

DockFormer: Affinity Prediction and Flexible Docking with Pair Transformer

Ben Shor^{*} and Dina Schneidman-Duhovny[†]

Department of Computer Science, Hebrew University of Jerusalem, Jerusalem, Israel



(Received 23 March 2025; accepted 11 August 2025; published 18 September 2025)

Protein-small molecule interactions, or receptor-ligand interactions, are essential for understanding biological processes and advancing drug design. Despite advancements, existing prediction models of these interactions still lack the capabilities and accuracy needed to replace traditional screening. In this paper, we introduce DockFormer, a method that leverages multimodal learning to predict both the binding affinity and structure of these interactions. DockFormer employs fully flexible docking, where no part of the receptor remains rigid, by adapting the AlphaFold2 architecture. Instead of relying on protein sequences and multiple sequence alignments, DockFormer uses predicted receptor structures as input. This modification enables the model to concentrate on ligand docking prediction, rather than protein folding, while preserving full receptor flexibility. The streamlined design also reduces the model size to just eight layers, compared to AlphaFold2's 48 layers, greatly accelerating the inference process and making it more efficient for large-scale screening. When evaluated on affinity benchmarks such as CASF-2016, PLINDER, and the recently released CASP16 ligand screening benchmark, DockFormer performs comparably to or better than state-of-the-art methods, which typically rely on templates or bound structures as input. On structural benchmarks such as PoseBusters and PLINDER, DockFormer demonstrated success rates of 20% and 15%, respectively.

DOI: [10.1103/1862-rs2k](https://doi.org/10.1103/1862-rs2k)

I. INTRODUCTION

Interactions between proteins and small molecules, or receptor-ligand interactions, are critical for biological functions and drug design [1–3]. A major challenge is the screening process, in which a huge database is searched for ligands that bind to a specific target receptor with high affinity [4]. Currently, this screening process heavily relies on wet lab experiments and physics-based computations [5], both of which are expensive and time consuming. The importance of these interactions stresses a need for accurate *in silico* screening prediction methods.

The interaction prediction task is defined as the prediction of structure, prediction of affinity, or both. The prediction of bound interaction structure given the receptor holo or apo structure is generally known as the docking of a ligand in the receptor. Docking methods can be physics based, such as GLIDE [6] and AutoDock [7], or deep-learning based [8,9], such as DiffDock [10]. These methods use the receptor backbone structure as rigid and change the position and conformation of the ligand and possibly the side chains. However, the performance of these methods deteriorates if only apo structure is available. Prediction of both protein structure and the interaction is commonly referred to as co-folding. Co-Folding methods, while potentially more accurate, are

resource intensive and complex [11,12]. For many years, a major obstacle to co-folding methods was the lack of accurate methods to predict individual protein structures.

The release of AlphaFold2 [13] marked a significant advancement in protein structure prediction, utilizing a transformer-based architecture called EvoFormer to process single and pairwise sequence representations. In its successor, AlphaFold3, a more lightweight variation named Pairformer was used [11], where multiple sequence alignments (MSAs) are preprocessed separately. Although the success of AlphaFold2 was partly due to MSAs, studies have shown that even without MSAs, the model can predict protein interaction interfaces [14,15], stressing the competence of this architecture. Despite these advances, structural models generated are not ideal for docking ligands, as evidenced by the reduced accuracy of methods like GLIDE (44% to 15%) and DiffDock (38% to 21%) when using AlphaFold2 predictions [5,16]. This is because small-molecule binding can alter the protein conformation in mechanisms known as induced fit or population shift [17,18]. Change can be small, for example, only in the side chains, or significant when binding to the cryptic allosteric sites [19].

In addition to structure prediction methods, there are methods designed to predict affinity given the bound structure [20–25]. There are also approaches for predicting affinity without the bound structure [26,27]. These methods can work with either the protein structure or just its sequences. While they perform well when trained on a specific protein, they struggle to generalize to new proteins outside their training set. This limitation often arises because, without the bound structure, the models cannot learn the specific interactions influencing affinity and instead tend to memorize small molecule features and their impact on affinity in specific proteins [28].

^{*}Contact author: ben.shor@mail.huji.ac.il

[†]Contact author: dina.schneidman@mail.huji.ac.il

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

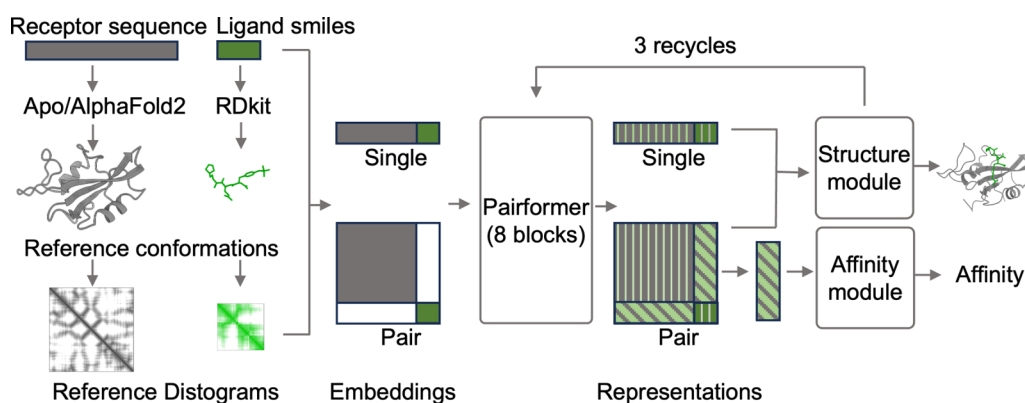


FIG. 1. Model architecture. The inputs to DockFormer are receptor and ligand structure distograms. The structures are not required to be in the holo conformation; they can be an apo-conformation, experimental structure, or a predicted structure. For the ligand, a random reference conformation is generated. The distogram of the structures and the types of amino acids and atoms are then used to create a one-dimensional embedding for each input token referenced as single embedding and a two-dimensional input embedding for each pair of tokens referenced as pair embedding. These embeddings are the input for the deep-learning-based model. The model consists of eight Pairformer blocks that each output a single and pair representation. Both representations are used by the structure module to generate a structure. Only the parts in the pair representation that represent a pair of a ligand atom and a receptor amino acid are considered in the affinity module.

