Scrutinization on Docking Against Individually Generated Target Pockets for Each Ligand

Rui Zhan [1,2,3 #], Wenyi Zhang [1,2,3 #], and Jing Huang [1,2,3 *]

[1] Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou, 310024, China

[2] Westlake AI Therapeutics Lab, Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, 310024, China

[3] Institute of Biology, Westlake Institute for Advanced Study, Hangzhou, 310024, China

# Equal contribution

*Corresponding author: Jing Huang, huangjing@westlake.edu.cn

Abstract: The longstanding challenge of limited exploration in protein receptor conformational space continues to constrain the precision of molecular docking. Ensemble docking, which employs methods such as molecular dynamics simulations to generate multiple receptor conformations for docking, has improved accuracy but remains limited by incomplete sampling and an inability to fully account for ligand-induced fit. To address these limitations, we introduce the concept of individual docking, a novel approach that involves docking against receptor conformations generated individually for each ligand in the docking library. This approach has only very recently become feasible due to advances in protein structure prediction, in particular end-to-end protein-ligand complex prediction technologies exemplified by AlphaFold3 and NeuralPLexer. In this study, we performed individual docking on 27 targets from the DUD-E dataset, using a two-step protocol that integrates NeuralPLexer's inference as the receptor sampling step with subsequent physics-based docking. Our results reveal that individual docking leads to approximately 24% reduction in the enrichment factors compared with standard docking, yet it recovers different sets of active ligands. Detailed analyses of pocket and ligand conformations suggest several potential incompatibilities between deep learning-based and physics-based virtual screening tools.

Keywords: Virtual Screening, Structure Prediction, Ensemble Docking, Individual Docking

Introduction

High-throughput screening has long been the standard method for identifying biologically active compounds, but its application is often limited by high costs. Virtual screening (VS), which simulates compound-target interactions or evaluates compound and pharmacophore similarities computationally, has emerged as a powerful alternative [1-3]. VS has been widely adopted in drug discovery, not only for accelerating early-stage drug discovery but also for applications such as drug repurposing [4], compound optimization [5], and target identification [6]. With the advancement of protein 3D structure determination methods, structure-based virtual screening, such as molecular docking and free energy perturbation calculations, has gained increasing attention [7].

Molecular docking computationally positions molecules within a selected binding pocket through search algorithms, generating binding poses by considering factors such as pocket properties, ligand rotations, translations, and internal degrees of freedom. By employing empirical scoring functions, it further provides a docking score to estimate the complementarity between ligands and targets, enabling the effective ranking of millions of small molecules in virtual libraries. Docking is probably the most popular and successful computational approaches in drug discovery [8]. Early research on molecular docking focused on improving scoring functions and ligand sampling methods. Advancements in scoring functions have sought to enhance docking accuracy by incorporating key structural and physicochemical properties [9], while improvements in sampling techniques have involved iterative random search algorithms such as Monte Carlo and evolutionary algorithms [10,11]. However, a remaining limitation of docking is that the protein structure is kept fixed, which does not account for protein dynamics. Even in flexible docking, only a limited number of side chains can be rotated.

One way to partially account for protein flexibility is ensemble docking [12-14], which involves performing docking calculations against multiple receptor conformations simultaneously. To generate a diverse and representative set of receptor conformations, molecular dynamics (MD) simulations or normal mode analysis of the apo-protein are often employed. While the conformational change induced by ligand binding is often described using the induced fit model, the conformational selection model can also be considered, which suggests that a protein populates all binding-necessary conformations even in its unbound state [15-17]. However, obtaining a convergent protein conformational landscape remains highly challenging even with

enhanced sampling techniques [18]. Moreover, selecting a limited number of conformations from the landscape to perform ensemble docking for a given screening library is equally difficult.

Artificial intelligence (AI), in particular deep learning (DL) methods, offer exciting new approaches to addressing challenges in structure-based drug design such as the scarcity of high-quality protein structures and the limited exploration of protein conformational space [19-24]. Most notably, AlphaFold2 enables highly accurate end-to-end protein structure prediction using its Evoformer module within the Transformer architecture, combined with geometric modeling. AlphaFold1 [25], AlphaFold2 [26], and contemporary methods [27-32] have primarily focused on predicting the apo structures of single-domain proteins. In the 2022 Critical Assessment of Structure Prediction (CASP15) competition [33], protein-ligand complex structure prediction was introduced for the first time. The winning method, CoDock [34], combined template-based docking on AlphaFold2-predicted apo structures with deep learning-based scoring to assess ligand binding predictions.

Diffusion models, inspired by thermodynamics modeling, represent the current state-of-the-art in DL generative models. By learning an estimated score function, diffusion models effectively reduce the dependence on large training datasets. Diffusion models further advance structure prediction by enabling more accurate predictions of biomolecular complexes, including protein interactions with small molecules, metals, nucleic acids, and other biomolecules [35-37]. A wide range of diffusion-based generative models has emerged for predicting the structure of protein-ligand complexes, including AlphaFold3 [38], RoseTTAFold AA [39], NeuralPLexer [40,41], Chai-1 [42], and Boltz-1 [43]. Unlike physics-based docking methods, these DL models perform blind predictions without requiring prior knowledge of the binding pocket and generate fully induced-fit conformations of the target proteins. However, DL models fail to differentiate between binders and non-binders, as they always predict complex structures based on given inputs and lack the capability to directly predict binding affinities. The integration of structure prediction, docking, and virtual screening within a unified computational or DL framework has emerged as a central focus in advancing ligand-protein interaction prediction [44,45]. To this end, effective scientific validation is critical for assessing and understanding the compatibility between physics-based and DL methods, ensuring their synergistic integration.

In this study, we propose an individual docking pipeline that leverages a DL model to predict

protein-ligand complex structures and then performs molecular docking using the predicted protein conformations. To evaluate the performance of this approach, we conducted virtual screening on 27 systems from the DUD-E dataset [46], encompassing a total of 445,730 predicted protein-ligand complex structures generated by the NeuralPLexer model. To assess the flexibility of the binding pockets, we used a 3D pocket structure alignment tool to evaluate the generated conformations. In addition, we analyzed the docking score distribution and compared the predicted ligand poses from NeuralPLexer with the post-docking poses. Through this comprehensive evaluation, we provide insights into the compatibility between DL-based structure prediction and physics-based docking methods, highlighting the potential advantages and limitations of integrating DL models with conventional molecular docking techniques for virtual screening.

Results

Docking calculations have to be performed on individually generated pockets

The 27 systems from the DUD-E dataset were used to investigate ligand-protein interactions with individually generated pockets for each ligand, as used in a previous study benchmarking AlphaFold2-predicted structures for hit discovery [47]. For individual docking, we employed NeuralPLexer to predict ligand-protein complex structures, taking the protein sequence and ligand graph structures as inputs. Subsequently, one can directly compute the docking score using an empirical scoring function, such as Vina [48], to evaluate the interaction based on the ligand-protein conformation predicted by NeuralPLexer. Despite retaining most of the information from the structure prediction, direct scoring performed poorly, with most Vina scores being positive, indicating unreasonable binding as assessed by the Vina scoring function. Among 445,730 cases, 96.42% of the docking scores exceeded 0, and 10.89% were greater than 100 kcal/mol (Fig. S1A). Notably, only 1275 complex structures directly generated by NeuralPLexer achieved docking scores within the typically accepted range of below -6 kcal/mol, accounting for a mere 0.29% (Fig. S1B). Direct calculation of the docking scores suggests a lack of stringent physical constraints in the predicted complex structures, such that some atomistic distances between the protein and ligand can be excessively close and may lead to steric clashes. Consequently, docking calculations, which involve sampling ligand binding poses in the pocket, must be performed to obtain valid docking conformations for evaluating

ligand-protein interactions using docking scores.

