zPoseScore model for accurate and robust protein-ligand docking pose scoring in CASP15

Liangzhen Zheng[1], Tao Shen[1], Fuxu Liu[1], Zechen Wang[2], Jinyuan Sun[3], Yifan Bu[1], Jintao Meng[4], Weihua Chen[1], Yuguang Mu[5], Weifeng Li[2], Guoping Zhao[4], Sheng Wang[6], and Wei Yanjie[4]

[1]Shanghai Zelixir Biotech Company Ltd
[2]Shandong University
[3]Institute of Microbiology Chinese Academy of Sciences
[4]Chinese Academy of Sciences Shenzhen Institutes of Advanced Technology
[5]Nanyang Technological University School of Biological Sciences
[6]Affiliation not available

April 17, 2023

Abstract

We introduce a deep learning-based ligand pose scoring model called zPoseScore for predicting protein-ligand complexes in the 15th Critical Assessment of Protein Structure Prediction (CASP15). Our contributions are three-fold: firstly, we generate six training and evaluation datasets by employing advanced data augmentation and sampling methods. Secondly, we redesign the "zFormer" module, inspired by AlphaFold2's Evoformer, to efficiently describe protein-ligand interactions. This module enables the extraction of protein-ligand paired features that lead to accurate predictions. Lastly, we develop the zPoseScore framework with zFormer for scoring and ranking ligand poses, allowing for atomic-level protein-ligand feature encoding and fusion to output refined ligand poses and ligand per-atom deviations. Our results demonstrate excellent performance on various testing datasets, achieving Pearson's correlation R = 0.783 and 0.659 for ranking docking decoys generated based on experimental and predicted protein structures of CASF-2016 protein-ligand complexes. Additionally, we obtain an averaged lDDT = 0.558 of *AIchemy_LIG2* in CASP15 for de novo protein-ligand complex structure predictions. Detailed analysis shows that accurate ligand binding site prediction and side-chain orientation are crucial for achieving better prediction performance. Our proposed model is one of the most accurate protein-ligand pose prediction models and could serve as a valuable tool in small molecule drug discovery.

ORIGINAL ARTICLE
Journal Section

zPoseScore model for accurate and robust protein-ligand docking pose scoring in CASP15

Tao Shen*[1] | Fuxu Liu*[1] | Zechen Wang[2] | Jinyuan Sun[3] | Yifan Bu[1] | Jintao Meng[4] | Weihua Chen[1] | Yuguang Mu[5] | Weifeng Li[2] | Guoping Zhao[4] | Sheng Wang†[1] | Yanjie Wei†[4] | Liangzhen Zheng†[1,4]

[1]Shanghai Zelixir Biotech Company Ltd., Shanghai 200030, China

[2]Shandong University, Jinan 250100ShandongChina

[3]Institute of Microbiology, Chinese Academy of Sciences, Chinese Academy of Sciences Beijing 100101 China

[4]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Chinese Academy of Sciences, Shenzhen 518055, Guangdong, China

[5]School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551

Correspondence
Liangzhen Zheng, Yanjie Wei, Sheng Wang
Email: astrozheng@gmail.com, yj.wei@siat.ac.cn, wangsheng@zelixir.com

Funding information

We introduce a deep learning-based ligand pose scoring model called zPoseScore for predicting protein-ligand complexes in the 15th Critical Assessment of Protein Structure Prediction (CASP15). Our contributions are three-fold: firstly, we generate six training and evaluation datasets by employing advanced data augmentation and sampling methods. Secondly, we redesign the "zFormer" module, inspired by AlphaFold2's Evoformer, to efficiently describe protein-ligand interactions. This module enables the extraction of protein-ligand paired features that lead to accurate predictions. Lastly, we develop the zPoseScore framework with zFormer for scoring and ranking ligand poses, allowing for atomic-level protein-ligand feature encoding and fusion to output refined ligand poses and ligand per-atom deviations. Our results demonstrate excellent performance on various testing datasets, achieving Pearson's correlation R = 0.783 and 0.659 for ranking docking decoys generated based on experimental and predicted protein structures of CASF-2016 protein-ligand complexes. Additionally, we obtain an averaged lDDT = 0.558 of *Alchemy_LIG2* in CASP15 for de novo protein-ligand complex structure predictions. Detailed analysis shows that accurate ligand binding site prediction and side-chain orientation are crucial for achieving better prediction performance. Our proposed model is one of the most accurate protein-ligand pose prediction models and could serve as a valuable tool in small molecule drug discovery.

KEYWORDS
CASP15, protein-ligand prediction, pose scoring, attention-based model

1 | INTRODUCTION

The prediction of protein-ligand binding represents a crucial technology in drug discovery [1, 2]. The binding pattern is a fundamental information for understanding the interaction between the target protein and a drug or a small molecule, providing visual and structural insights into the molecular mechanisms underlying relevant biological actions. Accurate prediction of the protein-ligand binding pattern enhances the rational design of new drugs, allowing for the development of pharmacologically relevant molecules with desired properties.

† Co-corresponding authors.
* Equally contributing authors.

Therefore, improving the accuracy of such binding pattern prediction is a vital goal in the field of computer-aided drug discovery (CADD). Additionally, understanding the substrate-enzyme binding pattern can facilitate the rational design of industrial enzymes or bio-sensors in the realm of enzyme engineering and bio-sensor design [3, 4, 5]. Molecular docking, molecular dynamics simulations, and artificial intelligence-based approaches represent the three main strategies employed for protein-ligand binding pose prediction.

Molecular docking is a widely used computational method CADD that predicts the energetically favorable conformation of a ligand bound to a protein [1, 2]. The aim is to determine the optimal pose of a ligand preferably in a specific binding pocket of the protein to ensure accuracy. However, the accuracy of predicting protein-ligand interactions is limited when protein side-chain flexibility is considered. In order to overcome this limitation, various molecular docking methods have been developed, such as AutoDock [6], AutoDock Vina [7, 8], Smina, Glide [9], GOLD [10], and LeDock [11]. These methods utilize specific sampling methods (including Monte Carlo sampling or genetic algorithm and other sampling strategies) to modify the ligand pose and adjust the protein side-chain orientations, as well as to evaluate the binding energies through corresponding scoring functions. The resulting docking poses are then scored and clustered to provide final outcomes [7, 6]. While some challenges remain, including the requirement for the near-native sampling of poses and accurate ranking of poses according to their binding energies, molecular docking-based methods have been successful in previous protein-ligand challenges (such as D3R grand challenge) [12].

Molecular dynamics (MD) simulations represent a powerful computational methodology for the prediction of protein-ligand interactions with an unprecedented degree of accuracy at the atomic level [13, 14, 15]. In essence, MD simulations deal with fully flexible biomolecules, solvents, and ions under the influence of Newton's laws of motion. For a protein-ligand complex system, the solvation free energy of the protein and the ligand could be calculated, and the nonbonded interactions between them could also be calculated, lastly, the conformational entropy could also be estimated [16, 17, 18, 19]. Consequently, the corresponding binding free energies can be estimated. However, one should note that the accurate prediction of such interactions via molecular dynamics simulations hinges on the use of a structurally plausible protein-ligand complex model as an initial input, which consequently restricts its applicability to direct protein-ligand binding pose prediction tasks. Furthermore, it is vital to acknowledge that molecular dynamics simulations are computationally demanding [15].

