Automated benchmarking of combined protein structure and ligand conformation prediction

Xavier Robin[1], Michèle Leemann[1], Ander Sagasta[1], Jerome Eberhardt[1], Torsten Schwede[1], and Janani Durairaj[1]

[1]Universitat Basel Department Biozentrum

May 11, 2023

Abstract

The prediction of protein-ligand complexes (PLC), using both experimental and predicted structures, is an active and important area of research, underscored by the inclusion of the Protein-Ligand Interaction category in the latest round of the Critical Assessment of Protein Structure Prediction experiment CASP15. The prediction task in CASP15 consisted of predicting both the 3-dimensional structure of the receptor protein as well as the position and conformation of the ligand. This paper addresses the challenges and proposed solutions for devising automated benchmarking techniques for PLC prediction. The reliability of experimentally solved PLC as ground truth reference structures is assessed using various validation criteria. Similarity of PLC to previously released complexes are employed to judge the novelty and difficulty of a PLC as a prediction target. We show that the commonly used PDBBind time-split test-set is inappropriate for comprehensive PLC evaluation. Finally, we introduce a fully automated pipeline that predicts PLC and evaluates the accuracy of the protein structure, ligand pose, and protein-ligand interactions.

Authors: Michèle Leemann[1,2], Ander Sagasta[1,2], Jerome Eberhardt[1,2], Torsten Schwede[1,2], Xavier Robin[1,2,*], Janani Durairaj[1,2, *]

Affiliations

[1] Biozentrum, University of Basel, Basel 4056, Switzerland,

[2] SIB Swiss Institute of Bioinformatics, Basel 4056, Switzerland,

* Equal contributions

Keywords : CASP15; protein structure; 3D structure prediction; protein-ligand complexes

Abstract

The prediction of protein-ligand complexes (PLC), using both experimental and predicted structures, is an active and important area of research, underscored by the inclusion of the Protein-Ligand Interaction category in the latest round of the Critical Assessment of Protein Structure Prediction experiment CASP15. The prediction task in CASP15 consisted of predicting both the 3-dimensional structure of the receptor protein as well as the position and conformation of the ligand. This paper addresses the challenges and proposed solutions for devising automated benchmarking techniques for PLC prediction. The reliability of experimentally solved PLC as ground truth reference structures is assessed using various validation criteria. Similarity of PLC to previously released complexes are employed to judge the novelty and difficulty of a PLC as a prediction target. We show that the commonly used PDBBind time-split test-set is inappropriate

for comprehensive PLC evaluation. Finally, we introduce a fully automated pipeline that predicts PLC and evaluates the accuracy of the protein structure, ligand pose, and protein-ligand interactions.

1 Introduction

The latest round of the Critical Assessment of Protein Structure Prediction experiment CASP15, held in 2022, introduced a novel category for protein-ligand interaction prediction (CASP-PLI), aiming to evaluate cutting-edge methodologies on a blind target set of experimentally resolved complexes. In contrast to typical ligand docking benchmark experiments like Teach Discover Treat (TDT)[1], Continuous Evaluation of Ligand Prediction Performance (CELPP)[2], Drug Discovery Data Resource (D3R)[3–6], or Community Structure-Activity Resource (CSAR)[7,8], the prediction task in CASP consisted of predicting both the structure of the receptor protein as well as the position and conformation of the ligand, hereafter referred to as protein-ligand complex (PLC) prediction. The evaluation results of this experiment are presented elsewhere in this issue[9], as well as the technical details and challenges encountered during the establishment of the new category as part of CASP[10]. These challenges include, (1) PLC with incomplete ligands or suboptimal quality to be used as ground truth ligand poses, (2) the need for extensive manual verification of data input and prediction output, and (3) the lack of suitable scoring metrics that consider both protein structure and ligand pose prediction accuracy, which necessitated the development of novel scores.

By integrating the insights and developments from the CASP-PLI experiment, automated systems for the continuous benchmarking of combined PLC prediction can be established. We discuss challenges and insights associated with the development of two complementary approaches for PLC benchmarking: a continuous evaluation of newly released PLC in the Protein Data Bank PDB[11], as implemented in Continuous Automated Model EvaluatiOn (CAMEO, https://beta.cameo3d.org/)[12], and a comprehensive evaluation of PLC prediction tools based on a diverse, curated, and annotated benchmark dataset of PLC.

CAMEO is a benchmarking platform conducting fully automated blind evaluations of three-dimensional protein prediction servers based on the weekly prerelease of sequences of those structures, which are going to be published in the upcoming release of the Protein Data Bank[13–15]. Since 2012, the 3D structure prediction category has been assessing the accuracy of single-chain predictions. Additional assessment categories have been implemented over time to serve the structural bioinformatics community, in particular around the assessment of quality estimates (QE). Recently, efforts were made towards the assessment of protein-protein complexes (quaternary structures) and protein-ligand pose prediction[12].

While CAMEO allows for continuous validation of newly developed methods, it is dependent on the distribution of PLC released in the PDB in a given period. Thus, CAMEO evaluation in a given time period may not be representative of the entire PLC space and method developers may not have immediate access to problem cases or specific sets of PLC where their algorithm under or overperforms. This suggests a second, complementary angle to automated benchmarking, namely the creation of a diverse dataset of PLC with representative complexes from across protein-ligand space, which would allow both global comparative scoring as well as pinpointing cases that method developers would need to address to improve their global performance. While many recent deep-learning docking methods train and validate their approach on the time-split PDBBind set[16] of PLC (where 363 protein-ligand pockets are used for benchmarking), we demonstrate that this approach has shortcomings arising from the lack of crystal structure quality verification and the lack of consistent redundancy removal.

Previous research has shown that the quality of experimentally resolved structures can vary significantly[17]. Efforts have been made to establish criteria for assessing the quality of such structures, like the Iridium criteria[18]. Comparing prediction results to lower quality structures can skew the perception of their performance, an especially important consideration when assessing deep learning-based tools which have been trained to reproduce results seen in experimentally resolved structures. Additionally, many crystal structures with ligands contain missing atoms or missing residues in the binding site, complicating their use as ground

truth.

Even in the era of deep learning, determining the difficulty of predicting a PLC still relies, to some degree, on previously experimentally resolved structures. This was exemplified in this year's CASP-PLI results[9], where template-based docking methods outperformed others due to the availability of previously solved highly similar PLC for many of the targets. Thus, incorporating the novelty of a PLC into automated benchmarking setups is crucial for a fair and comprehensive evaluation. For CAMEO, this consists of filtering out "easy" targets based on sequence and ligand information available in the PDB pre-release. For the generation of a representative benchmark set, one can additionally look at the novelty of the binding site and ligand pose on a structural level.

