HAVE PROTEIN-LIGAND CO-FOLDING METHODS MOVED BEYOND MEMORISATION?

Peter Škrinjar
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
peter.skrinjar@unibas.ch

Jérôme Eberhardt
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
jerome.eberhardt@unibas.ch

Janani Durairaj
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
janani.durairaj@unibas.ch

Torsten Schwede
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
torsten.schwede@unibas.ch

ABSTRACT

Deep learning has driven major breakthroughs in protein structure prediction, however the next critical advance is accurately predicting how proteins interact with other molecules, especially small molecule ligands, to enable real-world applications such as drug discovery and design. Recent deep learning all-atom methods have been built to address this challenge, but evaluating their performance on the prediction of protein-ligand complexes has been inconclusive due to the lack of relevant benchmarking datasets. Here we present a comprehensive evaluation of four leading all-atom cofolding deep learning methods using our newly introduced benchmark dataset Runs N' Poses, which comprises 2,600 high-resolution protein-ligand systems released after the training cutoff used by these methods. We demonstrate that current co-folding approaches largely memorise ligand poses from their training data, hindering their use for *de novo* drug design. This limitation is especially pronounced for ligands that have only been seen binding in one pocket, whereas more promiscuous ligands such as cofactors show moderately improved performance. With this work and benchmark dataset, we aim to accelerate progress in the field by allowing for a more realistic assessment of the current state-of-the-art deep learning methods for predicting protein-ligand interactions.

1 Introduction

Being able to accurately predict and characterise protein-ligand interactions (PLI) is a cornerstone of modern drug discovery, enabling the identification of small molecules that can modulate protein function with high specificity (1). The process of discovering and optimising such modulators into viable lead compounds is a critical aspect of early-stage pharmaceutical development and lead optimisation. Structure-based drug design (SBDD), a component of computer-aided drug design (CADD), relies on three-dimensional structures of the target protein and the protein-ligand interface to inform medicinal chemistry decisions (2; 3; 4). Therefore, to achieve accurate predictions of how ligands bind and their binding affinity, a deep understanding of the fundamental principles governing PLIs is essential (5). The combination of different computational techniques has become indispensable in both industrial and academic drug discovery efforts, helping to reduce costs and accelerate development timelines (6). However, despite significant advances in computational methods, accurately modeling protein-ligand binding at atomic resolution remains a significant challenge (7; 8).

Over the years, there have been numerous community initiatives to assess the state of the field in predicting the structure of ligand poses binding to proteins. More recently, CASP15 had a pilot experiment in the evaluation of protein-ligand complex prediction (9), inspired by previous efforts such as Teach Discover Treat (TDT) (10), Continuous Evaluation of Ligand Prediction Performance (CELPP) (11), Drug Discovery Data Resource (D3R) (12; 13; 14; 15), and Community Structure–Activity Resource (CSAR) (16; 17), with the added complexity that the prediction task consisted of predicting

both the structure of the receptor protein as well as the position and conformation of one or more ligands. Conclusions from CASP15 (18) and parallel benchmarking efforts (19) were threefold: (1) The joint protein-ligand prediction task requires accuracy metrics that take both protein and ligand positioning into account, such as binding-site superposed RMSD and local difference distance test (LDDT) of protein-ligand interactions (LDDT-PLI) (18); (2) Experimental quality is a more acute issue for PLI prediction compared to protein structure prediction, due to the atom-level reliability required and the fact that many PDB entries fail to meet even relaxed X-ray quality standards (19; 20) to enable their use as reliable ground truth; and (3) Many complexes with ligands released in the PDB are either very similar to previously released complexes or not relevant for drug discovery applications, or both. While CASP16 addressed the latter issue by soliciting unreleased PLI complexes from pharmaceutical companies, the scarcity of targets and the availability of very similar templates in the PDB make it difficult to draw conclusions specifically about deep learning methods that qualify for this task, most of which also did not participate.

Deep learning methods such as AlphaFold2 (21) have revolutionised protein structure prediction, inspiring the development of all-atom co-folding methods like RosettaFold All-Atom (22), AlphaFold3 (23), Chai-1 (24), Boltz-1 (25), Protenix (26), NeuralPLexer3 (27), for example. While these methods show promise, evaluating their performance is difficult due to the lack of established benchmarking techniques for joint protein-ligand binding in the deep learning domain. Existing benchmarks often fail to account for similarities between the training data and test sets, leading to inflated performance estimates (19; 20; 28; 29). This is particularly problematic for PLI systems, where the scarcity and bias in experimental structures can exacerbate both the risk of memorisation and our inability to detect it.

To address this, we present a comprehensive benchmark of four leading all-atom co-folding deep learning methods with highly similar architectures and training paradigms (AlphaFold3 (23), Chai-1 (24), Protenix (26), and Boltz-1 (25)) on 2,600 high-resolution PLI systems released after their training cutoff (30 September 2021). We explore the impact of training data similarity on prediction accuracy, revealing a critical limitation: current co-folding methods struggle to generalise beyond ligand poses seen in their training data. We also compared the top-ranked and best-scored models, observing a slight performance improvement, yet the overall trend of limited generalisation remained unchanged.

This overfitting issue, independent of model confidence scores, highlights the need for improved architectures and data augmentation strategies. Notably, we observe better generalisation for cofactor, amino acid, and nucleotide analogs, suggesting the importance of abundant training data. We provide a benchmark dataset (Runs N' Poses) with varying degrees of training set similarity, along with resources to analyse the relationship between data leakage and performance. With this work, we aim to facilitate progress and enable identification of genuine breakthroughs in the critical field of protein-ligand complex prediction.

2 Results

We aimed to explore the dependence of prediction accuracy to the similarity of the ground truth to the training set of the four deep learning co-folding methods, based on our benchmark dataset of 2,600 systems covering 3,047 proper ligands. Our analysis shows a clear correlation between the performance of the top-ranked model for each method and the similarity of the system to the training set, depicted in Figure 1A. Here, we define the similarity to the training set as the product of binding pocket coverage (29) and Combined Overlap Score (SuCOS) of the ligand pose, which measures the extent to which ligand poses overlap in both volume and placement of chemical features (30; 31). Thus, a training system is only considered similar if it has a similar ligand pose in a similar protein pocket to the test system.

In classical docking benchmark studies, success is typically defined solely by ligand RMSD (< 2Å), as the protein structure is generally treated as rigid and assumed to be correct. However, for co-folding methods, it is essential to assess the accuracy of both the ligand and the protein structure. Therefore, we define the success rate as a combination of LDDT-PLI (> 0.8) (18) and ligand pose RMSD (< 2 Å) (see Section 7.3 for further details on this choice). Across all four methods, similar correlations are observed, with AlphaFold3 showing a slight performance advantage. However, there is no subset where AlphaFold3 consistently outperforms while the other three methods fail, as indicated by the gray line representing the best-scored of the top-ranked models across all methods. While Figure 1 shows the common subset of obtained predictions across the four methods (2,077 systems covering 2,311 scored ligands), Supplementary Figure S2 shows the same, including any system predicted by any method, with failures not counted for each method. See Table 1 and Section 4 for more details on the method-specific failures.

Further stratification reveals slightly improved performance on ligands with significant amounts of training data. This includes cofactors and their analogs, as these are found across a wide variety of protein families and thus protein-ligand structures and pockets in the PDB. It also includes analogs of amino acids and nucleotides, likely a benefit of training on protein-protein complexes and protein-nucleotide complexes. To set these apart from more drug-like molecules, we classify them based on their frequency of occurrence in the PDB, independent of the protein and pocket. Specifically, for each ligand in the test set we counted the number of times it or any ligand with >0.9 ligand Morgan fingerprint Tanimoto

similarity (32) to it appears in the training data and separated those with more than 100 occurrences. Supplementary Figure S4 depicts these molecules, hereafter referred to as promiscuous. Removing these ligands from consideration shows a more linear trend of similarity to success rate (Figure 1B), demonstrating that the successes in the lower similarity bins are primarily due to such cases (Supplementary Figure S5A-B). This indicates that data scarcity is a major driver of memorisation, as drug-like molecules typically do not have a lot of experimental structures resolved in different protein pockets.

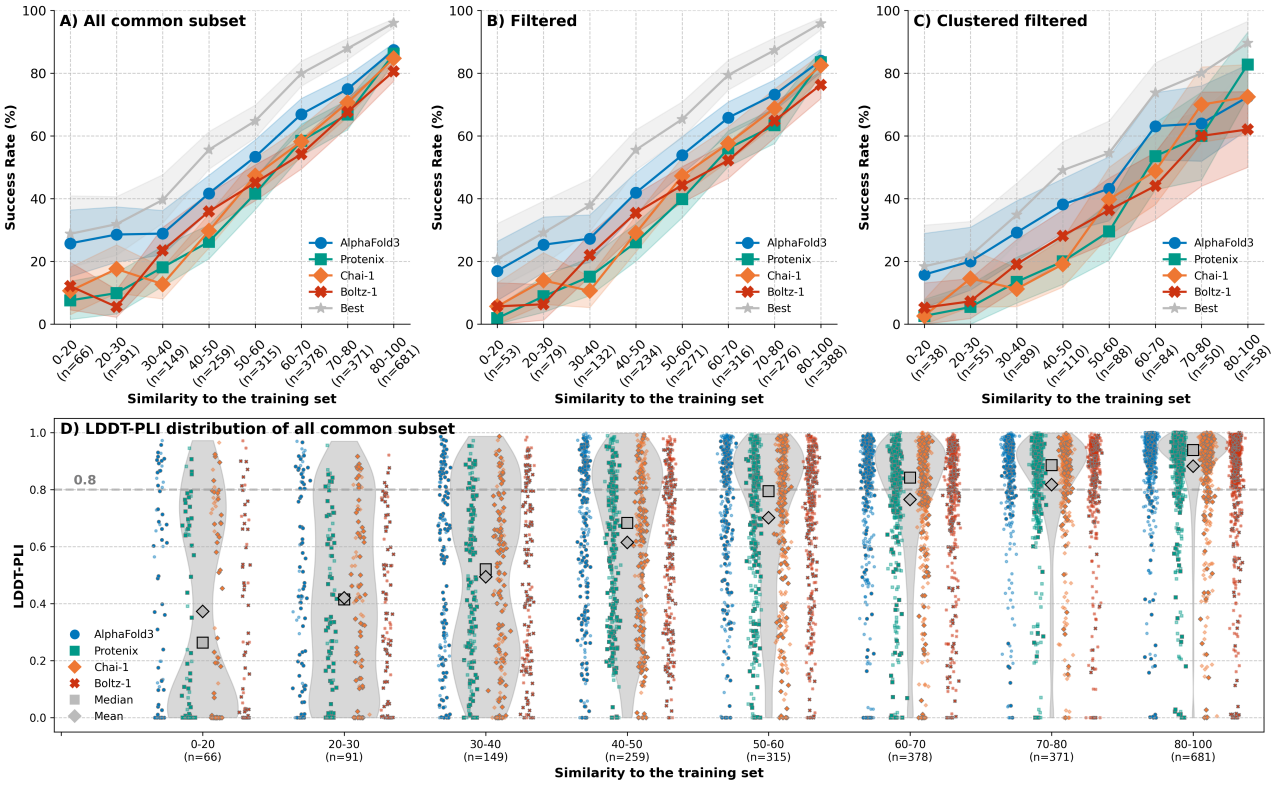


Figure 1: **Prediction accuracy vs. training set similarity.** The success rate, defined as the percentage of system ligands with <2Å RMSD and >0.8 LDDT-PLI, for system ligands in each similarity bin across the four methods and when selecting the best-scored model from the four top-ranked, for **A)** all system ligands in the common subset, **B)** only system ligands having <100 training systems containing analogs, and **C)** only cluster representatives from B. Shaded regions correspond to the 95% confidence interval, calculated from 1,000 bootstrap samples for each bin and method. **D)** The distribution of LDDT-PLI values for all methods displayed as a gray violin plot, with each individual method being shown as colored scatter points where cluster representatives have a gray outline.