Current research in the field of Artificial Intelligence (AI)-based approaches has been focused on predicting protein-ligand binding affinity and docking pose scoring [20, 21, 22, 23, 24]. Previous models, such as random-forest [25], convolutional neural network (CNN) [26, 27, 28, 29], and graph-based models [27, 28, 29], were designed to predict ligand binding affinity based on a given protein-ligand native poses. The early models primarily focused on scoring native poses as opposed to docking poses, thus limiting their use in real drug design or virtual screening tasks [30]. Recent advancements in the field of Machine Learning (ML) and Deep Learning (DL) have improved performance on ligand binding pose scoring, leading to more accurate docking pose selection [20, 31, 32, 33] or optimization [34]. For instance, 3D CNN-based pose prediction models [20] have demonstrated the potential of Graph Neural Networks (GNNs) as a viable alternative for docking pose prediction and virtual screening tasks [35, 36]. Additionally, the DL docking approach Gnina has successfully integrated 3D CNN-based scoring functions to score and rank the Vina scoring function guide poses [37]. More recent works, such as the use of graph models to generate ligand poses from the protein surfaces [38, 39], have shown promising performance in blinded docking scenarios. However, these methods are not yet practical enough for structure-based drug discovery, especially in cases where the binding pocket is predefined. Meanwhile, generating high-quality docking poses with protein side-chain flexibility remains an open research question.

In this study, a novel scoring model, called zPoseScore, is introduced and its performance on testing datasets and CASP15 tasks is reported. The zPoseScore model incorporates three key features: (1) utilization of both experimental structures and fastAF2 predicted protein structures and predicted pockets for large-scale pose generation, (2) introduction of a "zFormer" network module for protein-ligand pairwise interactions with iterative representation updating, and (3) incorporation of per-atom deviations similar to plDDT [40] in AlphaFold2 for predicting atomic level ligand pose quality. Moreover, the protein-ligand interaction atoms sampling strategy enhances the accuracy of the model. Results show that zPoseScore outperforms the traditional scoring function Vinascore and other DL-based scoring functions on different testing datasets. Additionally, the model's performance on CASP15 tasks, namely, *Alchemy_LIG*, *Alchemy_LIG2*, and *Alchemy_LIG3*, is summarized. It is also observed that the binding sites of the ligand and the orientations of the pocket residue side-chains could affect zPoseScore's performance in protein-ligand pose predictions. Overall, zPoseScore presents a robust scoring model for docking pose scoring, and it could potentially serve as a valuable tool for future drug discovery and drug design.

2 | MATERIALS AND METHODS

2.1 | Training data preparation

Firstly, we collected the 285 protein-ligand complex pairs from CASF-2016 core set (*dataset* 3) [30]. Then we selected the protein-ligand complexes released after the year 2019 from PDBBind2020 and only kept the protein-ligand complexes dissimilar to complexes in the general set of PDBBind2020 to form *dataset* 5. The similarity was defined as the product of the protein sequence similarity and ligand fingerprint similarity [20], and a cutoff = 0.6 was used to filter the similar data. The protein-ligand complexes from RCSB PDB were collected (Date: 2022-Mar-24) to form *dataset* 1. For all pdb structures, the molecules were split by chains, and protein-ligand pairs were generated. In this process, only ligands with heavy atom numbers between 10-100 were considered, and the ligand should interact with only one protein chain

(heavy atom contacts between the ligand and the protein chain should be larger than 10), otherwise, the protein-ligand pair was discarded. For all protein-ligand pairs, the pairs were also discarded if the protein amino acid sequence length was less than 30 or greater than 1500. For one experimental protein-ligand complex structure, if multiple same ligands exist, only one copy was kept. In the meanwhile, the similar protein-ligand complexes (similarity higher than 0.6) to those in *datasets* 3 and 5 were removed.

For all these pairs (from experimental complexes), the protein sequences were extracted and then used for protein structure prediction using the fastMSA-based AlphaFold2 protocol described in this paper [41]. Later, the predicted structures were structurally aligned to their corresponding native structures by DeepAlign [42]. This way, we have three more artificial protein-ligand complexes datasets (*datasets* 2, 4, and 6). The test sets (*datasets* 3 and 4) are designed to compare the docking pose scoring performance with other methods [34, 20]. And to more strictly evaluate the performance of the model, *datasets* 5 and 6 are designed by collecting the most recent protein-ligand experimental complexes and these complexes are quite different from the complex structures in PDBBind 2019 general set, which is often used as training data in many DL or ML based scoring functions for docking pose scoring [37, 31, 20, 34]. Therefore, *datasets* 5 and 6 are the most strict testing datasets for docking pose scoring evaluation and performance comparison.

In this research, we try to collect and generate datasets that could mimic the real applications in CASP15 tasks, where no receptor protein structure and ligand initial conformer are provided. By predicting the protein structure from scratch, defining the possible ligand binding sites, and docking the ligand into multiple pockets, the generated docking poses could be used to train models more suitable for CASP15 tasks by minimizing the bias in protein structures and binding sites between model training and inference.

2.2 | The pose scoring and generation pipeline

For all pairs in Table 1, the ligands are re-docked into the protein structures by various docking protocols. For experimental protein structures, the ligands were firstly docked into their original pockets (using the ligands' geometry centers as pocket centers). While for the complexes with experimental or predicted protein structures, the binding pockets were predicted by PointSite [43], and then the ligands were docked into the most probable predicted pockets as described in [31] for reverse docking. Multiple docking tools (such as AutoDock Vina [7], iDock [44], Qvina2 [45] and Smina [46] were used to generate various docking poses, while the three scoring functions (ad4 [6], dkoes [46], and vina [7]) were used in Smina docking. For each pair, the ligand was repeatedly docked back into the defined pocket with pocket sizes (in three dimensions) of 15 to 30 *Å* (randomly defined in 25 docking repeats to explore various binding areas) to generate no more than 500 decoys. Default docking parameters were used in all docking calculations. For each decoy, the decoy qualities were calculated as the root mean square deviation (RMSD) for overall conformation and each atom using DockRMSD developed by Zhang Lab [47]. The decoys whose overall RMSDs to native pose were higher than 15 *Å* were then removed.

2.3 | The networks of the zPoseScore single models

The zPoseScore single model is a transformer-based model designed to depict protein-ligand interactions which can be used to rank docking poses by the predicted RMSD (pRMSD) values. The model utilizes a complex structure encoder and a backbone network called zFormer and a scoring module.

The complex structure encoder we designed is to capture the structural features of input conformations. The approach to managing input features is simple and can be described as follows: firstly, the atomic-level features of proteins and ligands are extracted separately, including one-hot encoding of element names and residue names. In addition to atom features, we also encode the 3D coordinates of the conformation as pair representations by calculating the Euclidean distances between atoms. We did not employ extra features but utilized only element names, residue names, and inter-atomic distance information, encouraging the neural network to learn complex interactions based on these fundamental properties. In section 3.1, we discovered that such a configuration is sufficient to achieve high-precision predictions.

After obtaining the atom and pair representations of proteins and ligands, we constructed a backbone network for subsequent feature updates. This backbone network, inspired by the Evoformer module in AlphaFold2 [48], designs an interaction module specifically for protein-ligand interactions, which we named zFormer. Unlike the Evoformer, zFormer focuses on iteratively updating atom-wise representations of protein-ligand complexes, while the Evoformer addresses protein-only amino acid-level features. Furthermore, the Evoformer employs axial attention [49] to handle columns and rows in two-dimensional multiple sequence alignment (MSA) [50]. In our model, however, atomic features are one-dimensional, so we utilize a standard attention mechanism [51] instead of axial attention [52]. Moreover, the parameters in each layer of zFormer share weights, reducing the number of model parameters and enhancing the model's expressive power.