Proteins are inherently flexible, exhibiting a range of conformations in line with their functions. Not every observed conformation is compatible with ligand binding, and this can significantly impact the accuracy of docking predictions even when using high quality experimentally resolved structures[19,20]. These factors are further complicated by the use of computationally predicted protein structures, as previous studies indicate that even state-of-the-art methods for structure prediction are not always suited for the task of ligand docking, due to inaccuracies in conformations and side-chain positioning[21]. Moreover, some ligands have highly flexible regions that mainly interact with the solvent, where evaluating the conformation of the flexible part may not be as meaningful as the parts of the ligand forming crucial interaction with protein residues. Thus, it is necessary to develop and employ evaluation metrics that extend beyond rigid ligand pose assessments.

2 Results

2.1 Is the ground truth good enough?

To assess the distribution of high quality crystal structures of PLC in the Protein Data Bank (PDB)[22], we extracted protein, ligand, and binding pocket (defined as a 6Å radius around the ligand) information from PDB validation reports from 114,973 PLC entries in the PDB solved by X-ray crystallography for which Electron-Density Server (EDS) validation information is made available in the PDB[23–25], and which contain at least one protein chain (polymer entity) and at least one non-polymer entity (small molecule ligand or ion). Ligands present in the BioLIP artifact list were excluded[26]. This list contains 463 frequent crystallization artifacts such as solvents and buffers. It may also filter out a few biologically relevant ligands, however this is rare and we considered the trade-off acceptable for this study.

We analyzed 236,538 small molecule pockets across 75,065 PLC PDB entries and 32,273 unique small-molecule ligands, and 798,651 ion pockets across 84,215 PLC and 138 unique ions. In total, this corresponds to over a million pockets.

The authors of the Iridium dataset defined a highly stringent set of criteria regarding the quality of crystal structures, with emphasis on the suitability for pose prediction, virtual screening and binding affinity estimation[18]. These include criteria on the protein (resolution [?] 3.5A, $R < 0.4$, $R_{free} < 0.45$, absolute difference between R and R_{free} [?] 0.05) as well as ligand and pocket criteria (full density with RSR [?] 0.1 and RSCC [?] 0.9, full atom occupancies and no alternative configurations for ligand atoms and protein atoms within 6A of ligand.)

We applied the Iridium criteria to the binding pockets within our set of PLC. Only 0.3% (721) of small molecule pockets across 504 PLC and 0.98% (315) of unique small molecule ligands, and 0.66% (5,248) of ion pockets across 3,379 PLC and 35.51% (49) of unique ion ligands passed the criteria. In total, 0.58% of all pockets are acceptable according to the Iridium criteria, across 3.21% (3,686) of PLC and 1.12% (364) of unique ligands..

Thus this criteria is too stringent for both of the applications we explore. For continuous evaluation methods

such as CAMEO which runs on a weekly basis, the majority, if not all PLC would be discarded. Similarly, restricting to such a small fraction of the PDB is incompatible with creating a diverse and representative dataset of PLC for comprehensive benchmarking. We suggest alternative "relaxed" criteria with RSCC > 0.8 and >90% protein residues within 6A of ligand with RSCC > 0.8, with the remaining criteria the same as Iridium. The threshold of 0.8 for RSCC is in accordance with the widely accepted rule of thumb that 0.8 < RSCC < 0.95 are generally ok, RSCC > 0.95 indicate a very good fit, and RSCC < 0.8 indicate that the experimental data may not accord with the ligand placement[27]. Having such a set of relaxed criteria could be used as a post-filter step in the CAMEO setting and, in the latter case, the stringent Iridium criteria could be used to create the starting set with more PLC being added based on their novelty and the relaxed criteria.

Figure 1 shows the distribution of validation data values across all binding pockets as well as the selected relaxed thresholds for four criteria: resolution (Figure 1A), absolute difference between R and R_{free} (Figure 1B), RSCC (Figure 1C) and percentage of protein residues within 6A of the ligand with RSCC > 0.8 (Figure 1D). The most stringent criterion is by far the absolute difference between R and R_{free} , which removes almost 15% of the pockets.

We applied these relaxed criteria to the dataset of binding pockets. We found that 44.96% (106,357) of small molecule pockets across 36,959 PLC and 51.34% (16,568) of unique small molecule ligands passed the relaxed criteria. Similarly, 48.73% (389,217) of ion pockets across 55,594 PLC and 89.86% (124) of unique ion ligands passed. Thus, the criteria retains 47.87% (495,574) of all pockets, spread across 62.38% (71,720) of PLC and 51.50% (16,692) of unique ligands.

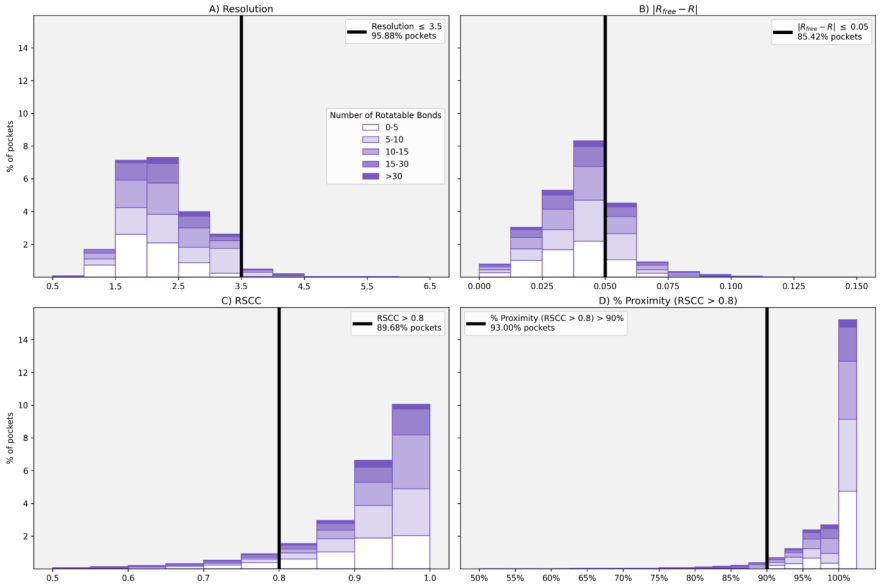

Figure 1: Distributions across PLC pockets of A) experimental resolution, B) Difference between R and R_{free}, C) Ligand RSCC, and D) The percentage of protein atoms within 6Å of the ligand, which have an RSCC > 0.8. Pockets are divided into categories depending on the number of rotatable bonds of the ligands they contain. In each panel, the black line shows the suggested threshold, and the percentage of pockets passing this criterion is displayed.

