# Protein language models are performant in structure-free virtual screening

Authors: Hilbert LAM Yuen In [A, B], GUAN Jia Sheng [A], ONG Xing Er [B], Robbe PINCKET [C], MU Yuguang [A, B]

**[A] - School of Biological Sciences (SBS), Nanyang Technological University (NTU), Singapore, Singapore, Republic of Singapore**

**[B] - MagMol Pte. Ltd., Singapore, Republic of Singapore**

**[C] - Heliovision, Leuven, Kingdom of Belgium**

* Correspondence should be addressed to:

A/Prof. Mu Yuguang
Email: ygmu@ntu.edu.sg

## 1.1 Abstract

Hitherto virtual screening has been typically performed using a structure-based drug design paradigm. Such methods typically require the use of molecular docking on high-resolution three-dimensional structures of a target protein - a computationally-intensive and time-consuming exercise. This work demonstrates that by employing protein language models and molecular graphs as inputs to a novel graph-to-transformer cross-attention mechanism, a screening power comparable to state-of-the-art structure-based models can be achieved. The implications thereof include highly expedited virtual screening due to the greatly reduced compute required to run this model, and the ability to perform early stages of computer-aided drug design in the complete absence of 3D protein structure.

## 2.1 Introduction

One major pillar of rational drug design has always been the use of virtual screening through docking in what is commonly known as structure-based drug design (SBDD) [1]. In docking, molecular ligands are typically conformationally explored in a protein pocket either through the use of biophysically defined constraints or machine learning (ML) methods, and a best pose with its corresponding computed binding affinity reported. A quintessential virtual screening pipeline will iteratively perform docking through a library, usually consisting of millions to billions of unique chemical compounds, and rank the ligands based on the derived affinity - the top scored ligands will then proceed onto the next phase of drug development, either through computational means such as molecular dynamics (MD) simulations or through experimental validation [2].

In order to enhance the accuracy of SBDD and virtual screening (VS), significant progress has been made in developing docking tools and rescoring functions - the latter of which are typically deep learning methods that, given a docked ligand, output a correction term or a new score

entirely. These re-scoring methods have shown large promise, increasing the screening power in benchmarks [3-5]. Screening power is defined as the ability of a given model to differentiate between what will and will not bind to a target experimentally [6], and is largely considered the ultimate test of any virtual screening model. By increasing screening power, a model can more effectively discover leads, and overall improve the efficacy of computer-aided drug design (CADD).

However, docking itself and the addition of a rescoring term greatly increases the computational time and expense required to score each ligand. With very large libraries being used and an estimated chemical space of $10^{63}$ unique compounds [7], this increased compute in already time-consuming VS means that either cost would have to go up per target or the amount of ligands that can be screened per target would have to go down, inadvertently compromising the comprehensiveness of a virtual screening pipeline on the chemical space explored.

Aggravatingly, recent work also suggests that many SBDD deep learning models merely memorise the ligands, resulting in poor generalisability and outcomes [8]. In the case of models predicting the binding affinity directly, studies have shown that many of these models consider less of the protein-ligand interaction but more so of quantitative structure analysis relationship (QSAR) of the ligand alone in making their predictions [9]. Intrinsic biases have also been found in common datasets used for training SBDD deep learning models [9]. There is also the issue of limited data, with common datasets such as the PDBBind+ dataset [10] consisting of 22,920 ligand-protein pairs, further complicating the issue of training performative and generalisable models due to data scarcity. SBDD also has a pitfall in which only a single snapshot of a pocket is considered during docking and subsequent rescoring - and although these limitations have been addressed through many works throughout the years [11], the fundamental flexible and dynamic nature of the protein is still, unfortunately, largely given little consideration [12]. The performance of molecular docking can also vary between different conformational states of a protein (i.e. the apo or holo state), and hence docking predominantly suffers from considering the induced fitting of ligands [13].

To address this, some have resorted to instead predicting drug-target affinity (DTA), an experimentally derived figure such as the half-maximal inhibitory concentration ($IC_{50}$) or dissociation constants. Such work includes SSM-DTA [14], a sequence-based deep learning method that achieves state-of-the-art performance in DTA prediction. However, in this work, the authors demonstrate that DTA-only models do not necessarily have the highest screening power.

Therefore, to tackle the issue of limited data in SBDD, the single snapshot issue, the lack of consideration of the entire protein during SBDD, induced fitting model and different conformations, this work combines protein language models (PLMs), a form of large language model (LLM) together with graph neural networks to represent ligands to perform virtual screening without inputting structural information (Fig. 1). The fundamental idea is that any structural and dynamic information of proteins is implicit in any PLMs pre-trained through

self-supervised means such as Evolutionary Scale Modelling 2 (ESM-2) and this theorem has been further demonstrated in works such as ESMFold [15]. This work further introduces a novel graph-to-transformer cross-attention block, which essentially treats every single node in a graph as a token and allows it to query a separate sequence. The combined model, named **B**inding **IN**teraction **D**etermination (BIND), achieves comparable performance to state-of-the-art SBDD models with a fraction of the compute and time required and with only protein sequence and ligand information as inputs. Primarily, this work differs from previous work in its use of a decoy/true binder classifier and application to datasets typically used to benchmark SBDD models such as DUD-AD [16]. Through the use of this model, a state-of-the-art reverse screening performance was also achieved when compared with SBDD models on AlphaFold2-predicted structures. This work also showed that when trained with the multi-objective classification and DTA as used in BIND, the model's DTA prediction also outperforms state-of-the-art DTA-only models in the screening power domain. Lastly, the trained model in this work can be deconstructed to determine which residues of a protein contribute most to the protein-ligand interaction, despite never being trained to do so.

## 3. Results

**3.1 Screening power of BIND comparable to top models.** The screening power of the model (Fig. 2) is similar to that of top SBDD models for all benchmarks performed (Table 1-5). Overall, a high enrichment factor is observed for all the benchmarks tested. In the DEKOIS 2.0, DUD-AD, DUD-E and LIT-PCBA datasets, the BIND model appears to have the highest enrichment factor at 1% cutoff with respect to reviewed literature (Fig. 3c-f). However, the model falls short in the CASF-2016 forward screening dataset (Fig. 3a-b), achieving higher enrichment than Glide-SP and Glide-XP but is overall deficient in the success rate and when compared to other very recent machine learning methods. Furthermore, the model's enrichment factor is higher across all datasets tested when compared to using the $pIC_{50}$ prediction from SSM-DTA, a fully DTA prediction model (Fig. 1) . The training and validation losses of the evaluated model are described in Supplementary Fig. 1, and individual benchmark scores for each protein are given in Supplementary File 1. Further training experiments to optimise the model and their corresponding results are also available in the supplementary material. In zero-shot protein settings, where the sequences in the training set containing > 90% homology compared to the evaluation datasets and their associated ligands are removed, the zero-shot BIND model still achieves good performance, outperforming some contemporary score functions in the datasets tested. The individual zero-shot BIND results are given in Supplementary File 2.

