families accross different assays is shown in Fig. 5. We see that the biochemical assay has a larger proportion of phosphatases, kinases and enzymes, while the cell assay has more nuclear receptors and GPCRs. There is also a larger number of proteins in the cell assay, for which we could not parse family information.

A.3.2 Ligands

We used `RDKit` (Landrum et al., 2025) to compute a range of chemical descriptors, as summarized in Table 2. The values capture the statistics of the unique ligands in the dataset.

It is worth noting that we did *not* perform any filtering on the dataset on the basis of things like "drug-likeness" of the ligands, for example filtering samples with ligands below a certain molecular weight. This is to avoid losing useful and chemically-meaningful data points of biologically-relevant species, such as ionic cofactors or small organic fragments that can teach a model protein-small-molecule interaction chemistry. It is our expectation that users will choose to filter samples based on ligand characteristics according to their use case.

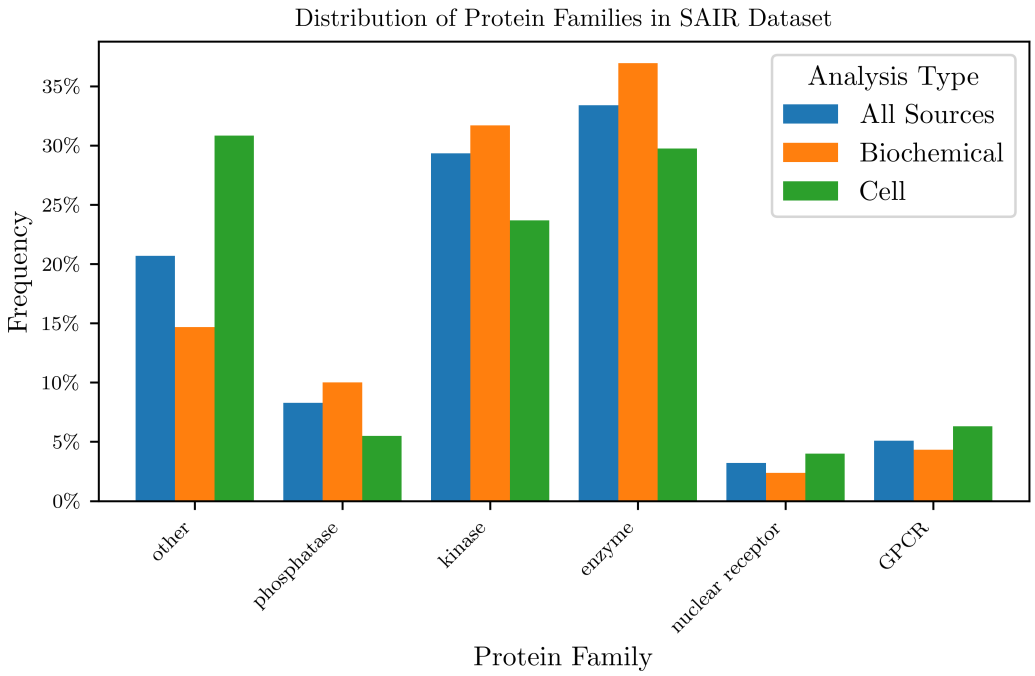

Figure 5: Distribution of protein families across all unique entries in the dataset. We could not parse the protein family information for a fraction of the proteins, which are shown under 'other'.

Table 2: Chemical descriptors across the dataset (aggregation of unique ligand entries).