To further investigate the impact of NeuralPLexer sampling on virtual screening, we designed two groups of individual docking experiments with progressively decreasing levels of information retention from NeuralPLexer: individual rigid docking (ID-rigid) and individual docking (ID). In ID-rigid, Vina was used to resample the ligand docking orientations while both the protein and ligand conformations were fixed. In contrast, ID preserved only the receptor conformation and allowed the ligand conformations and docking poses to be resampled during the docking calculations. The binding site was determined based on the ligand's predicted position from NeuralPLexer. Standard docking (SD) was performed as the baseline using the crystal structures of the receptors. Virtual screening (VS) performance of individual docking was benchmarked and compared with SD across 27 systems from the DUD-E dataset.

Comparison of VS performance

The enrichment factors (EFs) at 1% cutoff are shown in Table 1, comparing ID with SD. In 9 out of 27 systems studied, for example ACES and AKT2, individual docking led to increased enrichment in active hits when selecting the top 1% of compounds. In another 16 systems, ID showed worse performance compared with SD, while in 2 systems (AMPC and HSP90A) neither ID nor SD identified any active hits at the 1% cutoff. The median enrichment factors for active compounds at the top 1%, 5%, and 10% ranges were 2.35, 2.16, and 2.05 for ID, slightly but consistently lower than the corresponding values for SD, which were 3.08, 2.77, and 2.65, respectively (data for the 5% and 10% cutoff are provided in Table S1). In addition to the EFs, we also calculated the area under the ROC curve (AUC) values (Fig. 1). The average AUC value for ID across all 27 systems studied was 0.63, which was 0.04 lower than the AUC value of 0.67 for SD. The ROC curves aligned with the 1% EF results, showing that SD generally had better VS performance. However, in a few systems, such as TRY1 and VGFR2, ID demonstrated a stronger capability to select active hits. We also note that while in some systems (e.g., MAPK2 and THRB) the ID and ID-rigid pipelines produced almost identical results, in most cases, their performance differed significantly.

Table 1. Virtual screening performance of individual docking evaluated at 1% cutoff.

System	Enrichment Factor		# Active Hits (SD/ID/Shared)	Overlap Rate	Consensus Ranking	
	SD	ID			Docking Score-Based	Borda Count-Based
ACES	3.31	4.41	15/20/2	10.9%	4.19	4.85
AKT2	1.69	4.22	2/5/0	14.1%	0.84	3.38
AMPC	0.00	0.00	0/0/0	23.3%	0.00	0.00
BACE1	4.51	2.05	22/10/2	16.0%	2.05	3.08
BRAF	15.46	10.70	39/27/17	25.0%	15.86	14.27
CDK2	8.62	3.37	69/27/14	24.0%	3.87	6.37
DPP4	0.74	0.65	8/7/0	16.7%	0.83	0.37
EGFR	3.36	0.24	28/2/1	9.4%	3.36	1.08
FA10	8.17	1.76	65/14/3	10.8%	8.05	4.90
FABP4	18.73	1.70	11/1/1	20.0%	20.44	10.22
GRIA2	3.02	2.35	9/7/3	15.3%	2.35	3.69
HSP90A	0.00	0.00	0/0/0	3.9%	0.00	0.00
IGF1R	3.08	0.00	7/0/0	9.3%	0.00	3.52
ITAL	0.43	1.28	1/3/0	11.1%	0.00	0.00
MAPK2	6.74	1.93	14/4/2	27.7%	3.37	6.26
MET	1.64	2.45	4/6/0	0.9%	2.05	2.45
MK10	1.59	1.06	3/2/1	24.3%	1.59	2.12
MK14	1.64	2.73	15/25/2	15.0%	2.29	1.96
PPARD	2.42	0.35	7/1/0	8.1%	2.07	1.38
PPARG	5.12	3.87	37/28/12	16.5%	4.42	6.22
PTN1	7.07	4.42	16/10/2	14.3%	5.75	7.07
RXRA	28.90	1.84	47/3/2	5.1%	18.45	4.30
TGFR1	11.69	6.02	33/17/10	25.6%	10.63	9.92
THRB	3.60	3.95	31/34/7	23.4%	3.37	3.95
TRY1	2.24	3.29	17/25/4	17.2%	2.89	2.10
TRYB1	1.17	3.35	2/6/0	11.4%	3.50	2.92
VGFR2	2.73	4.66	17/29/2	2.3%	4.18	3.86
AVERAGE	**5.74**	**2.70**	--	14.9%	**4.68**	**4.08**
MEDIAN	**3.08**	**2.35**	--	15.0%	**3.36**	**3.52**

 # Active Hits: The number of active hits identified by SD and ID, as well as the number identified by both pipelines. **Overlap Rate**: The percentage of hits identified by both the SD and ID pipelines within the top 1% ranking. **Consensus Ranking**: VS pipelines that rank compounds by combining SD and ID results. **Docking Score-Based** ranking uses the lower value of the SD and ID scores, while **Borda Count-Based** ranking selects compounds by taking the order of ranks in an alternating way from the SD and ID pipelines.