We also verified that the observed trend is not driven by an over-representation of certain protein families, which could bias the results. Figure 1C displays the same trend correlating success vs. training set similarity for just the cluster representatives of drug-like systems. This is based on clustering at SuCOS-pocket similarity > 50. This leads to a total of 789 clusters in the common subset out of 1,017. Supplementary Table S1 details the sizes, cluster representative details, and similarity bin coverage of the ten biggest clusters. The first consists of 171 crystal structures of SARS-CoV2 main protease (MPro) in complex with different small molecules from independent efforts such as the SARS-CoV2 moonshot series (33) and others (34), while the second consists of 136 protein kinases complexed with various inhibitors. Thus SuCOS-pocket clustering puts together protein-ligand complexes where similar molecules bind similar proteins in the same pocket and with similar three-dimensional poses. For example, the HCov-NL63 MPro complexes in cluster #4

have different binding modes than cluster #1. While Figure 1C shows that the trend is independent of redundancy, such congeneric series, i.e. ligands sharing the same core structure and differing only in some substituent groups, are of high relevance to drug discovery applications and accuracy across these series is an important target to aim for.

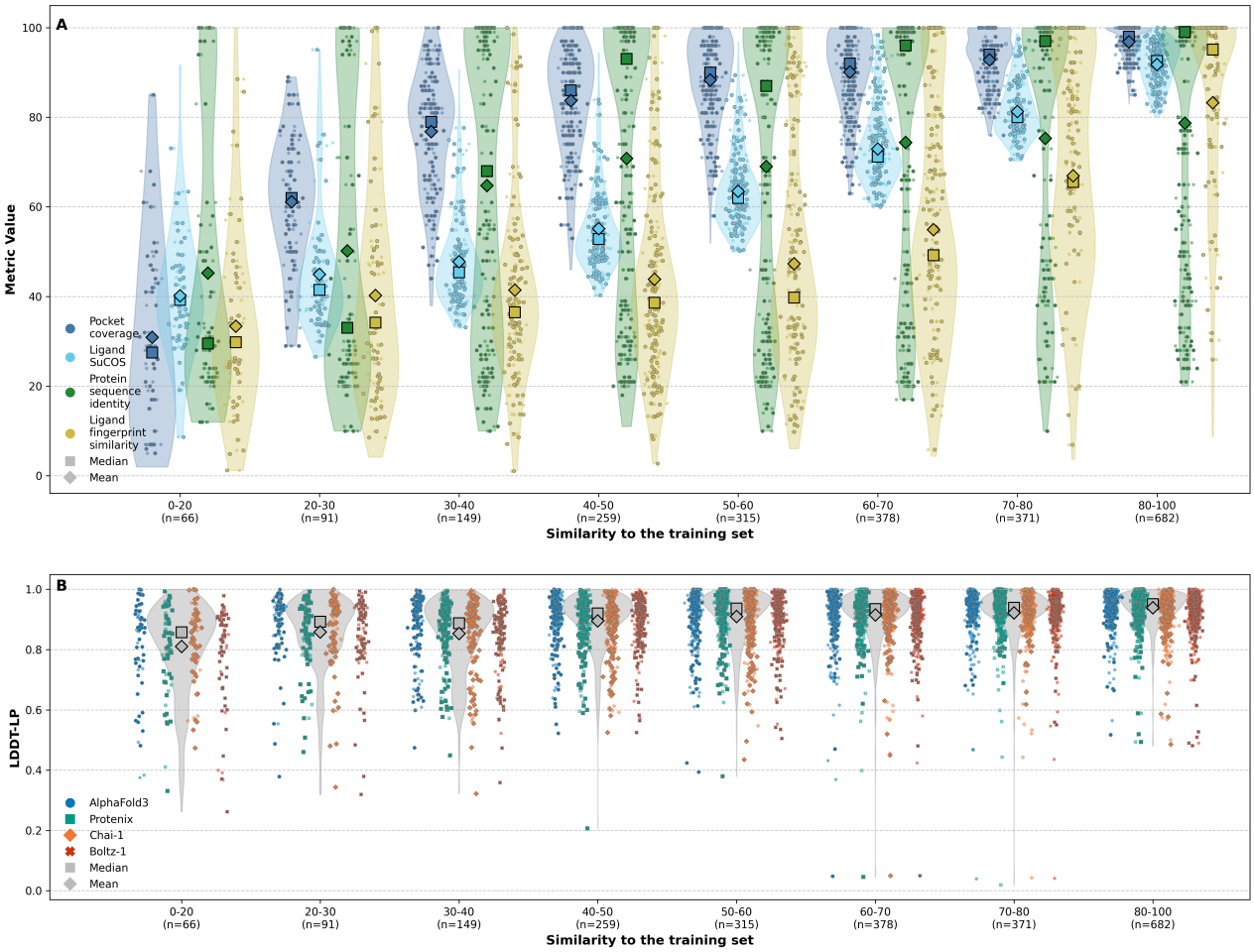


Figure 2: **Comparision with other similarity and accuracy metrics. A)** The distribution of pocket coverage (shown as percentage), ligand SuCOS similarity, protein sequence identity, and ligand Morgan fingerprint Tanimoto similarity (32), for the closest training system found by SuCOS-pocket similarity. **B)** Distribution of LDDT-LP values across similarity bins for all methods (gray violin) and each method separately (colored scatter). Percentage of systems with LDDT-LP > 0.8, for all systems: AlphaFold3: 92%, Protenix: 91%, Chai-1: 90%, Boltz-1: 90% For cluster representatives (gray outlines): AlphaFold3: 87%, Protenix: 83%, Chai-1: 82%, Boltz-1: 83%

The LDDT-PLI distribution for each method also consistently increases with the similarity to the training set (Figure 1D). Similarly, defining success based solely on RMSD (< 2Å) and examining the RMSD distributions confirms that our success rate criteria do not influence the observed trend (Supplementary Figure S3). Looking at different subsets of the dataset, such as single-ligand, multi-ligand and multi-protein systems only, shows again the same pattern, demonstrating the effectiveness of this dataset for benchmarking these sub-tasks (Supplementary Figure S5D-F). We also show that the methods are not biased in performance towards smaller or larger ligands, as seen in Supplementary Figure S5G-H.

While some bias toward larger binding pockets exists (Supplementary Figure S5I), it appears to stem from a higher prevalence of test systems with larger pockets in the higher training similarity bins (Supplementary Figure S6).

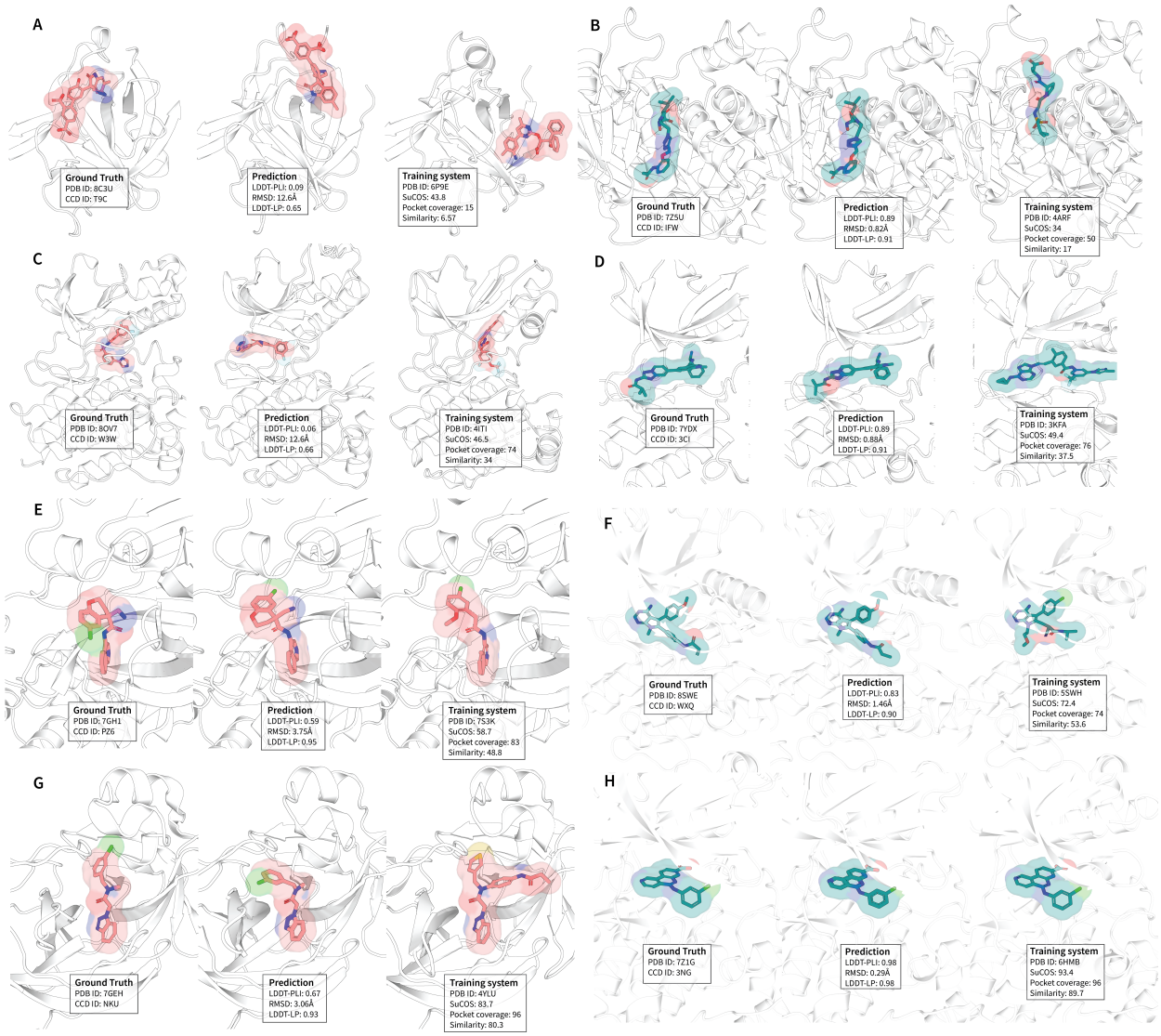


Figure 3: **Examples of incorrect and correct predictions across different similarity bins.** Predicted models from AlphaFold3 (middle of each panel) compared to the ground truth experimental structure (left) and the closest training system by SuCOS-pocket similarity (right). For each panel, the PDB ID and CCD code of the ground truth, the RMSD, LDDT-PLI and LDDT-LP of the prediction, and the PDB ID, SuCOS score, pocket coverage (percentage), and SuCOS-Pocket similarity score of the closest training system are labelled. We depict an incorrect prediction (red, left panels) and a correct prediction (green, right panels) from similarity bins 0-20 (**A-B**), 20-40 (**C-D**), 40-60 (**E-F**) and 80-100 (**F-G**).

To explore how much the memorisation trend is influenced by similarity at protein, pocket and ligand levels, Figure 2A shows how multiple similarity metrics, such as pocket coverage, ligand SuCOS, protein sequence identity, and ligand fingerprint similarity, evolve across similarity bins, reflecting the relationship between each system and its closest training system as determined by SuCOS-pocket similarity. Here, the relationship between the SuCOS score and pocket coverage becomes clearer, as the lowest bin (SuCOS-pocket similarity between 0 and 20) includes examples with high SuCOS but low pocket coverage, and vice versa. In contrast, the protein sequence identity and the ligand fingerprint similarity vary more broadly, but also tend to increase with increasing SuCOS-pocket similarity.