In an automated benchmarking setting such as CAMEO, where quality information is not available at the time the targets are selected, filtering out even half of the data after predictions have been generated would be unfortunate, indicating that even the relaxed criteria are too stringent as a post-filter. An alternative would be to take quality into account in the scoring process, and downweight low quality regions of a

structure in aggregate scores, without removing the entire target. Ideally an atom-level weighting would be used, especially for larger ligands that can display variable levels of quality within the residue itself. Unfortunately the PDB does not make atom-level quality information available in the validation reports at the time of writing, and the only information that would be available is occupancy numbers which are part of the structural data.

However, analyzing and incorporating validation data is a critical step towards creating a representative dataset for other benchmarking settings. For example, of the 255 small molecule pockets in the PDBBind time-split test-set, 105 do not pass the relaxed criteria, which could bias the results seen in recent benchmarking efforts using this test set. Previous efforts have been made to create high-quality subsets of PDBBind specifically for evaluation purposes[28]. However, these produced very small test sets, unlikely to be representative of the entire protein-ligand space. The stringent Iridium criteria, the suggested relaxed criteria, and the assessment of novelty and diversity described in the next section form the basis for the creation of a representative benchmark dataset. Indeed, similar efforts to create benchmark sets for PLC are ongoing in the ELIXIR 3D-BioInfo community[29]. The results of that initiative could be incorporated in this assessment once they are available.

2.2 Is a protein-ligand complex target interesting to assess?

In the context of large scale structural databases, such as the PDB, it is possible to encounter several very similar PLC or complexes with the same protein and ligand that have been crystallized in different experimental conditions or resolved by means of different experimental methods. When it comes to automated benchmarking of PLC prediction, besides the quality of the structure, an important aspect to consider is the novelty of the PLC to assess.

The CASP15 CASP-PLI assessment[9]highlighted the superiority of template-based methods to model PLC accurately. While most top predictions were produced by human groups rather than automated methods, it is likely that automated methods will in the future also leverage template information to predict PLC. Therefore, when generating a benchmarking dataset for PLC prediction, we need to ensure that PLC are not already represented in the PDB. For a challenge such as CAMEO, the exact protein conformation and the pose of the ligand within the protein complex is unknown. Thus, we will use the sequence as a proxy for protein novelty. As very similar ligands can have striking differences in their poses, and we would like to retain as many PLC as possible in the CAMEO pre-filtering setting, we use ligand names as a proxy for the novelty of the ligand pose. To that end, we investigated the novelty of the 236,538 small molecule pockets across 75,065 PLC and 32,273 unique small-molecule ligands described in section 1.1.

We assessed the novelty of PLC released every year in the PDB by verifying whether a particular combination of polymer entities and ligands was present in previously released structures. For that purpose, we performed sequence-based clustering of all polymer entities followed by the assignment of an identifier to each PLC entry, consisting of the sequence cluster identifiers of each entity and the chemical component code of the ligands present in the PLC. Using different minimum sequence identity thresholds helps reveal the level of novelty between the entities of a PLC compared to previously seen PLC. Similarly, even for PLC with identical proteins, the combination of ligands seen may differ. The distribution of sequence clusters and ligand combinations seen per year is shown in Figure 2, along with the fraction of PLC that pass the relaxed quality criteria from Section 1. For example, the four different bars for the 70-90% cluster in the year 2022 represent, in order,**(1)** all PLC released in 2022 where every entity in the PLC has 70-90% identity to every entity in a matching PLC from a previous year but the ligands are not all the same, **(2)** same as (1) but only the PLC passing the relaxed quality criteria from Section 1 **(3)**all PLC released in 2022 where every entity has 70-90% identity to every entity in a matching PLC from a previous year and the ligands are all the same, and **(4)** same as (3) but only the PLC passing the relaxed quality criteria from Section 1.

We see that, from the protein perspective, 78.85% of PLC (and 71.83% of valid PLC) released in 2022 have at least 30% sequence identity to a matching PLC from previous years (across all entities). However, most

of these (79.14%) still have different combinations of ligands, indicating that they may still be interesting to assess for PLC prediction. We consider two different minimum sequence identity thresholds, 30% for creating a diverse dataset and 90% for PLC prediction in CAMEO, and define a PLC as novel if the minimum sequence identity between any of its entities is less than the threshold in all matching PLC, or at least one ligand in the PLC is not seen in matching PLC. With this classification criteria, we found that out of all the PLC released in 2022, 4515 (83.55%) PLC were novel and 889 were redundant at a threshold of 30%, and 4833 (89.43%) PLC were novel and 571 were redundant at a threshold of 90%. Hence, even at 30% sequence identity, 83.55% of all released structures contained some kind of novelty, with at least one previously unseen protein(entity)-ligand combination. Among the PLC that passed the validation criteria, 2202 (86.76%) PLC were novel and 336 were redundant at a threshold of 30%, and 2360 (92.99%) PLC were novel and 178 were redundant at a threshold of 90%.

Thus, most newly-released PLC are novel from either the protein or the ligand perspective. However, every year some redundant PLC are also released in the range of 10-20% redundant structures per year, out of which more than half are highly redundant structures (90-100% sequence identity and same ligands). The PDBBind time-split test-set also suffers from a high degree of redundancy, with 62% of the test-set proteins having >90% sequence identity with other test-set proteins and 59% having >90% identity to proteins in the training-set. This indicates that this set would not be able to accurately represent protein-ligand space, even if all the ligands were chemically dissimilar, which is not the case.

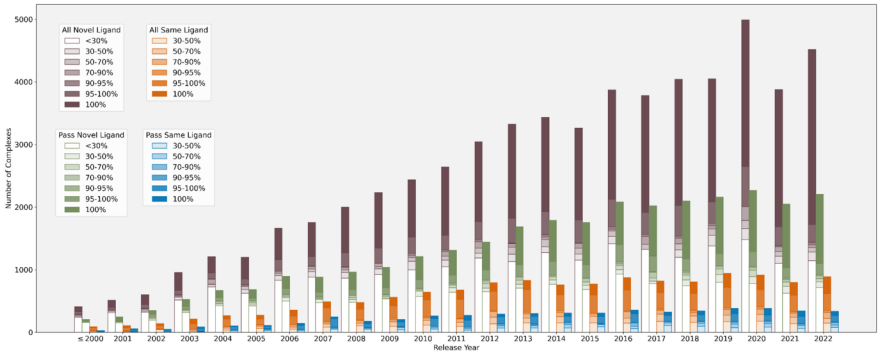

Figure 2 : Protein-ligand complexes (PLC) released per year (in brown and orange) and those passing the relaxed quality criteria (in green and blue), divided according to sequence identity to PLC seen in previous years. The left two bars of each year (in brown and green) are PLC with ligand combinations which differ from previous PLC, and the right two bars (in orange and blue) are PLC containing the same set of ligands as a matching PLC at that sequence identity.