Multimodal learning refers to a machine learning approach where a model processes and integrates multiple types of data (modalities) to improve learning and prediction and previously had some validation for receptor-ligand prediction tasks [29–31]. Here we apply multimodal learning to train a multi-task model, DockFormer, that performs two prediction tasks of different data types: structure and affinity. In computer vision, it has been shown that if tasks are similar enough we could get improved accuracy in all tasks by using a single model, as opposed to different specialized models [32]. In the field of protein folding, multimodal methods such as ESM3 [33] have also been proven to be successful in obtaining state-of-the-art performance on multiple tasks such as structure prediction, structure design, and masked position prediction.

Here we introduce DockFormer, a multimodal model designed to predict both receptor-ligand affinity and structure. DockFormer leverages the Pairformer architecture to identify biological and physical interfaces without relying on MSAs. Instead, it predicts receptor-ligand contacts based solely on geometric and physicochemical complementarity.

To balance the trade-off between docking and co-folding, DockFormer adopts a unique approach: While it cofolds the receptor, it requires the backbone distogram, representing $C\alpha$ - $C\alpha$ distances from a predicted or experimentally determined apo structure, as input. Since the backbone distogram serves only as an optional structural constraint, the model retains the ability to cofold receptor-ligand complexes while avoiding the complexities of full protein folding. This design choice simplifies inference and reduces computational costs.

II. METHODS

A. DockFormer architecture and key modifications

The DockFormer architecture builds upon AlphaFold2 [13] architecture, with several key modifications to support ligand modeling and enhance efficiency (see Fig. 1). The most notable change is the replacement of MSA and template inputs with an input receptor structure. Rather than inferring contacts

from evolutionary information in the MSA, the model directly receives approximated contact data. This modification simplifies the Evoformer, making it more similar to the Pairformer architecture introduced in AlphaFold3 [11]. The number of layers in the Evoformer portion of the model is reduced to 8 (instead of 48 in AlphaFold2). Another major adaptation is the incorporation of tokens representing ligand atoms alongside amino acid tokens. Finally, DockFormer incorporates an affinity module, enabling multimodal training for both docking and affinity prediction. To the best of our knowledge, this is the first adaptation of the AlphaFold2 architecture for flexible docking and the first adaptation of AlphaFold2 pairwise representations for affinity prediction.

The first step is the input embedding. The input for DockFormer is an AlphaFold2 prediction, or an experimentally determined apo conformation, of the receptor structure and a reference structure of the ligand generated by RDKit [34]. The output of the embedding stage is a single embedding and a pair embedding. The single embedding is a one-dimensional vector containing an embedding vector for each token, either an amino acid or a ligand atom. The pair embedding is a two-dimensional vector that contains an embedding for each pair of tokens. The features that are used for embedding protein tokens are only the amino acid types. For ligand tokens, the features used are atom types, charge, chirality, and bond types between atoms. Additionally, the input structures are converted into intraprotein and intraligand distograms, that is, the distances between each pair of tokens. For proteins, $C\alpha$ - $C\alpha$ distances are discretized into 15 bins evenly distributed between 3.25 and 20.75 Å, similar to the recycling process in AlphaFold2. For ligands, the interatomic distances are assigned to one of the ten bins, evenly distributed between 0.75 and 9.75 Å. The encoding of each matched bin is added to the embedding in the pair embedding.

After generating single and pair embedding, those are passed through a transformer-based model, with eight blocks, where each block is similar to a Pairformer block. Recycling

is also applied similarly to AlphaFold2 with three recycles. The final single and pair representations that are outputted by the last block are passed to the structure module as implemented in AlphaFold2. For ligand tokens, the frame rotations are ignored, and only translations of frames are used as the atom positions. An additional optional step is to replace the predicted ligand positions with the closest conformation from a set of 1000 conformations by RDkit. This step prevents impossible bond distances or conformations.

B. Affinity prediction module

In addition to the structure module, there is also an affinity module, which predicts the affinity as a classification task. The classification task is defined as 30 bins for affinities between 0 and 15, where the number represents the negative logarithm of the nanomolar measured affinity. The final predicted affinity is a weighted average of the bins, weighted by the probability predicted for each bin. To predict the probability of each affinity bin, we test two approaches: The first is interface based and the second is based on a classification token.

In the first approach, we apply a linear layer to every vector in the pair representation that represents a pair of a ligand atom and a receptor amino acid. Each such vector is weighted by a layer designated to weigh vectors for affinity relevance and also by considering the logits predicted by the intercontact auxiliary head. The weighted sum is used to predict the probability for each affinity bin.

In the second approach, we add a classification token to the model input, similar to what is common in natural language processing models [35]. The vector corresponding to the affinity token in the single representation is passed through a linear layer to predict the affinity using regression. Although the interface-based approach is more complicated, it leverages data learned from structures, in the form of receptor-ligand contacts and encourages biological restraints on affinity prediction.

C. Training and model variants

We have implemented several new auxiliary heads and loss modifications. The intercontact auxiliary head is trained to predict, given a pairwise representation, the binary classification task of whether the amino acid C α atom and ligand atom have a distance of at most 5 Å. The binding site prediction head predicts for a single representation of a receptor token whether this token is within at most 5 Å of the ligand. We have also added a variation of frame-aligned point error loss that aligns frames of the receptor, but only considers the error of ligand atoms, and by weighing this loss, we have forced the model to emphasize the contact between receptor and ligand.

We trained three versions of the model using different training datasets: DockFormer-PDBBind, DockFormer-Screen, and DockFormer-PLINDER.

DockFormer-PDBBind was trained on 18 101 samples from the PDBBind 2020 dataset. To enable fair evaluation against CASF-2016, we filtered out 479 samples from the original PDBBind dataset that were highly similar to CASF-2016, following the approach of previous methods [20]. Specifically, we removed any sample where the lig-

and's ECFP4 fingerprint had a Tanimoto coefficient above 0.90 compared to any ligand in CASF-2016 or where a protein chain had over 90% sequence identity with a chain in CASF-2016.

DockFormer-Screen was designed to enhance the model's ability to generalize for ligand screening, where a single protein interacts with many ligands. Since PDBBind contains diverse protein-ligand complexes, direct training on it does not optimize for this scenario. To address this, we curated a dataset of 12 374 samples across only 119 proteins (averaging 104 ligands per protein) and fine-tuned DockFormer-PDBBind on it. Affinity data were sourced from BindingDB [36], and receptor-ligand structures were predicted using Boltz-1 [37].