**3.2 Protein language models can perform reverse screening.** Reverse screening was performed on the benchmark as described by Luo *et al*., 2023 [17]. In summary, 90 selected ligands were scored against 12,195 human proteins representative of the human proteome. From the results, BIND is comparable to that of top models with rescoring such as OnionNet-SFCT with Glide-SP docking on AlphaFold2-predicted structures and PointSite or SiteMap for pocket determination; BIND being able to cumulatively rank the highest number of correct pairs for the top 2-1000 pairs (Fig. 4a-b). Furthermore, the reverse screening is

expedient, taking only approximately eight hours on a Ryzen Threadripper Pro 5975WX CPU to run through all 1,097,550 protein-ligand combinations. The reverse docking performance of BIND was also evaluated on the Astex Diver dataset, of which in the top 1 scoring outperformed DOCK, Glide, and Autodock Vina, with comparable performance on the top 5 metrics (Fig. 4c).

**3.3 Model can perform drug-target affinity prediction.** The model is able to perform on the DTA prediction datasets DAVIS, despite not being specifically trained on this dataset but only BindingDB (Table 6). Furthermore, on the BindingDB test split, it achieved state-of-the-art RMSE in $pK_d$ prediction specifically, but falls short in the other metrics of $pK_i$, $pIC_{50}$ and $pEC_{50}$ (Table 7).

**3.4 Drug-target affinity prediction alone has lower screening power compared to classification.** When ranked using $pK_i$, $pK_d$, $pIC_{50}$ or $pEC_{50}$, the enrichment factors across benchmark datasets dropped compared to classification alone (Fig. 5a - e). Furthermore, the classifier also has the strongest enriching effects in the AlphaFold2 reverse docking benchmark (Fig. 5f). Lastly, in spite of state-of-the-art DTA predictors being more accurate in predicting DTA compared to BIND (Table 6), BIND still achieves higher screening power on CASF-2016, DEKOIS 2.0, DUD-AD, DUD-E and LIT-PCBA using its $pIC_{50}$ prediction compared to SSM-DTA's $pIC_{50}$ (Supplementary Fig. 2).

**3.5 Attention weights can pinpoint residues important for binding despite the model not being trained for such.** 187 single-chain entries in the CASF-2016 core set were put into the model and the residues with the highest attention scores extracted, with attention scores maxed across all four cross-attention blocks. The model is able to identify the binding residues (Fig. 6), and hence determine potential binding pockets in a protein with unknown structure. At a 10 Å cutoff, the model is able to identify approximately 30% of binding pocket residues. However, the attention weights are still patchy across the entire protein (Fig. 6b).

## 4. Discussion

The use of protein language models in predicting DTA is not new. Previous work has delivered very accurate DTA predictors using PLMs such as SSM-DTA [14]. However, there has been far less application of this to determine screening power - arguably the most important goal in CADD. Previous work by Tsubaki *et al*., 2019, used a one-dimensional n-gram convolutional neural network and cross-attentions to do similar work, achieving good discriminative power between real and decoy ligands when trained and evaluated on the DUD-E dataset using fivefold cross-validation [18]. This work builds upon that foundation and uses the pre-trained ESM-2 to transfer pre-learned knowledge about the protein language into the virtual screening pipeline.

Primarily, BIND serves as a proof of concept that even in the complete absence of protein structure or binding pocket information, it is possible to achieve screening power similar to that of and sometimes even exceed certain top SBDD models with PLMs by training the model to

discriminate between real and decoy ligands selected from the same dataset. The results show that by using a model such as the proposed BIND, it may be possible to perform virtual screening for drug discovery, or to elucidate the identity of protein targets which bind to a specific ligand as per reverse docking. However, from the CASF-2016 results, it appears that BIND is not the best at ranking the very top and most potent binders as apparent in its poorer success rate in a much smaller set of ligands as seen in CASF-2016. From this, and with the comparable performance and occasionally state-of-the-art enrichment factor scores in the other larger datasets, it can be said that BIND has rather demonstrated its utility in enrichment of large set; and given the high speed at which BIND can screen compounds, it can be pragmatically deployed in initial screening stages to filter out a large number of non-binders (i.e. from millions to thousands), of which refinement can then be done with other score functions such as GenScore, PIGNet2 or OnionNet-SFCT to pull out only the top binders. The strong enriching effects of BIND in reverse screening also indicates that BIND could be better at finding a target protein from a large number of proteins given a ligand than performing docking on AlphaFold2 structures. This is especially useful as, in many reverse docking pipelines, the experimental structure of all proteins is typically not available, and reverse docking may be performed on non-binding protein conformations [19].

The authors posit from the results that BIND is able to achieve the reported screening power by borrowing the pre-learned context from PLMs such as ESM-2, which are able to understand the implicit structure and underlying language of proteins. This includes understanding biochemical properties, dynamic movement of proteins, and interdomain interactions.The performance of the zero-shot model also shows BIND has sufficient generalisation capability in screening unseen proteins, potentially obtaining this capability from its ESM-2 pre-training which allowed it to see a much more diverse set of protein sequences. This premise of stronger performance from transfer learning is supported by that of literature [20], and previous work such as ESMFold have demonstrated that structure can be inferred from sequence alone with the help of PLMs. Therefore, it may not be far-fetched that the flexibility and even how a protein interacts with other molecules such as ligands be implicitly learnt through the masked language modelling and other self-supervised objectives of PLMs in pre-training. PLMs may be able to better model molecular interactions with disordered regions in proteins and even decipher cryptic pockets which evade the typical SBDD virtual screening pipeline due to their transient nature [21]. There remains much to be explored in this domain.