Descriptor	**Min**	**Max**	**Mean**	**Stddev.**
Molecular weight	17.0	1.25e+03	4.46e+02	1.25e+02
Heavy atom count	1.00	94.0	31.7	8.88
Hetero atom count	0.00	46.0	8.55	3.31
H-bond acceptor count	0.00	36.0	5.89	2.33
H-bond donor count	0.00	25.0	1.89	1.58
Topological polar surface area	0.00	6.40e+02	92.9	43.3
Wildman-Crippen LogP	-13.1	19.2	3.80	1.72
QED (Drug-likeness)	0.00684	0.948	0.498	0.204
Rotatable bond count	0.00	53.0	6.11	3.69
Fraction Csp3	0.00	1.00	0.321	0.177
Aliphatic carbocycle count	0.00	19.0	0.328	0.720
Aliphatic heterocycle count	0.00	20.0	0.754	0.828
Aliphatic rings count	0.00	21.0	1.08	1.08
Aromatic carbocycle count	0.00	20.0	1.55	0.935
Aromatic heterocycles count	0.00	11.0	1.45	1.10
Aromatic ring count	0.00	20.0	3.00	1.16
Bridgehead atom count	0.00	20.0	0.110	0.598
Spiro atom count	0.00	6.00	0.0392	0.210

B Extra Results

B.1 Boltz Confidence Metrics

Given that we have access to experimental binding potencies for the corresponding complexes, we assessed whether Boltz-1x's confidence metrics correlate with binding affinity. Prior work has demonstrated that AlphaFold confidence scores correlate with binding affinity in protein–protein interactions (Zambaldi et al., 2024) (PPIs). Here, we explore whether similar correlations exist in

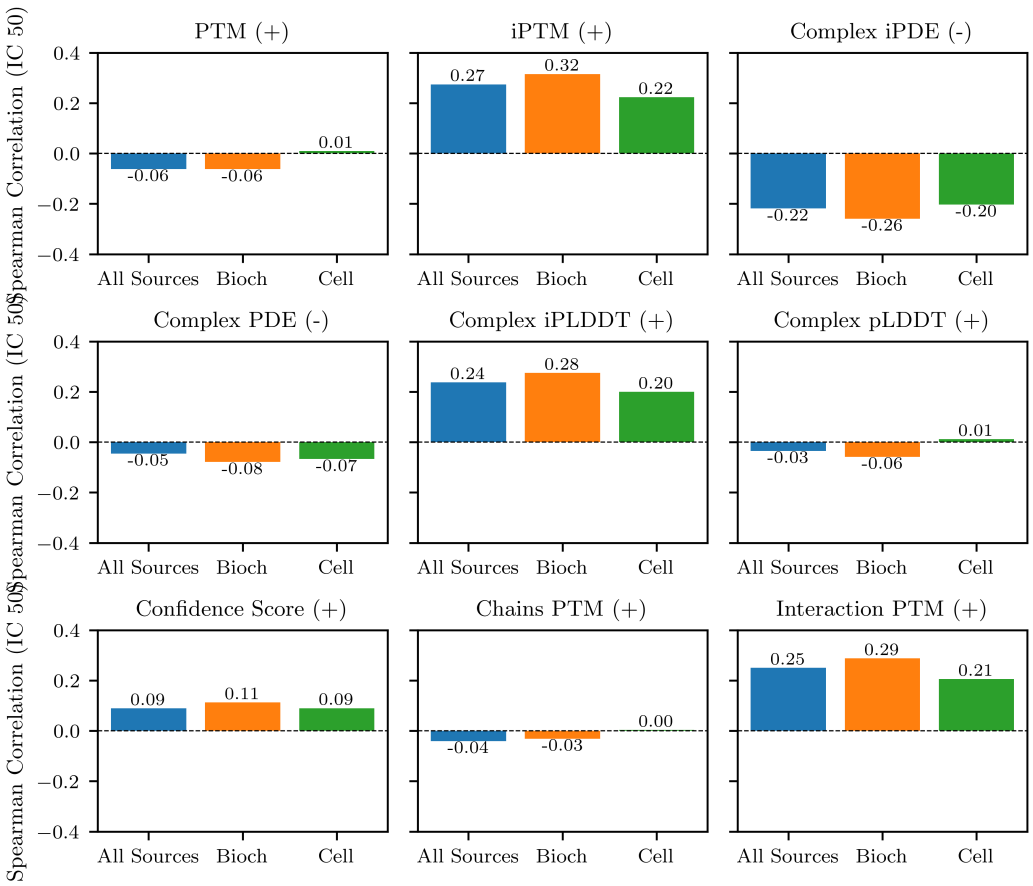

Figure 6: Comparison of the Spearman correlation r_s between different Boltz-1x confidence metrics, and the experimental IC50 activity, by assay type. The signs shown in the title of each panel indicate the expected direction of correlation: negative for PDE and iPDE (as they represent distances), and positive for all other metrics.