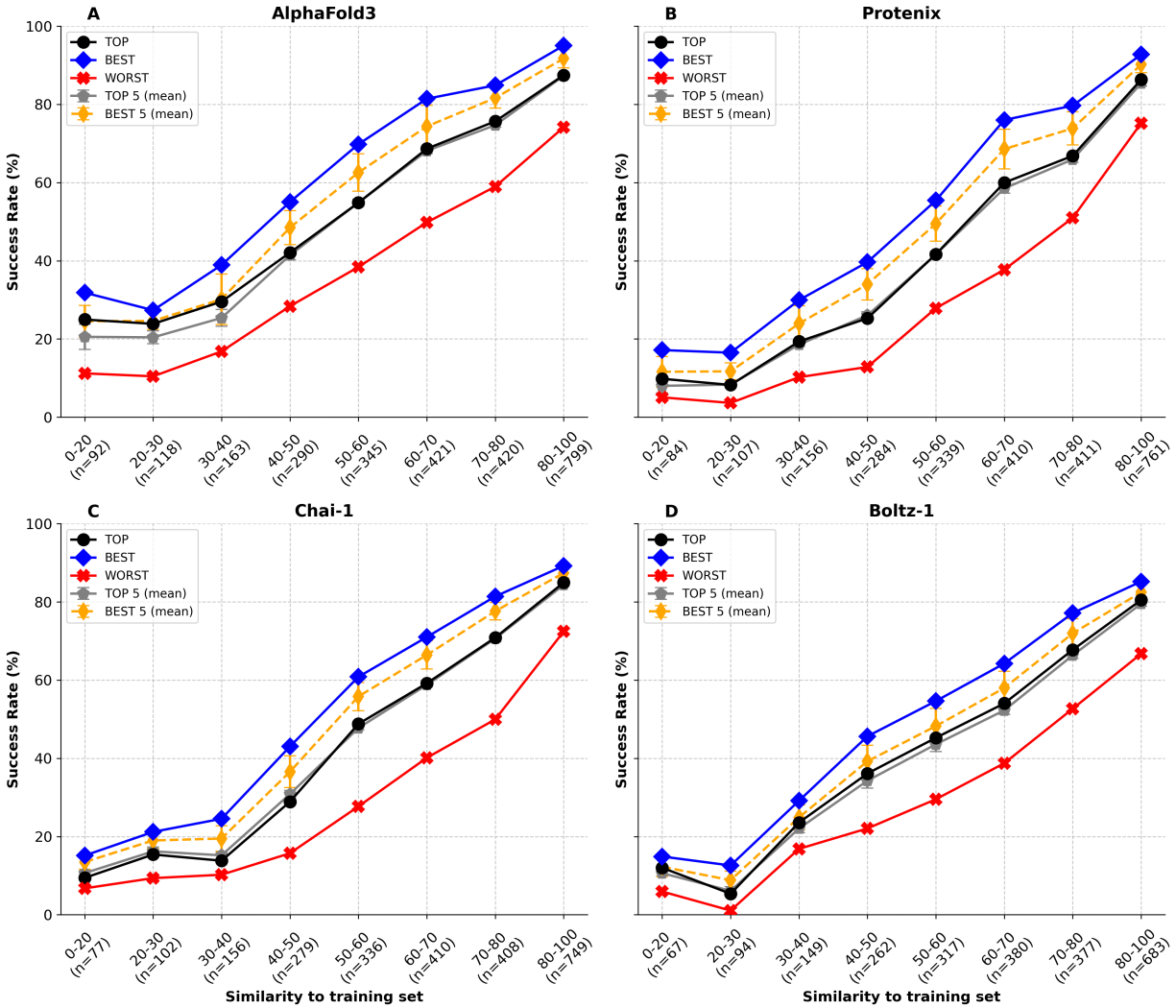


Figure 4: **Comparision of top-ranked and best-scored predicted models.** For each method, the success rate across similarity bins for the top-ranked model (black circles, with the highest ranking score), best-scored model (blue diamonds, best LDDT-PLI across 25), worst-scored model (red crosses, worst LDDT-PLI across 25), top-ranked across 1 seed (gray pentagons, with standard deviation as whiskers across the 5 seeds), and best-scored across 1 seed (yellow diamonds).

Interestingly, already from the third bin (SuCOS-pocket similarity between 30 and 40) we see that the protein and pocket of a test system share more than 60% sequence identity and 80% pocket coverage to the closest training system on average, indicating that the lack of generalisation is not just for lesser-studied proteins or novel pockets but for any kind of protein-ligand complex where that particular ligand pose has not been seen in training. Figure 2B shows the per-similarity bin distribution of LDDT-LP, i.e. the accuracy of the relative positioning of residues in the predicted binding pocket, and 2C shows the RMSD of the pocket backbone atoms. Although there is some increasing trend across the bins, the majority of predictions have an LDDT-LP > 0.8, again indicating that the protein is well-modelled and the pocket is almost always found. Thus, despite our dataset mainly containing well-studied proteins and pockets, co-folding methods are not able to predict poses of ligands within them unless a very similar ligand pose is already seen in that pocket during training.

Figure 3 shows some examples of successful and failed predictions from different similarity bins along with the ground truth complex and closest training system, found based on the SuCOS-pocket similarity score. As the similarity to the training systems increases, we observe, as expected, a growing overlap in terms of molecular shape and common substituents between the target ligand and the one found in the closest training system. However, in some cases, important chemical features are overlooked during inference, as exemplified in Figure 3E. Despite being explicitly

defined in the input SMILES, all the methods place the 5-chlorobenzofuran-3-aminomethyl group with incorrect chirality, perhaps mimicking the 5-chloro-2-methoxyphenyl group found in the closest training system. Regarding proteins, globally, we observe that their structures and binding pockets are accurately predicted across all similarity bins, as depicted in 2B. This is not always the case, as Figure 3C shows an example where also the protein is mispredicted - the ground truth is a protein kinase in inactive DFG-out conformation (35), however the predicted structure adopts an inactive DFG-in conformation. Although Figure 3 provides some useful visual cues on when these current methods are likely to perform well, it also paints a clear picture of their inherent limitations in generating accurate predictions for drug-like molecules beyond the training data.

To de-correlate the effect of predicting poses versus ranking them, we look at the accuracies for all 25 models (5 seeds, 5 diffusion samples) for each system across the four methods. Figure 4 depicts for each method and SuCOS-pocket similarity bin the success rates of the best-scored and worst-scored models among the 25 (i.e. those closest and furthest from the ground truth), the top-ranked model selected based on each method's ranking scores, and the distribution of top-ranked and best-scored models across the 5 seeds, to demonstrate the impact of using multiple seeds. While the top-ranked model is better than the worst-scored model generated (black vs. red), there is still a gap to the best possible model (in blue). Additionally, we observe that using multiple seeds is currently not useful at inference time for any of the methods as the ranking picks similar models when using 1 or 5 seeds (as seen by the black line overlapping with the gray line). Further analysis of using chain pair iPTM scores for filtering models is presented in Appendix 8.1.

3 Discussion

Overall, our benchmark shows that the current generation of deep learning co-folding methods still have a long way to go to accurately model unseen protein-ligand complexes. While all methods achieve higher accuracy at modelling the protein and its pocket, ligand pose prediction remains the primary challenge. We demonstrate that the performance of current approaches strongly correlates with the similarity to their training data, regardless of the metric used to define success or the subsets considered. As seen in Figure 1, most released complexes have a high similarity to the training set (75% system ligands have >50 SuCOS-pocket similarity), demonstrating that difficult test systems are underrepresented in general, and the commonly-used time split approaches would not capture these. The results on promiscuous ligands found in multiple training systems across diverse pockets indicate that data scarcity remains a major challenge for these methods. Given the vastness of chemical space, we are far from the optimal data regime required for deep learning models to generalise effectively. Moreover, these methods sometimes fail to account for crucial information during inference, such as chirality, underscoring the need for post-processing validation of their predictions. Finally, our analysis suggests that both sampling and ranking contribute to the limitations of these methods. In the cases where the prediction failed, the correct pose is either never sampled, or even when it is, the methods fail to select it during ranking, further highlighting the need for improvement in both aspects.

Data augmentation strategies such as cross-docking (36; 37), direct usage of underlying electron densities (38; 39; 40), or using small molecule data in isolation (41; 42) may help improve the diversity and coverage of training data across chemical space as well as across different pairs of pockets and ligands. In addition, a large number of experimentally resolved protein-ligand complexes, binding assays useful for cross-docking, and negative binding data are available within pharmaceutical companies and closed-source settings. A collective effort to make such industry data available for training, perhaps through infrastructures for distributed privacy-preserving federated learning such as the MELLODDY project (43) could greatly improve the data scarcity situation.

However, even with improved data strategies and increased dataset diversity, it remains unclear whether the optimal architecture has been developed for this task or if current methods would fully exploit these advancements. While protein structure prediction methods benefited from co-evolutionary signals extracted from billions of sequenced proteins to learn an approximate energy landscape (44; 45), all-atom co-folding methods do not currently rely on any such external non-structural data for encoding and representing ligand interactions. Incorporating physics-based terms to more accurately model protein-ligand interactions, potentially from simulations, conformational ensembles, or other sources, are likely needed to achieve more exciting results in this field. Our results are also in line with recent work demonstrating, through adversarial examples, significant divergences from expected physical behaviors in protein-ligand predictions (46). This highlights the need to integrate robust physical and chemical priors to enhance method generalisation.

Interestingly, AlphaFold3 still seems to have a slight edge over the alternatives, despite all following roughly the same architecture and training paradigms. This may also potentially be due to differences in training times, dataset sampling strategies, validation sets, and other methodological choices between the four methods. Chai-1 additionally makes use of ESM embeddings for protein featurisation, though this does not seem to impact protein-ligand structure prediction. Another difference is that AlphaFold3 is the only method out of four to make use of templates by default for the protein

modelling. We aim to re-run AlphaFold3 without templates in the next update of this manuscript, although we don't expect this to have a substantial impact on the outputs, since ligand information from the template is not utilised by the method. Finally, in order to directly compare methods to each other, the main analysis focused on the common subset of systems that were predicted by all four methods. To make full use of the dataset and increase this common subset, we aim to re-run modelling failures caused due to wrong input specifications, timeouts, and out-of-memory issues, but would still have cases where certain methods fail due to varying steps in their data processing and prediction pipelines.

All in all, our findings stress the importance of specialised benchmark datasets and evaluation techniques tailored to deep learning PLI prediction. In fact, of the four benchmarked methods only the authors of Protenix list memorisation artifacts as an expected limitation, further underscoring the difficulty in uncovering such biases without extensive benchmarking combining measures of prediction performance with measures of data leakage. In this regard, Runs N' Poses builds upon existing efforts, particularly PLINDER, which was designed to provide resources for training and evaluating deep learning PLI prediction methods. The current PLINDER-MLSB split prioritises generalisability from the outset, as reflected in validation and test sets that feature different proteins, pockets, and ligands than the training set. However, this split is not readily applicable to deep learning co-folding methods due to the need for costly retraining. Additionally, from a benchmarking perspective rather than purely method development, a dataset that includes known proteins and pockets can help the community track progress at a more granular and graded scale. Therefore, we see the Runs N' Poses dataset as complementary to the PLINDER-MLSB split: the latter is more useful during method development to identify innovations with better generalisation capabilities, while the former serves to benchmark the current state-of-the-art until more generalisation capabilities can be demonstrated. We note that the only requirement to apply the Runs N' Poses benchmark on a new method is a structural training cutoff of 30 September 2021.

The future of benchmarking for deep learning methods, as we move toward more complex multidimensional tasks such as co-folding, requires different measures to assess leakage and difficulty, whether for protein-protein interactions (PPIs), protein-ligand interactions (PLIs), or protein-nucleotide complexes. This is also particularly crucial for tasks with even more limited data, such as those involving covalent bonds and modified residues. In this context, we look forward to input from the community on how to best support innovation in this field and strive towards more accurate methods, along with fairer ways of assessing their capabilities.

4 Limitations and Future Plans

We attempted to perform this benchmark in the broadest and most inclusive way possible, to include enough systems, criteria, and evaluation strategies to enable exploration and use of this data by future methods while still remaining relevant for downstream applications such as drug discovery and design. Here we list a few unavoidable limitations that remain and future analyses that will be added to this work to increase its applicability and re-usability.

First, selecting the Runs N' Poses dataset had a number of considerations influenced by the available data. We were restricted to use only a fraction of PDB entries released after the training cutoff in a large part due to the X-ray validation criteria required to ensure reliable ground truth for evaluation (19). While this criteria is still not as strict as some recommendations (47), it could be further loosened in future efforts by combining electron density information in evaluation to enable using more complexes for benchmarking, especially those with novel binding modes, to better explore what these methods learn. For some systems, the SMILES for the ligand in the PDB does not contain chiral information while the ground truth has at least one chiral center, i.e the molecule given as input to each method may not be the molecule expected as the output. This is not a straightforward issue to fix in an automated fashion, as the ground truth may not have all the atoms resolved and so cannot be the source of truth for extracting the input SMILES. While technically in Runs N' Poses we ensured that all atoms are resolved, this is not true for many PDB entries, indicating that these methods are also trained on such cases where the chirality information in the input does not match the output. A further 212 system ligands from 202 systems could not be analysed as they could not be read with RDKit, and thus SuCOS scores could not be obtained. Finally, as the dataset was constructed in an automated fashion, there may be a few complexes which do not qualify as biologically relevant. We expect that careful inspection and continued use will reveal such cases, but since we demonstrate the robustness of the trends discussed across various stratifications, we do not expect this to alter the conclusions.

Second, despite our best efforts, we cannot be sure that we are always able to accurately calculate similarity to the training set, as our methods rely on a combination of alignment and superposition which may fail in edge cases. Also, we only search for similarity against PLI systems, excluding only-artifact systems and those with oligopeptides, oligosaccharides or oligonucleotides over 10 in length, thus potentially missing cases of training set proteins complexed with oligo, protein or nucleotide chains which resemble small molecule binding. However, this latter similarity would indicate transfer of binding modes from different modalities and perhaps may not be considered memorisation, at least to the extent that we see here.