DockFormer-PLINDER was trained on the PLINDER [38] training set to assess generalization on its strictly separated test set. To maximize the utility of the dataset, we divided it using provided structural clusters and sampled data during training inversely proportional to cluster size. We excluded clusters representing different binding pockets of the same ligand (i.e., same protein and ligand but different structural clusters), as these could introduce ambiguity. Additionally, we capped the number of times each ligand could appear in the training set at 5 to prevent bias toward specific ligands. After filtering, we retained 6979 clusters with a total of 138 402 samples. Finally, to enhance affinity prediction, we fine-tuned the model on all 36 594 affinity-labeled samples in the training set.

All models were trained with a batch size of 16, and after 25 000 optimization steps, we introduced an additional loss term on interatomic distances within the ligand to improve structural learning.

III. RESULTS

A. CASF-2016 affinity benchmark

The CASF-2016 benchmark [39] is widely used to evaluate affinity prediction models, particularly those trained on PDBBind. It consists of 57 proteins, each with five ligands for which experimentally measured affinities are available. This benchmark allows for assessing a model's ability to predict binding affinities and separately its ability to rank different ligands for the same protein, making it suitable for evaluating both general affinity prediction and ligand screening performance.

To assess model performance, we use different correlation metrics depending on the nature of the prediction task. Pearson correlation measures the linear relationship between predicted and experimental affinity values and is typically used when evaluating global affinity prediction, where receptor-ligand pairs are unrelated. However, when comparing screening performance, where multiple ligands must be ranked relative to a single protein, Kendall's Tau is the preferred metric [40]. Kendall's Tau assesses the consistency of ligand ranking, regardless of absolute affinity values, making it a more appropriate measure of a model's ability to differentiate strong from weak binders.

Using CASF-2016, DockFormer-PDBBind achieved a Pearson correlation of 0.803 when using the interface-based affinity module and 0.817 when using the classification-

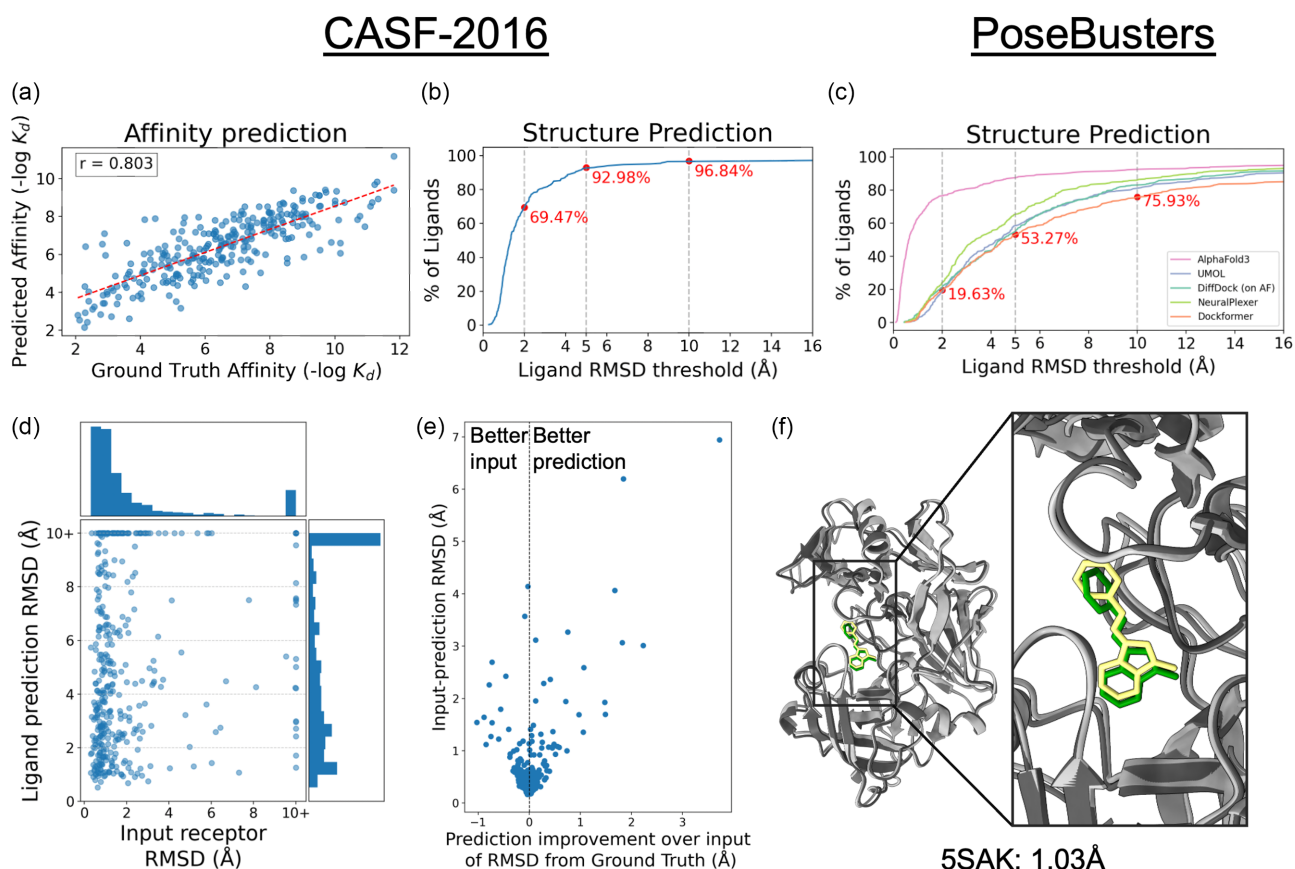


FIG. 2. Results on CASF-2016 and PoseBusters benchmarks. (a) Predicted affinity compared to experimentally measured affinity in the CASF-2016 dataset ($n = 285$). Axes are the negative logarithm of the nanomolar K_i or K_d affinity. The Pearson correlation is 0.803. (b) Structure prediction results for samples in the CASF-2016 dataset under continuous threshold. (c) Structure prediction results for samples in the PoseBusters dataset ($n = 428$) under continuous threshold. (d) RMSD between predicted ligand positions when pockets aligned, compared to RMSD between ground truth receptor structure and input receptor structure. (e) The y axis is the RMSD between the input receptor structure and the predicted receptor structure. The x axis is the result of subtracting the RMSD between the input structure and the ground truth structure from the RMSD of the predicted structure and the ground truth structure. Positive x values represent an improvement in the accuracy of DockFormer prediction compared to the input structure. (f) Example of accurate prediction by DockFormer from the PoseBusters benchmark: predicted receptor (light gray) and ligand (green) structure vs experimental receptor (dark gray) and ligand (yellow) structure.

token-based affinity module [Fig. 2(a)]. These results are comparable to state-of-the-art methods (Table I).