The backbone network is followed by a scoring module that predicts per-atom deviations and pRMSD for ligands. By calculating the average of these predicted deviations, we can assess errors and rank conformations generated by the docking process. The key component of the scoring module is the Invariant Point Attention [48] (IPA). The IPA module leverages geometric perception attention operations to effectively address rotation or translation invariance issues. In our model, the inputs for IPA include atom representation, pair representation, and three-dimensional coordinates of each atom; the output comprises updated atom features. Employing the rich feature representation obtained from the IPA module, the scoring module can predict the deviation (distance) between each atom and the corresponding native 3D positions. To enable the model to learn structural information more directly, we introduce coordinate refinement to enhance model

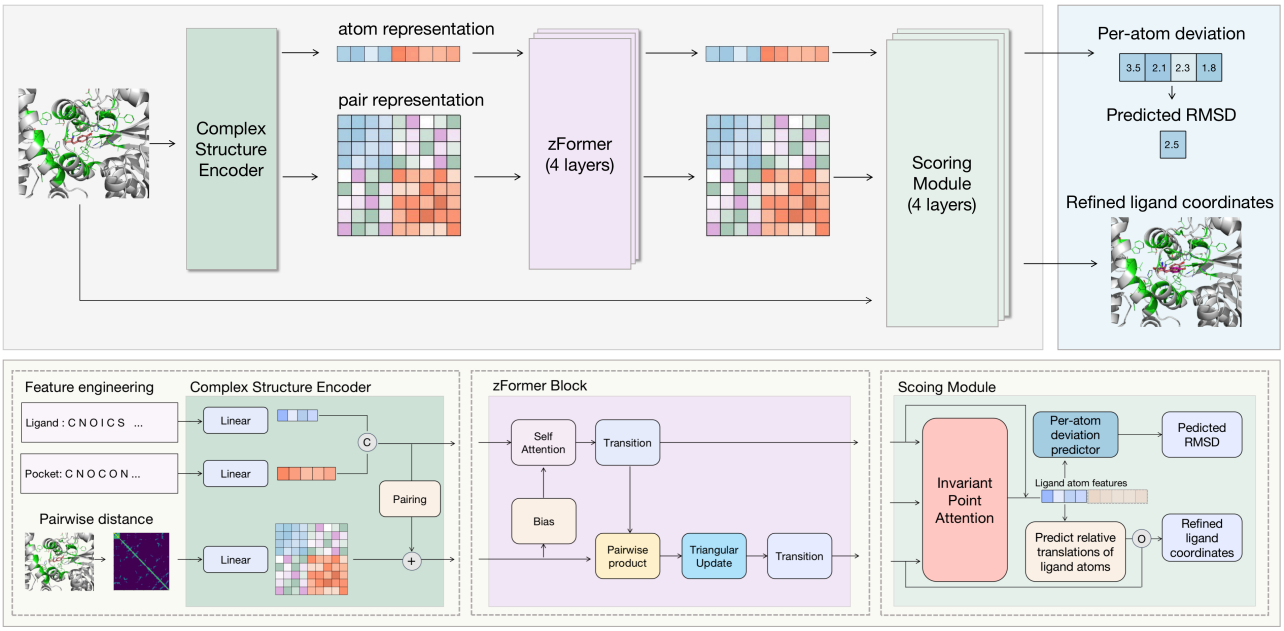

FIGURE 1 Overview of zPoseScore Models: the zPoseScore single model is a transformer-based model for depicting protein-ligand interactions and ranking docking poses using predicted RMSD (pRMSD) values. It consists of a complex structure encoder, a zFormer backbone network, and a scoring module. The encoder extracts atomic-level features and calculates inter-atomic distances. The zFormer network updates atom-wise representations of protein-ligand complexes using attention mechanisms. Next, the scoring module predicts per-atom deviations (with regard to native pose) and pRMSD (the root mean squared value of the per-atom deviations) for ligands. The single model also introduces coordinate refinement and self-supervised loss to enhance the pose-scoring performance.

performance. With the features processed by the IPA module, the model predicts the translations of each atom on the ligand in 3D space.

For CASP15 tasks, 6 single models were assembled to make the zPoseScore model (a meta-predictor) for all ligand pose scoring by calculating the average pRMSD values by the single models.

2.4 | The CASP15 protein-ligand prediction protocol

Given a protein sequence and a ligand's SMILES code, the following protocol was adopted for protein-ligand complex structure prediction. Firstly, the protein structure was predicted using the fastMSA-based AlphaFold2 protocol [41], and then the possible ligand binding regions were predicted by PointSite [43]. For each protein structure, no more than 5 pockets were defined as defined in a previous study [31]. Then the ligand was modeled with Rdkit [53] and its conformer optimization procedure and then docked into the ligand binding pockets. Multiple docking tools (such as AutoDock Vina [7], iDock [44], Qvina2 [45], and Smina [46] were adopted for docking pose generation. The docking pose generation protocol was exactly the same as used in *datasets* 2, 4, and 6 (Table 1), as described in Section 2.2. For each protein-ligand pair, no more than 500 decoys were generated and scored by the pose ranking model. The top-ranked 50 poses were then selected and clustered into 5 clusters by K-means algorithm [54], and the representative poses of the 5 clusters were visually re-ranked according to their binding patterns. The 5 poses then were submitted under group *AIchemy_LIG* (see Figure S1).

The selected five poses by clustering and visual inspection thus were further optimized using DeepRMSD [34] guided by DeepRMSD+Vina scoring scheme for no more than 100 steps. The last frames of the optimized poses thus were submitted under group *AIchemy_LIG2* (see Figure S1). For group *AIchemy_LIG3*, the 5 poses of *AIchemy_LIG2* were further scored by DeepRMSD and submitted.

2.5 | The zPoseScore training protocol

A multi-task learning strategy was employed for the training of zPoseScore single models. The loss function for zPoseScore optimization consisted of three parts: scoring loss, coordinate refinement loss, and self-supervised loss.

Scoring loss L_{pRMSD} is the primary loss in our optimization. It aims to train models as an RMSD evaluator that predicts the RMSD of the input ligand poses against the ground truth structures. The RMSD value is discretized into 10 bins with an interval of 1.0 Å. A cross-entropy loss is employed as L_{pRMSD} to calculate if the predicted RMSD falls in the ground truth bin. Coordinate refinement loss L_{refine} is designed to supervise the zPoseScore to learn the structural information of protein-ligand complexes directly. It calculates the Mean Squared Error (MSE) loss between the refined coordinates output by the scoring module and the native ones. We believe that this auxiliary atom-wise structural supervision can be beneficial to the training of the scoring function. Self-supervised loss $L_{self-supervised}$ is achieved by predicting the atomic

TABLE 1 The protein-ligand complex datasets.

Dataset	Protein structure	Group	pairs	Datasource
dataset 1	experimental	training + validating	33632	All PDB
dataset 2	predicted	training + validating	30674	All PDB
dataset 3	experimental	testing	283	CASF-2016 coreset
dataset 4	predicted	testing	250	CASF-2016 coreset
dataset 5	experimental	testing	247	PDBBind2020 new data after 2021
dataset 6	predicted	testing	175	PDBBind2020 new data after 2021

TABLE 2 The performance of single models with different network dimensions, pocket atom selection, and sampling strategies.

SN	Description	Network hyperparameters	Training data	*dataset 5*		*dataset 6*	
				R (PCC[a])	*R* (SCC[b])	*R* (PCC[a])	*R* (SCC[b])
1	model dimension	*dim64*, *prot64*, *distance*	*dataset 1*	0.439	0.405	0.262	0.268
2	model dimension	*dim96*, *prot64*, *distance*	*dataset 1*	0.458	0.423	0.292	0.272
3	model dimension	*dim128*, *prot64*, *distance*	*dataset 1*	0.467	0.434	0.311	0.278
4	number of pocket atoms	*dim64*, *prot64*, *distance*	*dataset 1*	0.439	0.405	0.262	0.228
5	number of pocket atoms	*dim64*, *prot96*, *distance*	*dataset 1*	0.454	0.410	0.290	0.261
6	number of pocket atoms	*dim64*, *prot128*, *distance*	*dataset 1*	0.470	0.427	0.335	0.311
7	training data	*dim64*, *prot128*, *distance*	*datasets 1 and 2*	0.449	0.376	0.511	0.470
8	atom sampling	*dim64*, *prot128*, *pointsite*	*datasets 1 and 2*	0.467	0.389	0.509	0.457
9	atom sampling	*dim64*, *prot128*, *distance+pointsite*	*datasets 1 and 2*	0.483	0.406	0.511	0.467

a. PCC indicates Pearson's correlation coefficient.

b. SCC indicates Spearman's correlation coefficient.

types of masked atoms. More specifically, in the training process, we generate random masks on a subset of atoms in both proteins and ligands and require the model to predict the masked atom types. The purpose of this self-supervised loss is to endow the model with better generalization capabilities.