We did not manage to obtain prediction results from all methods across all 2,600 systems in the dataset. This was due to varying reasons, some fixable and some not. 128 systems were unsuccessful across all four methods due to (1) incorrect specification of modified amino acids, (2) due to reaching memory limits in A100 40G GPUs, and (3) failures in the conformer generation step. The first two cases will be rerun with corrected inputs and higher memory GPU nodes respectively. The last case will be inspected to see if there are workarounds or if these systems would need to be dropped from the dataset. In addition, the different methods failed on different subsets of systems, with Boltz-1 having the highest number of missing predictions. These will also be inspected and re-run. Thus, the common subset is expected to increase in later revisions, allowing for more thorough comparisons.

Though our main focus so far was on the accuracy of the predicted coordinates, we aim to expand our evaluation with PoseBusters (20) physicochemical plausibility checks and protein-ligand interaction fingerprint recovery (48), to provide a more comprehensive benchmark taking all aspects relevant to downstream applications into account. Additionally, we plan to examine the accuracy when predicting non-ligand bound (*apo*) and ligand-bound (*holo*) structures with different degrees of conformational changes, thus focusing on the effect of context information on protein modelling (49).

Based on these results and preliminary findings from additional analyses, we aim to include more experiments in different directions. First, we already see some improvement of the best-scored model compared to the top-ranked one in Figure 4, and would like to explore the extent of this improvement by increasing the number of seeds for predictions, quantifying the gap between prediction and ranking. In addition, we would like to see whether the ranking score or chain-pair iPTM score is also biased by the similarity of the predicted pose to an existing pose in the training set, and thus will add SuCOS-pocket similarities of each prediction pose to the closest training system (as opposed to the ground truth) as an additional metric for analysis. Finally, comparing SuCOS-pocket similarities of RDKit-generated initial conformers to the training set will also reveal potential dependency of prediction performance on the conformer generation step. While SuCOS similarity proved most relevant for this analysis as it combines both volume and chemical features of the small molecule, it would also be interesting to disentangle these two aspects and inspect how much these methods make use of simple volume overlap over chemical similarity. Additionally, in cases where there are multiple training systems containing the same protein and ligand but with the ligand binding in different conformations, it would be interesting to know if these methods choose one pose to output or if they combine substructures from different poses.

Finally, to keep track of both established methods and progress in the field, we aim to include template-based and physics-based docking baselines. This is not straightforward, as physics-based methods used for this task often consider receptor protein structures to be rigid, and providing the ground-truth protein structure to these methods (a.k.a "re-docking") would not be a fair comparison to co-folding. Therefore, we will perform template-based and physics-based docking using similar systems from before the training cutoff as templates for homology modelling. While this template reliance may still lead to a marginally increasing trend across similarity bins, we expect this baseline to have a much flatter slope in Figure 1, as physics-based docking is typically not reliant on training sets (19).

5 Data availability

We provide the entire dataset for benchmarking use at doi.org/10.5281/zenodo.14794785, including the sequences, SMILES, and MSAs used as input for each prediction, and the ground truth CIF and SDF files for each system. For each method, we provide CSV files containing the prediction accuracies (ligand RMSD, LDDT-PLI, pocket backbone RMSD, and LDDT-LP), and confidence scores (ranking score, ligand-protein and protein-ligand chain pair iPTM) for all 25 models across all systems. Further, we provide a CSV file annotating systems with all similarity metrics to the training system with the highest SuCOS-pocket similarity, and the similarity bin that the system falls into, along with a Parquet file with all the similarities found to every training system. Scripts to recreate the figures in this manuscript, prepare inputs for the different methods, and evaluate predictions can be found at github.com/plinder-org/runs-n-poses. All predicted models along with associated metadata will be made available shortly in ModelArchive (50). An ML-ready version of the benchmark can be found on Polaris Hub (51) at https://polarishub.io/benchmarks/plinder-org/runs-n-poses.

6 Acknowledgements

We thank the SWISS-MODEL development team and other members of the Schwede group for technical support, useful suggestions, and text revisions, Vladas Oleinikovas, Lukas Jarosch, Seohyun (Chris) Kim for insightful discussions and comments, Cas Wognum for the Polaris set-up, and sciCORE at the University of Basel (https://scicore.unibas.ch/) for providing computational resources and storage space. This work was supported by funding from the SIB Swiss Institute of Bioinformatics (https://www.sib.swiss/), the Biozentrum of the University of Basel (https://www.biozentrum.unibas.ch/). This work uses outputs from AlphaFold3, subject to the AlphaFold 3

Output Terms of Use (version 3.0.0 [commit 2ffe43f], accessed on 15 Nov. 2024). All AlphaFold3 calculations and analyses were carried out by and using the weights provided to Jérôme Eberhardt.

7 Methods

7.1 Dataset curation

We ran the PLINDER ingestion pipeline (29) on all entries in the Protein Data Bank (PDB) (52) released after 30 September 2021. The resulting systems were further filtered to pass the following criteria:

- Solved by X-ray crystallography
- From the first bioassembly
- With at most 5 ligand chains defined within the system and at most 5 protein chains within the bioassembly
- No systems with covalent ligands
- No systems with all ligands marked as cofactors or oligo-peptides/nucleotides/saccharides
- Passing the following X-ray validation criteria (as defined for the PLINDER benchmark dataset) to ensure that they are reliable ground truth for evaluation:
  - Resolution $\leq$ 3.5Å
  - R-factor $\leq$ 0.4
  - $R_{free} \leq 0.45$
  - $R - R_{free} \leq 0.05$
  - Coordinates exist for all heavy atoms in the ligand and pocket
  - No alternative configurations in the ligand and pocket
  - No clash outliers for the ligand
  - No crystal contacts detected for the ligand
- Passing the following statistics criteria to ensure that they are relevant ligand interactions:
  - Having between 3 and 50 detected PLIP (53) interactions
  - Having between 5 and 100 protein residues within 6Å of the ligands in the system
  - The biggest ligand in the system having a molecular weight between 200 and 800

Systems passing these filters were further clustered to 80% sequence identity, and the system redundancy cluster was defined as the combination of the sequence cluster and the set of CCD (54) codes of the "proper" ligands in the system, i.e ligands not labelled as ions or artifacts. Only one system per redundancy cluster was chosen, the one with the maximum number of PLIP (53) interactions detected.

These filtering steps resulted in 2,600 systems used as inputs for the four prediction methods (Table 1). For systems where there were polymer chains present in the bioassembly which were not part of the system definition, we used all protein chains in the bioassembly and all ligand chains part of any PLINDER system involving that bioassembly as input for the four methods. While all ligands in the pocket were used for modelling (including ions), the accuracy metrics of only on proper ligands were analysed, only for the ligand and protein chains defined for the ground truth PLINDER system, taking the best-scored when multiple mappings were found. Results are presented per proper ligand and no aggregation was done for multi-ligand systems, resulting in 3,047 proper ligands across all the analysed systems.

For each system, the number of training systems with ligand analogs was calculated by counting systems before the training cutoff containing ligands with >0.9 ligand Morgan fingerprint Tanimoto similarity (32).

7.2 Running all-atom prediction methods

All four methods (AlphaFold3, Boltz-1, Chai-1 and Protenix) were run on one A100 40 GPU node each with 10 recycling steps, 200 diffusion sampling steps, 5 seeds and 5 diffusion samples per seed, which resulted in a total of 25 models per system for each method.

The input files were prepared according to the instructions for each method. Scripts for input file preparation can be found in our associated GitHub repository. To ensure that the prediction outcomes had no reliance on the depth or quality of MSAs generated by different pipelines, we used the same MSAs as input for all four methods. The standard AlphaFold3 MSA generation pipeline was run to obtain the `non_paired` and `paired` MSAs for each system. These

were given as is to Protenix in .a3m format with the `pairing` and `non_pairing` options, with `pairing_db` set to `uniprot` as this is the format AlphaFold3 uses for their paired MSA. For Chai-1, both paired and non-paired a3m files were used to generate the aligned.pqt using the provided `a3m-to-pqt` function, with taxonomy identifiers extracted from UniProt for the paired MSA. The same pairing keys were used to generate the custom MSA CSV files for Boltz-1. For unpaired sequences, pairing keys were set to -1. Thus, each method uses the same MSAs and pairing strategies as input. In addition, Chai-1 predictions were made with the ESM embedding option enabled.

In total we had 229,887 predictions across the four methods. Table 1 details the number of systems which could be modelled by the different methods, resulting in a common subset of 2,311 proper ligands (3,288 including ions) across 2,077 systems.

No. of	Total	AlphaFold3	Protenix	Chai-1	Boltz-1
PDB IDs	2,585	2,415	2,329	2,300	2,145
Systems	2,600	2,429	2,343	2,313	2,159
Ligands (incl. ions and artifacts)	4,282	3,906	3,772	3,730	3,418
Proper ligands	3,047	2,730	2,629	2,593	2,420
Multi-ligand systems	401	296	281	275	256
Multi-protein systems	790	739	703	687	627

Table 1: **Counts of the full Runs N' Poses dataset and the obtained predictions from the four methods.**

7.3 Accuracy scoring

To evaluate the accuracy of predicted models, we use the `compare-ligand-structures` action of OpenStructure (55) (version 2.8.0), which returns the binding-site superposed symmetry-corrected ligand RMSD (hereafter referred to just as RMSD), the local difference distance test of the ligand pocket (LDDT-LP), and the local difference distance test of protein-ligand interactions (LDDT-PLI). For LDDT-PLI calculation we enable the "added model contacts" flag which further penalises protein-ligand contacts present in the model which are not in the reference. Note that, in case of multiple identical ligands in a system, the mapping between predicted and ground truth ligands with the best scores may be different for RMSD calculation and LDDT-PLI calculation. All of these scores are calculated per-ligand using the receptor CIF and ligand SDF files saved by PLINDER as the ground truth for each system. Only the scores for ligands labelled as proper ligands (i.e excluding ions and artifacts) were analysed, and multi-ligand systems form multiple data points in our analysis, one per proper ligand. We provide code to extract relevant scores from the OpenStructure output JSON files in our associated GitHub repository.

As we are working with co-folding methods, where both the protein pocket and ligand pose have to be accurately modelled, we redefine the commonly used "Success rate" measure as the fraction of systems with <2Å RMSD and >0.8 LDDT-PLI. Figure 5A shows the distribution of RMSDs and LDDT-PLIs across all predictions. This addresses two kinds of outliers on the top right and the bottom left.

The top right has 1.6% of predictions with LDDT-PLI > 0.8 but RMSD > 2Å, sometimes consisting of cases such as the example shown in Figure 5B. While the global superposition here would have resulted in a low RMSD, and the LDDT-PLI is high (0.86) due to many of the protein-ligand distances with one domain being maintained, the binding site superposed RMSD is high 3.72Å as the two binding site residues shown in sticks are also used to perform the superposition and the top domain is incorrectly modelled. This area also has some cases where a part of the ligand not interacting with the protein is flexible and therefore its orientation is not biologically relevant nor well-resolved. While RMSD might correct the overoptimistic results of LDDT-PLI in the former case, in the latter case LDDT-PLI might be a more meaningful measure on its own. However, as <2% of cases across all 229,887 models fall into this area and we cannot automatically differentiate between the two cases, we opted to treat these as failures.

The bottom left has 8.3% of the predictions with RMSD < 2Å but LDDT-PLI < 0.8, consisting of cases where the pose is correct but the pocket is wrongly modelled. The example in Figure 5C shows the global superposition of the two complexes, which would lead to a very low RMSD value. While the binding-site superposed RMSD shows worse results (1.63 Å), it would still be considered as a success if only using RMSD <2Å as the success criteria. LDDT-PLI on the other hand is 0.43, correctly reflecting that the protein-ligand interactions are not maintained.