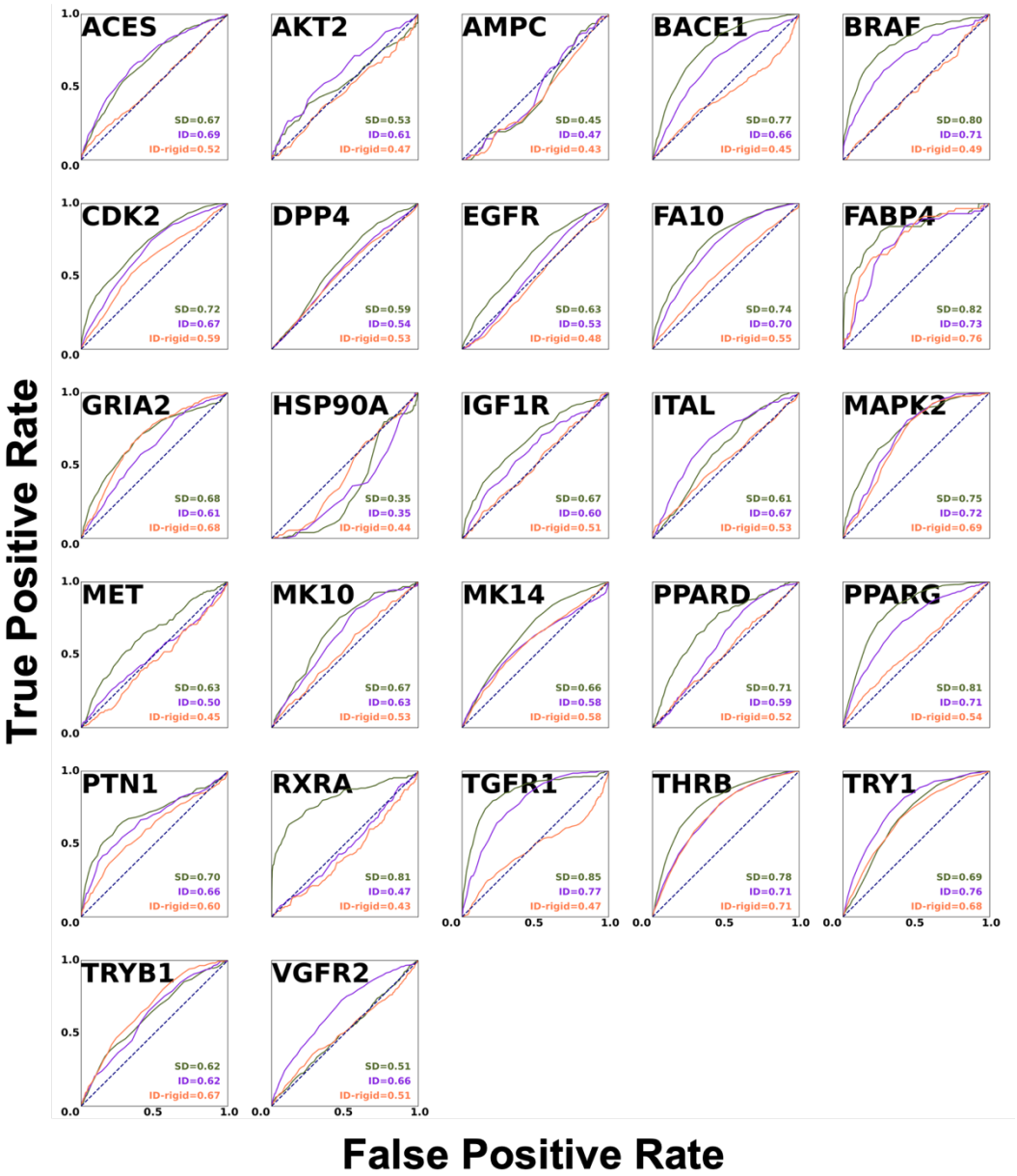

Figure 1. ROC curves and AUC values for standard and individual docking pipelines.

We also examined the active hits identified from the top 1% ranked compounds using ID and SD and found significant differences in the compounds predicted by the two pipelines (Table 1). For example, in the ACES system, SD identified 15 active hits and ID identified 20, with only 2 compounds overlapping. In DPP4, SD identified 8 active hits and ID found 7, but none of these 15 hits overlapped. The BRAF system had the highest number of shared active hits, with SD identifying 39 active hits and ID detecting 27, 17 of which were shared. Notably, this still means that individual docking identified 10 unique active hits compared to standard docking. This echoes a recent observation that docking calculations using AlphaFold2-predicted target conformations achieve similar VS performance compared to using

experimentally determined structures, but the identified active hits are markedly different [49]. Taking decoys into account, the overlap between the top 1% SD and the top 1% ID compounds ranged from 0.9% to 27.7%, with a median of 15.0%. Such a low overlap rate indicates that, even with identical docking settings, using either a single experimental target conformation or a library of individually generated conformations results in the prioritization of very different compounds. The origin of this discrepancy will be analyzed in the following subsections.

Pocket similarity and flexibility

We evaluated the differences between the predicted and crystal pockets using the pocket shape alignment tool, PPS-align [50]. The PPS score, ranging from 0 to 1, reflects the similarity between the predicted and crystal pockets, with a score closer to 1 indicating a high resemblance and a score closer to 0 suggesting greater pocket flexibility. In total, we used NeuralPLexer to predict ligand-protein complexes for 27 systems from the DUD-E dataset, generating 445,730 conformation samples. The computed pocket similarity scores (PPS scores) and root mean square deviations (RMSDs) for each sample are shown in Fig. 2. The average and median PPS scores were 0.59 and 0.61, respectively, while the average and median RMSD values were 1.08 Å and 1.06 Å, respectively. Of these, 24,219 pockets had PPS scores below 0.4 and RMSD values greater than 1.5 Å. Notably, the majority of these pockets were identified in the DPP4 system, which contributed 17,428 cases, followed by VGFR2, with 1,932 samples.

Interestingly, for these systems, the enrichment factors increased with individual docking, either directly (e.g., VGFR2, which increased from SD 2.73 to ID 4.66) or indirectly (e.g., DPP4, where SD and ID identified completely different active hits such that consensus ranking could be beneficial; see below). Most of the remaining systems produced approximately 10-150 conformation samples with similar characteristics, suggesting relatively low levels of pocket flexibility. We also calculated the Pearson's correlation coefficient (PCC) between the PPS score and the number of heavy atoms in the ligand. The PCC value was 0.08, indicating that the dynamic pocket generation in NeuralPLexer was not influenced by ligand size. This result highlights that the DL-based model can effectively handle pocket flexibility, adapting to variations in pocket geometry regardless of ligand size.

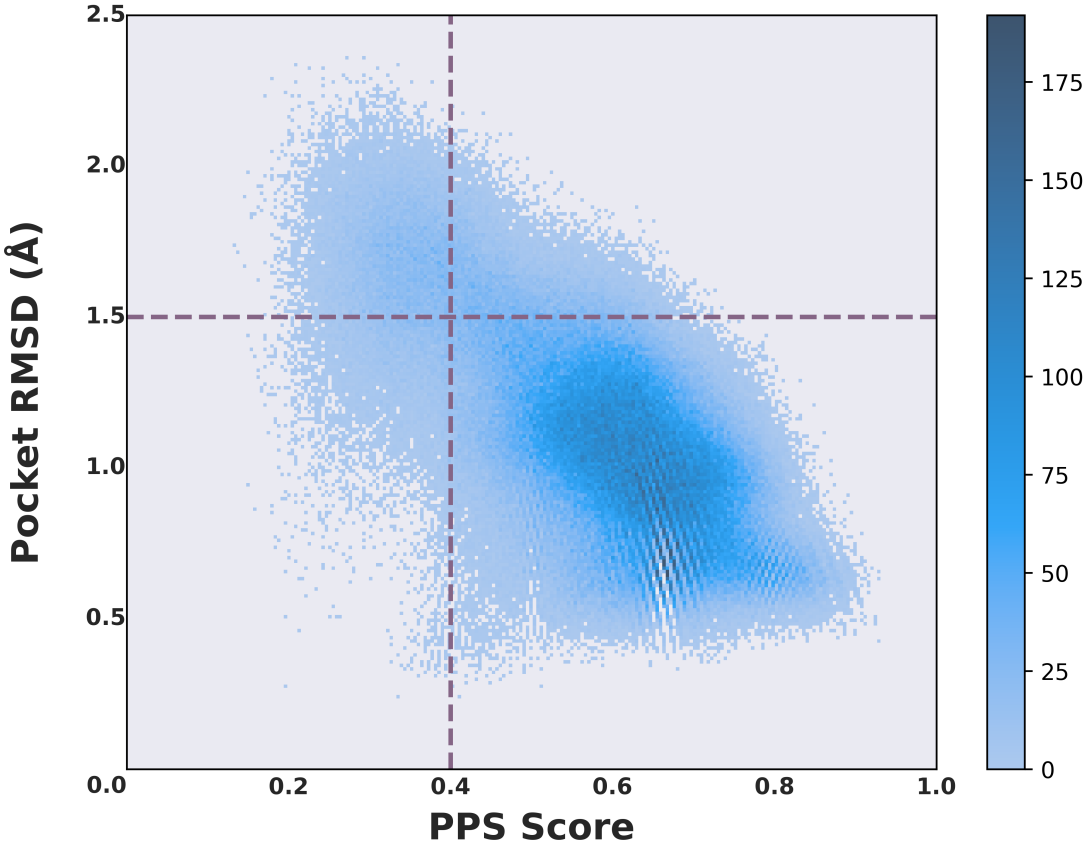

Figure 2. Distribution of PPS scores and pocket RMSD values for pockets generated by NeuralPLexer. The PPS scores were calculated using the PPS-align method, which aligns the generated pockets with the corresponding crystal pocket.