Furthermore, it is interesting that the BIND model's DTA prediction has lower screening power and enriching effects compared to discriminating between true and false binders (i.e. the classifier in BIND) alone. It is highly plausible that this is due to internal biases of the training dataset - as all of the training protein-ligand pairs are true binders, the algorithm is misled into thinking that every protein-ligand pair that is given is always a binder and hence predicts a binders' score regardless to prevent heavy loss penalties, as this is reflective of what it has seen and points raised by previous work. This result reinforces the indication that robust augmentation very likely goes hand-in-hand with screening power. The use of sole DTA models in screening may also not be appropriate in all cases, as DTA datasets, including BindingDB,

which derives its affinity values from multiple sources, have been shown to be noisy across multiple assays [22] - this underscores that it may be more practical to predict a binary label as to whether or not a drug binds than a quantitative experimental affinity that is subject to many other confounding and unforeseen experimental variables, which could also explain BIND's higher screening power compared to SSM-DTA although BIND achieves significantly worse $R^2$ and MSE values in $IC_{50}$ and other DTA predictions across the board.

The authors acknowledge several limitations of this work: (1) The BindingDB dataset is mainly obtained from ChEMBL and other sources, and contains overlapping information with the CASF-2016, DEKOIS 2.0, DUD-AD, DUD-E and LIT-PCBA evaluation datasets, in the process inflating the results for the forward screening for the non-zero-shot model. Although this is an issue, it is difficult to directly and objectively quantify the overlaps and individually eradicate the overlaps. Furthermore, by removing overlaps, a subjective defined stringency is required (i.e. Tanimoto similarity, protein sequence similarity), in the process getting rid of a large, diverse and representative chunk of protein-ligand binding data in *Homo sapiens* and other key protein families (which is the entire rationale of the benchmark datasets). This would prejudice BIND in evaluation as other score functions are typically trained with knowledge of that protein previously (i.e. from the PDBBind dataset [10], which has directly overlapping PDB structures with the evaluation dataset and are typically not removed during model training). By removing direct ligand-protein matches but not proteins during evaluation, the evaluation cannot be fairly performed with other score functions using the same dataset as the dataset would be modified and the enrichment factors cannot be directly compared. To mitigate this limitation and to assess the true screening power of BIND, the zero-shot model was attempted in which sequences containing > 90% homology on the proteins were removed from the training set, and the model still showed generalisation capability compared to contemporary score functions; (2) the explainability of BIND is limited - although the cross-attentions are generally able to pinpoint the important residues in a protein, they are far from perfect and are still largely uninterpretable in-line with post-hoc explainability analyses performed on transformer models; (3) much like previous work such as SSM-DTA, the training expense of this model is high. To compensate with limited resources, the authors used a large number of gradient accumulation steps, in the process dragging out the training time. The model may also further benefit in the DTA prediction criteria from a scaled-up training regime, which could not be achieved in this work due to limited computational resources; (4) No wet lab experimental validation was performed for this work - although BIND performs reasonably well on benchmarks as explored in this work, many factors ultimately affect the efficacy of a drug and experimental validation is needed.

Overall, this work proposes the use of PLMs as a potential alternative for SBDD virtual screening - showing that, even without structural information, is comparable to top-of-the-line SBDD models. The authors reinforce that this methodology be termed sequence-based drug discovery (SeBDD) [23]. SeBDD could be useful in work in which the protein structure cannot be accurately resolved, or when the binding pockets are unknown. In current screening pipelines, SeBDD can be used as an initial filter in a VS campaign, in which the top molecules from SeBDD are selected for standard SBDD docking and MD simulations. Due to the simplicity of

SeBDD compared with SBDD (no need for docking, preparing the Gasteiger partial charges of ligands and targets, etc.), pipelines can also be greatly accelerated by its use.

Future work in this domain likely involves simulated annealing of molecules and integrating SeBDD with fragment-based drug discovery (FBDD), the use of synthons [24], and machine learning-based chemical spaces to explore very large amounts of chemical spaces quickly and accurately.

## 5. Methods

**5.1 Model training, architecture, and training dataset.** The BIND model was written and evaluated in Python 3.11 using PyTorch Geometric 2.4.0 with a PyTorch 2.2.0+cu121 backend and HuggingFace Transformers 4.37.2. The full model, along with its trained weights, is disclosed in the GitHub repository. Three models were trained with the one with the lowest validation loss selected and benchmarked in this study, with each model trained on a single Nvidia RTX A6000 for approximately 28 days. The whole model is described in Fig. 2. The implementation code for the cross-attention graph block is also in the same GitHub repository - in essence, each node in the graph is unwrapped and treated as an individual token, before a standard cross-attention similar to that of the original transformer decoder was used - the graph is then reconstructed afterwards. Five prediction heads are used in the final layer of the model, four of which are for regression objectives - to predict the $pK_i$, $pK_d$, $pIC_{50}$, and $pEC_{50}$ respectively. A classification objective is also added to predict which ligands are true and which are decoys. The logits of the classification head are multiplied by the natural exponential of a trainable temperature parameter, $\tau$, initialised at 0.07 as per previous work [25]. After scaling, logits are clipped from -100 to 100 to ensure training stability. The loss functions used are Huber loss with a $\delta = 2.0$ and binary cross entropy for regression and classification heads respectively. Training was performed using the AdamW optimiser with a learning rate of 1e-4 and a weight decay of 1e-3. A dropout of 0.1 was applied to the multi-headed attention layer, and a leaky rectified linear unit (ReLU) activation of $\alpha = 0.05$ is used unless otherwise specified. Learning rate was scheduled using cosine decay with no warm up. The 650 million parameter ESM-2 (facebook/esm2_t33_650M_UR50D) model's weights were retrieved using the HuggingFace Transformers 4.37.1 library and are frozen during training. Latents from the ESM-2 layers 1, 11, 21, 31 are cross-attended to. Convolutions for the molecular graph are GATv2 [26] for all instances, and pooling used for the interaction network is Learnable Commutative Monoid (LCM) Aggregation [27]. Model training was done in automatic mixed precision and FlashAttention-2 [28] was used where applicable. The training dataset used is BindingDB [29], with a total of 2,469,626 (90%) protein-ligand pairs used for training, 54,902 (2%) for validation, and 224,738 (8%) for testing. A batch size of 1 was used during training together with 256 gradient accumulation steps, for a total of 100,000 full iterations (effectively around 10.37 epochs' worth). To optimise for computation, all protein sequences exceeding 2,048 amino acids were removed. In total, 1,202,465 unique SMILES strings as part of the BindingDB dataset were used. Since there are multiple prediction heads, for entries in which one or more of the

regression values were unavailable, the local loss was set to zero, meaning that there is no penalty for that term regardless of what the model predicts. For decoy ligands specifically, the loss for $K_i$, $K_d$, $IC_{50}$ and $EC_{50}$ terms are set to zero. All affinity values were normalised with $-\log_{10}(x / 1e9)$, where x is in nM, consistent with all previous work [14]. For the reported training loss, the zeroised terms are disregarded in averaging. Approximately 50% of ligands used during training are decoys as part of a data-balanced regime.