the context of protein–ligand interactions. Results are shown in Fig. 6.Focusing on the blue bars (Spearman correlation averaged across all assay types), we observe a significant correlation between certain confidence metrics, particularly those involving the protein-ligand interface —namely iPTM, complex iPDE, and complex iPLDDT—and experimental potency. These findings suggest that Boltz-1x's structural confidence metrics provide some predictive signal for protein–ligand binding affinity. Notably, the strength of the correlation varies by assay type: it is highest for biochemical assays and weakest for cell assays. We hypothesize that this is caused by the great specificity and accuracy of biochemical assays, while cellular and homogenate assays may introduce additional confounding factors such as off-target binding, permeability, and intracellular dynamics. To further probe protein–ligand interaction quality, we introduce a new metric — interaction PTM — defined as the average of the off-diagonal values in the `pair_chains_ptm` confidence head. This metric captures the confidence of the protein with respect to the ligand, and vice versa, and is analogous to the "interaction PAE" described in Zambaldi et al. (2024). We find that interaction PTM exhibits strong correlation with binding affinity ($r_s = 0.25$), ranking second only to iPTM ($r_s = 0.27$) in predictive power across our dataset.

We can furthermore look at the similarity between generated protein chains and protein chains from the training dataset. We find a high degree of correlation (Spearman correlation of 0.47) between the global PTM confidence and the highest TM-score (normalized by query length) to structures in the training set. This metric is independent of the ligand, which partly explains the poor correlation to the binding affinity. It is better suited for evaluating the global shape of the protein.

B.2 Pocket diversity

We can use our model to gain insight into the effect that changing the input ligand has in the generated protein conformation. If we give Boltz-1x sufficiently distinct ligands, is the model able to detect different appropriate binding sites, or will it re-use the pockets it has seen during training?

To address this, we need to define a pocket. First, we define pocket residue as every residue that has a non-hydrogen atom within 6 to the closest ligand atom, a cutoff commonly adopted in the field. The set of pocket residues defines a pocket, and two pockets are considered similar if

$$\frac{|p_1 \cap p_2|}{\min(|p_1|, |p_2|)} \geq \text{threshold}, \tag{1}$$

where we defined the threshold to be 0.8. That allows us to cluster the detected pockets per protein into groups, such that no group shares a similar pocket.

Fig. 7 shows both the diversity in pockets over the five generated structures per protein/ligand complex (left panel), and the diversity for a given protein as we change ligands (right). AlphaFold3-like models are known to generate similar conformations for different diffusion samples, which we also find when looking at the pocket diversity, with most complexes generating ligands in the same pocket for all five generated samples.

However when looking at a protein chain, with different ligands, we find a significant fraction of systems with a variable set of pockets. The most extreme example of this is protein P10636, where we found well over a thousand different potential binding sites for 345 different ligands, two of which are visualized in figure 8.

This shows a potential reason for the fat tail in the number of distinct pockets per protein. When Boltz-1x is uncertain about the structure or if the protein is very flexible, then we find a very diverse set of protein conformations. The pocket-residues will similarly change a lot, and there is no well defined binding site.