This approach can be used in CAMEO to select the set of PLC to send out for prediction, without sacrificing too many PLC and ensuring that predictors do not waste resources on previously seen PLC or those with very similar templates. However, this approach has some shortcomings mainly due to the limited information available to CAMEO when selecting targets, namely the unique protein sequences and ligand chemical identities.

First, highly redundant regions or pockets in a PLC might be classified as novel due to the presence of other novel pockets in different areas of the complex. On the other hand, small molecule binding poses, even for the same or very similar chemical compounds, can vary significantly even within the same protein due to different protein conformations or a small number of mutations in crucial binding regions. This cannot be accounted for in the CAMEO pre-filtering step but is useful information for evaluation and highly necessary for representative dataset creation. Therefore, utilizing structure and binding pocket clustering from the protein side and 3D ligand conformation clustering from the small molecule side is recommended. The same considerations apply to the oligomeric state of each entity and the stoichiometries of each ligand in a PLC,

information that is not available from the PDB pre-release. These factors are particularly important when the same ligand is present in different protein pockets or in cases where a ligand is involved in protein oligomerization. Therefore, this information must be incorporated for assessment and when creating a representative benchmark dataset, and will be explored in future efforts.

2.3 Can we automatically score predicted protein-ligand complexes?

We developed an automated benchmarking workflow, consisting of two components: **(1)** Preprocessing, input preparation, set-up and running of five PLC prediction tools (Autodock Vina[30,31], SMINA [32], GNINA [33], DiffDock[34], and TankBind[35]) with different input parameters, and **(2)** Assessment of PLC prediction results using different scoring metrics. The workflow is implemented using Nextflow[36] to enable efficient parallelization and distributed execution, making it well-suited for handling large datasets and computationally intensive tasks. Each process is encapsulated in a module, with dependency management controlled using Conda[37] or Singularity[38] . The resources for each step in the pipeline are defined individually, ensuring that only the required resources are reserved and failed processes are automatically restarted with increased resources. Upon completion, all the predicted binding poses are collected and a summary of scores is created, along with reporting on resource usage across the evaluated tools.

We run this workflow using the PDBBind time-split test-set of 363 protein-ligand pockets. As the two most recent deep learning tools in our set, TankBind and DiffDock, are trained on the remaining proteins in PDBBind, this is the most fair set to use for their evaluation at the current time. However, it is important to emphasize that the aim of this experiment is to demonstrate the feasibility of an automated benchmarking workflow, and not a comprehensive evaluation of the tools, due to the issues in this test set already discussed in the previous sections.

As these tools already take a protein structure as input and we are interested in extending this to settings where also the structure may be computationally modeled or in a different conformation, we also evaluated PLC prediction results on 256 AlphaFold[39]structures of monomeric proteins from the same test-set. 77% (197) of the AlphaFold models are within 2 Å RMSD of the crystal structure.

In order to demonstrate the workflow in different input settings, we use P2Rank [40]to detect pockets in each protein in the test set and report results in two scenarios: *Blind docking* , which is considered the worst-case scenario for docking tools where no indication is provided about the possible location of the ligand, and *Best pocket docking* , representing the best-case scenario where the correct binding pocket is known and used to define the docking search space. P2Rank was able to predict the center of the correct binding pocket for 89.2% (324) of the receptors within 8 Å distance of the true binding site center, defined as the mean coordinate of the ligand in the pocket. On the other hand, for the AlphaFold modeled receptors, the percentage was 81.1% (206), where the ground truth pocket is defined by structural superposition of the model with the reference structure. For the evaluation of Best pocket docking, the P2Rank pocket that had the smallest distance from the true binding site center was considered the best pocket.

The reporting workflow utilizes BiSyRMSD (referred to as RMSD) and lDDT-PLI scoring to evaluate the predicted ligand structures generated by the different docking methods. Both of these are novel scoring metrics developed for the CASP15 CASP-PLI experiment[9] that consider both predicted protein structure and predicted ligand conformation. In addition, lDDT-PLI focuses on the interactions between protein and ligand atoms. Table 1 and Table 2 display the outcomes for PLC prediction using the 363 receptors from the PDBbind test-set and the 256 AlphaFold modeled receptors respectively. The full results are available as Supplementary Table 1 and 2 for the experimentally solved and AlphaFold modeled receptors, respectively. The highest ranked pose (top-1) and the best scored pose out of the top-5 ranked poses (where the ranking is an output of each tool) are assessed for blind docking where the entire protein is employed to define the search box. Furthermore, for all tools except DiffDock where this option is not present, the same assessment is carried out for the best-case scenario using the best pocket for defining the search box. Figure 3 depicts the distributions of these scores for the top-1 and best out of top-5 poses for experimental and modeled

receptors for both docking scenarios.

Table 1 : Prediction of small molecule binding to crystallized protein structures from the PDBbind testset containing 363 PLC. For some PLC the pipeline did not complete successfully. Shown are the number of PLC (n), the success rate (SR) defined as the percentage of predictions with RMSD < 2 Å, the median RMSD, the mean lDDT-PLI, and the standard deviation of lDDT-PLI. DiffDock does not use a pocket definition. TANKBind gives only one prediction per search box.