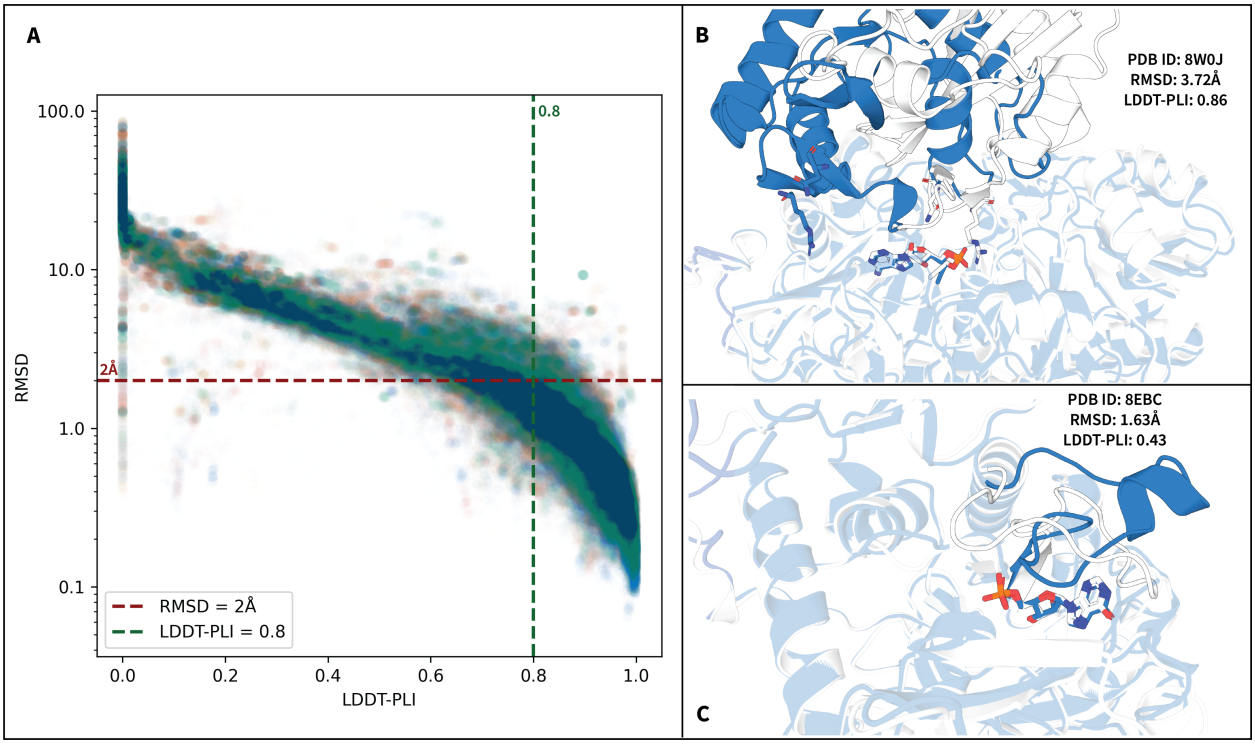


Figure 5: **LDDT-PLI and RMSD scores across all predictions. A)** The distribution of binding site superposed RMSD (log-scale) and LDDT-PLI for all 229,887 predictions with 2Å RMSD and 0.8 LDDT-PLI shown as dashed lines. **B)** An example with high RMSD and high LDDT-PLI, the binding site residues used for performing the superposition are shown as sticks in the predicted structure (blue) and ground truth (gray). **C)** An example with low RMSD and low LDDT-PLI, prediction in blue and ground truth in gray.

7.4 Similarity scoring

PLINDER calculates 14 metrics covering protein sequence (similarity and identity), protein structure (CA-LDDT); pocket coverage, identity and structure; and the similarity of PLIP interactions (29). Building on observations that the Combined Overlap Score (SuCOS), which measures the overlap of volume and chemical features between two ligand poses, has a better recovery rate than RMSD and PLI-based scores (31), we calculated three more metrics working in concert with the protein-based scores.

As SuCOS calculation requires a superposition between ligand poses, we align the ligand poses of a query and target system using RDKit's rdShapeAlign.AlignMol (56) command which returns shape and color scores from the alignment. The superposed molecules were then used to calculate the SuCOS score (30), as implemented in https://github.com/susanhleung/SuCOS. These three scores are further multiplied by the pocket coverage (`pocket_qcov`), which measures the fraction of residues within 6Å of the ligand in the query system which align with residues within 6Å of the ligand in the target system, to obtain relevant scores measuring the similarity of ligand poses within the same pocket. While approaches combining protein and pocket-level superposition were attempted, they turned out to be too sensitive to the efficacy of rigid protein superposition.

The main similarity score used in this manuscript to define training set similarity, find the closest training system as well as cluster similar systems, unless otherwise specified, is the SuCOS multiplied by the `pocket_qcov` (SuCOS-pocket). To obtain these values for each system in our set, we perform a Foldseek (57) search against the entire PDB and calculate similarities between any pair of test and train systems where any protein chain has an alignment.

Using PLINDER's `pli_qcov` scores, which counts the fraction of shared protein-ligand interaction types detected by PLIP (53) on aligning residues, gave similar results (Supplementary Figure S5C) - however, as we suspected shape-based memorisation, we wanted to account for cases where the SuCOS similarity of the ligand pose remains high but pocket residue conformational changes result in different PLIs being detected (i.e low `pli_qcov` similarity, high

SuCOS-pocket similarity), as well as cases where the part of the ligand interacting with the protein remains the same, but other regions of the ligand change in conformation (i.e high `pli_qcov` similarity, low SuCOS-pocket similarity).

The dataset was clustered using community clustering on the graph connecting systems having SuCOS-pocket similarity scores above 50. Cluster representatives are chosen as those with the lowest similarity to the training set.

Figure 6 shows examples of the kinds of ligands found in each training set similarity bin.

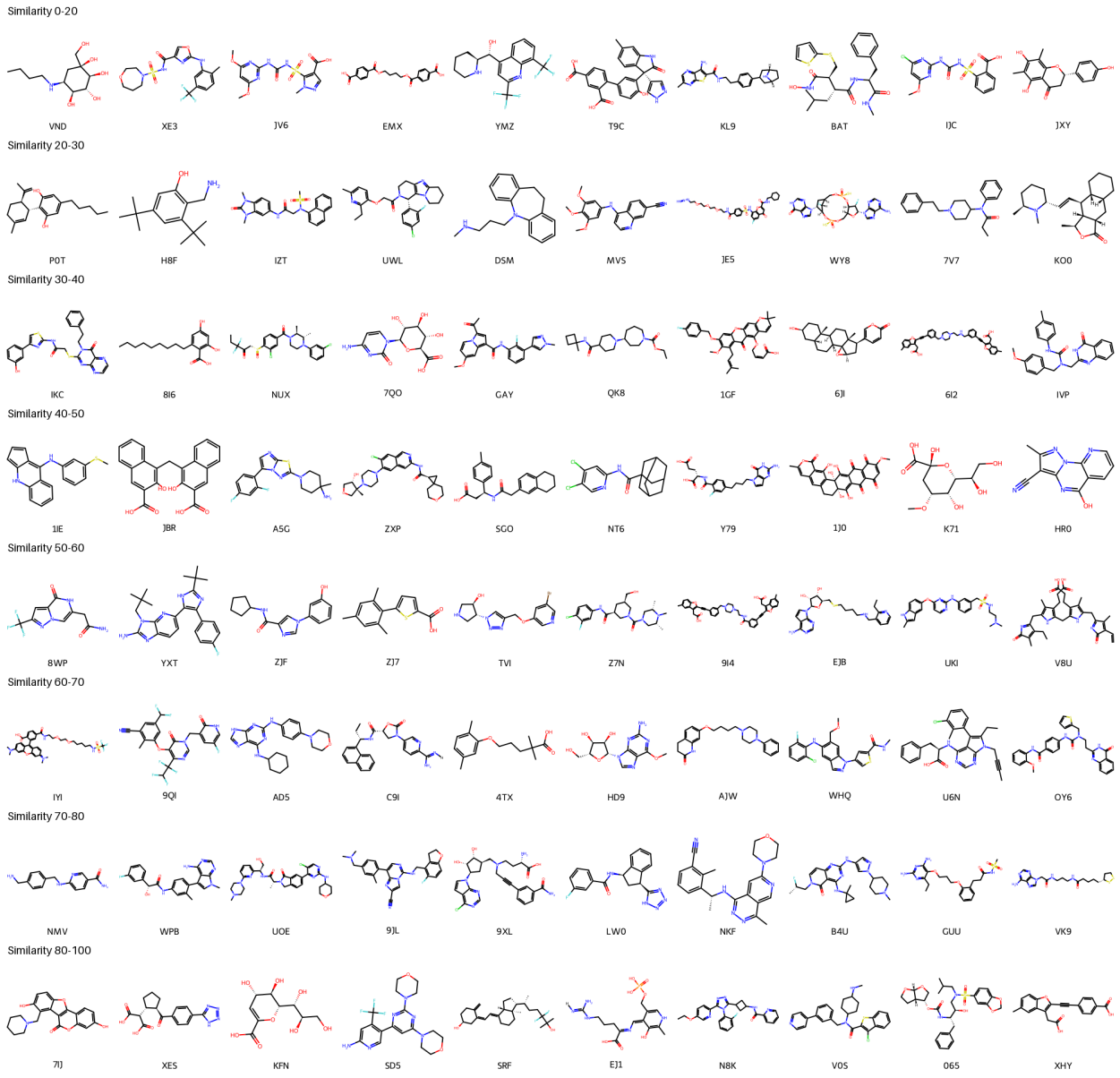


Figure 6: **Ten examples of small molecules found in each training set similarity bin.**

References

[1] Andrew R Leach. *Molecular modelling: principles and applications*. Pearson education, 2001.

[2] William L Jorgensen. The many roles of computation in drug discovery. *Science*, 303(5665):1813–1818, 2004.

[3] Ingo Muegge, Andreas Bergner, and Jan M Kriegl. Computer-aided drug design at Boehringer Ingelheim. *Journal of computer-aided molecular design*, 31(3):275–285, 2017.

[4] Alexander Hillisch, Nikolaus Heinrich, and Hanno Wild. Computational chemistry in the pharmaceutical industry: from childhood to adolescence. *ChemMedChem*, 10(12):1958–1962, 2015.

[5] Caterina Bissantz, Bernd Kuhn, and Martin Stahl. A medicinal chemist's guide to molecular interactions. *Journal of medicinal chemistry*, 53(14):5061–5084, 2010.

[6] Paweł Śledź and Amedeo Caflisch. Protein structure-based drug design: from docking to molecular dynamics. *Current opinion in structural biology*, 48:93–102, 2018.

[7] Peter Kolb and John J Irwin. Docking screens: right for the right reasons? *Current topics in medicinal chemistry*, 9(9):755–770, 2009.

[8] Hossam Nada, Nicholas A Meanwell, and Moustafa T Gabr. Virtual screening: hope, hype, and the fine line in between. *Expert Opinion on Drug Discovery*, (just-accepted), 2025.

[9] Andriy Kryshtafovych, Maciej Antczak, Marta Szachniuk, Tomasz Zok, Rachael C Kretsch, Ramya Rangan, Phillip Pham, Rhiju Das, Xavier Robin, Gabriel Studer, et al. New prediction categories in CASP15. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1550–1557, 2023.

[10] Johanna M Jansen, Wendy Cornell, Y Jane Tseng, and Rommie E Amaro. Teach–Discover–Treat (TDT): collaborative computational drug discovery for neglected diseases. *Journal of Molecular Graphics and Modelling*, 38:360–362, 2012.

[11] Jeffrey R Wagner, Christopher P Churas, Shuai Liu, Robert V Swift, Michael Chiu, Chenghua Shao, Victoria A Feher, Stephen K Burley, Michael K Gilson, and Rommie E Amaro. Continuous evaluation of ligand protein predictions: a weekly community challenge for drug docking. *Structure*, 27(8):1326–1335, 2019.

[12] Symon Gathiaka, Shuai Liu, Michael Chiu, Huanwang Yang, Jeanne A Stuckey, You Na Kang, Jim Delproposto, Ginger Kubish, James B Dunbar, Heather A Carlson, et al. D3R grand challenge 2015: evaluation of protein–ligand pose and affinity predictions. *Journal of computer-aided molecular design*, 30:651–668, 2016.

[13] Zied Gaieb, Shuai Liu, Symon Gathiaka, Michael Chiu, Huanwang Yang, Chenghua Shao, Victoria A Feher, W Patrick Walters, Bernd Kuhn, Markus G Rudolph, et al. D3R Grand Challenge 2: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *Journal of computer-aided molecular design*, 32:1–20, 2018.

[14] Zied Gaieb, Conor D Parks, Michael Chiu, Huanwang Yang, Chenghua Shao, W Patrick Walters, Millard H Lambert, Neysa Nevins, Scott D Bembenek, Michael K Ameriks, et al. D3R Grand Challenge 3: blind prediction of protein–ligand poses and affinity rankings. *Journal of computer-aided molecular design*, 33:1–18, 2019.