The overall loss function is:

$$L = L_{pRMSD} + L_{refine} + L_{self-supervised} \quad (1)$$

During training, we also performed online data augmentation by randomly rotating and translating ligands' 3D positions. We employed the Adam optimizer [55] with a learning rate of 1e-3 and utilized a cosine annealing scheduler to adjust the learning rate [56]. The entire training process consisted of 40,000 steps with a batch size of 512. We used 8 32GB-V100 GPUs, and the training took 9 hours to finish.

3 | RESULTS AND DISCUSSIONS

3.1 | The performance of different models with various hyper-parameters

Based on the network architecture defined in Figure 1, several single models were trained with various hyper-parameters. In these single models, the atoms around the protein-ligand interface are encoded by the structure encoder module with various encoding dimensions (see Table 2). Three dimension options (64, 96, and 128) were tested and the results indicate that a higher dimension could not guarantee better performance, and dimension = 64 was adopted for further hyper-parameter tuning (models 1-3). In many graph-based DL models [35, 57], an arbitrary distance cut-off is used to select the protein atoms or residues to describe the interaction area, however, it is not clear whether such a setting is suitable for our single models. Meanwhile, very large graphs would slow down the model training and inference, so in these single models, only a limited number of protein atoms are selected for encoding. Here, the distance-ranked and binding-pocket probability-based [43] protein atom sampling strategies are applied. In each protein-ligand complex (either native complex or docking complex), the protein atoms are ranked by their minimal distances to ligand-heavy atoms and only a fixed number of atoms (64, 96, and 128 in models 4-6 in Table 2) are then selected for protein atom encoding. It is obvious that more protein atoms could contribute to higher prediction

accuracy in both *datasets* 5 and 6. It is worth trying to explore the possibilities of using more protein atoms for binding pocket encoding, but exponential resources would be required for model training. In CASP15, we used 128 protein atoms for pocket encoding in the single models.

Models 1-6 were all trained with training *dataset* 1, where the experimental protein structures were used for docking decoy generation. To ensure robust performance for CASP15 tasks, the protein-structure-bias should be avoided, therefore training *dataset* 2 was composed by using the predicted protein structures and predicted binding sites for docking decoys generation. With the same model architecture and hyperparameters, model 7 shows better prediction accuracy (Pearson's R = 0.511 and Spearman's R = 0.47) on testing *dataset* 6 which is also based on prediction protein structures, but model 6 (Pearson's R = 0.47 and Spearman's R = 0.437) is more accurate than model 7 (Pearson's R = 0.449 and Spearman's R = 0.376) on testing *dataset* 5, which is based on experimental protein structures.

Finally, we posit that employing computed binding site probabilities as criteria for protein atom sampling could enhance the accuracy of pose scoring. In model 8, instead of utilizing minimum distances between protein and ligand atoms for protein atom selection, we adopted the protein atoms with maximum predicted probabilities generated by PointSite for encoding. Our findings indicate that the integration of PointSite-based atom sampling improves the performance (Pearson's correlation R = 0.467 and Spearman's correlation R = 0.389) on *dataset 5*, which comprises experimental protein structures, in contrast to model 7. However, the performance of model 8 is inferior to that of model 7 on *dataset 6*, consisting of predicted protein structures. The original study posits that PointSite, which is a point-cloud-based deep learning model trained on experimental protein-ligand complexes for identifying binding sites, may suffer from decreased accuracy when predicting protein structures [43]. Moreover, the pocket residue side-chains predicted by AlphaFold2 may obstruct ligand binding, thus making PointSite-based protein atom sampling suboptimal for assessing *dataset* 6 and *dataset* 4. Section 3.4 dives further into the impact of pocket residue side-chain orientations. Moreover, combining distance-based and PointSite-based protein atom sampling (model 9) enhanced pose scoring accuracy for both *datasets* 5 and 6.

In essence, it is evident that disparate hyper-parameters associated with the network would undeniably impact the capacity of algorithms to accurately estimate the pose score of ligand molecules. Specifically, augmenting the number of atoms that constitute the binding site of the ligands during the process of pocket encoding and exploring varying pocket atom sampling strategies could effectively improve the pose-scoring abilities. Notably, amalgamating models to produce a meta-predictor, referred to as the zPoseScore model, has the potential to bolster the prediction performance. A more comprehensive discussion on the zPoseScore model is presented in Section 3.2.

3.2 | The ensemble zPoseScore model performance

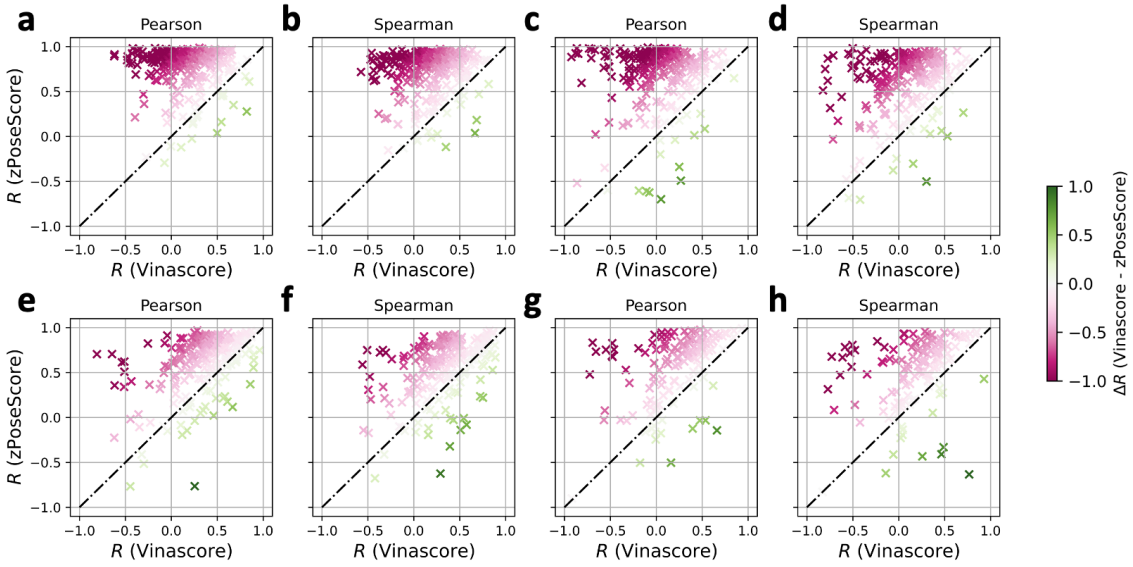

FIGURE 2 The per-target ranking abilities of zPoseScore and Vinascore on *dataset* 3 (a and b), *dataset* 4 (c and d), *dataset* 5 (e and f), and *dataset* 6 (g and h). The color bars indicate the performance difference between the two scoring methods Vinascore and zPoseScore, brick red dots, therefore, suggest zPoseScore greatly outperforms Vinascore for different protein-ligand complex systems.