CASF-2016 can also be used as a ranking power benchmark, where Kendall's Tau is computed separately for each of the 57 proteins by comparing the predicted rankings of their five ligands. Using this metric, DockFormer-

PDBBind achieved an average Kendall's Tau of 0.60, while DockFormer-Screen improved this to 0.62. This suggests that fine-tuning on screening-specific data enhances the model's ability to rank ligands within the same protein, as opposed to a model trained solely on general receptor-ligand affinity predictions. These results highlight DockFormer's strong

TABLE I. Comparison of affinity prediction performance across two benchmarks (CASF-2016 and PLINDER) using DockFormer and competing methods. Results are reported for two DockFormer affinity modules: interface-based (I) and token-based (T). Performance is evaluated using root-mean-square error (RMSE) and Pearson correlation. Lower RMSE and higher Pearson correlation indicate better performance.

Method	CASF-2016		PLINDER	
	RMSE ↓	Pearson correlation ↑	RMSE ↓	Pearson correlation ↑
ConPlex [26]				0.290
Pafnucy [21]	1.42	0.78	1.322	0.569
PLANET [20]	1.247	0.824	1.742	0.347
DockFormer (I)	1.299	0.803	1.199	0.656
DockFormer (T)	1.262	0.817	0.960	0.753

performance in both global affinity prediction and ligand screening, reinforcing its effectiveness in structure-based drug discovery applications.

B. Structure prediction on CASF-2016 and PoseBusters

For structural evaluation, we assessed ligand root-mean-square deviation (RMSD) relative to the experimental structure, aligning based on the C α atoms of the protein pocket amino acids, amino acids with at least one atom within 8 Å of any of the ligand atoms. A ligand RMSD below 2 Å is considered a successful prediction, while an RMSD under 10 Å generally indicates correct binding pocket identification. DockFormer accurately predicted ligand structures with an RMSD less than 2 Å for 69% of complexes in the CASF-2016 dataset and correctly identified the binding pocket (RMSD less than 10 Å) for 93% of complexes [Fig. 2(b)].

Despite filtering the training data to remove complexes similar to those in CASF-2016, it appears that the dataset still contains structurally related complexes. To further evaluate the generalization ability of DockFormer for structural prediction, we used PoseBusters, a widely adopted benchmark for protein-small molecule interactions [41]. This dataset consists of 428 interactions, none of which are included in PDBBind. On this benchmark, DockFormer-PDBBind successfully predicted 20% of complexes with ligand RMSD less than 2 Å, a performance comparable to competing methods [10,16,42], except for AlphaFold3 and its variants [11,37,43] [Figs. 2(c) and 2(f)].

A particularly interesting comparison is between DockFormer and Umol [16], where DockFormer achieved 20% success, slightly outperforming Umol's 18% success rate. Since both models are based on the AlphaFold2 architecture, the main distinctions are DockFormer's reduced number of layers and its removal of MSA from the pipeline. This suggests that using input structures instead of co-folding from scratch can maintain performance while significantly reducing model complexity.

C. Impact of input receptor structure accuracy

To analyze the relationship between input receptor structure accuracy and docking performance, we examined the C α RMSD between the input and ground truth structures in the PoseBusters test set. In 73% of cases, the input receptor structure had a C α RMSD less than 2 Å, and only 8% of samples exceeded a 10-Å RMSD [Fig. 2(d)]. We also examined how accurately DockFormer predicted the binding pocket, which is the most critical region for protein-ligand interactions. In 93% of cases, the pocket RMSD was less than 2 Å, and in 99% it was less than 5 Å. These findings support the central premise of DockFormer, that leveraging apo structures for flexible docking rather than predicting the protein structure itself from scratch using MSA is a valid and effective approach.

We also investigated protein flexibility during docking by comparing the input and output protein structures. Excluding 38 cases where the input structure was either highly inaccurate (greater than 10-Å RMSD) or too large (greater than 1300 residues), we found that in 97% of cases, the receptor structure remained largely unchanged [C α RMSD less than 1 Å,

Fig. 2(e)]. However, in cases where significant changes occurred, the predicted structure was often closer to the ground truth than the input, suggesting that DockFormer refines receptor structures to better fit the ligand.

To test whether DockFormer depends on an accurate receptor conformation, we evaluated its performance using holo-input structures instead of AlphaFold2-predicted receptor structures. This resulted in a slight performance improvement from 20% to 21%, indicating that DockFormer is not strongly dependent on receptor structure accuracy.

D. Generalization on PLINDER

It has been observed that structure prediction accuracy correlates with the similarity between test samples and training data [41]. The PLINDER dataset [38] was recently introduced with stricter train-test separation criteria to address this concern. To benchmark generalization, we trained DockFormer-PLINDER exclusively on the PLINDER training set and tested it on the PLINDER test set.

On PLINDER, DockFormer's structure prediction success rate (ligand RMSD less than 2 Å) dropped from 20% to 15% compared to PDBBind training. This highlights the importance of avoiding data leakage while also demonstrating DockFormer's ability to generalize.

For affinity prediction, we evaluated 112 test set samples with experimentally measured affinities. Using the interface-based affinity module, DockFormer achieved a Pearson correlation of 0.656, which increased to 0.753 when using the classification-token affinity module. This aligns with the minor drop in structure prediction accuracy, while still maintaining strong overall performance. Notably, competing methods achieved a maximum Pearson correlation of 0.569 (Table I).

Additionally, some of these competing methods [20,21] require the experimental bound structure to predict affinity, whereas DockFormer operates fully independently. Furthermore, all competing methods were trained on datasets without strict train-test separation, raising the possibility of data leakage.

E. Validation on CASP16 ligand targets

To further validate DockFormer, we evaluated its performance in the CASP16 Pharmaceutical Protein-Ligand Challenge, which includes two target proteins (chymase and autotaxin) with multiple ligands and experimentally measured affinities. We compared DockFormer-Screen to the top-performing groups in the challenge for both structure and affinity prediction, using the interface-based affinity module for affinity estimation.