[15] Conor D Parks, Zied Gaieb, Michael Chiu, Huanwang Yang, Chenghua Shao, W Patrick Walters, Johanna M Jansen, Georgia McGaughey, Richard A Lewis, Scott D Bembenek, et al. D3R grand challenge 4: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *Journal of computer-aided molecular design*, 34:99–119, 2020.

[16] Heather A Carlson, Richard D Smith, Kelly L Damm-Ganamet, Jeanne A Stuckey, Aqeel Ahmed, Maire A Convery, Donald O Somers, Michael Kranz, Patricia A Elkins, Guanglei Cui, et al. CSAR 2014: a benchmark exercise using unpublished data from pharma. *Journal of chemical information and modeling*, 56(6):1063–1077, 2016.

[17] Richard D Smith, Kelly L Damm-Ganamet, James B Dunbar Jr, Aqeel Ahmed, Krishnapriya Chinnaswamy, James E Delproposto, Ginger M Kubish, Christine E Tinberg, Sagar D Khare, Jiayi Dou, et al. CSAR benchmark exercise 2013: evaluation of results from a combined computational protein design, docking, and scoring/ranking challenge. *Journal of chemical information and modeling*, 56(6):1022–1031, 2016.

[18] Xavier Robin, Gabriel Studer, Janani Durairaj, Jerome Eberhardt, Torsten Schwede, and W Patrick Walters. Assessment of protein–ligand complexes in CASP15. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1811–1821, 2023.

[19] Michèle Leemann, Ander Sagasta, Jerome Eberhardt, Torsten Schwede, Xavier Robin, and Janani Durairaj. Automated benchmarking of combined protein structure and ligand conformation prediction. *Proteins: Structure, Function, and Bioinformatics*, 91(12):1912–1924, 2023.

[20] 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.

[21] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *nature*, 596(7873):583–589, 2021.

[22] Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S Morey-Burrows, Ivan Anishchenko, Ian R Humphreys, et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science*, 384(6693):eadl2528, 2024.

[23] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, pages 1–3, 2024.

[24] Chai Discovery, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhnikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, pages 2024–10, 2024.

[25] 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*, pages 2024–11, 2024.

[26] xinshi chen, yuxuan zhang, chan lu, wenzhi ma, jiaqi guan, chengyue gong, jincai yang, hanyu zhang, ke zhang, shenghao wu, et al. Protenix-Advancing Structure Prediction Through a Comprehensive AlphaFold3 Reproduction. *bioRxiv*, pages 2025–01, 2025.

[27] Zhuoran Qiao, Feizhi Ding, Thomas Dresselhaus, Mia A Rosenfeld, Xiaotian Han, Owen Howell, Aniketh Iyengar, Stephen Opalenski, Anders S Christensen, Sai Krishna Sirumalla, et al. NeuralPLexer3: Physio-Realistic Biomolecular Complex Structure Prediction with Flow Models. *arXiv preprint arXiv:2412.10743*, 2024.

[28] Daniel Kovtun, Mehmet Akdel, Alexander Goncearenco, Guoqing Zhou, Graham Holt, David Baugher, Dejun Lin, Yusuf Adeshina, Thomas Castiglione, Xiaoyun Wang, et al. PINDER: The protein interaction dataset and evaluation resource. *bioRxiv*, pages 2024–07, 2024.

[29] 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*, pages 2024–07, 2024.

[30] Shipra Malhotra and John Karanicolas. When does chemical elaboration induce a ligand to change its binding mode? *Journal of medicinal chemistry*, 60(1):128–145, 2017.

[31] Susan Leung, Michael Bodkin, Frank von Delft, Paul Brennan, and Garrett Morris. SuCOS is Better than RMSD for Evaluating Fragment Elaboration and Docking Poses. 2019.

[32] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[33] Melissa L Boby, Daren Fearon, Matteo Ferla, Mihajlo Filep, Lizbé Koekemoer, Matthew C Robinson, COVID Moonshot Consortium‡, John D Chodera, Alpha A Lee, Nir London, et al. Open science discovery of potent noncovalent SARS-CoV-2 main protease inhibitors. *Science*, 382(6671):eabo7201, 2023.

[34] Andreas Luttens, Hjalmar Gullberg, Eldar Abdurakhmanov, Duy Duc Vo, Dario Akaberi, Vladimir O Talibov, Natalia Nekhotiaeva, Laura Vangeel, Steven De Jonghe, Dirk Jochmans, et al. Ultralarge virtual screening identifies SARS-CoV-2 main protease inhibitors with broad-spectrum activity against coronaviruses. *Journal of the American Chemical Society*, 144(7):2905–2920, 2022.

[35] Vivek Modi and Roland L Dunbrack Jr. Kincore: a web resource for structural classification of protein kinases and their inhibitors. *Nucleic acids research*, 50(D1):D654–D664, 2022.

[36] Matthew Ragoza, Joshua Hochuli, Elisa Idrobo, Jocelyn Sunseri, and David Ryan Koes. Protein–ligand scoring with convolutional neural networks. *Journal of chemical information and modeling*, 57(4):942–957, 2017.

[37] 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.

[38] Stephanie A Wankowicz, Ashraya Ravikumar, Shivani Sharma, Blake Riley, Akshay Raju, Daniel W Hogan, Jessica Flowers, Henry van den Bedem, Daniel A Keedy, and James S Fraser. Automated multiconformer model building for X-ray crystallography and cryo-EM. *eLife*, 12:RP90606, 2024.

[39] Samuel E Hoff, F Emil Thomasen, Kresten Lindorff-Larsen, and Massimiliano Bonomi. Accurate model and ensemble refinement using cryo-electron microscopy maps and Bayesian inference. *PLOS Computational Biology*, 20(7):e1012180, 2024.

[40] Minhuan Li, Kevin Dalton, and Doeke Hekstra. SFCalculator: connecting deep generative models and crystallography. *bioRxiv*, pages 2025–01, 2025.

[41] Frank H Allen and Robin Taylor. Research applications of the Cambridge structural database (CSD). *Chemical Society Reviews*, 33(8):463–475, 2004.

[42] Ayesh Madushanka, Renaldo T Moura Jr, and Elfi Kraka. QM40, Realistic Quantum Mechanical Dataset for Machine Learning in Molecular Science. *Scientific Data*, 11(1):1376, 2024.

[43] Wouter Heyndrickx, Lewis Mervin, Tobias Morawietz, Noé Sturm, Lukas Friedrich, Adam Zalewski, Anastasia Pentina, Lina Humbeck, Martijn Oldenhof, Ritsuya Niwayama, et al. Melloddy: Cross-pharma federated learning at unprecedented scale unlocks benefits in qsar without compromising proprietary information. *Journal of chemical information and modeling*, 64(7):2331–2344, 2023.

[44] Zhidian Zhang, Hannah K Wayment-Steele, Garyk Brixi, Haobo Wang, Dorothee Kern, and Sergey Ovchinnikov. Protein language models learn evolutionary statistics of interacting sequence motifs. *Proceedings of the National Academy of Sciences*, 121(45):e2406285121, 2024.

[45] Liwei Chang and Alberto Perez. AlphaFold2 knows some protein folding principles. *bioRxiv*, 2024.

[46] Matthew R. Masters, Amr H. Mahmoud, and Markus A. Lill. Do Deep Learning Models for Co-Folding Learn the Physics of Protein-Ligand Interactions? *bioRxiv*, 2024.

[47] Gregory L Warren, Thanh D Do, Brian P Kelley, Anthony Nicholls, and Stephen D Warren. Essential considerations for using protein–ligand structures in drug discovery. *Drug Discovery Today*, 17(23-24):1270–1281, 2012.

[48] David Errington, Constantin Schneider, Cédric Bouysset, and Frédéric A Dreyer. Assessing interaction recovery of predicted protein-ligand poses. *arXiv preprint arXiv:2409.20227*, 2024.

[49] Vít Škrhák, Marian Novotnỳ, Christos P Feidakis, Radoslav Krivák, and David Hoksza. CryptoBench: cryptic protein–ligand binding sites dataset and benchmark. *Bioinformatics*, 41(1):btae745, 2025.

[50] Gerardo Tauriello, Andrew M Waterhouse, Juergen Haas, Dario Behringer, Stefan Bienert, Thomas Garello, and Torsten Schwede. Modelarchive: A Deposition Database for Computational Macromolecular Structural Models. *Available at SSRN 5042665*.

[51] Cas Wognum, Lu Zhu, Hadrien Mary, Julien St-Laurent, Larissa, Andrew Quirke, Honoré Hounwanou, Kun Zhu, Shawn Whitfield, Jackson Burns, Kira Howe (McLean), Kira Howe (McLean), and felix. polaris-hub/polaris: 0.11.8, February 2025.

[52] Stephen K Burley, Helen M Berman, Gerard J Kleywegt, John L Markley, Haruki Nakamura, and Sameer Velankar. Protein Data Bank (PDB): the single global macromolecular structure archive. *Protein crystallography: methods and protocols*, pages 627–641, 2017.

[53] Sebastian Salentin, Sven Schreiber, V Joachim Haupt, Melissa F Adasme, and Michael Schroeder. PLIP: fully automated protein–ligand interaction profiler. *Nucleic acids research*, 43(W1):W443–W447, 2015.

[54] John D Westbrook, Chenghua Shao, Zukang Feng, Marina Zhuravleva, Sameer Velankar, and Jasmine Young. The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. *Bioinformatics*, 31(8):1274–1278, 2015.

[55] Marco Biasini, Tobias Schmidt, Stefan Bienert, Valerio Mariani, Gabriel Studer, Jürgen Haas, Niklaus Johner, Andreas Daniel Schenk, Ansgar Philippsen, and Torsten Schwede. OpenStructure: an integrated software framework for computational structural biology. *Acta Crystallographica Section D: Biological Crystallography*, 69(5):701–709, 2013.

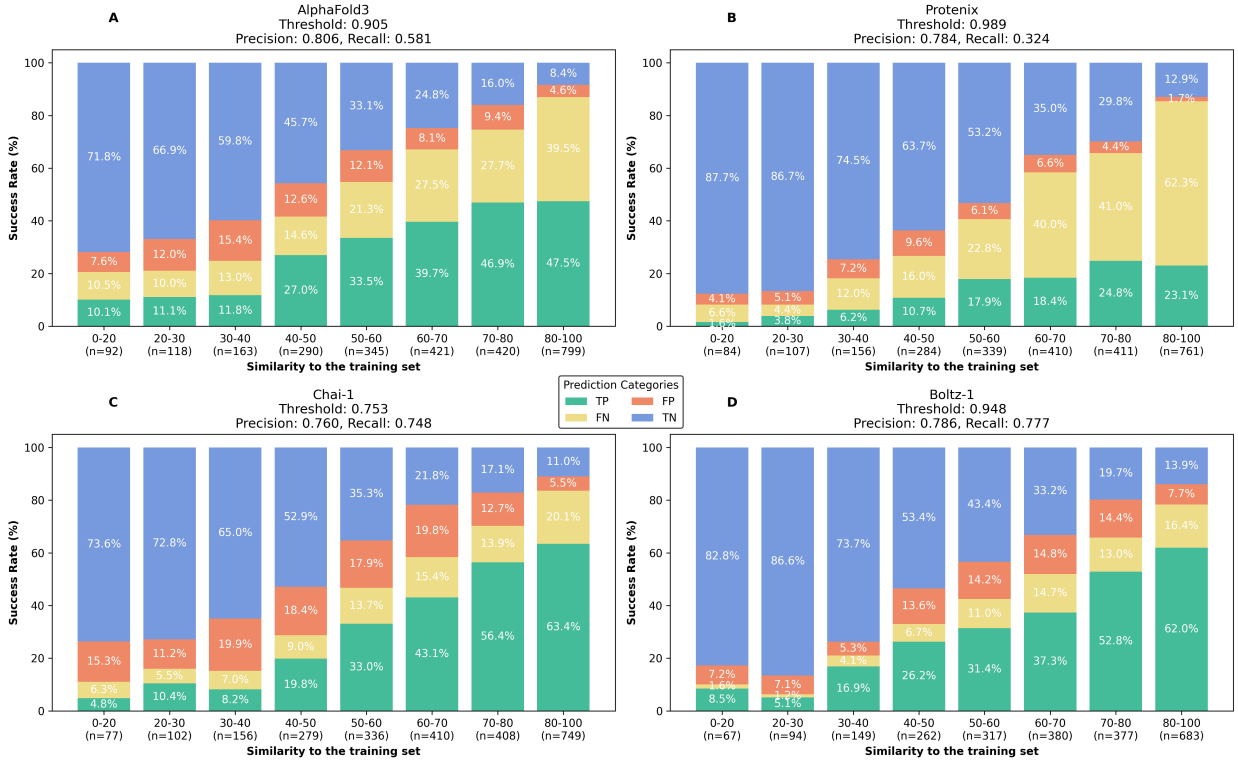
[56] Greg Landrum et al. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8(31.10):5281, 2013.