To improve the performance on the CASP15 tasks, we generated a meta-predictor (called zPoseScore model) by ensembling several single models (models 2,3, 6, 7, 8 and 9 in Table 2). To evaluate the performance of the zPoseScore model and the widely used docking tool AutoDock Vina [7] for docking pose ranking, four testing protein-ligand docking decoys datasets in Table 1 were used. In AutoDock Vina, the docking poses are sampled through Monte Carlo sampling and local energy minimization using the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) algorithm guided by Vinascore. This scoring method (Vinascore) is a representative traditional scoring function

TABLE 3 Performance (averaged per-target correlation) comparison of different scoring methods.

	dataset 3		*dataset 4*		*dataset 5*		*dataset 6*	
Method	*R* (PCC[a])	*R* (SCC[b])	*R* (PCC[a])	*R* (SCC[b])	*R* (PCC[a])	*R* (SCC[b])	*R* (PCC[a])	*R* (SCC[b])
zPoseScore	0.783	0.729	0.659	0.593	0.604	0.535	0.554	0.507
Vinascore	0.162	0.186	0.035	0.071	0.364	0.356	0.209	0.192
DeepBSP	0.561	0.539	0.401	0.375	0.543	0.507	0.451	0.418
RTMscore[c]	0.736	0.668	0.632	0.576	-	-	-	-
DeepRMSD	0.62	0.584	0.463	0.43	0.285	0.246	0.261	0.247
DeepRMSD+Vina[d]	0.449	0.449	0.248	0.29	0.405	0.362	0.299	0.287

a. PCC indicates Pearson's correlation coefficient.

b. SCC indicates Spearman's correlation coefficient.

c. the complexes in *dataset 5* and *dataset 6* are part of the train set of RTMscore. The values multiplied by -1.0 are used for RMSD prediction.

d. the weight for Vinascore is 0.5.

[30, 34].

In this work, we demonstrate that our zPoseScore model outperforms both Vinascore and other DL-based scoring methods [20, 57, 34] in terms of ligand pose ranking and RMSD prediction tasks, based on experimental or predicted protein structures. The use of predicted protein structures and optimized pocket side-chain orientations for docking pose generation and ranking is more practical in the context of the rapidly growing structural proteome study and functional annotations of predicted structures in AlphaFold DB [58]. Moreover, cross-docked docking poses are crucial for model training for real-world docking applications; similarly, the predicted structure and related pockets serve as the non-native docking pockets.

In Figure 2, the Pearson's correlation (or per-target correlation) between predicted pose scores and true pose RMSDs is shown for each protein-ligand complex system, using no more than 500 docking poses generated with various docking protocols based on experimental or predicted protein structures. A higher correlation value indicates better pose ranking ability. For all four test sets, zPoseScore outperforms Vinascore in pose ranking tasks. For instance, for the protein-ligand complex 6UFO (see Figure S3), zPoseScore effectively predicts the true RMSD values of the docking poses, while Vinascore failed to predict the docking pose RMSD ranking.

The performance of different DL-based methods is also summarized in Table 3. The three deep learning models (DeepBSP [20], RTMscore [57], and DeepRMSD [34]) show reasonable accuracy on dataset 3 and dataset 4. Among them, RTMscore achieves an averaged per-target Pearson's correlation coefficient of 0.736, indicating its potential to serve as a good scoring function for docking pose ranking in dataset 3. When equipped with Vinascore, DeepRMSD shows improved performance in *dataset* 5 and *dataset* 6. Nonetheless, DeepRMSD is trained with experimental re-docking poses, which makes it susceptible to overfitting to the training data and unsuitable for predicting pose RMSD for *dataset* 4 and *dataset* 6, where predicted protein structures are used for docking pose generation. Although the training sets of these scoring functions are not identical, the results still indicate the potential advantages and practical usability of our zPoseScore model.

3.3 | The CASP15 prediction performance

In CASP15, protocols were designed to predict the protein-ligand complexes, by predicting the protein structure, predicting the binding sites, followed by generating a group of docking poses with certain variability, and then scoring and ranking the poses, and lastly clustering the poses. Optionally, DL-based optimization was performed to ensure more fine-grained docking poses. For all these predictions, the standard alone performance of each step, however, is not easy to evaluate. Rather, we summarize the overall prediction accuracy of the three groups (*AIchemy_LIG*, *AIchemy_LIG2* and *AIchemy_LIG3*).

The CASP15 performance of different groups is illustrated in Figure 3. The groups (*AIchemy_LIG*, *AIchemy_LIG2* and *AIchemy_LIG3*) achieved average lDDT scores (Figure 3, panel a) of 0.541, 0.558, and 0.549. The lDDT score represents the average ratio of local protein-ligand contacts located within certain distance ranges. It is a commonly used metric in protein structure predictions and is more fine-grained and alignment-free than the global distance test (GDT), which is defined by calculating the largest set of alpha carbon atoms within several predefined distance cutoffs after structure superimposition [59]. Higher lDDT scores indicate more accurately predicted local protein-ligand heavy atom contacts. In computing the lDDT scores, all unique ligands for a given target were merged and evaluated as a single task, and only the best-performed first poses were considered.

Meanwhile, around 50.9%, 54.4%, and 52.6% protein-ligand pairs could be successfully predicted (Figure 3, panel a) by the three groups (*AIchemy_LIG*, *AIchemy_LIG2* and *AIchemy_LIG3*). Based on the averaged protein-ligand interaction-based lDDT scores, *Zou*, *CoDock* and *AIchemy_LIG2* are the top-ranked 3 groups, while if evaluated by the successfully predicted ratio, the three groups achieve the same perfor-

mance. Here, the ratio is defined as the ratio of the protein-ligand pairs whose top-ranked poses are near-native poses (global RMSD less than 2 *Å* to the native ligand poses with the predicted proteins aligned to the native protein structures ahead of the RMSD calculation). Different from lDDT score, the RMSD values should be computed after structure superimposition based on overall protein structure or ligand binding site residues, thus introducing potential uncertainties in the calculation. The metric calculated here is quite similar to the docking successful rate (or so-called docking power) defined in [30, 31], but the difference is that in CASF-2016, the protein structures are extracted from experimental structures.

According to various assessment criteria, the performance outcomes of our predictions, as generated by *AIchemy_LIG*, *AIchemy_LIG2*, and *AIchemy_LIG3*, rank among the top performers in the field [**?**].

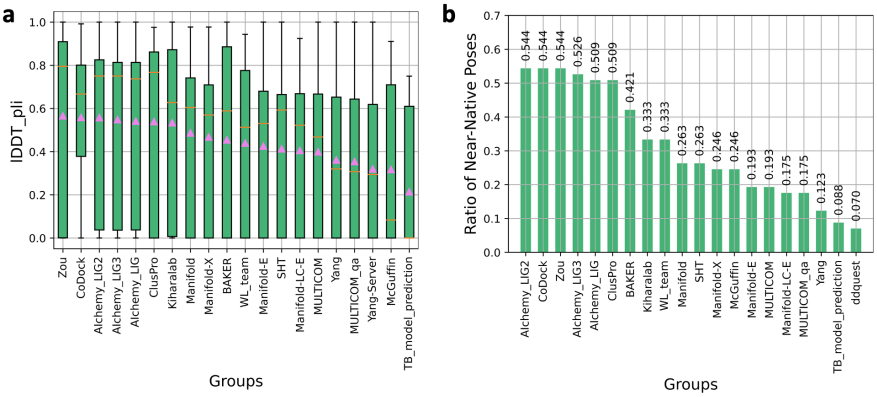

FIGURE 3 The CASP15 performance of different groups. a, the distribution of the top-ranked pose protein-ligand interaction lDDT scores of the protein-ligand pairs in CASP15 targets; the orange lines indicate the median values, and the violet triangles indicate the mean values. b, the ratio of the successfully predicted protein-ligand pairs.

3.4 | Protein pocket predictions affect the model quality and ligand pose prediction

In CASP15, only protein sequence and ligand SMILES information are provided. In current solutions, the protein structure is firstly predicted with AlphaFold2 [48] or other structure prediction models. We built extensive training sets to mimic the real usage scenario by predicting the protein structure, followed by binding pocket prediction, ligand pose generation, and pose scoring.