**5.2 Graph node and edge generation.** The construction of the molecular graph is identical to that done in previous work [30], with the exception that the PySmiles library is now set to not reinterpret aromatic bonds due to a software bug. PySmiles 1.1.2 and NetworkX 3.2.1 were used to generate the graphs used by the model.

**5.3 Decoy ligand determination.** Decoys were determined at training by randomly selecting another ligand from the loaded SMILES, and programmatically checking if there is another entry in the BindingDB data that tags that SMILES to the same protein sequence.

**5.4 Evaluation datasets.** The DEKOIS 2.0 [31], DUD-E [32], DUD-AD [16], LIT-PCBA [33] and CASF-2016 [6] datasets were used to determine screening power. In all instances, ligands were converted into SMILES using OpenBabel 3.1.0. For all datasets, the UniProt sequences were obtained which corresponded to each RCSB PDB entry. In instances where multiple sequences relating to the target were present, the sequences were concatenated before inputting into the model - protein sequences that were unrelated to the test protein such as hirudin in thrombin and nuclear cofactors in the oestrogen receptors were removed. All protein sequences used in evaluation are available in Supplementary File 3. DTA evaluation was performed on the DAVIS [34] dataset obtained from TDCommons [35], and binding affinity prediction was also performed on a test set of BindingDB split prior to training. Unlike training, there is no cutoff on protein length used during evaluation/testing.

**5.5 Dataset splitting and zero-shot protein performance evaluation.** As there are overlaps between the BindingDB dataset and the screening power evaluation datasets, sequence-ligand pairs in BindingDB which contained > 90% overlap in the protein sequence compared to the UniProt sequences in the evaluating datasets were removed for the zero-shot model. The percentage alignment was calculated using Biopython 1.78's pairwise2 align module, with the UniProt sequence used as the reference length in computing the final homology percentage. The zero-shot model was re-trained on this dataset with the exact same parameters as stated in 5.1.

**5.6 Evaluation with SSM-DTA.** The SSM-DTA [14] $pIC_{50}$ model (BindingDB_IC50/checkpoint_best_20221021) was used on the same datasets described in 5.4, with the same protein sequence input and SMILES as per Supplementary File 3, after canonicalisation and tokenisation as provided by the SSM-DTA source code.

**5.7 Metric calculation.** BEDROC [36] and Enrichment Factors were calculated using RDKit

2023.09.5, whereas AUROC was calculated using the SciPy 1.11.0 library. Reported scores are averaged across all entries in the respective evaluation datasets (57 for CASF-2016, 81 proteins for DEKOIS 2.0, 102 for DUD-AD and DUD-E, 15 for LIT-PCBA). All decoys and ligands are used in benchmarking as per the original datasets.

**5.8 Reverse screening.** Reverse screening was performed as per Luo *et al.*, 2023 [17]. In essence, 90 ligands were selected and reverse screened on a trimmed down human proteome of 12,195 proteins from UniProt using BIND. For the reverse docking performed using SBDD models done by Luo *et al.*, 2023, structures obtained from the AlphaFold Protein Structure Database [37] were used. PointSite [38] or SiteMap [39] were used to determine potential binding pockets. Reverse screening was also done on the Astex Diverse Set using the same procedure as with other standard docking tools described by Hartshorn *et al.*, 2007 [40]. Summarily, the 84 ligands found in the dataset were docked against the 85 proteins to identify the correct protein target for the ligand and ranked, with the percentage of correct pairs identified in the top 1 and top 5 rankings as per literature [41].

**5.9 Attention mapping and scoring.** Attention weights were extracted per layer and averaged over all heads as per the MultiHeadedAttention class in PyTorch. Subsequently, they are averaged across the four cross-attention graph blocks in BIND, and the maximum weight per residue selected across all atoms, creating a single attention weight score per residue between 0-1.

**5.10 Attention scores binding pocket overlap determination.** The CASF-2016 core set was filtered for single chains. Of which, 187 out of 285 proteins fulfilled aforementioned criteria. Open-source PyMOL 2.5.0 was then used to determine all residues within 8 Å, 10 Å, 12 Å and 15 Å of the native ligand as provided by CASF-2016 respectively. Residues are counted so long as any part of the amino acid (including its sidechains) are within the cutoff distance to any of the ligand atoms using the “X within Y of Z” command in the PyMOL API. Attention scores from BIND were determined as previously described, and the top number of residues corresponding to the total number of residues selected with respect to the docked ligand (through PyMOL) were extracted and overlaps counted. In essence, this metric measures the proportion of most attended residues to the actual binding pocket. Generally, the larger the cutoff value, the easier it is to get a higher metric score.

## 6. Acknowledgements

The authors thank Saxena Shikhar for the fruitful discussions as part of this work. We further would like to express our appreciation to Qing Luo and Jingjing Guo of Macao Polytechnic University for supplying their full reverse docking data from their previous work.

## 7. Code Availability

The code is available at: https://github.com/Chokyotager/BIND.