Comparison of docking scores

We further analyzed the docking scores obtained with SD, ID and ID-rigid, which include approximately 440,000 docking cases for each pipeline. Significant differences in the distribution of docking scores were observed, as shown in the histogram plot in Fig. 3A. Interestingly, ID-rigid resulted in a significantly larger number of cases with very favorable binding (docking scores below -11 kcal/mol, Fig. 3B), as well as a greater number of cases with very unfavorable binding (positive docking scores). Essentially, ID-rigid adjusts the predicted binding pose through rigid-body translation and rotation of the ligand, while ID further allows for ligand deformation and internal rotation along rotatable bonds. A large number of positive docking scores associated with ID-rigid indicates cases where the docking program was unable to identify a reasonable binding pose given the conformation of the protein receptor. It's worth noting that, in contrast, ID resulted in much fewer unfavorable binding cases compared to SD, although in general ID and SD had similar docking score distributions, with ID showing a

slightly higher probability in the [-14, -12] and [-7, -4] kcal/mol ranges.

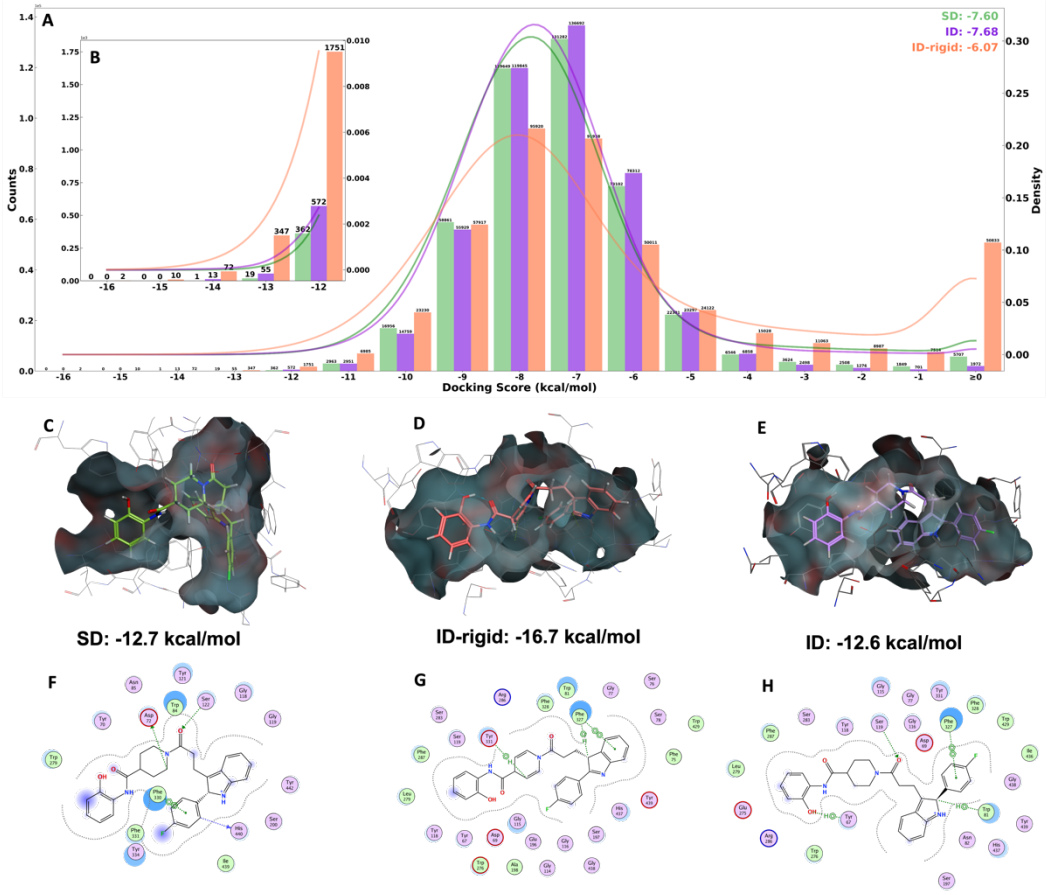

Figure 3. Comparison of docking score distributions for the SD, ID, and ID-rigid pipelines. A: Overall docking score distribution for SD, ID, and ID-rigid. B: Docking score distribution for scores below -11 kcal/mol. C-E: Docking poses and interactions of ZINC63182445 targeting ACES, obtained with SD, ID-rigid, and ID, respectively. F-H: Detailed interaction diagrams for SD, ID-rigid, and ID, respectively.

A zoom-in on the favorable binding range is shown in Fig. 3B. ID-rigid computed 9,167 cases with docking scores below -11 kcal/mol, approximately 2.6 times more than SD (3,345 cases) and ID (3,591 cases). Some ID-rigid docking scores were lower than -15 kcal/mol, while no such cases were observed with SD or ID, suggesting that the DL model's sampling process could optimize docking scores for certain ligands. Specifically, we selected the case with the most favorable docking score with ID-rigid for conformational analysis, as shown in Figures 3C-H. This case involves the docking result of the 22,573rd decoy ligand (ZINC63182445) targeting ACES. The docking scores were -12.7 kcal/mol for SD, -12.6 kcal/mol for ID, and -16.7 kcal/mol for ID-rigid.

By comparing the docking poses of SD and ID-rigid, we found that while SD identified four hydrogen bond interactions, the ligand underwent significant bending to fit the pocket shape, resulting in a high-energy conformation and consequently poorer docking scores. In contrast, the ID-rigid docking pose demonstrated a significant adjustment in pocket shape due to NeuralPLexer sampling (PPS score: 0.64, pocket RMSD: 1.15 Å), creating a more compatible fit with the ligand conformation. This adjustment facilitated an aromatic interaction network with the ligand through the critical residue F330. Despite having fewer hydrogen bond interactions than SD, ID-rigid achieved a substantial improvement in docking score, showcasing the advantage of DL-sampled protein pocket conformations in improving docking performance. Moreover, docking into the same pocket resulted in overall similar ligand-protein binding modes between ID and ID-rigid. However, a twist at one end of the ligand, with the fluorobenzene moiety making contact with F330 instead of the indole moiety, effectively reset the docking score to -12.6 kcal/mol (Fig. 3G and 3H). This example demonstrates the sensitivity of docking scores to subtle changes in ligand orientation and interactions within the given pocket.

To further investigate the significance of more favorable docking scores in VS performance, we analyzed the proportion of active ligands in a subset of 9,167 cases identified by ID-rigid with docking scores below -11 kcal/mol. Among these, only 309 were active ligands, accounting for a mere 3.4%, which corresponds to an enrichment factor of 1.31. This is notably lower than the 1% EFs achieved by any of the three pipelines. Therefore, the improvement in docking scores associated with individual rigid docking does not translate into enhanced virtual screening performance. We also compared the docking poses and NeuralPLexer-predicted poses of compounds in this subset using the DockRMSD method. DockRMSD is a convenient tool for comparing ligand structures without requiring RMS alignment of the target conformation, allowing efficient calculation of the RMSD values between two ligand sets [51]. Relatively large DockRMSD values, with an average of 5.81 Å and a median of 7.74 Å, were observed. This indicates that, despite preserving receptor and ligand conformations from DL predictions in ID-rigid, most cases with improved docking scores still relied on docking algorithms to search for significantly different ligand orientations.