Therefore, the ligand binding site prediction is a fundamental requirement for correct ligand pose prediction [30, 31, 60]. For example, target T1181 is a tetramer structure of the tail fiber proteins (gene name: *gp66*) from Escherichia phage, and four small molecules and four ions are supposed to bind the tetramer structure (Figure 4, panel a and b). For this target, AIchemy_LIG and the other two groups both failed in predicting the right binding sites using PointSite [43], resulting in very large ligand pose RMSDs with regard to the native ligand poses. Similarly, the ligand poses predicted by other teams also are quite different from the native ligand poses (with RMSDs larger than 5 *Å*). A more robust ligand binding sites prediction method should be developed to solve the situations where protein structure and binding site are unknown.

The side chains or sometimes the backbones of the binding pocket residues are however often quite divergent to the native orientations [62], and the scoring model zPoseScore and traditional scoring function Vinascore are less able to accurately predict the docking poses' RMSD. In *dataset* 5 (Figure S2 left panel) and *dataset* 6 (Figure S2 right panel), in more cases, the ranking abilities are much stronger for the poses generated based on experimental protein structures (higher per-target correlations in lower triangle regions). For the cases with docking poses generated based on predicted structures, the per-target correlations of nearly half of the protein-ligand complexes are not satisfactory (Pearson's correlation less than 0.25) for Vinascore. Although zPoseScore is more tolerant to the predicted protein structure-based docking poses, around 1/3 of the protein-ligand complex systems are still badly predicted (Pearson's correlation less than 0.25).

Sometimes, the ligand binding site could be accurately predicted either by deep learning methods [43, 63, 64], geometry-based methods [65], or even by the template-based hypothesis [66], however, the orientations of the binding site residues' side-chains are required to be precisely predicted to generate ligand binding poses with atomic-level accuracies. AlphaFold2 [48] predicts high accuracy of protein overall structures, but the prediction accuracies of the side-chain orientations, especially the ligand binding sites, are not satisfactory [62, 67]. Several studies [68, 69] adopt the AlphaFold2 predicted structures for ligand docking and virtual screening, and conclude that the predicted structures could achieve similar enrichment performance when compared to the *apo*-form receptors, but by optimizing the binding site site-chains orientations (using Glide induced-fit protocol [9]), the enrichment performance could be greatly improved.

Similar behavior is observed in CASP15 protein-ligand predictions. For example, in CASP15, T1124 is a dimeric tyrosine methyltransferase (MfnG) from *Streptomyces drozdowiczii* (Figure 4, panel c and d). We first predicted the binding patterns of the cofactor S-adenosyl-L-Homocysteine (SAH) by protein structure prediction and binding site prediction. By searching similar structures in RCSB PDB using DeepAlign

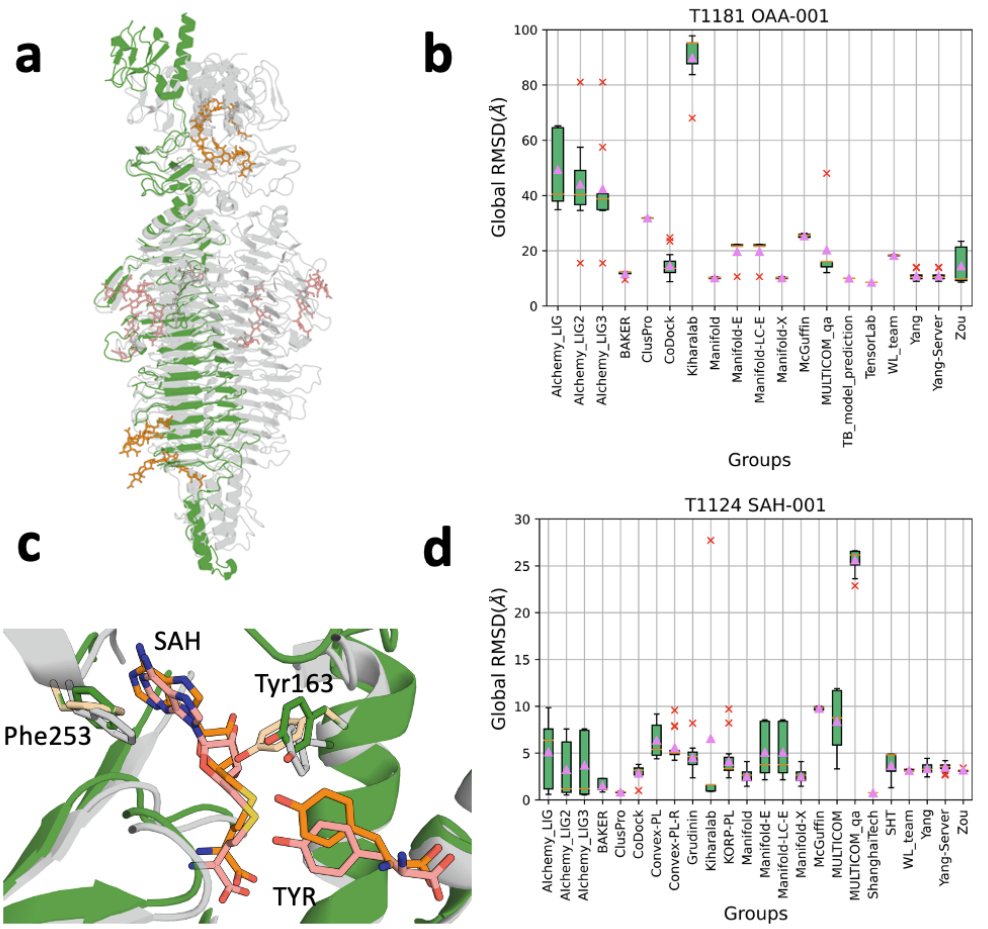

FIGURE 4 The binding patterns and RMSD distributions of CASP15 targets T1181 (a and b) and T1124 (c and d). For a and c, the predicted ligand poses are orange, the native ligand poses are pink, the green color structure is a monomer of the predicted protein, and the gray color structure is the native protein structure. For b and d, the orange lines indicate the median values and the violet triangles indicate the mean values, and the outliers are marked by red crosses. In panel c, the yellow color side-chains (Tyr163 and Phe253) are optimized by PyMol [61].

[42], similar structures with SAH binding were identified (such as PDB IDs 6C5B and 4A6E). Detailed inspection suggested that several side-chains in the binding site may hinder the SAH binding, thus the two residues (Tyr163 and Phe253) were optimized by PyMol [61] mutation tool to avoid the potential spatial clashes with the cofactor SAH. Based on the optimized ligand binding pocket, the predicted SAH binding poses are quite close to the native poses (best lDDT=0.884 and RMSD=0.585 *Å* by *Alchemy_LIG2*). Meanwhile, the average RMSDs of most of the predicted poses are higher than 2 *Å*.

4 | CONCLUSION

To enhance the accuracy of predicting protein-ligand binding poses for the CASP15 tasks, we have created large-scale protein-ligand docking decoy datasets. These datasets are composed of multiple docking decoys generated by various docking tools based on experimental and predicted protein structures and DL-based binding site predictions. We have used these datasets for training and testing purposes to ensure objective and reliable model inference. We have developed several single models for this purpose. These models incorporate protein and ligand atoms encoding, iterative information updating through zFormer, and a ligand pose optimization and scoring output, therefore they could predict both optimized ligand poses (not utilized for CASP15 submissions) and score per-atom deviations of the docking poses. With the methodology of scoring and ligand pose optimization by DeepRMSD [34], our model demonstrates relatively robust performance when applied to various testing datasets and CASP15 protein-ligand predictions. We have identified that accurate prediction of ligand binding sites and side-chain orientation is crucial for improved prediction capabilities. In the future, the zPoseScore model will be expanded to include fully flexible protein-ligand modeling with protein side-chain orientation optimization. Additionally, adapting the zPoseScore model to become a more lightweight scoring method would be advantageous for large-scale ligand virtual screening and structural-based molecule generation with reliable binding poses and high affinities.

references

[1] Macalino SJY, Gosu V, Hong S, Choi S. Role of computer-aided drug design in modern drug discovery. Archives of pharmacal research 2015;38:1686–1701.