DockFormer's results were comparable to the top-scoring groups in both structure and affinity prediction (Fig. 3). Notably, while most CASP16 participants utilized templates and manual refinements, DockFormer generated fully automated predictions with no manual intervention. It is also important to emphasize that we compare against only the top-performing groups among at least 32 competitors, representing state-of-the-art methods in receptor-ligand interaction prediction.

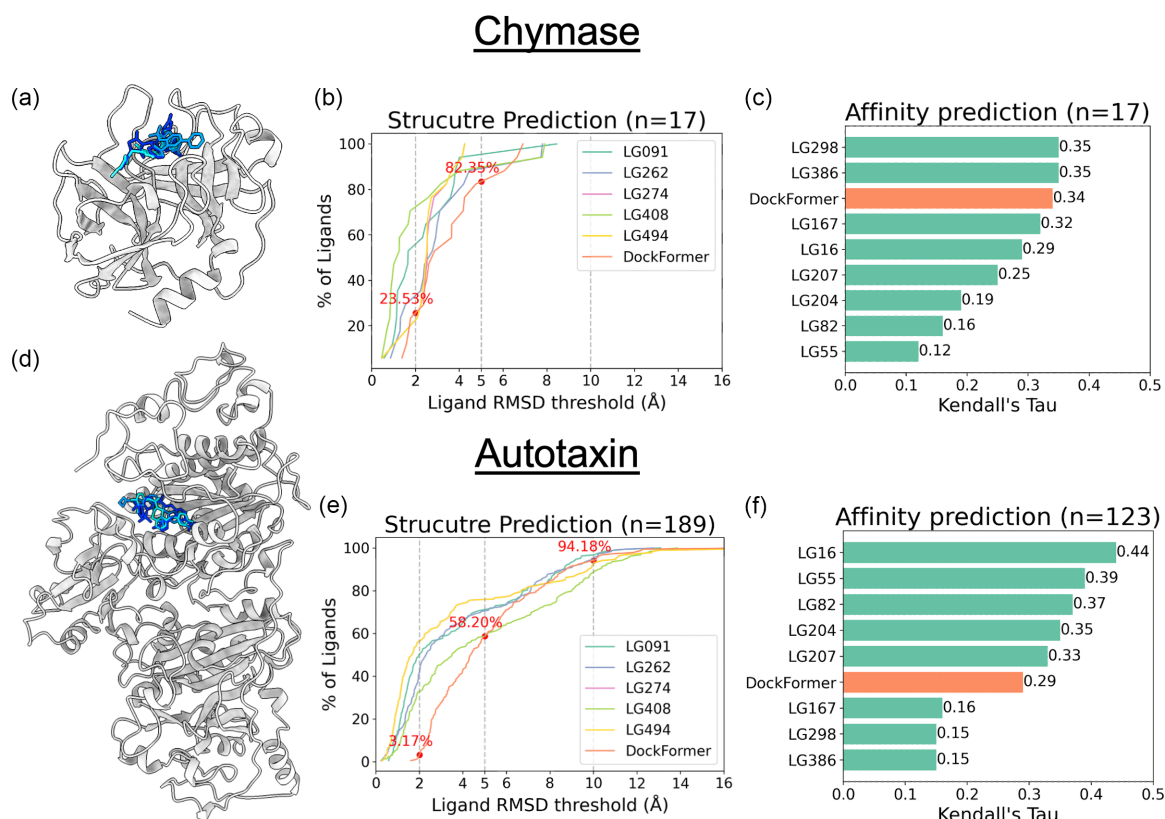


FIG. 3. Results on the CASP16 benchmark. (a) Experimental structure of chymase (gray) with three of the ligands imposed onto it (shades of blue). (b) Structure prediction accuracy for 17 ligands of chymase structure under a continuous RMSD threshold. Accuracy is shown for DockFormer and for top-performing groups in structure prediction of ligands in CASP16. (c) Kendall's Tau of DockFormer-Screen interface-based affinity predictions and experimentally measured affinities for 17 ligands bound to chymase, compared to top-performing groups in affinity prediction in CASP16. (d) Experimental structure of autotaxin (gray) with three of the ligands imposed onto it (shades of blue). (e) Structure prediction accuracy for 189 ligands of autotaxin structure under a continuous RMSD threshold. Accuracy is shown for DockFormer and for top-performing groups in structure prediction of ligands in CASP16. (f) Kendall's Tau of DockFormer-Screen interface-based affinity predictions and experimentally measured affinities for 123 ligands bound to autotaxin (93 of them also had structure), compared to top-performing groups in affinity prediction in CASP16.

While the main advantage of DockFormer compared to its competitors in CASP16 is the fully automated pipeline, an important observation is that different teams excelled on the different protein targets. This can stem from the use of domain-specific knowledge or custom optimizations tailored to particular proteins or from compatibility between certain approaches to specific protein targets. DockFormer performed consistently well across both targets without relying on such prior knowledge, highlighting its generalizability. This also illustrates a broader challenge in the field: For any given protein, it remains difficult to predict in advance which method will perform best. As such, DockFormer can be used as a decent start point, or it could be effectively integrated into an ensemble of models, either by identifying the top-performing method per target or by combining outputs into a consensus prediction.

When comparing DockFormer-PDBBind to DockFormer-Screen, we observe no substantial difference in structure prediction, but a clear improvement in affinity prediction for DockFormer-Screen. For chymase, DockFormer-PDBBind achieves a Kendall's Tau of 0.24 (compared to 0.34 with DockFormer-Screen), while for autotaxin, the correlation is 0.18 (compared to 0.29 with DockFormer-Screen). In con-

trast, DockFormer-PLINDER performs significantly worse in both structure and affinity prediction, achieving only 0.04 for chymase and 0.05 for autotaxin in affinity correlation.

Additionally, we compared the classification-token-based affinity module to the interface-based module. For autotaxin, the classification-token module yielded slightly better performance (0.30 vs 0.29), whereas for chymase performance dropped significantly (0.06 vs 0.34), highlighting the model's sensitivity to specific proteins, suggesting that certain affinity modules may be more suited to particular receptor-ligand interactions.