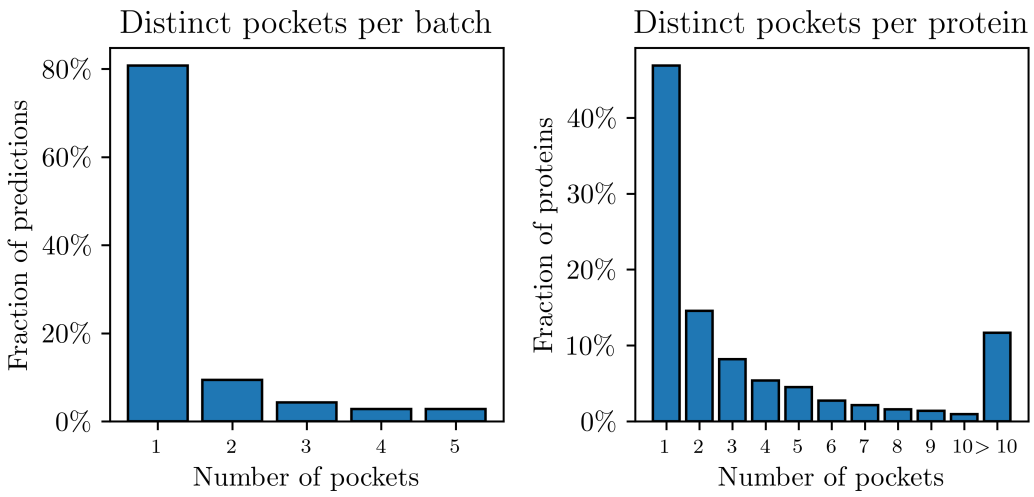

Figure 7: Pocket diversity for different samples and different ligands. Pocket similarity was determined using Eq. (1). **Left**: The diversity of pockets in the five generated structures per protein/ligand complex. **Right**: The diversity of pockets for different proteins.

We can also study pocket similarity to the training set. Fig. 9 shows the result of performing a similarity search of all generated structures against the Boltz-1x training data, the pocket-LDDT score, as defined in Durairaj et al. (2024). The pocket-LDDT is defined by structurally aligning predicted structures to ground truth structures, and calculating the average LDDT over the backbone carbon atoms in the aligned pocket residues. In our case, this score is not a measure of correctness, but more a measure of similarity to the training dataset.

We find no correlation between the pocket-LDDT and the interface confidence (Spearman correlation of -0.02), but we do find most pockets to be highly similar to the training data. We have previously

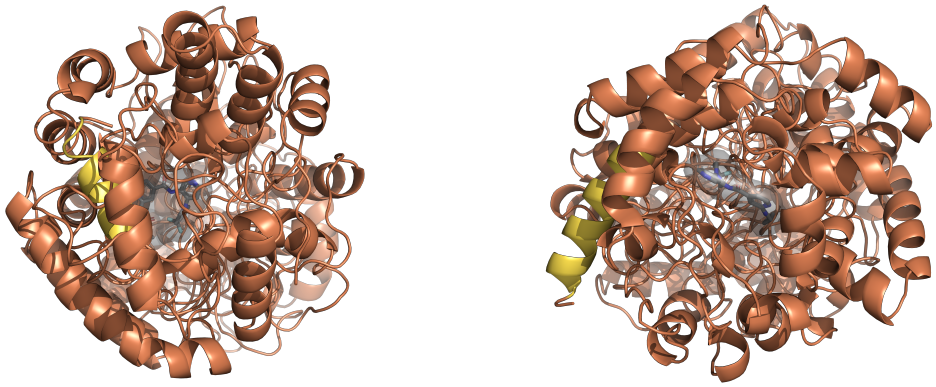

Figure 8: Diverse binding sites in protein P10636.