## 8. References


1. Ferreira, L.G., et al., *Molecular docking and structure-based drug design strategies.* Molecules, 2015. **20**(7): p. 13384-421.
2. Zhang, B., et al., *Molecular docking-based computational platform for high-throughput virtual screening.* CCF Trans High Perform Comput, 2022. **4**(1): p. 63-74.
3. Wojcikowski, M., P.J. Ballester, and P. Siedlecki, *Performance of machine-learning scoring functions in structure-based virtual screening.* Sci Rep, 2017. **7**: p. 46710.
4. Zheng, L., et al., *Improving protein-ligand docking and screening accuracies by incorporating a scoring function correction term.* Brief Bioinform, 2022. **23**(3).
5. Shen, C., et al., *Boosting Protein-Ligand Binding Pose Prediction and Virtual Screening Based on Residue-Atom Distance Likelihood Potential and Graph Transformer.* J Med Chem, 2022. **65**(15): p. 10691-10706.
6. Su, M., et al., *Comparative Assessment of Scoring Functions: The CASF-2016 Update.* J Chem Inf Model, 2019. **59**(2): p. 895-913.
7. Bohacek, R.S., C. McMartin, and W.C. Guida, *The art and practice of structure-based drug design: a molecular modeling perspective.* Med Res Rev, 1996. **16**(1): p. 3-50.
8. Powers, A.S., et al., *Geometric Deep Learning for Structure-Based Ligand Design.* ACS Cent Sci, 2023. **9**(12): p. 2257-2267.
9. Libouban, P.Y., et al., *The Impact of Data on Structure-Based Binding Affinity Predictions Using Deep Neural Networks.* Int J Mol Sci, 2023. **24**(22).
10. Wang, R., et al., *The PDBbind database: methodologies and updates.* J Med Chem, 2005. **48**(12): p. 4111-9.
11. Andrusier, N., et al., *Principles of flexible protein-protein docking.* Proteins, 2008. **73**(2): p. 271-89.
12. Lexa, K.W. and H.A. Carlson, *Protein flexibility in docking and surface mapping.* Q Rev Biophys, 2012. **45**(3): p. 301-43.
13. Fan, H., et al., *Molecular docking screens using comparative models of proteins.* J Chem Inf Model, 2009. **49**(11): p. 2512-27.
14. Pei, Q., et al., *Breaking the barriers of data scarcity in drug-target affinity prediction.* Brief Bioinform, 2023. **24**(6).
15. Lin, Z., et al., *Evolutionary-scale prediction of atomic-level protein structure with a language model.* Science, 2023. **379**(6637): p. 1123-1130.
16. Chen, L., et al., *Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening.* PLoS One, 2019. **14**(8): p. e0220113.
17. Qing Luo, S.W., Hoi Yeung Li, Liangzhen Zheng, Yuguang Mu, Jingjing Guo, *Benchmarking Reverse Docking through AlphaFold2 Human Proteome.* bioRxiv, 2023.
18. Tsubaki, M., K. Tomii, and J. Sese, *Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences.* Bioinformatics, 2019. **35**(2): p. 309-318.
19. Pinzi, L. and G. Rastelli, *Molecular Docking: Shifting Paradigms in Drug Discovery.* Int J Mol Sci, 2019. **20**(18).
20. Bugnon, L.A., et al., *Transfer learning: The key to functionally annotate the protein*

*universe.* Patterns (N Y), 2023. **4**(2): p. 100691.
21. Bloore, D.A.K., J. C.; Kapoor, K.; Chen, E.; Gao, K.; Wang, M., H M.-H., *Protein Language Models Enable Accurate Cryptic Ligand Binding Pocket Prediction.* arXiv, 2024.
22. Landrum, G.A. and S. Riniker, *Combining IC(50) or K(i) Values from Different Sources Is a Source of Significant Noise.* J Chem Inf Model, 2024. **64**(5): p. 1560-1567.
23. Chen, L., et al., *Sequence-based drug design as a concept in computational drug design.* Nat Commun, 2023. **14**(1): p. 4217.
24. Sadybekov, A.A., et al., *Synthon-based ligand discovery in virtual libraries of over 11 billion compounds.* Nature, 2022. **601**(7893): p. 452-459.
25. Wu, Z.X., Y; Yu, S.; Lin, D., *Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination.* arXiv, 2018.
26. Brody, S.A., U., Yahav, E., *How Attentive are Graph Attention Networks?* arXiv, 2021.
27. Ong, E.V., P., *Learnable Commutative Monoids for Graph Neural Networks.* arXiv, 2022.
28. Dao, T., *FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning.* arXiv, 2023.
29. Liu, T., et al., *BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities.* Nucleic Acids Res, 2007. **35**(Database issue): p. D198-201.
30. Lam, H.Y.I.P., R.; Han, H.; Ong; X. E.; Wang, Z., Hinks, J.; Wei, Y.; Li, W.; Zheng, L.; Mu, Y., *Application of variational graph encoders as an effective generalist algorithm in computer-aided drug design.* Nature Machine Intelligence, 2023.
31. Bauer, M.R., et al., *Evaluation and optimization of virtual screening workflows with DEKOIS 2.0--a public library of challenging docking benchmark sets.* J Chem Inf Model, 2013. **53**(6): p. 1447-62.
32. Mysinger, M.M., et al., *Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking.* J Med Chem, 2012. **55**(14): p. 6582-94.
33. Tran-Nguyen, V.K., C. Jacquemard, and D. Rognan, *LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening.* J Chem Inf Model, 2020. **60**(9): p. 4263-4273.
34. Davis, M.I., et al., *Comprehensive analysis of kinase inhibitor selectivity.* Nat Biotechnol, 2011. **29**(11): p. 1046-51.
35. Huang, K., et al., *Artificial intelligence foundation for therapeutic science.* Nat Chem Biol, 2022. **18**(10): p. 1033-1036.
36. Truchon, J.F. and C.I. Bayly, *Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem.* J Chem Inf Model, 2007. **47**(2): p. 488-508.
37. Varadi, M., et al., *AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models.* Nucleic Acids Res, 2022. **50**(D1): p. D439-D444.
38. Yan, X., et al., *PointSite: A Point Cloud Segmentation Tool for Identification of Protein Ligand Binding Atoms.* J Chem Inf Model, 2022. **62**(11): p. 2835-2845.
39. Schrödinger Inc., *SiteMap*.
40. Hartshorn, M.J., et al., *Diverse, high-quality test set for the validation of protein-ligand docking performance.* J Med Chem, 2007. **50**(4): p. 726-41.
41. Luo, Q., et al., *The scoring bias in reverse docking and the score normalization strategy to improve success rate of target fishing.* PLoS One, 2017. **12**(2): p. e0171433.
42. Shen, C., et al., *A generalized protein-ligand scoring framework with balanced scoring, docking, ranking and screening powers.* Chem Sci, 2023. **14**(30): p. 8129-8146.