F. Ablation study

To assess the contributions of different components of the DockFormer architecture, we conducted ablation studies using three model variants. In one variant, we entirely removed the structural component, excluding both structural inputs and the structure prediction task, to evaluate how well the Pairformer architecture alone can predict affinities. The other two variants differed in the number of Pairformer layers: a smaller version with two layers and a larger one with 16 layers. For consistency and computational feasibility, all variants, along

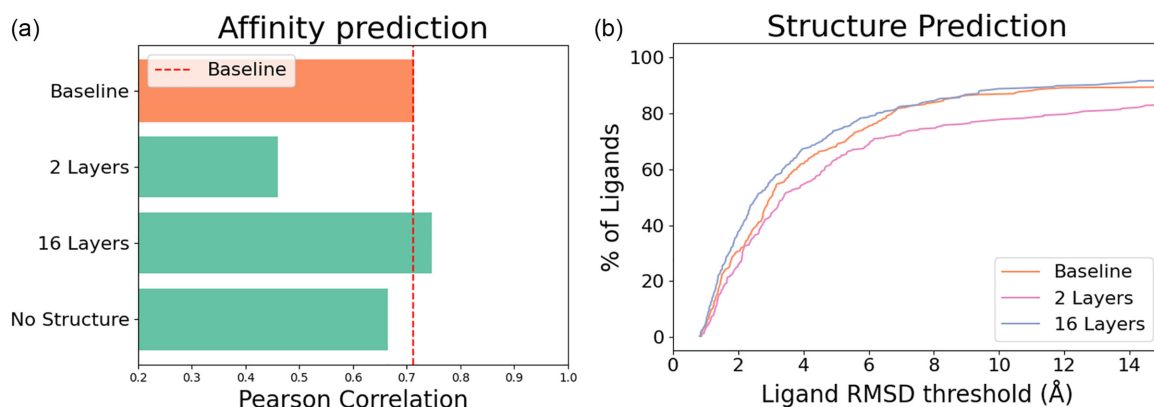


FIG. 4. Ablation tests on CASF-2016. (a) Predicted affinity to ground truth affinity Pearson correlation for each variation of the model. (b) Structure prediction accuracy, for its model variation, except for the variation that does not predict structures, under a continuous threshold.

with a baseline model using the default configuration, were trained for 50 000 optimization steps with a reduced crop size of 256 on the PDBBind dataset.

Results show that the baseline and large models achieve comparable affinity prediction performance [Pearson correlations of 0.71 and 0.75, respectively, Fig. 4(a)], while the smaller model performs significantly worse (0.46). This suggests that affinity prediction is nontrivial and requires sufficient model capacity. The benefit of increased capacity is also evident in the structure prediction task, where accuracy improves with model size [Fig. 4(b)]. Notably, the variant without any structural loss still performs reasonably well in affinity prediction (Pearson correlation of 0.66), though it falls short of the baseline. This indicates that while the Pairformer architecture is intrinsically effective for modeling affinity, incorporating structural information further enhances performance.

G. Further analysis

A key strength of DockFormer is its lightweight architecture, enhancing prediction efficiency. On the PoseBusters dataset, the median prediction time was 2.67 s, with an average of 6.11 s. Although these runtimes do not account for the generation of the input receptor structure, they remain highly relevant for screening processes where the same receptor is tested against multiple ligands, making the receptor structure generation a negligible factor.

One could expect that the accurate prediction of affinity will be dependent on the accurate prediction of structure, as affinity is highly dependent on the protein-ligand contacts. However, it is clear that in many cases DockFormer was not able to predict an accurate structure, but was able to predict an accurate affinity. A similar pattern was observed during training, where the affinity prediction loss converged much faster compared to structure prediction losses. This is not completely nonintuitive, as there are many known models for affinity prediction that are not based on structure. These findings suggest that the Pairformer architecture can learn to predict the interaction properties, even without predicting the structure itself.

Raw results used to create the figures and analysis in this section are available as Supplemental Material [44].

IV. DISCUSSION

We have presented DockFormer, a method for predicting structure and affinity of receptor-ligand interactions that integrates two key concepts into the AlphaFold2 [13] architecture. First, the method uses an approximated receptor structure, either apo conformation or prediction, as an input, enabling the model to specialize in the docking task over the protein folding task, while keeping the receptor structure flexible to capture conformational changes. This also enables reducing the model architecture to provide efficiency in training and inference. Second, by employing multitask learning for simultaneous affinity and structure prediction, the network gains additional information that can further enhance its accuracy.

Our results demonstrate that DockFormer achieves competitive structure prediction performance while excelling in affinity prediction and ligand ranking. While DockFormer successfully predicted ligand poses with reasonable RMSD values, its structural accuracy remains lower than that of diffusion-based models such as AlphaFold3 [11] and Boltz-1 [37]. One potential reason for this difference is that diffusion architectures explicitly model fine-grained ligand-receptor interactions and optimize side-chain flexibility, whereas DockFormer treats side chains implicitly, which may limit structural accuracy. Despite this, DockFormer demonstrated strong generalization on the PoseBusters and PLINDER benchmarks while maintaining high efficiency, making it well suited for high-throughput virtual screening where speed is crucial.

Beyond structure prediction, DockFormer demonstrated state-of-the-art affinity prediction performance, achieving high Pearson correlations across multiple benchmarks, including CASF-2016 and PLINDER. Its ability to rank ligands within the same protein, as seen in CASF-2016 scoring benchmarks and CASP16, further underscores its effectiveness in screening scenarios. Interestingly, even in cases where structural accuracy was limited, DockFormer often produced accurate affinity predictions, suggesting that the Pairformer architecture effectively captures binding interactions without relying on perfect structure prediction, as also evident by the ablation study. These findings suggest that DockFormer is a powerful and scalable approach to structure-based drug discovery, with potential for further refinement through

architectural enhancements, such as diffusion-based modeling, to improve structural accuracy while maintaining its efficiency and affinity prediction accuracy.

While DockFormer is significantly faster than co-folding methods, its inference time of several seconds per ligand remains a bottleneck for ultra-large-scale screening involving billions of ligands [45]. One potential solution is to train a lightweight pocket-specific model that predicts ligand affinity, using DockFormer as an oracle in a teacher-student setup. In this framework, DockFormer provides high-quality supervision on a smaller set of ligands, enabling the student model to screen vast libraries efficiently, with DockFormer reserved for final validation. An alternative strategy is to evaluate multiple ligands in a single DockFormer inference with a single receptor structure, retaining only those that bind the pocket, in addition to having high predicted affinity. While this can offer significant speedup, it may require fine-tuning and is still less scalable than the teacher-student approach.