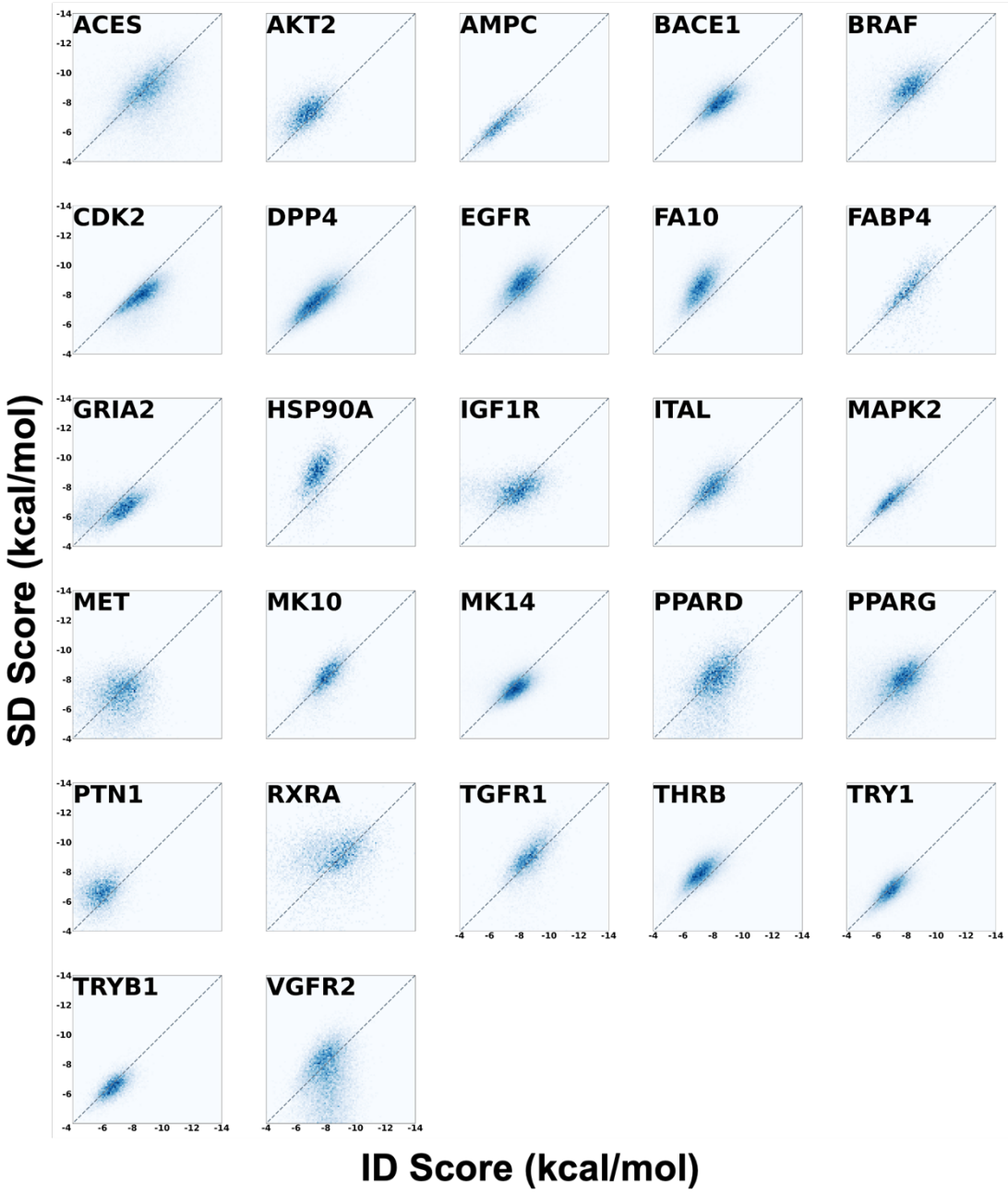

Figure 4. Comparison of docking scores obtained with ID and SD pipelines.

We also performed a direct comparison of docking scores for each ligand between SD and ID, as illustrated by the scatter plots in Figure 4. Only docking scores below -4 kcal/mol were considered. For most systems, docking scores obtained using two different types of target conformations correlated well, with correlation coefficients ranging from 0.15 (VGFR2) to 0.87 (AMPC). The median correlation coefficient was 0.57. Furthermore, among the 427,094 cases included in the analysis, 210,129 (49.2%) had docking scores that favored ID. This observation aligns with our findings in Figure 3A, where the two methods exhibited comparable docking score distributions. Furthermore, among the four systems where ID scores were consistently

more negative than SD scores (BACE1, CDK2, GRIA2 and MK14), only MK14 demonstrated an improvement in EF with the ID pipeline. This is consistent with our analysis of the ID-rigid favorable subset, where more favorable docking scores did not correlate with an enhancement in virtual screening capability, which requires differentiating actives from decoys.

Evaluation of consensus ranking pipelines

An important observation from our study is that individual docking is able to identify distinct active compounds compared to standard docking. The top 1% ranked compounds using ID and SD also showed a low overlap rate of 15%. We further explore the possibility to combine results from both pipelines to further enhance the VS enrichment. Two such consensus ranking methods were evaluated: one based on docking scores and the other utilizing the Borda Count approach. In the first method, each compound was assigned the smaller of the SD and ID docking scores as its final score, based on which compounds were ranked. In the second consensus ranking method, each compound was assigned a rank in both the SD and ID pipelines based on its respective docking scores. The rank m of a compound was converted into a score of $(N-m)$ where N is the total number of compounds in the VS library. This conversion ensures that the score is inversely proportional to the rank, assigning higher scores to higher-ranked compounds. The scores from both pipelines were then summed to generated a final score used for ranking.

The EFs at the 1% cutoff were calculated after re-ranking compounds using these methods (Table 1). In the docking score-based consensus ranking, 9 systems achieved a higher EF than SD, 4 systems were identical (including 2 systems where both EFs were 0), and 14 systems had a lower EF compared to SD. Similarly, in the Borda Count-based consensus ranking, 10 systems outperformed SD, 3 systems had the same EF (including 2 with EF=0), and 14 systems performed worse. Notably, the median EF across the 27 systems improved for both consensus methods compared to SD, increasing slightly from 3.08 to 3.36 and 3.52, respectively. These results demonstrate that the additional information generated by individual docking could potentially be used to enhance VS workflows.

Impact of docking on DL-sampled ligand poses

It's interesting to analyze the impact of docking on NeuralPLexer-predicted ligand binding poses. Figure 5A shows the distribution of DockRMSD values with ligand kept rigid (ID-rigid) or flexible (ID) during the docking calculations. The average DockRMSD values were 6.41 Å

and 6.47 Å for ID and ID-rigid, with median values of 6.40 Å and 6.60 Å, respectively. These significant deviations indicate substantial changes in ligand conformations, sometimes even resulting in reversals of binding poses within the pocket. Such results suggest that binding pose information provided by NeuralPLexer is hardly recognized or utilized by Vina, highlighting the incompatibility between DL-based predictions and physics-based docking calculations.

We further analyzed a minority of cases where DockRMSD values were less than 2 Å, hypothesizing that Vina might be able to effectively utilize the protein-ligand information captured by NeuralPLexer in these instances. We extracted this subset from either ID or ID-rigid to analyze the distribution of docking scores (Fig. 5B and 5C). In the ID pipeline, there were 1,396 cases with ligand RMSD less than 2 Å, of which 823 (59.0%) had docking scores lower than those from SD. ID-rigid showed an increased portion, with 4,653 of 6,971 cases (66.7%) having more negative docking scores compared to SD ones.

To demonstrate how the docking scores of these cases improve through individual docking pipelines, we selected a representative case to conduct conformational analysis, as shown in Figure 5. This case involved the docking result of the 77th active ligand (CHEMBL1202881) targeting ACES. The ligand DockRMSD values for ID and ID-rigid between post-sampling and post-docking stages were 1.19 Å and 0.69 Å, respectively, indicating that both pipelines successfully preserved the deep model's understanding of the ligand binding pose. Consequently, the conformations in SD, ID, and ID-Rigid reflected an increasing trend in retaining information from the DL model. A comparison with the crystal binding mode (Fig. 5D and 5H) further highlighted the advantages of DL model sampling in improving docking scores.