43. Moon, S.H., SY.; Lim, J.; Woo, Y. K., *PIGNet2: a versatile deep learning-based protein–ligand interaction prediction model for binding affinity scoring and virtual screening.* Digital Discovery, 2023. **3**: p. 287-299.
44. Wang, Z., et al., *A fully differentiable ligand pose optimization framework guided by deep learning and a traditional scoring function.* Brief Bioinform, 2023. **24**(1).
45. Oscar Méndez-Lucio, M.A., Ehecatl Antonio del Rio-Chanona, Jörg Kurt Wegner, *A geometric deep learning approach to predict binding conformations of bioactive molecules.* Nature Machine Intelligence, 2021. **3**: p. 1033 - 1039.
46. Korb, O., T. Stutzle, and T.E. Exner, *Empirical scoring functions for advanced protein-ligand docking with PLANTS.* J Chem Inf Model, 2009. **49**(1): p. 84-96.
47. Wang, C. and Y. Zhang, *Improving scoring-docking-screening powers of protein-ligand scoring functions using random forest.* J Comput Chem, 2017. **38**(3): p. 169-177.
48. Schrödinger Inc., *Glide*.
49. Verdonk, M.L., et al., *Improved protein-ligand docking using GOLD.* Proteins, 2003. **52**(4): p. 609-23.
50. Trott, O. 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, 2010. **31**(2): p. 455-61.
51. Zhang, X., et al., *Efficient and accurate large library ligand docking with KarmaDock.* Nat Comput Sci, 2023. **3**(9): p. 789-804.
52. McGibbon, M., et al., *SCORCH: Improving structure-based virtual screening with machine learning classifiers, data augmentation, and uncertainty estimation.* J Adv Res, 2023. **46**: p. 135-147.
53. Li, Y., et al., *DyScore: A Boosting Scoring Method with Dynamic Properties for Identifying True Binders and Nonbinders in Structure-Based Drug Discovery.* J Chem Inf Model, 2022. **62**(22): p. 5550-5567.
54. McNutt, A.T., et al., *GNINA 1.0: molecular docking with deep learning.* J Cheminform, 2021. **13**(1): p. 43.
55. Sunseri, J. and D.R. Koes, *Virtual Screening with Gnina 1.0.* Molecules, 2021. **26**(23).
56. Marcou, G. and D. Rognan, *Optimizing fragment and scaffold docking by use of molecular interaction fingerprints.* J Chem Inf Model, 2007. **47**(1): p. 195-207.
57. Tran-Nguyen, V.K., G. Bret, and D. Rognan, *True Accuracy of Fast Scoring Functions to Predict High-Throughput Screening Data from Docking Poses: The Simpler the Better.* J Chem Inf Model, 2021. **61**(6): p. 2788-2797.
58. Desaphy, J., et al., *Encoding protein-ligand interaction patterns in fingerprints and graphs.* J Chem Inf Model, 2013. **53**(3): p. 623-37.
59. Stepniewska-Dziubinska, M.M., P. Zielenkiewicz, and P. Siedlecki, *Development and evaluation of a deep learning model for protein-ligand binding affinity prediction.* Bioinformatics, 2018. **34**(21): p. 3666-3674.
60. Zhou, H., H. Cao, and J. Skolnick, *FRAGSITE: A Fragment-Based Approach for Virtual Ligand Screening.* J Chem Inf Model, 2021. **61**(4): p. 2074-2089.
61. Zhang, W. and J. Huang, *EViS: An Enhanced Virtual Screening Approach Based on Pocket-Ligand Similarity.* J Chem Inf Model, 2022. **62**(3): p. 498-510.
62. Zhou, H., H. Cao, and J. Skolnick, *FINDSITE(comb2.0): A New Approach for Virtual Ligand Screening of Proteins and Virtual Target Screening of Biomolecules.* J Chem Inf Model, 2018. **58**(11): p. 2343-2354.
63. Jain, A.N., *Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine.* J Med Chem, 2003. **46**(4): p. 499-511.

64. Brocidiacono, M., et al., *BigBind: Learning from Nonstructural Data for Structure-Based Virtual Screening.* J Chem Inf Model, 2024. **64**(7): p. 2488-2495.
65. Durrant, J.D. and J.A. McCammon, *NNScore 2.0: a neural-network receptor-ligand scoring function.* J Chem Inf Model, 2011. **51**(11): p. 2897-903.
66. Quiroga, R. and M.A. Villarreal, *Vinardo: A Scoring Function Based on Autodock Vina Improves Scoring, Docking, and Virtual Screening.* PLoS One, 2016. **11**(5): p. e0155183.
67. Ozturk, H., A. Ozgur, and E. Ozkirimli, *DeepDTA: deep drug-target binding affinity prediction.* Bioinformatics, 2018. **34**(17): p. i821-i829.
68. Wang, J., et al., *ELECTRA-DTA: a new compound-protein binding affinity prediction model based on the contextualized sequence encoding.* J Cheminform, 2022. **14**(1): p. 14.
69. Pan, S., et al., *SubMDTA: drug target affinity prediction based on substructure extraction and multi-scale features.* BMC Bioinformatics, 2023. **24**(1): p. 334.
70. Huang, K., et al., *DeepPurpose: a deep learning library for drug-target interaction prediction.* Bioinformatics, 2021. **36**(22-23): p. 5545-5547.
71. Wu, H., et al., *AttentionMGT-DTA: A multi-modal drug-target affinity prediction using graph transformer and attention mechanism.* Neural Netw, 2024. **169**: p. 623-636.
72. Zhao, Q., et al., *AttentionDTA: Drug-Target Binding Affinity Prediction by Sequence-Based Deep Learning With Attention Mechanism.* IEEE/ACM Trans Comput Biol Bioinform, 2023. **20**(2): p. 852-863.
73. Karimi, M., et al., *DeepAffinity: interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks.* Bioinformatics, 2019. **35**(18): p. 3329-3338.
74. Li, M., et al., *BACPI: a bi-directional attention neural network for compound-protein interaction and binding affinity prediction.* Bioinformatics, 2022. **38**(7): p. 1995-2002.
75. Shuya Li, F.W., Hantao Shu, Tao Jiang, Dan Zhao, Jianyang Zeng, *MONN: A Multi-objective Neural Network for Predicting Compound-Protein Interactions and Affinities.* Cell Systems, 2020. **10**(4): p. 308-322.

**9. Funding.** This work is supported by the Singapore Ministry of Education (MOE) Tier 1 grant RG97/22. The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg) and the HADLEY supercomputer by Singapore Centre for Environmental Life Sciences Engineering (SCELSE). SCELSE is funded by Singapore's National Research Foundation, the Ministry of Education, Nanyang Technological University (NTU), and the National University of Singapore (NUS), and is hosted by NTU in partnership with NUS.