[57] Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with Foldseek. *Nature biotechnology*, 42(2):243–246, 2024.

8 Supplementary Material

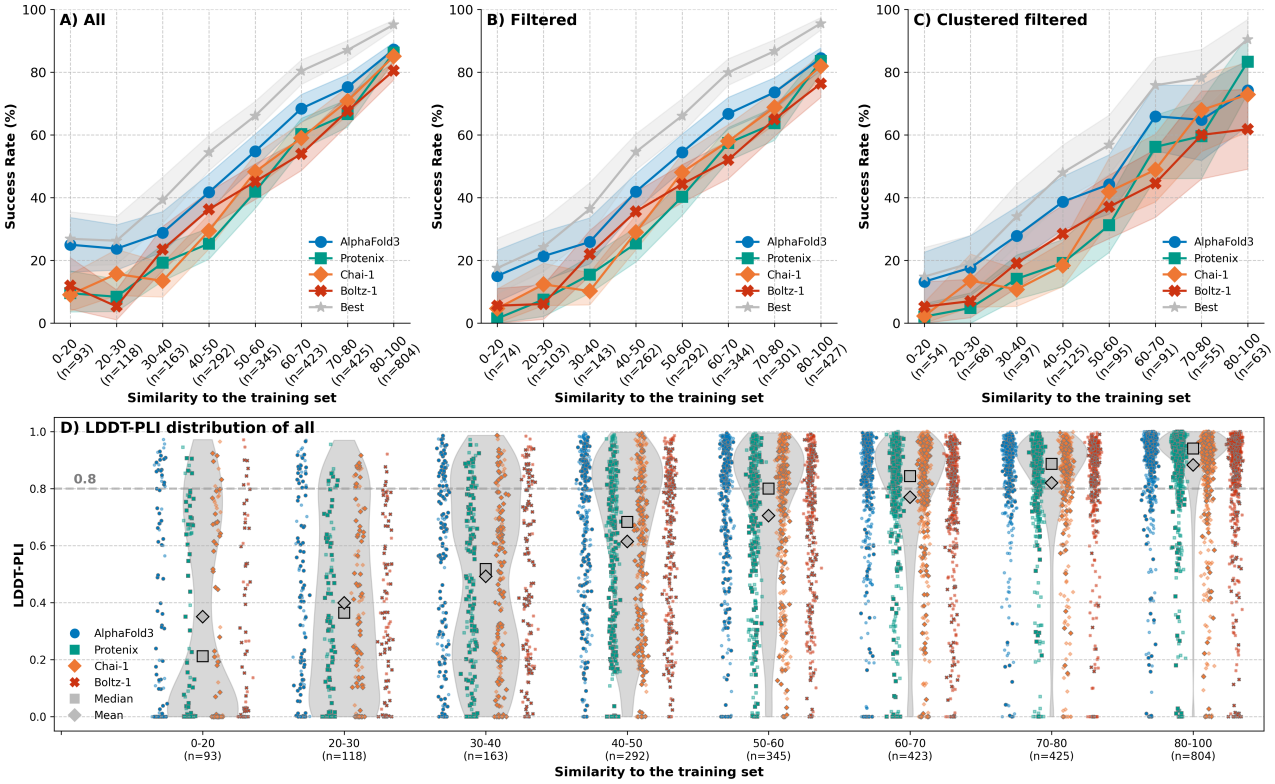
8.1 Appendix: The filtering power of chain pair iPTM

Figure S1 demonstrates the usability of the ligand-protein chain pair iPTM score for differentiating successes and failures across top-ranked models of different systems. For multi-protein systems, we calculate this score as the average of chain pair iPTM scores across all protein chains to the considered ligand chain. For each method, we independently determine the best threshold as the optimal operating point of the ROC-AUC curve across all predicted systems, representing the threshold with the best trade-off between sensitivity and specificity. This optimal threshold is different for each method, with Protenix being the most optimistic about its predictions and Chai-1 being the least. Thus, previously established TM-score thresholds cannot be used for model selection, and the confidence scores are not comparable across the methods. In addition, the precision and recall varies across the different similarity bins - for example, for the lowest bin the fraction of false positives that would be obtained varies from 46% of the selected models for Boltz-1 to 76% for Chai-1, while this drops to 7% (for Protenix) to 11% (for Boltz-1) when looking at the highest similarity bin where false negatives are more of an issue. We note that using the ranking score or using the chain pair iPTM score gives very similar results for selecting among the 25 predicted models for the same system, but, unlike the ranking score, the ligand-protein pair iPTM score can be compared across different systems.



Supplementary Figure S1: **Confusion matrices of filtering based on chain-pair iPTM scores.** For each method and each similarity bin, the confusion matrix obtained when using the chain pair iPTM score for filtering high confidence models. The optimal operating point of the ROC curve is used to determine the ligand-protein pair iPTM threshold for each method. The precision and recall across all systems is labelled.

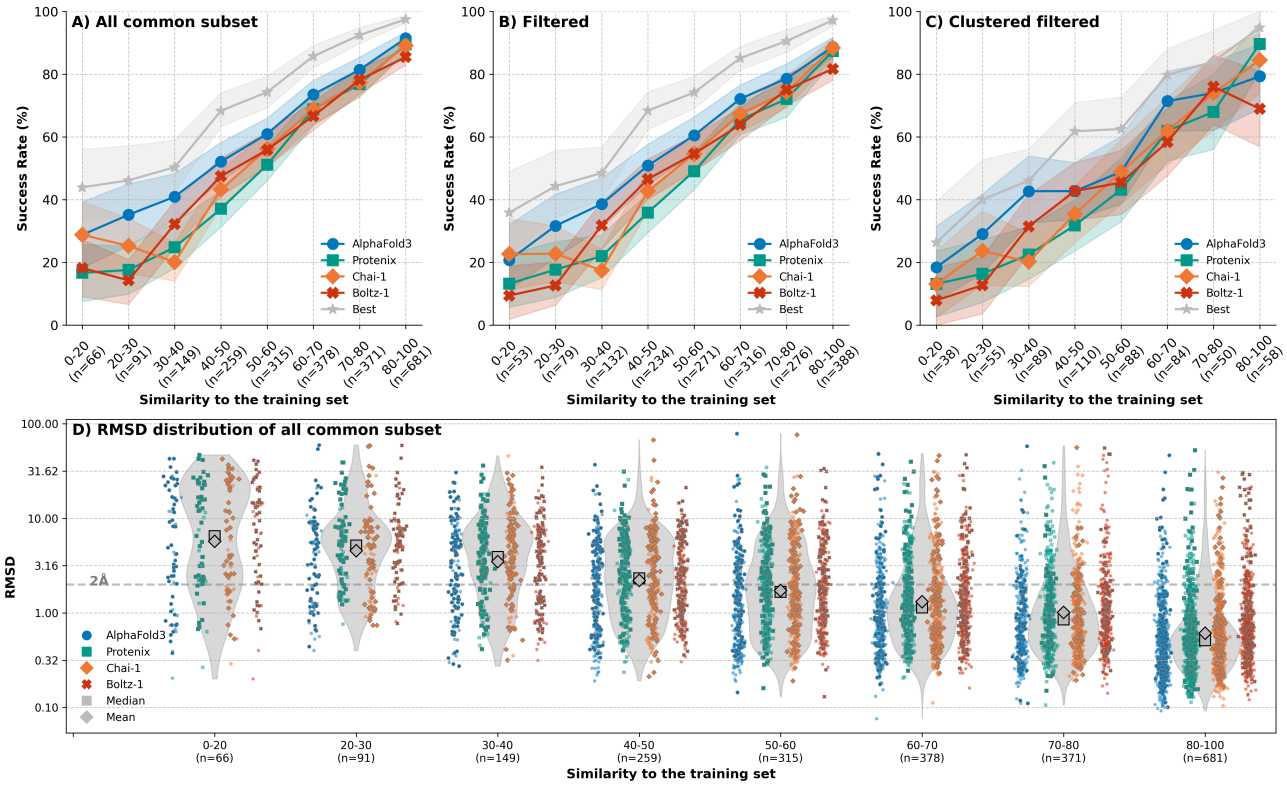
An additional peculiarity differentiating the four methods is the asymmetrical nature of the chain-pair iPTM matrix for Boltz-1 and Chai-1, where the protein-ligand chain pair iPTM is a different value than the ligand-protein chain pair iPTM. This is not the case for AlphaFold3 and Protenix, both of which have symmetrical matrices. The protein-ligand chain pair iPTM shows much lower ranking power, with optimal threshold, precision and recall of 0.53, 0.711, 0.812 for Boltz-1 and 0.595, 0.733, 0.598 for Chai-1, respectively.



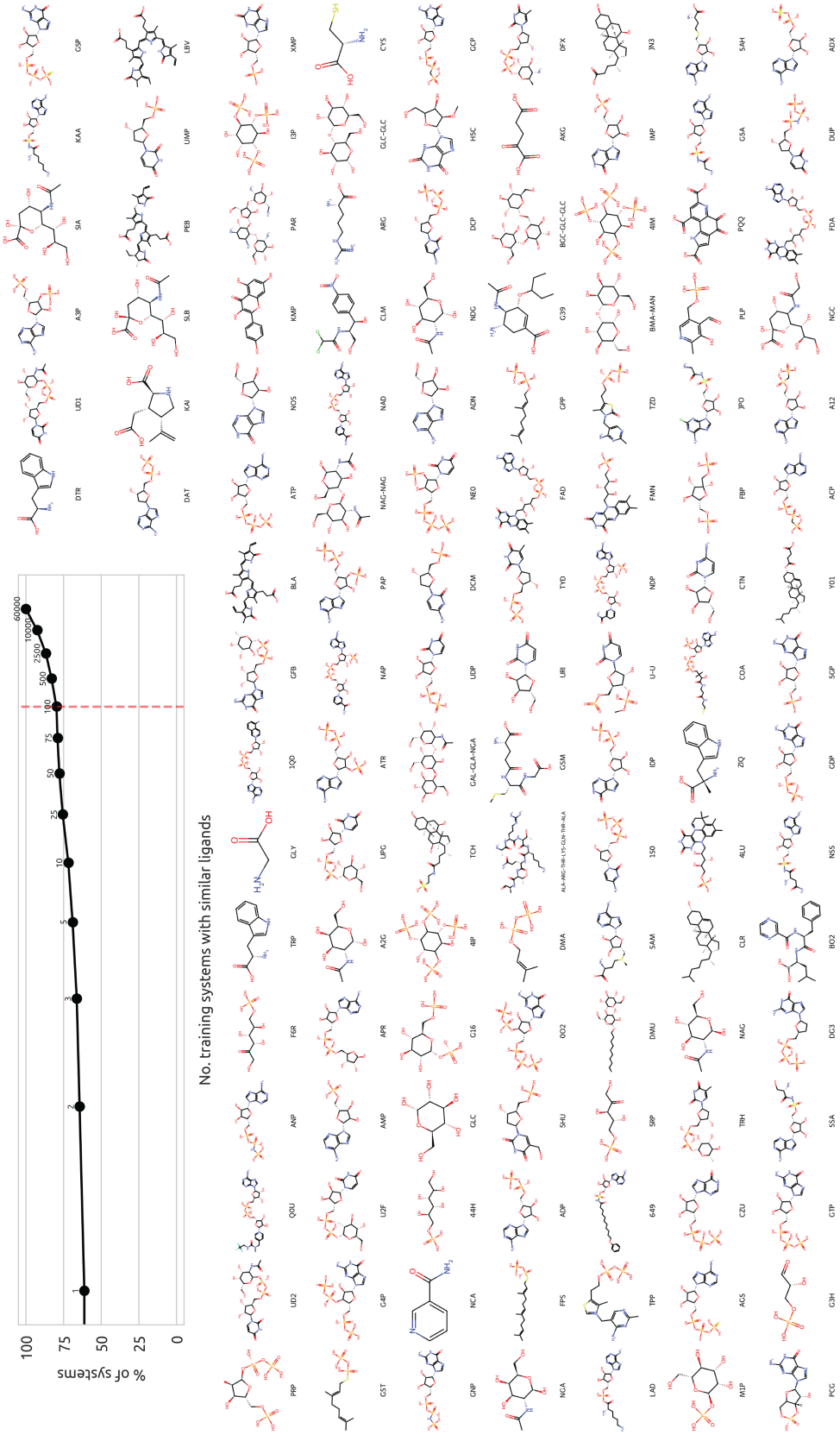
Supplementary Figure S2: **Prediction accuracy vs. training set similarity across all predicted system ligands.** The success rate, defined as the percentage of system ligands with <2Å RMSD and >0.8 LDDT-PLI, for system ligands in each similarity bin across the four methods and when selecting the best-scored model from the four top-ranked, for **A)** all system ligands, **B)** only system ligands having <100 training systems containing analogs, and **C)** only cluster representatives from B. Shaded regions correspond to the 95% confidence interval, calculated from 1,000 bootstrap samples for each bin and method. **D)** The distribution of LDDT-PLI values for all methods displayed as a gray violin plot, with each individual method being shown as colored scatter points where cluster representatives have a gray outline.