			Top-1	**Top-1**	**Top-1**	**Top-1**	**Top-5**	**Top-5**
	Method	**n**	**RMSD (Å) SR (%)**	**RMSD (Å) median**	**lDDT-PLI mean**	**lDDT-PLI std**	**RMSD (Å) SR (%)**	**RMSD median**
Blind docking	Autodock Vina	360	14.72	7.82	0.36	0.34	25.28	5.71
	SMINA	362	17.40	7.75	0.37	0.34	27.07	5.67
	GNINA	362	22.65	8.68	0.38	0.38	30.66	4.84
	TANKBind	363	9.09	6.39	0.35	0.27		
Best pocket	Autodock Vina	359	29.81	5.01	0.53	0.34	47.35	2.19
	SMINA	361	30.75	5.00	0.53	0.34	45.71	2.41
	GNINA	361	42.11	2.61	0.62	0.34	55.96	1.71
	TANKBind	362	11.60	5.09	0.44	0.25		
	DiffDock	361	37.67	3.24	0.59	0.31	44.04	2.55

Table 2 : Prediction of small molecule binding to AlphaFold predicted structures for 256 monomeric proteins from the PDBbind testset. For some PLC the pipeline did not complete successfully. Shown are the number of PLC (n), the success rate (SR) defined as the percentage of predictions with RMSD < 2 Å, the median RMSD, the mean lDDT-PLI, and the standard deviation of lDDT-PLI. DiffDock does not use a pocket definition. TANKBind gives only one prediction per search box.

			Top-1 pose	**Top-1 pose**	**Top-1 pose**	**Top-1 pose**	**Top-5 poses**	**Top-**
	Method	**n**	**RMSD (Å) SR (%)**	**RMSD (Å) median**	**lDDT-PLI mean**	**lDDT-PLI std**	**RMSD (Å) SR (%)**	**RMS med**
Blind docking	Autodock Vina	255	3.92	11.84	0.20	0.24	5.49	7.70
	SMINA	256	3.52	12.06	0.20	0.23	6.25	7.97
	GNINA	256	4.69	18.47	0.17	0.26	9.38	9.44
	TANKBind	256	4.30	7.65	0.29	0.25		
Best pocket	Autodock Vina	252	4.76	8.52	0.28	0.25	10.71	5.62
	SMINA	253	4.35	8.51	0.27	0.24	12.25	5.83
	GNINA	253	8.70	7.46	0.33	0.28	14.62	5.69
	TANKBind	253	5.53	6.15	0.36	0.26		
	DiffDock	252	21.03	4.15	0.48	0.29	32.94	3.07

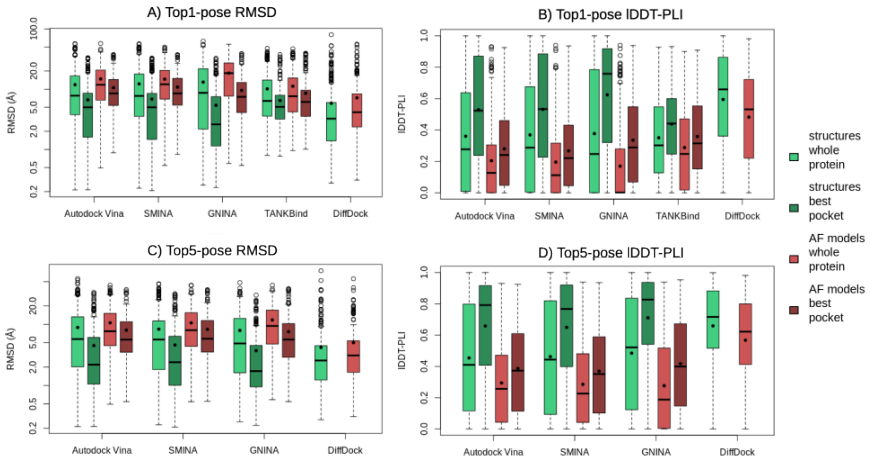

Figure 3: Distribution of the scores shown in Table 1 and Table 2 for A) Top1-pose RMSD, B) Top1-pose lDDT-PLI, C) Top5-pose RMSD, and D) Top5-pose lDDT-PLI. The lines and the black dots in the bars represent the median and the mean respectively.

As expected, the results for the Best pocket docking are better than Blind docking, as the search space is restricted. DiffDock performs well on this set despite only offering blind docking mode, as also reported by the authors along with the suggestion to use DiffDock as a ligand-specific pocket detector[34]. The difference between the two scoring metrics is especially seen when comparing blind docking results of GNINA and TankBind in Table 1. The median RMSD is worse for GNINA indicating more "severe" failures which bring up the RMSD, as it is an unbounded metric. In contrast, lDDT-PLI is bounded, and all PLC poses beyond the thresholds used are assigned a score of 0, and are less penalized by very bad predictions. In addition, lDDT-PLI does not penalize parts of the ligand which are floating in areas not in contact with the protein. All the tools have a significant performance decrease when using AlphaFold models as input. This is especially striking when considering Best pocket docking, where exact side-chain and conformation positioning seem to be crucial for obtaining the right ligand pose for physics-based docking tools, as seen in Figure 4, where the backbone RMSD of the AlphaFold model is 3.56 Å and it is clear that a rearrangement has pushed a helix into the binding pocket, preventing the correct ligand pose from being found. This trend is not as striking for the deep learning tool DiffDock, as its training has less reliance on side-chain atoms, although the performance is still lower than on crystal structures.

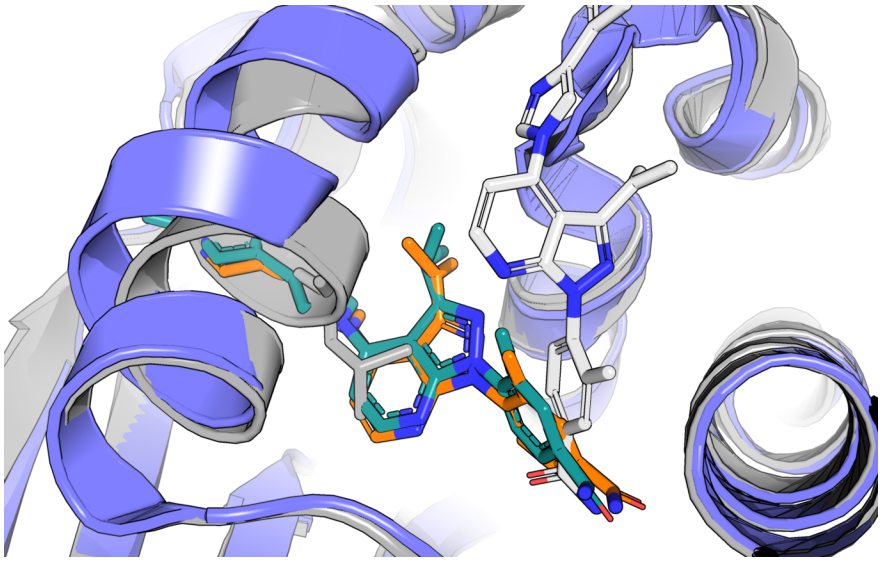

Figure 4: An example of GNINA docking results on the Hsp90 receptor in complex with ligand 9J0 (PDB ID: 5ZR3). The crystal structure of the receptor is shown in purple with the ground truth ligand in green. The AlphaFold model is shown in gray. The GNINA docked conformation using the AlphaFold model as input is in white and the docked conformation using the crystal structure as input is in orange.