In SD, which did not contain any NeuralPLexer information (Fig. 5E and 5I), the ligand's right-terminal phenyl group appeared in a position similar to that in the crystal structure and successfully reproduced the π-π stacking interaction with W84 and hydrogen bonding with H440. However, the larger SD ligand was forced to fit the protein pocket in an over-bending conformation, consistent with the observations from the ACES 22,573rd decoy case (Fig. 3C). In the ID binding mode, the protein pocket underwent significant conformational changes (PPS score: 0.52, pocket RMSD: 1.19 Å) induced by the ligand, allowing the ligand to adopt a conformation closer to its natural state. This adjustment also revealed that this larger ligand utilized its pyridazine group and aromatic amine to form new hydrogen bonds with G118 and

S122, enhancing ligand-protein interactions. However, ID failed to recognize W84 as a key interacting residue, which is crucial for real binding and a desirable feature to be reproduced in molecular docking.

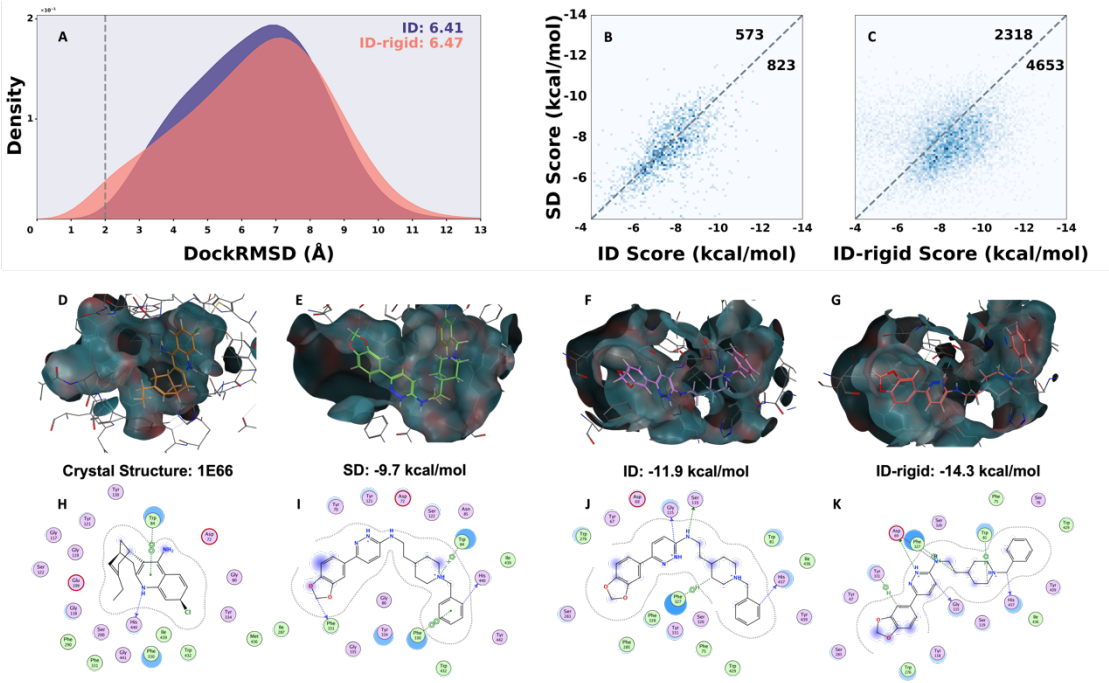

Figure 5. Comparison of SD, ID, and ID-rigid pipelines in ligand binding poses. A: Distribution of ligand DockRMSD values with the ID and ID-rigid pipelines. B-C: Comparison of docking scores for compounds within DockRMSD values less than 2 Å. D-G: Binding poses and interactions for crystal structure, SD, ID, and ID-rigid pipelines. H-K: Detailed interaction diagrams for crystal structure, SD, ID, and ID-rigid.

These limitations observed in SD and ID were effectively optimized in this case by ID-rigid (Fig. 5G and 5K). By fully retaining the DL-sampled information on both the protein pocket and ligand conformation, ID-rigid generated an energetically very favorable ligand conformation. It also successfully reproduced the key contributions of W84 and H440 to the binding mode. Moreover, ID-rigid identified additional interactions, including backbone and side-chain hydrogen bonds with the 1,3-benzodioxole and pyridazine groups of the ligand compared. Overall, these observations align with the trends noted in the conformational analysis of the ACES decoy ligand: DL models tend to sample protein pockets and ligand conformations that maximize ligand-protein interactions, which might also contribute to enhanced docking scores in some cases.

Subsequently, we calculated the EFs of these two subsets. In the ID subset, among 1396 ligands, 60 were active, yielding an equivalent enrichment factor of 1.67, which remains below the average level. Similarly, in the ID-rigid subset, no improvement in VS performance was observed, with an equivalent EF of 1.66 or 297 active hits out of 6,971 cases. This outcome aligns with our previous observation: even when the conformational sampling performed by the DL model can optimize docking scores for protein-ligand pairs (in the ID-rigid subset, the proportion is 2/3), it does not make active ligands bind more favorably relative stronger to decoy compounds. Instead, it may represent a non-discriminatory fitting approach that universally forms more ligand-protein interactions.

Discussion and Conclusions

The flexibility of protein receptor conformations is crucial for accurately predicting potential binding ligands. For example, Mobley and co-workers showed in a model system (13 ligands targeting the L99A mutant of T4 lysozyme) that performing energy minimization of ligand-bound proteins significantly improved the correlation between calculated absolute binding free energies and experimental values, increasing the correlation coefficient from -0.05 to 0.82 [52]. DL-based structure prediction methods, such as NeuralPLexer and AlphaFold3, provide a new and computationally efficient way to generate ligand-specific protein structures. The large-scale generation of ligand-protein complex structures creates an opportunity to mitigate the impact of assuming a fixed receptor conformation during docking. This study explored the concept of individual docking, in which complex structures were predicted for each individual ligand using DL models and the predicted protein structures were used for docking calculations. While showing the feasibility of ID, our findings reveal several critical limitations that hinder its VS performance.

Benchmarking the performance of the ID pipeline against SD using 27 systems from the DUD-E dataset showed general underperformance in metrics such as EFs and AUC. These results were consistent across different docking methods, including Vina (presented in the main text) and MOE (presented in SI), indicating that integrating DL-based predictions with physics-based scoring has not yet resulted in improvements in hit enrichment. For example, the median EF values for the top 1%, 5%, and 10% ranked compounds using MOE docking were 2.38, 1.87, and 1.67, respectively, as shown in Table S1 and Figure S2. Notably, an analysis of the top 1% of screened compounds revealed limited overlap between ID and SD, with only 15% of the ligands shared. This finding underscores the potential of ID to generate structurally distinct hit

compounds, broadening the diversity of VS results. The additional information obtained by ID also allows the design of consensus ranking methods, which showed improved median EF values compared to SD.

An in-depth analysis of pocket conformational changes using the PPS score demonstrated that NeuralPLexer is capable to generate pockets geometrically tailored to individual ligands, a unique capability not present in traditional VS approaches. However, this customization comes at the cost of reduced compatibility with the energetics (empirical docking scores) used for scoring and sampling in docking workflows. Variations in docking score distributions among SD, ID, and ID-rigid pipelines suggest that the DL model struggles to accurately model the subtle physical interactions essential for protein-ligand binding. Additionally, its predictions failed to reliably distinguish between actives and decoys, as evidenced by specific examples. This limitation might be partly due to NeuralPLexer prioritizing spatial arrangements based on distance constraints rather than the physical and chemical properties critical to binding interactions. Further analysis of the DockRMSDs between NeuralPLexer-generated poses and docking-refined poses revealed significant rearrangements of the ligand from the predicted structure, addressing atomic clashes and enhancing interactions between protein and ligand atoms according to docking energy evaluations.