**Table 1. Screening power results for top models and commonly used models on CASF-2016, with all scores represented as the mean.** Arrows indicate direction of better results.

| Score function / model | Method | Prediction target | $EF_{1\%}$ ↑ | Forward screening success rate 1% ↑ |
|---|---|---|---|---|
| GenScore [42] | Deep learning, structure-based | Distance likelihood | 28.20 | 71.9% |
| RTMScore [5] | Deep learning, structure-based | Distance likelihood | 28.00 | 66.7% |
| PIGNet2 [43] | Deep learning, physics-based, structure-based | Binding affinity | 24.90 | 66.7% |
| DeepRMSD [44] | Deep learning, structure-based | Δ binding affinity | 21.95 | 47.4% |
| DeepDock [45] | Deep learning, structure-based | Distance likelihood | 16.41 | 43.9% |
| OnionNet-SFCT [4] | Deep learning, structure-based | Δ binding affinity | 15.5 | 35.1% |
| BIND | Deep learning, sequence-based | Binary classification between true binders and decoys | 14.91 | 28.7% |
| ChemPLP@GOLD [46] | Physics-based, Structure-based | Binding affinity | 11.91 [6] | 35.1% [6] |
| $\Delta VinaRF_{20}$ [47] | Machine learning, structure-based | Δ binding affinity | 11.73 [6] | 42.1% [6] |
| Glide-SP [48] | Physics-based, structure-based | Binding affinity | 11.44 [6] | 36.8% [6] |
| Glide-XP [48] | Physics-based, structure-based | Binding affinity | 8.83 [6] | 26.3% [6] |
| ChemScore@GOLD [49] | Machine learning, structure-based | Binding affinity | 8.65 [6] | 28.1% [6] |
| Autodock Vina [50] | Physics-based, structure-based | Binding affinity | 7.70 [6] | 29.8% [6] |

| | | | | |
|---|---|---|---|---|
| Zero-shot BIND (90% homologous sequences removed) | Deep learning, sequence-based | Binary classification between true binders and decoys | 6.98 | 13.5% |
| SSM-DTA ($pIC_{50}$) [14] | Machine learning, sequence-based | Drug-target affinity | 3.67 [6] | 8.2% [6] |

* best model's results as reported by respective authors

**Table 2. Screening power results for DEKOIS 2.0, with all scores represented as the mean.** Arrows indicate direction of better results.

| **Model** | **$EF_{0.5\%}$ ↑** | **$EF_{1\%}$ ↑** | **AUROC ↑** | **BEDROC ($\alpha$ = 80.5) ↑** |
|---|---|---|---|---|
| BIND | 25.36 | 24.46 | 0.944 | 0.730 |
| SSM-DTA [14] | 21.21 | 18.49 | - | 0.504 |
| RTMScore [5] | 20.99 | 18.53 | - | 0.558 |
| GenScore [42] | 20.24 | 17.87 | 0.757 | 0.539 |
| PIGNet2 [43] | 20.00 | 18.60 | 0.812 | 0.544 |
| KarmaDock [51] | 19.41 | 16.29 | 0.783 | 0.519 |
| Zero-shot BIND (90% homologous sequences removed) | 16.84 | 15.19 | 0.832 | 0.453 |
| SCORCH [52] | 15.01 | 13.78 | - | - |
| DyScore [53] | - | 7.5 | 0.702 | 0.216 |
| Vina [50] | - | 4.8 [53] | 0.631 [53] | 0.140 [53] |

* best model's results as reported by respective authors

**Table 3. Screening power results for DUD-AD, with all scores represented as the mean.** Arrows indicate direction of better results.

| **Score function / model** | **$EF_{1\%}$ ↑** | **AUROC ↑** | **BEDROC (α = 80.5) ↑** |
|---|---|---|---|
| BIND | 24.14 | 0.952 | 0.698 |
| Zero-shot BIND (90% protein homology sequences removed) | 12.03 | 0.851 | 0.396 |
| SSM-DTA ($pIC_{50}$) [14] | 10.19 | - | 0.336 |
| DeepRMSD [44] | 6.04 | - | - |
| OnionNet-SFCT with Vina [4] | 5.22 | 0.549 | - |
| Gnina [54] | 2.33 [4] | - | - |
| Vina [50] | 1.90 [4] | - | - |

* best model's results as reported by respective authors

**Table 4. Screening power results for DUD-E, with all scores represented as the mean**

| **Score function / model** | **$EF_{0.5\%}$ ↑** | **$EF_{1\%}$ ↑** | **AUROC ↑** | **BEDROC (α = 80.5) ↑** |
|---|---|---|---|---|
| BIND | 51.88 | 46.35 | 0.944 | 0.733 |
| RTMScore [5] | 42.47 | 35.10 | 0.831 | 0.558 |
| GenScore [42] | 44.02 | 33.96 | 0.828 | 0.546 |
| DyScore (target-aware split) [53] | - | 32.20 | 0.858 | 0.502 |
| PIGNet2 [43] | 36.80 | 31.20 | 0.850 | 0.515 |

| | | | | |
|---|---|---|---|---|
| Zero-shot BIND (90% protein homology sequences removed) | 30.52 | 26.39 | 0.826 | 0.430 |
| SSM-DTA ($pIC_{50}$) [14] | 33.59 | 24.78 | - | 0.397 |
| Gnina [54] | - | 20.40 [55] | 0.795 [55] | - |
| OnionNet-SFCT with Vina [4] | - | 15.54 | 0.724 | - |
| DeepRMSD with Vina* [44] | - | 14.80 | - | - |
| DyScore (target-unaware) [53] | - | 12.10 | 0.755 | 0.216 |
| Vina [50] | - | 8.82 [4] | 0.697 [4] | - |

* best model's results as reported by respective authors

**Table 5. Screening power results for LIT-PCBA, with all scores represented as the mean**

| **Score function / model** | **$EF_{1\%}$ ↑** | **AUROC ↑** | **BEDROC ($\alpha$ = 80.5) ↑** |
|---|---|---|---|
| BIND | 10.93 | 0.597 | 0.112 |
| IFP [56] | 7.46 [57] | - | - |
| GRIM [58] | 6.87 [57] | - | - |
| GenScore [42] | 6.80 | - | - |
| DyScore-MF [53] | 5.92 | 0.594 | 0.071 |
| $\Delta VinaRF_{20}$ [47] | 5.38 [57] | - | - |
| Pafnucy [59] | 5.32 [57] | - | - |
| FRAGSITE [60] | 4.78 | - | - |
| Zero-shot BIND | 4.45 | 0.564 | 0.047 |

| (90% protein homology sequences removed) | | | |
|---|---|---|---|
| EViS [61] | 4.18 | - | - |
| FINDSITE$^{comb2.0}$ [62] | 3.03 [61] | - | - |
| SSM-DTA ($pIC_{50}$) [14] | 3.03 | - | 0.040 |
| Score2 [53] | 3.00 | 0.621 | 0.047 |
| Gnina [54] | 2.58 [55] | 0.616 [55] | - |
| Surflex [63] | 2.51 [57] | - | - |
| Vina [50] | 2.33 | 0.565 | 0.037 |
| BANANA [64] | 1.81 | 0.580 | - |
| NNScore-v2 [65] | 1.70 [53] | 0.557 [53] | 0.025 [53] |
| RFScore-4 [3] | 1.28 [53] | 0.600 [53] | - |
| DSX [53] | 1.00 | 0.523 | 0.025 |
| Vinardo [66] | 0.99 [55] | 0.577 [55] | - |
| RFScore-VS [3] | 0.73 [55] | 0.542 [55] | - |