While these results have proven valuable for testing our automated workflow, they are not meant to be a comprehensive evaluation of these PLC prediction tools especially in the context of the challenges and concepts discussed in the previous sections. In the small set of 363 PLC used: (1) 108 protein-ligand pairs have peptide and oligosaccharide ligands which are not ideal as most docking tools are not calibrated for these types of ligands[31]. (2) Only 104 out of the remaining 255 small molecule and ion pockets pass the relaxed validation criteria, and (3) the test set was created using a time-based split and thus contains redundant proteins within itself, indicating a biased representation of PLC space, as well as with the PDBBind training set, indicating an overestimation of prediction results for the tools trained on this set. Thus, it is critical to repeat this analysis on a diverse benchmarking dataset created with both structure quality and PLC diversity taken into account, and after ensuring that the PLC prediction tools based on machine learning or deep learning are trained on a dataset different from the benchmark set. This will both ensure a more reliable and comprehensive evaluation as well as allow for more specific pinpointing of problem cases for different tools to aid in their further development.

For four out of the 363 complexes the workflow failed due to issues with various steps in the process. The inability to generate conformers using RDKit for the stapled peptide ligand of 6q4q resulted in the failure of both DiffDock inference and the definition of the search box required to run Autodock Vina, SMINA, and GNINA. For the 6o0h protein-ligand pair DiffDock failed because the language model embeddings did not have the right length for the protein. In addition, 6uhu and 6rtn failed to run with Autodock Vina due to the presence of unsupported atoms. Furthermore, for the 6d07 receptor, P2Rank was unable to predict a binding pocket. During the analysis of the 256 AlphaFold modeled receptors, P2Rank failed to predict a binding pocket for three receptors (6d07, 6d08 and 6qlt). Further, complexes 6o0h and 6uhu suffered the same issues already mentioned above. In addition, DiffDock inference failed for three more complexes (6cjj, 6jib, and 6jid). These failures were automatically identified,reported and isolated by the workflow. Overall, we demonstrate that automated workflows can be employed for PLC preparation, prediction and assessment.

3 Methods

3.1 PLC validation criteria

PLC were obtained from the PDB, release 2023-03-15. The PDB Chemical Component Dictionary[41] was downloaded on March 17, 2023. X-ray validation information was extracted from the XML files provided by the PDB. Additional information including the entry ID to polymer entity ID mapping, release date and polymer composition for each entry as well as the canonical one-letter code sequence for each entity in the dataset was retrieved with the GraphQL-based API of the RCSB PDB Web Services[42] on 2023-03-28. 37 entries marked as obsolete in the API results were discarded.

Ligands were defined as any non-polymer entity. A PLC was defined as a PDB entry with at least one polymer and one non-polymer entity (ion or small molecule). PDB entries for which the "polymer composition" was one of "DNA", "RNA", "DNA/RNA", "NA-hybrid", "other type pair", "NA/oligosaccharide" or "other type composition", as well as any remaining entry containing DNA or RNA polymers were ignored.

Binding pockets were defined as the set of amino acid residues in the reference structure with at least one heavy atom within a 6 Å radius of any heavy ligand atom.

The filtering thresholds for the Iridium criteria were extracted from the original manuscript[18]. The suggestion to filter PLC where atoms from crystal packing are within 6 Å of any ligand atom was not used as this information could not easily be extracted from the PDB validation report.

3.2 PLC clustering and novelty assessment

For PLC clustering, the set of PLC described in section 3.1 was used. PLC were grouped together based on the cluster identifier of all the unique polymer entities and the chemical component 3-letter code of the ligands (i.e. identical ligands) they contained. Polymer entity cluster identifiers were obtained by performing sequence-based clustering of all polymer entities in the dataset with the cluster module from the MMseqs2 software (version 13.45111)[43]. Six different sequence-based clustering patterns were obtained as a result of clustering with minimum sequence identity thresholds of 100%, 95%, 90%, 70%, 50% and 30% respectively. For the sequence alignment, a coverage threshold of 90% (-c 0.9) of both the query and target sequences was used (–cov-mode 0). The sensitivity of the prefiltering was set to (-s 8.0). Clustering was performed with the connected component algorithm (–cluster-mode 1) with the option (–cluster-reassign) to reassign cluster members to other clusters if they no longer fulfill the clustering criteria after each iteration. Each PLC entry in the dataset was subsequently given an identifying string consisting of the cluster ids of the entities and the 3-letter code of the unique ligands present in the PLC.

The assessment of the novelty of a given PLC with respect to a different set of PLC, at a given minimum sequence identity threshold, was performed by comparing its PLC identifier to the set of all PLC identifiers of the other set.

3.3 Benchmarking state-of-the-art docking tools

A Nextflow[36]pipeline (20.10.0) was developed to run and assess 5 state-of-the-art PLC prediction tools. This is available at*https://github.com/PickyBinders/PickyBinder*

3.3.1 Benchmark dataset

The 363 PLC in the PDBBind time-split test-set that were not used as training data by TANKBind and DiffDock were used as a test set to demonstrate the automated benchmarking workflow[16]. To compare docking on experimental and predicted structures, AlphaFold v2.3.0 [39]was used to predict models for 256

monomeric proteins in this set, using the canonical one-letter code sequence, and default parameters and relaxation. Results are present on the best relaxed model (according to average pLDDT) for each protein.

3.3.2 Molecule preparation

Each ligand was prepared starting from the SMILES string. Ligands were first standardized by neutralizing the charges and re-adjusted for pH 7 using protonation rules. Explicit hydrogen atoms were then added. The 3D conformation was generated using the ETKDG method from RDKit[44], and stored in SDF format. For docking tools related to the AutoDock family, the Python package Meeko (v0.4.0) was used to generate the PDBQT input files [45].

3.3.3 PLC prediction tools

The predictions were run with the default parameters given by the tools unless stated differently below.