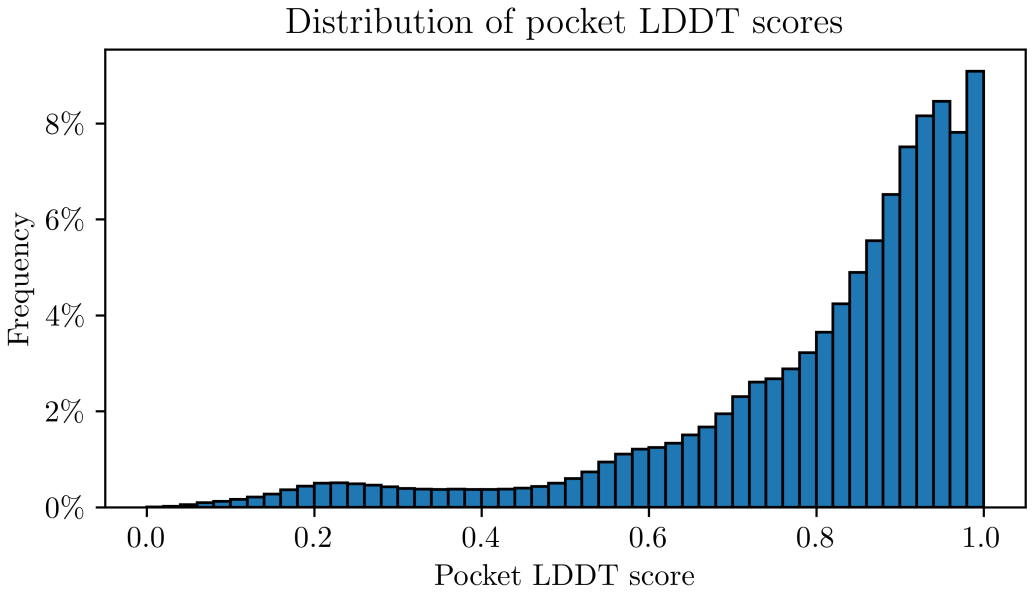

Figure 9: The distribution of pocket LDDT's calculated, comparing the generated structures to the training dataset.

seen that Boltz is able to generate multiple distinct binding poses, which together implies that Boltz successfully places ligand atoms in plausible looking pockets.

B.3 Evaluation Metrics

We use the following metrics to evaluate the performance of binding affinity models:

- **Spearman Correlation**: The Spearman correlation between the predicted and experimental binding affinity, defined as:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{2}$$

  where d_i is the difference between the ranks of the predicted and experimental binding affinities, and n is the number of samples.

- **Pearson Correlation**: The Pearson correlation between the predicted and experimental binding affinity, defined as:

$$r_p = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \tag{3}$$

  where x_i and y_i are the predicted and experimental binding affinities, respectively, and $\bar{x}$ and $\bar{y}$ are the means of the predicted and experimental binding affinities, respectively.

- **Kendall's Tau**: Kendall's Tau is a measure of the ordinal association between two quantities, defined as:

$$\tau = \frac{(n_c - n_d)}{\frac{1}{2}n(n-1)} \tag{4}$$

  where n_c is the number of concordant pairs, and n_d is the number of discordant pairs, and n is the number of samples.

- **Area Under the Curve (AUC)**: The AUC is a measure of the ability of a model to distinguish between positive and negative samples. It is defined as the area under the Receiver Operating Characteristic (ROC) curve, which is a plot of the true positive rate against the false positive rate. To calculate the AUC, we first need to define a threshold for the predicted binding affinity, and then compute the true positive rate (TPR) and false positive rate (FPR) for that threshold. We use a threshold of $100nM$, which is a common threshold for binding affinity prediction.

B.4 Posebusters results

A table summarizing posebusters results by protein family is shown in Table 3

Table 3: Summary of PoseBusters results by family

Family	Metric	Structures	Assemblies
Overall Analysis	Total failed	166,241	5,526
	Total	5,244,285	1,048,857
	Percentage failed	3.17%	0.53%
Enzyme	Total failed	59,169	1,944
	Total	1,699,025	339,805
	Percentage failed	3.48%	0.57%
Kinase	Total failed	34,160	1,639
	Total	1,531,115	306,223
	Percentage failed	2.23%	0.54%
Other	Total failed	46,248	1,200
	Total	1,170,445	234,089
	Percentage failed	3.95%	0.51%
Phosphatase	Total failed	12,181	329
	Total	418,895	83,779
	Percentage failed	2.91%	0.39%
GPCR	Total failed	10,304	188
	Total	267,650	53,530
	Percentage failed	3.85%	0.35%
Nuclear Receptor	Total failed	4,179	226
	Total	157,155	31,431
	Percentage failed	2.66%	0.72%