The key concepts introduced here, leveraging input structures and integrating affinity modules, can be applied to emerging high-performance architectures. These approaches can be implemented either by training a new streamlined model with fewer layers to improve efficiency or by fine-tuning existing models to incorporate these capabilities while benefiting from their already well-optimized weights. The concepts presented here hold significant potential for enhancing affinity prediction by leveraging the high accuracy achieved in structure prediction. In future work, we plan to explore both strategies and evaluate their impact on accuracy.

The foundational concepts behind DockFormer have the potential to be extended beyond receptor-ligand interactions. For instance, similar models could be developed for

protein-protein interactions, including specialized models for challenging cases like antibody-antigen binding or protein-peptide interactions. In those cases, the additional modality of the model can be specificity. By leveraging large databases such as STRING [46], which provide interaction labels, we can train models to predict whether proteins interact. The reduction DockFormer offers in model size may make specialized task-specific models more feasible. We provide below an initial implementation of protein-protein interaction prediction named DockFormerPP, which we intend to analyze and expand in future work.

DockFormer was implemented by building upon the OpenFold [47] implementation of AlphaFold2 [13]. The code for running and training DockFormer is available in a GitHub repository [48], and the weights of the three presented variations are available as code and as a server at HuggingFace [49]. Additionally, the variation DockFormer-PINDER for protein-protein interaction prediction is available at HuggingFace [50].

ACKNOWLEDGMENTS

D.S.-D. and B.S. are supported by the Israeli Science Foundation (ISF 737/23). Molecular graphics and analyses performed with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325.

DATA AVAILABILITY

The data that support the findings of this article are openly available [48,49].

-
- [1] A. D. Attie and R. T. Raines, Analysis of receptor-ligand interactions, *J. Chem. Educ.* **72**, 119 (1995).
 - [2] V. Cherezov, D. M. Rosenbaum, M. A. Hanson, S. G. F. Rasmussen, F. S. Thian, T. S. Kobilka, H.-J. Choi, P. Kuhn, W. I. Weis, B. K. Kobilka *et al.*, High-resolution crystal structure of an engineered human β_2 -adrenergic G protein-coupled receptor, *Science* **318**, 1258 (2007).
 - [3] A. M. De Vos, M. Ultsch, and A. A. Kossiakoff, Human growth hormone and extracellular domain of its receptor: Crystal structure of the complex, *Science* **255**, 306 (1992).
 - [4] T. B. Kimber, Y. Chen, and A. Volkamer, Deep learning in virtual screening: Recent applications and developments, *Int. J. Mol. Sci.* **22**, 4435 (2021).
 - [5] M. Karelina, J. J. Noh, and R. O. Dror, How accurately can one predict drug binding modes using AlphaFold models? *Elife* **12**, RP89386 (2023).
 - [6] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry *et al.*, Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy, *J. Med. Chem.* **47**, 1739 (2004).
 - [7] J. Eberhardt, D. Santos-Martins, A. F. Tillack, and S. Forli, AutoDock Vina 1.2.0: New docking methods, expanded force field, and Python bindings, *J. Chem. Inf. Model.* **61**, 3891 (2021).
 - [8] D. Cao, M. Chen, R. Zhang, Z. Wang, M. Huang, J. Yu, X. Jiang, Z. Fan, W. Zhang, H. Zhou *et al.*, SurfDock is a surface-informed diffusion generative model for reliable and accurate protein-ligand complex prediction, *Nat. Methods* **22**, 310 (2024).
 - [9] G. Zhou, Z. Gao, Q. Ding, H. Zheng, H. Xu, Z. Wei, L. Zhang, and G. Ke, Uni-Mol: A universal 3D molecular representation learning framework, in *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)* (2023).
 - [10] G. Corso, H. Stärk, B. Jing, R. Barzilay, and T. Jaakkola, Diffdock: Diffusion steps, twists, and turns for molecular docking, *Proceedings of the International Conference on Learning Representations (ICLR)* (2023).
 - [11] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick *et al.*, Accurate structure prediction of biomolecular interactions with AlphaFold 3, *Nature (London)* **630**, 493 (2024).
 - [12] R. Krishna, J. Wang, W. Ahern, P. Sturmfels, P. Venkatesh, I. Kalvet, G. R. Lee, F. S. Morey-Burrows, I. Anishchenko, I. R. Humphreys *et al.*, Generalized biomolecular modeling and