These findings highlight a fundamental incompatibility between traditional docking techniques and DL-predicted structures, emphasizing the need for methodological advancements. Docking primarily focuses on physical interactions, whereas DL-based methods aim to optimize the distances of key interaction pairs in a static manner. DL-based structure prediction methods appear to maximize the number of ligand-protein interactions, forming more interaction pairs compared to docking calculations. However, this comes at a cost, as unfavorable interactions and even physical clashes are often neglected. Consequently, the straightforward combination of docking and DL-predicted structures in the ID and ID-rigid pipelines does not enhance VS performance.

To improve the performance of AI-based ID, two key strategies hold promise. First, incorporating physical interaction constraints during AI model training could improve the structural realism and interpretability of predicted protein-ligand complexes, making them more compatible with physics-based binding free energy estimations [53]. Second, training or fine-tuning models to distinguish actives from decoys might increase the accuracy and

reliability of structure predictions by injecting deeper insights into ligand-protein interactions. We also note that the landscape of ligand-protein complex prediction models is advancing rapidly. For example, NeuralPLexer, the conformational sampling model used in this study, has already progressed to its third-generation version [41]. Although ID currently underperforms SD in conventional VS metrics, its ability to generate ligand-specific pockets and diversify virtual screening outcomes underscores its potential as a complementary strategy. Further methodological improvements could help ID evolve into a useful tool that supplements standard docking. The insights obtained in this study lay a foundation for more effective integration of DL-based and physics-based VS methods, advancing computational drug discovery.

Methods

Dataset

This study employs a subset of the DUD-E dataset, specifically DUD-E 27, as the benchmark for individual docking experiments. The dataset includes 27 protein targets and a total of 445,730 compounds, comprising 11,447 active compounds and 434,283 decoy compounds. All ligand data were downloaded from the official DUD-E website (https://dude.docking.org/) and formatted using Molecular Operating Environment (MOE2022) software [54]. Additionally, 3D structures for all proteins were retrieved in PDB format from the Protein Data Bank (PDB) website (https://www.rcsb.org/), in accordance with the descriptions provided by DUD-E.

Protocol of standard docking

For standard docking, both Vina and MOE [54] were used. In the Vina docking setup, receptor and ligand files were converted from PDB to PDBQT format using MGLTools [55]. The docking site was defined as a cube with a side length of 15 Å, with the center of mass of the ligand in each crystal structure serving as the central point. In the MOE docking setup, energy minimizations were performed to prepare the molecular 3D structures for docking. The docking site was automatically defined by selecting the 'ligand' option. Thirty docking poses were generated, and the pose with the best score was selected as the final result. The receptor was kept rigid during docking.

Protocol of individual docking

During the sampling process, the ligand in SDF format and the receptor in PDB format are used as inputs for NeuralPLexer. The 'batched_structure_sampling' function of NeuralPLexer [40]

predicts 16 complex structures, with the one having the highest pLDDT score selected as the final result. All unspecified parameters are set to their default values. Templates were not used during the sampling process to prevent generating receptor conformations similar to the templates, thus enabling ligand-induced conformational fitting. The inference process is conducted in parallel, with each parallel execution taking approximately 6 hours to predict about 1,000 complex structures on an NVIDIA L40 GPU core. In this study, version 2 of NeuralPLexer was used, unless otherwise specified.

The individual docking scoring pipeline utilizes the '--score_only' option in the Vina program. In the ID pipeline, flexible ligand conformations are enabled by setting the 'inactivate_all_torsions' option to 'False' during the processing step. In contrast, in the ID-rigid pipeline, this option is activated to preserve the ligand in the same pose as predicted by NeuralPLexer.

Assessment of pocket similarity and ligand RMSD

We use PPS-align [50] to assess the conformational differences between each sampled pocket predicted by NeuralPLexer and its corresponding crystal structure. Residues within 5 Å of the ligand atoms are defined as the pocket. DockRMSD [51] is then used to evaluate the ligand RMSD between the NeuralPLexer predictions and the post-docking results from both the ID and ID-rigid pipelines. This analysis aims to compare the ligand pose differences between NeuralPLexer predictions and the results after docking.

Acknowledgments

The work is supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2023C03109, 2024SSYS0036), the National Natural Science Foundation of China (32171247), the Zhejiang Provincial Natural Science Foundation of China (LQ23F020011), and the Westlake Education Foundation. We thank the Westlake University Supercomputer Center for computational resources and related assistance.

Reference

[1] C. Gorgulla et al., “An open-source drug discovery platform enables ultra-large virtual screens,” Nature, vol. 580, no. 7805, pp. 663–668, 2020.

[2] A. A. Sadybekov et al., “Synthon-based ligand discovery in virtual libraries of over 11 billion compounds,” Nature, vol. 601, no. 7893, pp. 452–459, 2022.

[3] W. Zhang and J. Huang, “EViS: An Enhanced Virtual Screening Approach Based on Pocket–Ligand Similarity,” J Chem Inf Model, vol. 62, no. 3, pp. 498–510, 2022.

[4] X. Lin et al., “Life beyond kinases: structure-based discovery of sorafenib as nanomolar antagonist of 5-HT receptors,” J Med Chem, vol. 55, no. 12, pp. 5749–5759, 2012.

[5] A. L. Kaplan et al., “Bespoke library docking for 5-HT2A receptor agonists with antidepressant activity,” Nature, vol. 610, no. 7932, pp. 582–591, 2022.

[6] A. Lee, K. Lee, and D. Kim, “Using reverse docking for target identification and its applications for drug discovery,” Expert Opin Drug Discov, vol. 11, no. 7, pp. 707–715, 2016.

[7] R. Qian, J. Xue, Y. Xu, and J. Huang, “Alchemical Transformations and Beyond: Recent Advances and Real-World Applications of Free Energy Calculations in Drug Discovery,” J Chem Inf Model, vol. 64, no. 19, pp. 7214–7237, 2024.

[8] J. Lyu et al., “Ultra-large library docking for discovering new chemotypes,” Nature, vol. 566, no. 7743, pp. 224–229, 2019.

[9] D. Jiang et al., “Interactiongraphnet: A novel and efficient deep graph representation learning framework for accurate protein–ligand interaction predictions,” J Med Chem, vol. 64, no. 24, pp. 18209–18232, 2021.

[10] W. Zhang, E. W. Bell, M. Yin, and Y. Zhang, “EDock: blind protein–ligand docking by replica-exchange monte carlo simulation,” J Cheminform, vol. 12, pp. 1–17, 2020.

[11] J. Meiler and D. Baker, “ROSETTALIGAND: Protein–small molecule docking with full side-chain flexibility,” Proteins: Structure, Function, and Bioinformatics, vol. 65, no. 3, pp. 538–548, 2006.

[12] A. Stank, D. B. Kokh, J. C. Fuller, and R. C. Wade, “Protein binding pocket dynamics,” Acc Chem Res, vol. 49, no. 5, pp. 809–815, 2016.

[13] R. E. Amaro et al., “Ensemble docking in drug discovery,” Biophys J, vol. 114, no. 10, pp. 2271–2278, 2018.