[2] Kapetanovic I. Computer-aided drug discovery and development (CADDD): in silico-chemico-biological approach. Chemico-biological interactions 2008;171(2):165–176.

[3] Kiss G, Çelebi-Ölçüm N, Moretti R, Baker D, Houk K. Computational enzyme design. Angewandte Chemie International Edition 2013;52(22):5700–5725.

[4] Kries H, Blomberg R, Hilvert D. De novo enzymes by computational design. Current opinion in chemical biology 2013;17(2):221–228.

[5] Khoshbin Z, Housaindokht MR, Izadyar M, Bozorgmehr MR, Verdian A. Recent advances in computational methods for biosensor design. Biotechnology and Bioengineering 2021;118(2):555–578.

[6] Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, et al. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. Journal of computational chemistry 2009;30(16):2785–2791.

[7] 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.

[8] Huey R, Morris GM, Forli S. Using AutoDock 4 and AutoDock vina with AutoDockTools: a tutorial. The Scripps Research Institute Molecular Graphics Laboratory 2012;10550(92037):1000.

[9] Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, et al. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. Journal of medicinal chemistry 2004;47(7):1750–1759.

[10] Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein–ligand docking using GOLD. Proteins: Structure, Function, and Bioinformatics 2003;52(4):609–623.

[11] Liu N, Xu Z. Using LeDock as a docking tool for computational drug design. In: IOP Conference Series: Earth and Environmental Science, vol. 218 IOP Publishing; 2019. p. 012143.

[12] Gathiaka S, Liu S, Chiu M, Yang H, Stuckey JA, Kang YN, et al. D3R grand challenge 2015: evaluation of protein–ligand pose and affinity predictions. Journal of computer-aided molecular design 2016;30:651–668.

[13] Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. Nature structural biology 2002;9(9):646–652.

[14] Rapaport DC, Rapaport DCR. The art of molecular dynamics simulation. Cambridge university press; 2004.

[15] Zheng L, Alhossary AA, Kwoh CK, Mu Y. Molecular dynamics and simulation 2019;.

[16] Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert opinion on drug discovery 2015;10(5):449–461.

[17] Wang E, Sun H, Wang J, Wang Z, Liu H, Zhang JZ, et al. End-point binding free energy calculation with MM/PBSA and MM/GBSA: strategies and applications in drug design. Chemical reviews 2019;119(16):9478–9508.

[18] Huang N, Kalyanaraman C, Bernacki K, Jacobson MP. Molecular mechanics methods for predicting protein–ligand binding. Physical Chemistry Chemical Physics 2006;8(44):5166–5177.

[19] Wang L, Berne B, Friesner RA. On achieving high accuracy and reliability in the calculation of relative protein–ligand binding affinities. Proceedings of the National Academy of Sciences 2012;109(6):1937–1942.

[20] Bao J, He X, Zhang JZ. DeepBSP—a machine learning method for accurate prediction of protein–ligand docking structures. Journal of Chemical Information and Modeling 2021;61(5):2231–2240.

[21] Zhang H, Liao L, Saravanan KM, Yin P, Wei Y. DeepBindRG: a deep learning based method for estimating effective protein–ligand affinity. PeerJ 2019;7:e7362.

[22] Meng Z, Xia K. Persistent spectral–based machine learning (PerSpect ML) for protein-ligand binding affinity prediction. Science advances 2021;7(19):eabc5329.

[23] Jiménez J, Skalic M, Martinez-Rosell G, De Fabritiis G. K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. Journal of chemical information and modeling 2018;58(2):287–296.

[24] Hassan-Harrirou H, Zhang C, Lemmin T. RosENet: improving binding affinity prediction by leveraging molecular mechanics energies with an ensemble of 3D convolutional neural networks. Journal of chemical information and modeling 2020;60(6):2791–2802.

[25] Ballester PJ, Mitchell JBO. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics 2010 03;26(9):1169–1175. https://doi.org/10.1093/bioinformatics/btq112.

[26] Wallach I, Dzamba M, Heifets A. AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXiv preprint arXiv:151002855 2015;.

[27] Wang Z, Zheng L, Liu Y, Qu Y, Li YQ, Zhao M, et al. OnionNet-2: a convolutional neural network model for predicting protein-ligand binding affinity based on residue-atom contacting shells. Frontiers in Chemistry 2021;p. 913.

[28] Zheng L, Fan J, Mu Y. Onionnet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. ACS omega 2019;4(14):15956–15965.

[29] Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. Bioinformatics 2018;34(21):3666–3674.

[30] Su M, Yang Q, Du Y, Feng G, Liu Z, Li Y, et al. Comparative assessment of scoring functions: the CASF-2016 update. Journal of chemical information and modeling 2018;59(2):895–913.

[31] Zheng L, Meng J, Jiang K, Lan H, Wang Z, Lin M, et al. Improving protein–ligand docking and screening accuracies by incorporating a scoring function correction term. Briefings in Bioinformatics 2022;23(3):bbac051.

[32] Wang C, Zhang Y. Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. Journal of computational chemistry 2017;38(3):169–177.

[33] Lu J, Hou X, Wang C, Zhang Y. Incorporating explicit water molecules and ligand conformation stability in machine-learning scoring functions. Journal of chemical information and modeling 2019;59(11):4540–4549.

[34] Wang Z, Zheng L, Wang S, Lin M, Wang Z, Kong AWK, et al. A fully differentiable ligand pose optimization framework guided by deep learning and a traditional scoring function. Briefings in Bioinformatics 2023;24(1):bbac520.

[35] Jiang D, Hsieh CY, Wu Z, Kang Y, Wang J, Wang E, et al. Interactiongraphnet: A novel and efficient deep graph representation learning framework for accurate protein–ligand interaction predictions. Journal of medicinal chemistry 2021;64(24):18209–18232.

[36] Jiang H, Wang J, Cong W, Huang Y, Ramezani M, Sarma A, et al. Predicting protein–ligand docking structure with graph neural network. Journal of Chemical Information and Modeling 2022;62(12):2923–2932.

[37] McNutt AT, Francoeur P, Aggarwal R, Masuda T, Meli R, Ragoza M, et al. GNINA 1.0: molecular docking with deep learning. Journal of cheminformatics 2021;13(1):1–20.

[38] Stärk H, Ganea O, Pattanaik L, Barzilay R, Jaakkola T. Equibind: Geometric deep learning for drug binding structure prediction. In: International Conference on Machine Learning PMLR; 2022. p. 20503–20521.

[39] Lu W, Wu Q, Zhang J, Rao J, Li C, Zheng S. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. bioRxiv 2022;p. 2022–06.

[40] Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 2013;29(21):2722–2728.

[41] Hong L, Sun S, Zheng L, Tan Q, Li Y. fastMSA: Accelerating Multiple Sequence Alignment with Dense Retrieval on Protein Language. bioRxiv 2021;https://www.biorxiv.org/content/early/2021/12/21/2021.12.20.473431.

[42] Wang S, Ma J, Peng J, Xu J. Protein structure alignment beyond spatial proximity. Scientific reports 2013;3(1):1–7.

[43] Yan X, Lu Y, Li Z, Wei Q, Gao X, Wang S, et al. PointSite: a point cloud segmentation tool for identification of protein ligand binding atoms. Journal of Chemical Information and Modeling 2022;62(11):2835–2845.