- design with RoseTTAFold All-Atom, *Science* **384**, ead12528 (2024).
- [13] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, Highly accurate protein structure prediction with AlphaFold, *Nature (London)* **596**, 583 (2021).
- [14] R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim *et al.*, Protein complex prediction with AlphaFold-Multimer, *bioRxiv*: 2021.10.04.463034.
- [15] J. P. Roney and S. Ovchinnikov, State-of-the-art estimation of protein model accuracy using AlphaFold, *Phys. Rev. Lett.* **129**, 238101 (2022).
- [16] P. Bryant, A. Kelkar, A. Guljas, C. Clementi, and F. Noé, Structure prediction of protein-ligand complexes from sequence information with Umol, *Nat. Commun.* **15**, 4536 (2024).
- [17] P. Csermely, R. Palotai, and R. Nussinov, Induced fit, conformational selection and independent dynamic segments: An extended view of binding events, *Trends Biochem. Sci.* **35**, 539 (2010).
- [18] K.-I. Okazaki and S. Takada, Dynamic energy landscape view of coupled binding and protein conformational change: Induced-fit versus population-shift mechanisms, *Proc. Natl. Acad. Sci. USA* **105**, 11182 (2008).
- [19] S. Vajda, D. Beglov, A. E. Wakefield, M. Egbert, and A. Whitty, Cryptic binding sites on proteins: Definition, detection, and druggability, *Curr. Opin. Chem. Biol.* **44**, 1 (2018).
- [20] X. Zhang, H. Gao, H. Wang, Z. Chen, Z. Zhang, X. Chen, Y. Li, Y. Qi, and R. Wang, Planet: A multi-objective graph neural network model for protein-ligand binding affinity prediction, *J. Chem. Inf. Model.* **64**, 2205 (2024).
- [21] M. M. Stepniewska-Dziubinska, P. Zielenkiewicz, and P. Siedlecki, Development and evaluation of a deep learning model for protein-ligand binding affinity prediction, *Bioinformatics* **34**, 3666 (2018).
- [22] Z. Cang, L. Mu, and G.-W. Wei, Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening, *PLoS Comput. Biol.* **14**, e1005929 (2018).
- [23] D. Jiang, C.-Y. Hsieh, Z. Wu, Y. Kang, J. Wang, E. Wang, B. Liao, C. Shen, L. Xu, J. Wu *et al.*, InteractionGraphNet: A novel and efficient deep graph representation learning framework for accurate protein-ligand interaction predictions, *J. Med. Chem.* **64**, 18209 (2021).
- [24] L. Zheng, J. Fan, and Y. Mu, OnionNet: A multiple-layer intermolecular-contact-based convolutional neural network for protein-ligand binding affinity prediction, *ACS Omega* **4**, 15956 (2019).
- [25] J. Jiménez, M. Skalic, G. Martinez-Rosell, and G. De Fabritiis, K_{DEEP} : Protein-ligand absolute binding affinity prediction via 3D-convolutional neural networks, *J. Chem. Inf. Model.* **58**, 287 (2018).
- [26] R. Singh, S. Sledzieski, B. Bryson, L. Cowen, and B. Berger, Contrastive learning in protein language space predicts interactions between drugs and protein targets, *Proc. Natl. Acad. Sci. USA* **120**, e2220778120 (2023).
- [27] M. Tsubaki, K. Tomii, and J. Sese, Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences, *Bioinformatics* **35**, 309 (2019).
- [28] M. Volkov, J.-A. Turk, N. Drizard, N. Martin, B. Hoffmann, Y. Gaston-Mathé, and D. Rognan, On the frustration to predict binding affinities from protein-ligand structures with deep neural networks, *J. Med. Chem.* **65**, 7946 (2022).
- [29] W. Lu, Q. Wu, J. Zhang, J. Rao, C. Li, and S. Zheng, Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction, in *Proceedings Advances in Neural Information Processing Systems 35*, edited by S. Koyejo, S. Mohamad, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (2022), pp. 7236–7249.
- [30] H. Lai, L. Wang, R. Qian, J. Huang, P. Zhou, G. Ye, F. Wu, F. Wu, X. Zeng, and W. Liu, Interformer: An interaction-aware model for protein-ligand docking and affinity prediction, *Nat. Commun.* **15**, 10223 (2024).
- [31] W. Treyde, S. C. Kim, N. Bouatta, and M. AlQuraishi, Quickbind: A light-weight and interpretable molecular docking model, in *Proceedings of the 19th Machine Learning in Computational Biology Meeting*, edited by D. A. Knowles and S. Mostafavi (PMLR, Cambridge, 2024), Vol. 261, pp. 129–152.
- [32] T. Standley, A. Zamir, D. Chen, L. Guibas, J. Malik, and S. Savarese, Which tasks should be learned together in multi-task learning? in *Proceedings of the 37th International Conference on Machine Learning*, edited by H. Daumé and A. Singh (JMLR, Brookline, 2020), pp. 9120–9132.
- [33] T. Hayes, R. Rao, H. Akin, N. J. Sofroniew, D. Oktay, Z. Lin, R. Verkuil, V. Q. Tran, J. Deaton, M. Wiggert *et al.*, Simulating 500 million years of evolution with a language model, *Science* **387**, 850 (2025).
- [34] G. Landrum, *RDKit: Open-source cheminformatics software*, https://github.com/rdkit/UGM_2016 (GitHub, San Francisco, 2016).
- [35] A. Cohan, I. Beltagy, D. King, B. Dalvi, and D. S. Weld, Pre-trained language models for sequential sentence classification, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, edited by S. Padó and R. Huang (Association for Computational Linguistics, Stroudsburg, 2019).
- [36] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities, *Nucl. Acids Res.* **35**, D198 (2007).
- [37] J. Wohlwend, G. Corso, S. Passaro, M. Reveiz, K. Leidal, W. Swiderski, T. Portnoi, I. Chinn, J. Silterra, T. Jaakkola, and R. Barzilay, Boltz-1: Democratizing biomolecular interaction modeling, *bioRxiv*:2024.11.19.624167.
- [38] J. Durairaj, Y. Adeshina, Z. Cao, X. Zhang, V. Oleinikovas, T. Duignan, Z. McClure, X. Robin, D. Kovtun, E. Rossi *et al.*, PLINDER: The protein-ligand interactions dataset and evaluation resource, *bioRxiv*:2024.07.17.603955.
- [39] M. Su, Q. Yang, Y. Du, G. Feng, Z. Liu, Y. Li, and R. Wang, Comparative assessment of scoring functions: The CASF-2016 update, *J. Chem. Inf. Model.* **59**, 895 (2019).
- [40] M. Gilson, J. Eberhardt, P. Škrinjar, J. Durairaj, X. Robin, and A. Kryshtafovych, Assessment of pharmaceutical protein-ligand pose and affinity predictions in CASP16. Proteins (2025).
- [41] M. Buttenschoen, G. M. Morris, and C. M. Deane, Posebusters: AI-based docking methods fail to generate physically valid

- poses or generalise to novel sequences, *Chem. Sci.* **15**, 3130 (2024).
- [42] 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.* **6**, 195 (2024).
- [43] ByteDance AML AI4Science Team, X. Chen, Y. Zhang, C. Lu, W. Ma, J. Guan, C. Gong, J. Yang, H. Zhang, K. Zhang *et al.*, Protenix—advancing structure prediction through a comprehensive AlphaFold3 reproduction, [bioRxiv:2025.01.08.631967](https://doi.org/10.1103/1862-rs2k).
- [44] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/1862-rs2k> for details.
- [45] F. Liu, O. Mailhot, I. S. Glenn, S. F. Vigneron, V. Bassim, X. Xu, K. Fonseca-Valencia, M. S. Smith, D. S. Radchenko, J. S. Fraser *et al.*, The impact of library size and scale of testing on virtual screening, *Nat. Chem. Biol.* **21**, 1039 (2025).
- [46] D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N. T. Doncheva, S. Pyysalo *et al.*, The STRING database in 2023: Protein–protein association networks and functional enrichment analyses for any sequenced genome of interest, *Nucl. Acids Res.* **51**, D638 (2023).
- [47] G. Ahdriz, N. Bouatta, C. Floristean, S. Kadyan, Q. Xia, W. Gerecke, T. J. O’Donnell, D. Berenberg, I. Fisk, N. Zanichelli *et al.*, OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization, *Nat. Meth.* **21**, 1514 (2024).
- [48] <https://github.com/ben-shor/DockFormer>.
- [49] <https://huggingface.co/spaces/benshor/DockFormer>.
- [50] https://huggingface.co/spaces/benshor/pinder_dockformer/tree/main.