(1) Autodock Vina version 1.2.3[30,31]docking was performed with exhaustiveness set to 64 within a Conda[37]environment containing the required python bindings. Meeko v0.4.0 was used to transform the PDBQT output file into an SDF file, to be used by the evaluation tools. (2) SMINA[32] was run within a Conda environment (v2020.12.10, conda-forge:b08c07c, based on AutoDock Vina 1.1.2) with exhaustiveness set to 64. (3) GNINA[33] was run using a Singularity image downloaded from*https://hub.docker.com/r/nmaus/gnina*(digest: 7087cbf4dafd, gnina v1.0.2 (master:0cb5eb8, built Sep 29 2022)) with exhaustiveness set to 64. (4) TANKBind[35], input preparation and inference was run according to the code provided at*https://github.com/luwei0917/TankBind*using a Singularity image for the dependencies downloaded from*https://hub.docker.com/r/qizhipei/tankbind_py38*. (5) DiffDock[34]inference was run using –samples_per_complex 40 –batch_size 10 –actual_steps 18 –no_final_step_noise within a Conda environment built according to the setup guide (master:2c7d438, built Mar 13 2023).

Each tool except DiffDock allows for the definition of a pocket center and grid size, within which the search space for ligand conformations is restricted. To assess predictions for different pockets, P2Rank[40] (v2.4) was used to predict and rank multiple binding pockets,with default parameters for experimental structures and -c alphafold option for AlphaFold predicted models. The box in which Autodock Vina,GNINA and SMINA search for binding poses was constructed around each predicted P2Rank pocket center. The diameter of the search box was the diameter of the ligand conformer generated by RDKit with an additional 10 Å on all 6 sides of the search box.Thus for each tool (p+1)*n predicted ligand poses were obtained as outputs, where p is the number of pockets predicted by P2Rank and n is the number of poses returned by the tool.

3.3.4 Scoring

BiSyRMSD (shortened to RMSD throughout this manuscript) and lDDT-PLI scores were calculated with OpenStructure version 2.4.0[46] with default parameters. The methods are identical to those described in the CASP15 CASP-PLI assessment paper[9]. Every ligand was scored separately and a summary CSV file containing scores for each ligand pose, pocket, and blind docking is generated.

4 Conclusion

With combined prediction of protein-ligand complexes forming the next frontier for deep learning in computational structural biology, we need approaches for independent, comprehensive and blind assessment of prediction methods to better assess the advantages and shortcomings of classical and novel approaches. Two complementary approaches can be employed for this purpose: weekly continuous evaluation of structures released in the PDB, and the creation of a representative, diverse dataset for benchmarking.

In this study, we examined three challenges essential for establishing such systems in an automated and unsupervised manner: determining whether an experimentally solved PLC can be used as ground truth,

assessing the interest or difficulty of a PLC for prediction, and automating the scoring of predicted PLC. In the process, we defined quality criteria for PLC pockets, assessed novelty in the PDB over the years, and developed an automated workflow for PLC prediction and assessment using newly developed scoring metrics. Ligand preparation is a known challenge in docking and throughout our research we faced obstacles in automating ligand preparation, in particular with molecule parsing and protonation.

The PDBBind dataset has been frequently utilized for training deep-learning based docking methods and evaluating their accuracy. Many deep learning methods retained 363 PDBBind PLC as a test set based on their release date after 2019. However, this selection is not ideal for benchmarking, as only half of the structures meet the quality criteria indicating unreliable ground truth, redundancy removal was not performed, and diversity was not considered when choosing the PLC. Consequently, there is a need for a representative dataset that follows the concepts presented in this study.

Acknowledgements

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 956137. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Italy, Sweden, Austria, Czech Republic, Switzerland. We are grateful for funding from the SIB Swiss Institute of Bioinformatics toward the development of CAMEO and OpenStructure, and from ELIXIR EXCELERATE to CAMEO (funded by the European Commission within the Research Infrastructures programme of Horizon 2020, grant agreement number 676559). Calculations were performed at the sciCORE (http://scicore.unibas.ch/) scientific computing center of the University of Basel.

Supplementary Data

Supplementary Table 1: full results for PLC prediction using the 363 receptors from the PDBbind test-set.

Supplementary Table 2: full results for PLC predictions using the 256 AlphaFold modeled receptors.

References

1. Jansen JM, Cornell W, Tseng YJ, Amaro RE. Teach-Discover-Treat (TDT): collaborative computational drug discovery for neglected diseases. *J Mol Graph Model*. 2012;38:360-362. doi:10.1016/j.jmgm.2012.07.007

2. Wagner JR, Churas CP, Liu S, et al. Continuous Evaluation of Ligand Protein Predictions: A Weekly Community Challenge for Drug Docking. *Structure*. 2019;27(8):1326-1335.e4. doi:10.1016/j.str.2019.05.012

3. Gathiaka S, Liu S, Chiu M, et al. D3R grand challenge 2015: Evaluation of protein-ligand pose and affinity predictions. *J Comput Aided Mol Des*. 2016;30(9):651-668. doi:10.1007/s10822-016-9946-8

4. Gaieb Z, Liu S, Gathiaka S, et al. D3R Grand Challenge 2: blind prediction of protein-ligand poses, affinity rankings, and relative binding free energies. *J Comput Aided Mol Des*. 2018;32(1):1-20. doi:10.1007/s10822-017-0088-4

5. Gaieb Z, Parks CD, Chiu M, et al. D3R Grand Challenge 3: blind prediction of protein-ligand poses and affinity rankings. *J Comput Aided Mol Des*. 2019;33(1):1-18. doi:10.1007/s10822-018-0180-4

6. Parks CD, Gaieb Z, Chiu M, et al. D3R grand challenge 4: blind prediction of protein-ligand poses, affinity rankings, and relative binding free energies. *J Comput Aided Mol Des*. 2020;34(2):99-119. doi:10.1007/s10822-020-00289-y

7. Carlson HA, Smith RD, Damm-Ganamet KL, et al. CSAR 2014: A Benchmark Exercise Using Unpublished Data from Pharma. *J Chem Inf Model*. 2016;56(6):1063-1077. doi:10.1021/acs.jcim.5b00523

8. Smith RD, Damm-Ganamet KL, Dunbar JB Jr, et al. CSAR Benchmark Exercise 2013: Evaluation of Results from a Combined Computational Protein Design, Docking, and Scoring/Ranking Challenge. *J Chem Inf Model*. 2016;56(6):1022-1031. doi:10.1021/acs.jcim.5b00387

9. Robin X, Studer G, Janani D, et al. Assessment of Protein-Ligand Complexes in CASP15.*Proteins*. (This issue).

10. Kryshtafovych A. New Prediction Categories In CASP15. *Proteins*. (this issue).

11. wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res*. 2019;47(D1):D520-D528. doi:10.1093/nar/gky949