[44] Li H, Leung KS, Wong MH. idock: A multithreaded virtual screening tool for flexible ligand docking. In: 2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) IEEE; 2012. p. 77–84.

[45] Alhossary A, Handoko SD, Mu Y, Kwoh CK. Fast, accurate, and reliable molecular docking with QuickVina 2. Bioinformatics 2015;31(13):2214–2216.

[46] Koes DR, Baumgartner MP, Camacho CJ. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. Journal of chemical information and modeling 2013;53(8):1893–1904.

[47] Bell EW, Zhang Y. DockRMSD: an open-source tool for atom mapping and RMSD calculation of symmetric molecules through graph isomorphism. Journal of Cheminformatics 2019;11(1):1–9.

[48] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596(7873):583–589.

[49] Wang H, Zhu Y, Green B, Adam H, Yuille A, Chen LC. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV Springer; 2020. p. 108–126.

[50] Shen T, Wu J, Lan H, Zheng L, Pei J, Wang S, et al. When homologous sequences meet structural decoys: Accurate contact prediction by tFold in CASP14—(tFold for CASP14 contact prediction). Proteins: Structure, Function, and Bioinformatics 2021;89(12):1901–1910.

[51] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems 2017;30.

[52] Ho J, Kalchbrenner N, Weissenborn D, Salimans T. Axial attention in multidimensional transformers. arXiv preprint arXiv:191212180 2019;.

[53] Landrum G, et al. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. Greg Landrum 2013;8.

[54] Ahmed M, Seraj R, Islam SMS. The k-means algorithm: A comprehensive survey and performance evaluation. Electronics 2020;9(8):1295.

[55] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980 2014;.

[56] Loshchilov I, Hutter F. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:160803983 2016;.

[57] Shen C, Zhang X, Deng Y, Gao J, Wang D, Xu L, et al. Boosting Protein–Ligand Binding Pose Prediction and Virtual Screening Based on Residue–Atom Distance Likelihood Potential and Graph Transformer. Journal of Medicinal Chemistry 2022;65(15):10691–10706.

[58] Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic acids research 2022;50(D1):D439–D444.

[59] Lüthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature 1992;356(6364):83–85.

[60] Wierbowski SD, Wingert BM, Zheng J, Camacho CJ. Cross-docking benchmark for automated pose and ranking prediction of ligand binding. Protein Science 2020;29(1):298–305.

[61] DeLano WL, et al. Pymol: An open-source molecular graphics tool. CCP4 Newsl Protein Crystallogr 2002;40(1):82–92.

[62] Xu G, Wang Q, Ma J. OPUS-Rota4: a gradient-based protein side-chain modeling framework assisted by deep learning-based predictors. Briefings in Bioinformatics 2022;23(1):bbab529.

[63] Zhao J, Cao Y, Zhang L. Exploring the computational methods for protein-ligand binding site prediction. Computational and structural biotechnology journal 2020;18:417–426.

[64] Krivák R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. Journal of cheminformatics 2018;10:1–12.

[65] Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC bioinformatics 2009;10(1):1–11.

[66] Brylinski M, Skolnick J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. Proceedings of the National Academy of sciences 2008;105(1):129–134.

[67] McPartlon M, Xu J. AttnPacker: An end-to-end deep learning method for rotamer-free protein side-chain packing. bioRxiv 2022;p. 2022–03.

[68] Scardino V, Di Filippo JI, Cavasotto CN. How good are AlphaFold models for docking-based virtual screening? iScience 2022;p. 105920.

[69] Zhang Y, Vass M, Shi D, Abualrous E, Chambers JM, Chopra N, et al. Benchmarking refined and unrefined alphafold2 structures for hit discovery. Journal of Chemical Information and Modeling 2022;.

Appendix A: Additional Figures and Tables

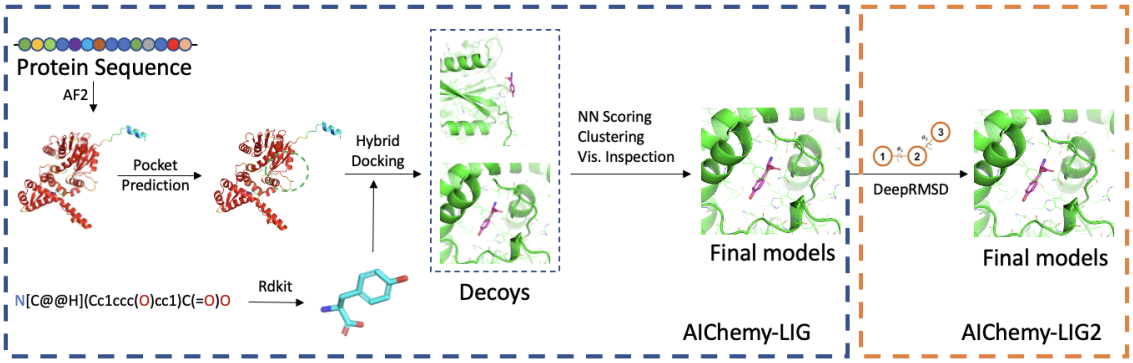

FIGURE S1 The protein-ligand prediction pipeline in CASP15 for group AIchemy-LIG and AIchemy-LIG2.

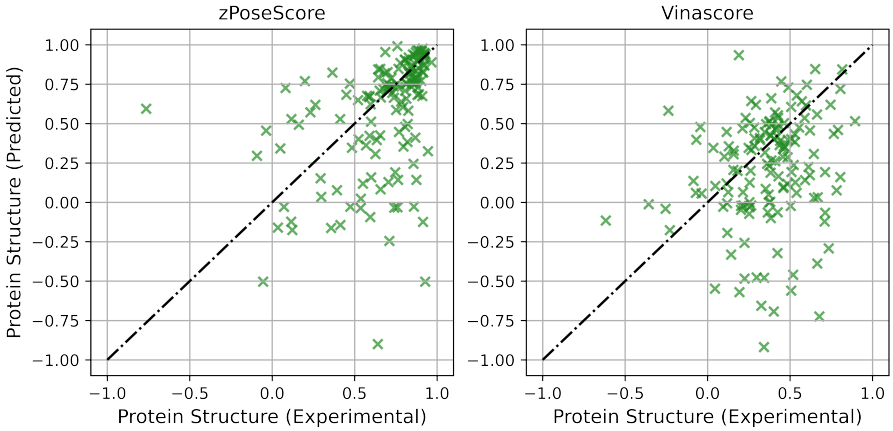

FIGURE S2 Scoring methods are less accurate for ranking docking poses based on predicted pockets.

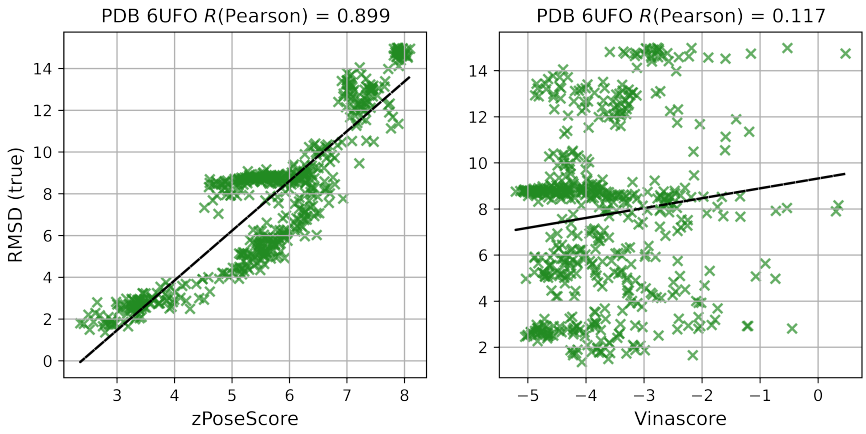

FIGURE S3 Scoring methods are less accurate for ranking docking poses based on predicted pockets.