Index	Cluster Size	PDB ID	Description	Similarity bins
1	171	7EN9	Crystal structure of SARS-CoV-2 3CLpro in complex with the non-covalent inhibitor WU-02	80-100 (79), 70-80 (46), 60-70 (25), 50-60 (15), 40-50 (6)
2	136	8OKU	Salt-Inducible Kinase 3 in complex with an inhibitor	60-70 (38), 70-80 (37), 80-100 (28), 50-60 (19), 40-50 (11), 30-40 (3)
3	100	5SH0	Crystal structure of human phosphodiesterase 10 in complex with IZT	50-60 (33), 40-50 (26), 30-40 (16), 60-70 (13), 80-100 (5), 20-30 (4), 70-80 (3)
4	92	8HUU	Crystal structure of HCoV-NL63 main protease with S217622	80-100 (43), 60-70 (20), 70-80 (20), 40-50 (5), 50-60 (4)
5	73	7S3Z	Structure of rat neuronal nitric oxide synthase heme domain in complex with V0P	80-100 (51), 70-80 (16), 60-70 (5), 50-60 (1)
6	43	8JOO	Crystal structure of cytochrome P450 IkaD from Streptomyces sp. ZJ306, in complex with the substrate ikarugamycin	80-100 (23), 60-70 (5), 70-80 (5), 30-40 (4), 40-50 (3), 50-60 (3)
7	35	7G00	Crystal structure of human FABP1 in complex with WII	60-70 (13), 50-60 (10), 40-50 (6), 70-80 (3), 80-100 (2), 30-40 (1)
8	28	7UBO	Crystal structure of the first bromodomain of human BRDT in complex with the inhibitor CCD-956	70-80 (7), 80-100 (7), 50-60 (6), 60-70 (5), 40-50 (2), 30-40 (1)
9	28	7S1S	PRMT5/MEP50 crystal structure with MTA and MRTX-1719 bound	80-100 (12), 40-50 (6), 50-60 (3), 60-70 (3), 30-40 (2), 70-80 (2)
10	26	7V0H	Crystal structure of putative glucose 1-dehydrogenase from *Burkholderia cenocepacia* in complex with NADP and a potential reaction product	80-100 (13), 60-70 (7), 70-80 (6)

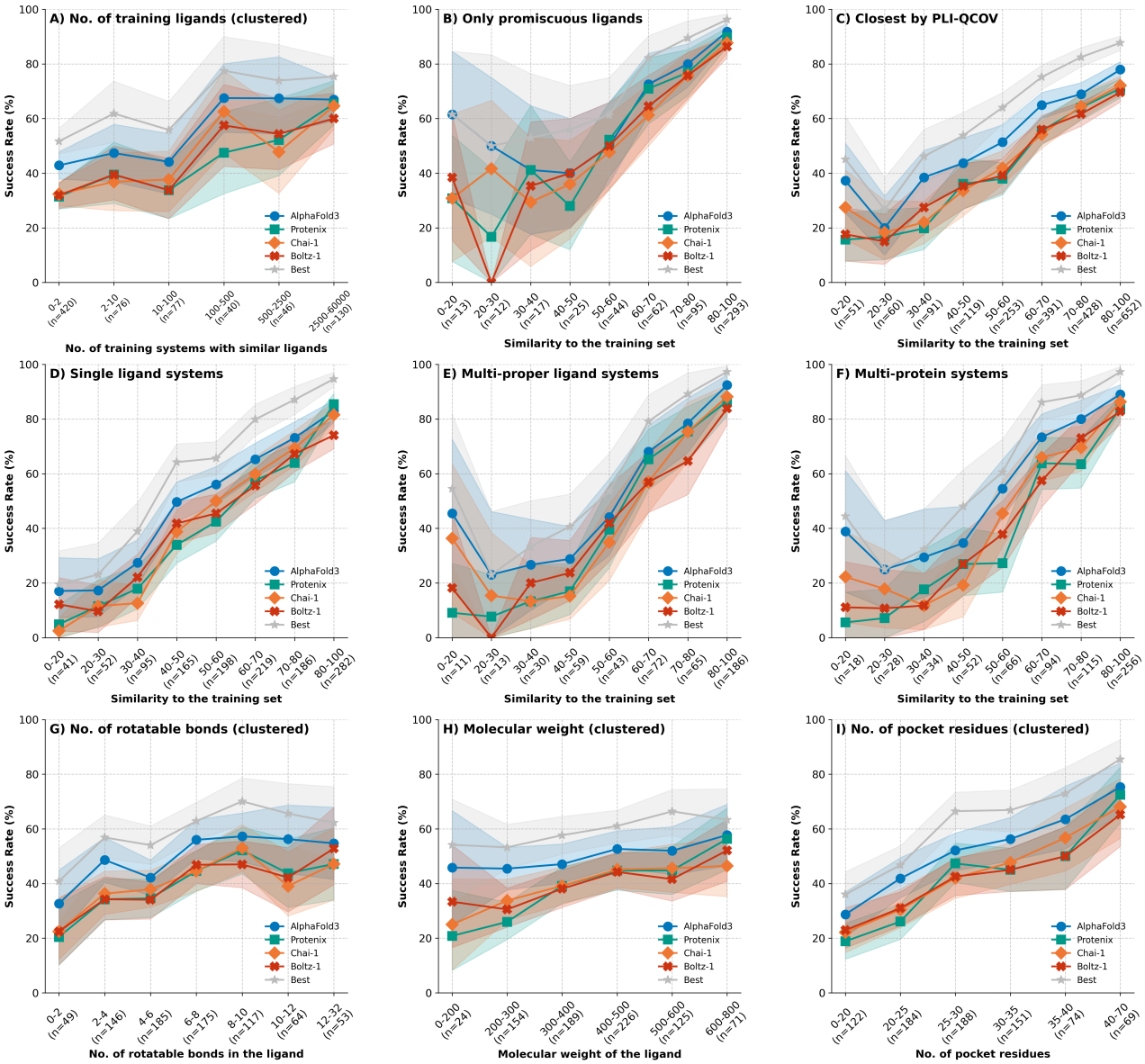
Supplementary Table S1: **Representatives of populated clusters.** Sizes of the ten most populated clusters, along with the PDB ID and description of the cluster representative with the lowest training set similarity for each, and distribution of the cluster across different training SuCOS-pocket similarity bins.



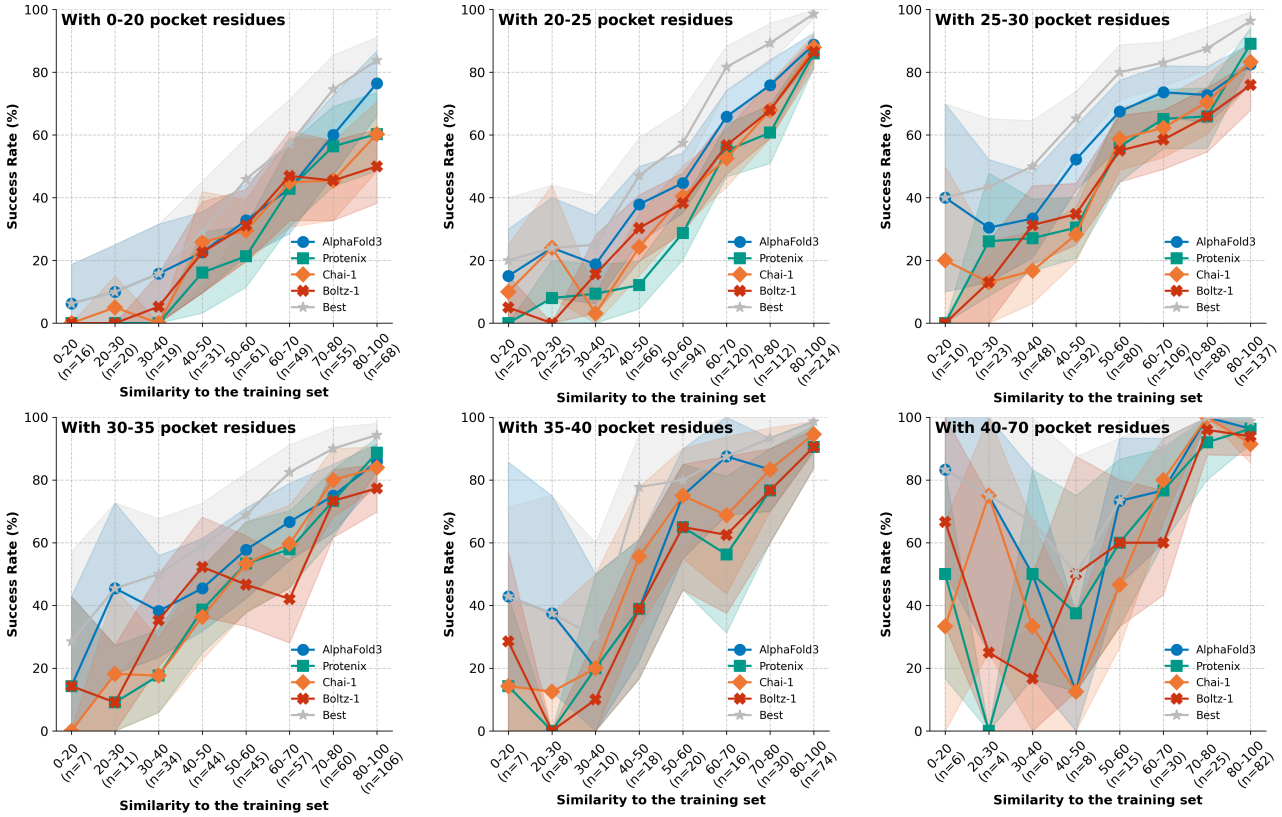
Supplementary Figure S3: **Prediction accuracy by RMSD vs. training set similarity.** The success rate, defined as the percentage of system ligands with <2Å RMSD, for system ligands in each similarity bin across the four methods and when selecting the best-scored model from the four top-ranked, for **A)** all system ligands in the common subset, **B)** only system ligands having <100 training systems containing similar ligands, and **C)** only cluster representatives from B. Shaded regions are defined by the 95% confidence interval of 1000 bootstrap samples for each bin and method. **D)** The distribution of RMSD values (log-scale) for the four methods and all together, with cluster representatives having a gray outline.



Supplementary Figure S4: **Promiscuous ligands.** The percentage of system ligands (Y-axis) having different numbers of training systems with analogous ligands (X-axis, log scale), and all the ligands within systems with >100 analogous ligand training systems.



Supplementary Figure S5: **Success rate distributions for different stratifications.** Success rates for the different methods and the best model of the four, **A)** for system ligands with differing numbers of analogous ligand training systems. **B)** across the different SuCOS-pocket similarity bins only for system ligands having >100 analogous ligand training systems. **C)** across different `pli_qcov` similarity bins when using `pli_qcov` similarity to find the closest training system. **D)** across the different SuCOS-pocket similarity bins only for systems having a single ligand. **E)** across the different SuCOS-pocket similarity bins only for systems having multiple proper ligands. **F)** across the different SuCOS-pocket similarity bins only for systems spanning multiple protein chains. **G)** for clustered system ligands with differing numbers of rotatable bonds. **H)** for clustered system ligands with differing numbers of residues in the binding pocket. **I)** across the different SuCOS-pocket similarity bins only for systems having 20-30 residues in the binding pocket.



Supplementary Figure S6: **Prediction accuracy vs. training set similarity for systems with different numbers of pocket residues.** The success rate for system ligands in each similarity bin across the four methods and when selecting the best-scored model from the four top-ranked, for system ligands having binding pockets defined by **A** 0-20, **B)** 20-25, **C)** 25-30, **D)** 30-35, **E)** 35-40, and **F)** >40 residues respectively.