[14] W. Evangelista Falcon, S. R. Ellingson, J. C. Smith, and J. Baudry, “Ensemble docking in drug discovery: how many protein configurations from molecular dynamics simulations are needed to reproduce known ligand binding?,” J Phys Chem B, vol. 123, no. 25, pp. 5189–5195, 2019.

[15] K. Henzler-Wildman and D. Kern, “Dynamic personalities of proteins,” Nature, vol. 450, no. 7172, pp. 964–972, 2007.

[16] R. Nussinov and C.-J. Tsai, “Allostery in disease and in drug discovery,” Cell, vol. 153, no. 2, pp. 293–305, 2013.

[17] P. Ayaz et al., “Structural mechanism of a drug-binding process involving a large conformational change of the protein target,” Nat Commun, vol. 14, no. 1, p. 1885, 2023.

[18] A. J. Clark et al., “Prediction of protein–ligand binding poses via a combination of induced fit docking and metadynamics simulations,” J Chem Theory Comput, vol. 12, no. 6, pp. 2990–2998, 2016.

[19] G. Monteiro da Silva, J. Y. Cui, D. C. Dalgarno, G. P. Lisi, and B. M. Rubenstein, “High-throughput prediction of protein conformational distributions with subsampled AlphaFold2,” Nat

Commun, vol. 15, no. 1, p. 2464, 2024.
[20] G. Janson, G. Valdes-Garcia, L. Heo, and M. Feig, "Direct generation of protein conformational ensembles via machine learning," Nat Commun, vol. 14, no. 1, p. 774, 2023.
[21] T. Xie and J. Huang, "Can Protein Structure Prediction Methods Capture Alternative Conformations of Membrane Transporters?," J Chem Inf Model, vol. 64, no. 8, pp. 3524–3536, 2024.
[22] T. Xie, Z. Song, and J. Huang, "Conditioned protein structure prediction," PRX Life, vol. 2, no. 4, p. 043001, 2024.
[23] H. K. Wayment-Steele et al., "Predicting multiple conformations via sequence clustering and AlphaFold2," Nature, vol. 625, no. 7996, pp. 832–839, 2024.
[24] N. Raouraoua et al., "MassiveFold: unveiling AlphaFold's hidden potential with optimized and parallelized massive sampling," Nat Comput Sci, pp. 1–5, 2024.
[25] A. W. Senior et al., "Improved protein structure prediction using potentials from deep learning," Nature, vol. 577, no. 7792, pp. 706–710, 2020.
[26] J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," Nature, vol. 596, no. 7873, pp. 583–589, 2021.
[27] M. Baek, R. McHugh, I. Anishchenko, H. Jiang, D. Baker, and F. DiMaio, "Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA," Nat Methods, vol. 21, no. 1, pp. 117–121, 2024.
[28] M. Baek et al., "Accurate prediction of protein structures and interactions using a three-track neural network," Science (1979), vol. 373, no. 6557, pp. 871–876, 2021.
[29] G. Ahdritz et al., "OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization," Nat Methods, pp. 1–11, 2024.
[30] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, "ColabFold: making protein folding accessible to all," Nat Methods, vol. 19, no. 6, pp. 679–682, 2022.
[31] R. Wu et al., "High-resolution de novo structure prediction from primary sequence," BioRxiv, pp. 2022–2027, 2022.
[32] Z. Lin et al., "Evolutionary-scale prediction of atomic-level protein structure with a language model," Science (1979), vol. 379, no. 6637, pp. 1123–1130, 2023.
[33] A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, and J. Moult, "Critical assessment of methods of protein structure prediction (CASP)-Round XV," Proteins: Structure, Function, and Bioinformatics, vol. 91, no. 12, pp. 1539–1549, 2023.
[34] M. Pang et al., "CoDock-Ligand: Combined template-based docking and CNN-based scoring in ligand binding prediction," BMC Bioinformatics, vol. 24, no. 1, p. 444, 2023.
[35] J. B. Ingraham et al., "Illuminating protein space with a programmable generative model," Nature, vol. 623, no. 7989, pp. 1070–1078, 2023.
[36] J. L. Watson et al., "De novo design of protein structure and function with RFdiffusion," Nature, vol. 620, no. 7976, pp. 1089–1100, 2023.
[37] K. E. Wu et al., "Protein structure generation via folding diffusion," Nat Commun, vol. 15, no. 1, p. 1059, 2024.
[38] J. Abramson et al., "Accurate structure prediction of biomolecular interactions with AlphaFold 3," Nature, pp. 1–3, 2024.
[39] R. Krishna et al., "Generalized biomolecular modeling and design with RoseTTAFold All-Atom," Science (1979), vol. 384, no. 6693, p. eadl2528, 2024.

[40] Z. Qiao, W. Nie, A. Vahdat, T. F. Miller III, and A. Anandkumar, "State-specific protein–ligand complex structure prediction with a multiscale deep generative model," Nat Mach Intell, vol. 6, no. 2, pp. 195–208, 2024.

[41] Z. Qiao et al., "NeuralPLexer3: Physio-Realistic Biomolecular Complex Structure Prediction with Flow Models," arXiv preprint arXiv:2412.10743, 2024.

[42] C. Discovery et al., "Chai-1: Decoding the molecular interactions of life," bioRxiv, pp. 2010–2024, 2024.

[43] J. Wohlwend et al., "Boltz-1: Democratizing Biomolecular Interaction Modeling," bioRxiv, pp. 2011–2024, 2024.

[44] G. Corso, H. Stärk, B. Jing, R. Barzilay, and T. Jaakkola, "Diffdock: Diffusion steps, twists, and turns for molecular docking," arXiv preprint arXiv:2210.01776, 2022.

[45] W. Lu et al., "DynamicBind: Predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model," Nat Commun, vol. 15, no. 1, p. 1071, 2024.

[46] M. M. Mysinger, M. Carchia, J. J. Irwin, and B. K. Shoichet, "Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking," J Med Chem, vol. 55, no. 14, pp. 6582–6594, 2012.

[47] Y. Zhang et al., "Benchmarking refined and unrefined AlphaFold2 structures for hit discovery," J Chem Inf Model, vol. 63, no. 6, pp. 1656–1667, 2023.

[48] O. Trott and A. J. Olson, "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," J Comput Chem, vol. 31, no. 2, pp. 455–461, 2010.

[49] J. Lyu et al., "AlphaFold2 structures guide prospective ligand discovery," Science (1979), p. eadn6354, 2024.

[50] J. Hu, X. Zhou, D. Yu, G. Zhang, and Y. Zhang, "PPS-align: an user-friendly pocket structural alignment server," https://zhanglab.dcmb.med.umich.edu/PPS-align/.

[51] E. W. Bell and Y. Zhang, "DockRMSD: an open-source tool for atom mapping and RMSD calculation of symmetric molecules through graph isomorphism," J Cheminform, vol. 11, pp. 1–9, 2019.

[52] D. L. Mobley, A. P. Graves, J. D. Chodera, A. C. McReynolds, B. K. Shoichet, and K. A. Dill, "Predicting absolute ligand binding free energies to a simple model site," J Mol Biol, vol. 371, no. 4, pp. 1118–1134, 2007.

[53] W. Zhang, K. Zhang, and J. Huang, "A Simple Way to Incorporate Target Structural Information in Molecular Generative Models," J Chem Inf Model, vol. 63, no. 12, pp. 3719–3730, 2023.

[54] S. Vilar, G. Cozza, and S. Moro, "Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery," Curr Top Med Chem, vol. 8, no. 18, pp. 1555–1572, 2008.

[55] G. M. Morris et al., "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility," J Comput Chem, vol. 30, no. 16, pp. 2785–2791, 2009.