12. Robin X, Haas J, Gumienny R, Smolinski A, Tauriello G, Schwede T. Continuous Automated Model EvaluatiOn (CAMEO)-Perspectives on the future of fully automated evaluation of structure prediction methods. *Proteins*. 2021;89(12):1977-1986. doi:10.1002/prot.26213

13. Haas J, Gumienny R, Barbato A, et al. Introducing "best single template" models as reference baseline for the Continuous Automated Model Evaluation (CAMEO). *Proteins*. 2019;87(12):1378-1387. doi:10.1002/prot.25815

14. Haas J, Barbato A, Behringer D, et al. Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12.*Proteins*. 2018;86 Suppl 1(Suppl 1):387-398. doi:10.1002/prot.25431

15. Haas J, Roth S, Arnold K, et al. The Protein Model Portal–a comprehensive resource for protein structure and model information. *Database* . 2013;2013:bat031. doi:10.1093/database/bat031

16. Stärk H, Ganea OE, Pattanaik L, Barzilay R, Jaakkola T. EquiBind: Geometric Deep Learning for Drug Binding Structure Prediction. Published online February 7, 2022. Accessed May 8, 2023. http://arxiv.org/abs/2202.05146

17. Gao Y, Thorn V, Thorn A. Errors in structural biology are not the exception. *Acta Crystallographica Section D: Structural Biology*. 2023;79(3):206-211. doi:10.1107/S2059798322011901

18. Warren GL, Do TD, Kelley BP, Nicholls A, Warren SD. Essential considerations for using protein-ligand structures in drug discovery. *Drug Discov Today*. 2012;17(23-24):1270-1281. doi:10.1016/j.drudis.2012.06.011

19. McGovern SL, Shoichet BK. Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. *J Med Chem*. 2003;46(14):2895-2907. doi:10.1021/jm0300330

20. Rueda M, Bottegoni G, Abagyan R. Recipes for the selection of experimental protein conformations for virtual screening. *J Chem Inf Model*. 2010;50(1):186-193. doi:10.1021/ci9003943

21. Scardino V, Di Filippo JI, Cavasotto CN. How good are AlphaFold models for docking-based virtual screening? *iScience*. 2023;26(1):105920. doi:10.1016/j.isci.2022.105920

22. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res*. 2000;28(1):235-242. doi:10.1093/nar/28.1.235

23. Read RJ, Adams PD, Arendall WB 3rd, et al. A new generation of crystallographic validation tools for the protein data bank. *Structure*. 2011;19(10):1395-1412. doi:10.1016/j.str.2011.08.006

24. Gore S, Velankar S, Kleywegt GJ. Implementing an X-ray validation pipeline for the Protein Data Bank. *Acta Crystallogr D Biol Crystallogr*. 2012;68(Pt 4):478-483. doi:10.1107/S0907444911050359

25. Gore S, Sanz García E, Hendrickx PMS, et al. Validation of Structures in the Protein Data Bank.*Structure*. 2017;25(12):1916-1927. doi:10.1016/j.str.2017.10.009

26. Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* 2013;41(Database issue):D1096-D1103. doi:10.1093/nar/gks966

27. Smart OS, Horský V, Gore S, et al. Validation of ligands in macromolecular structures determined by X-ray crystallography. *Acta Crystallogr D Struct Biol.* 2018;74(Pt 3):228-236. doi:10.1107/S2059798318002541

28. Liu Z, Su M, Han L, et al. Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions. *Acc Chem Res.* 2017;50(2):302-309. doi:10.1021/acs.accounts.6b00491

29. Orengo C, Velankar S, Wodak S, et al. A community proposal to integrate structural bioinformatics activities in ELIXIR (3D-Bioinfo Community).*F1000Res.* 2020;9. doi:10.12688/f1000research.20559.1

30. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.*Journal of Computational Chemistry.* 2010;31(2):455-461. doi:10.1002/jcc.21334

31. Eberhardt J, Santos-Martins D, Tillack AF, Forli S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J Chem Inf Model.* 2021;61(8):3891-3898. doi:10.1021/acs.jcim.1c00203

32. Koes DR, Baumgartner MP, Camacho CJ. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J Chem Inf Model.* 2013;53(8):1893-1904. doi:10.1021/ci300604z

33. McNutt AT, Francoeur P, Aggarwal R, et al. GNINA 1.0: molecular docking with deep learning.*J Cheminform.* 2021;13(1):43. doi:10.1186/s13321-021-00522-2

34. Corso G, Stärk H, Jing B, Barzilay R, Jaakkola T. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. Published online October 4, 2022. Accessed April 14, 2023. http://arxiv.org/abs/2210.01776

35. Lu W, Wu Q, Zhang J, Rao J, Li C, Zheng S. TANKBind: Trigonometry-Aware Neural NetworKs for drug-protein binding structure prediction. *bioRxiv.* Published online June 6, 2022. doi:10.1101/2022.06.06.495043

36. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. *Nat Biotechnol.* 2017;35(4):316-319. doi:10.1038/nbt.3820

37. Anaconda Software Distribution. *Anaconda Documentation.* Accessed May 11, 2023. https://docs.anaconda.com/

38. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute.*PLoS One.* 2017;12(5):e0177459. doi:10.1371/journal.pone.0177459

39. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold.*Nature.* 2021;596(7873):583-589. doi:10.1038/s41586-021-03819-2

40. Krivák R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J Cheminform.* 2018;10(1):1-12. doi:10.1186/s13321-018-0285-8

41. Westbrook JD, Shao C, Feng Z, Zhuravleva M, Velankar S, Young J. The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank.*Bioinformatics.* 2015;31(8):1274-1278. doi:10.1093/bioinformatics/btu789

42. Rose Y, Duarte JM, Lowe R, et al. RCSB Protein Data Bank: Architectural Advances Towards Integrated Searching and Efficient Access to Macromolecular Structure Data from the PDB Archive. *J Mol Biol.* 2021;433(11):166704. doi:10.1016/j.jmb.2020.11.003

43. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. *Nat Commun.* 2018;9(1):2542. doi:10.1038/s41467-018-04964-5

44. Riniker S, Landrum GA. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J Chem Inf Model.* 2015;55(12):2562-2574. doi:10.1021/acs.jcim.5b00654

45. Meeko: Preparation of small molecules for AutoDock (Forli Lab, 2022). Published 2022. Accessed May 8, 2023. https://github.com/forlilab/Meeko

46. Biasini M, Schmidt T, Bienert S, et al. OpenStructure: an integrated software framework for computational structural biology. *Acta Crystallogr D Biol Crystallogr*. 2013;69(Pt 5):701-709. doi:10.1107/S0907444913007051