* best model's results as reported by respective authors

**Table 6. Drug-target affinity on benchmark datasets**

| **Dataset** | **DAVIS ($pK_d$)** | |
|---|---|---|
| **Model** | **MSE ↓** | **CI ↑** |
| SSM-DTA [14] | 0.219 | 0.875 |
| DeepDTA [67] | 0.262 | 0.870 |
| ELECTRA-DTA [68] | 0.238 | 0.897 |
| SubMDTA [69] | 0.218 | 0.894 |
| DeepPurpose [70] | 0.242 | 0.881 |
| AttentionMGT-DTA [71] | 0.193 | 0.891 |
| AttentionDTA [72] | 0.195 | 0.888 |
| BIND (trained on BindingDB only) | 0.409 | 0.747 |

* best model's results as reported by respective authors

**Table 7. Drug-target affinity on BindingDB**

| **Metric** | **$pK_i$** | | **$pK_d$** | | **$pIC_{50}$** | | **$pEC_{50}$** | |
|---|---|---|---|---|---|---|---|---|
| **Model** | **RMSE ↓** | **R ↑** | **RMSE ↓** | **R ↑** | **RMSE ↓** | **R ↑** | **RMSE ↓** | **R ↑** |
| DeepAffinity* [73] | 0.840 | 0.840 | - | - | 0.780 | 0.840 | - | - |
| SSM-DTA [14] | 0.792 | 0.863 | - | - | 0.712 | 0.878 | - | - |
| BACPI [74] | 0.800 | 0.860 | 1.080 | 0.730 | 0.740 | 0.860 | 0.780 | 0.850 |
| BIND | 0.965 | 0.798 | 0.935 | 0.798 | 0.924 | 0.808 | 0.891 | 0.822 |
| MONN [75] | - | - | - | - | 0.764 | 0.858 | - | - |

* best model's results as reported by respective authors

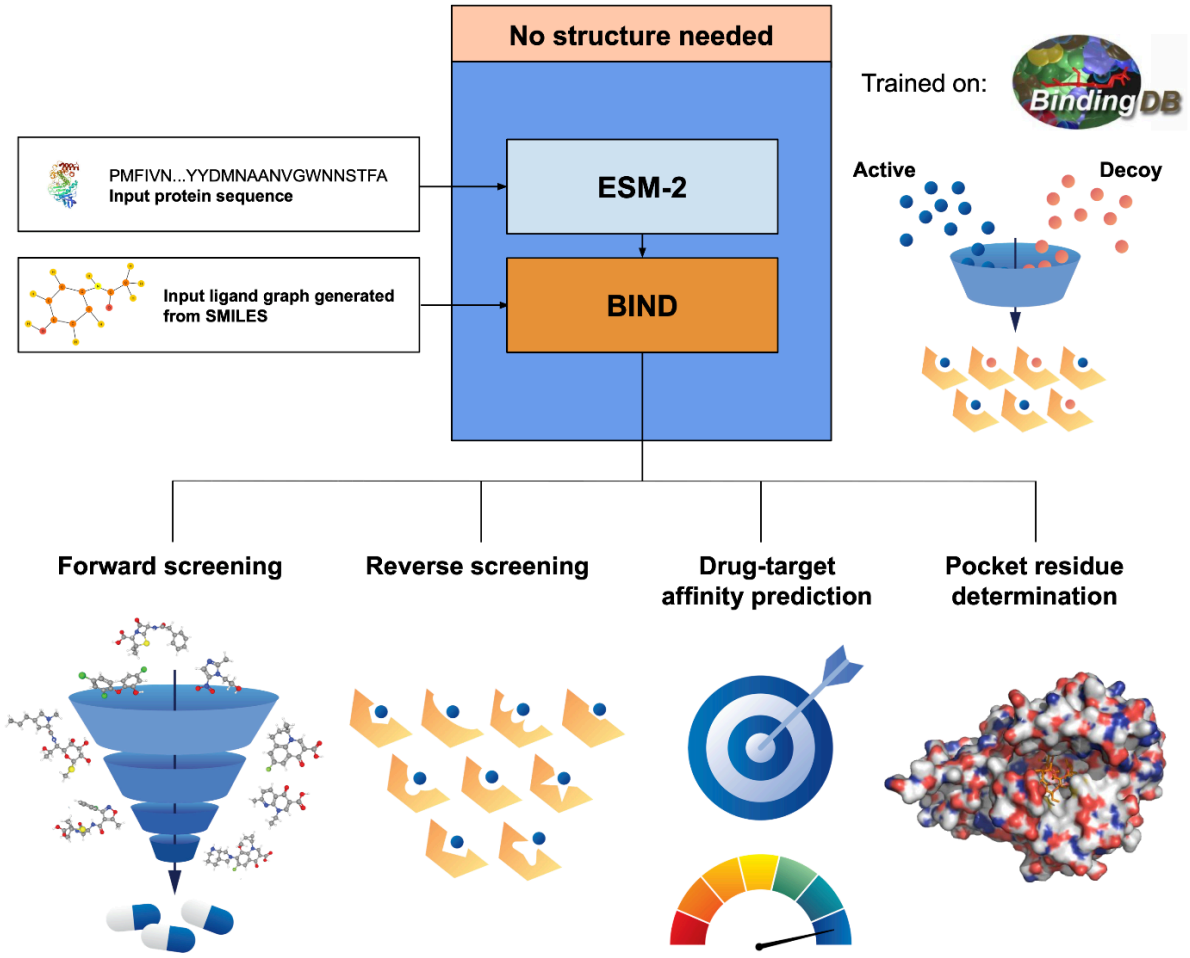



**Figure 1. BIND is able to perform forward screening, reverse screening, drug-target affinity prediction, and elucidate residues that make up the pocket, all without structural input.** BIND is trained on the BindingDB dataset only consisting of protein sequences and experimentally determined DTA values. By attaching BIND to a pre-trained ESM-2 protein language model and feeding in a protein sequence and SMILES molecular representation, which are then converted into graphs, the model can effectively discriminate between active and decoy ligands. This allows BIND to be used in forward and reverse screening, whilst predicting DTA values. Furthermore, residues in the binding pocket can be roughly elucidated using BIND, despite the model never being trained to do so.

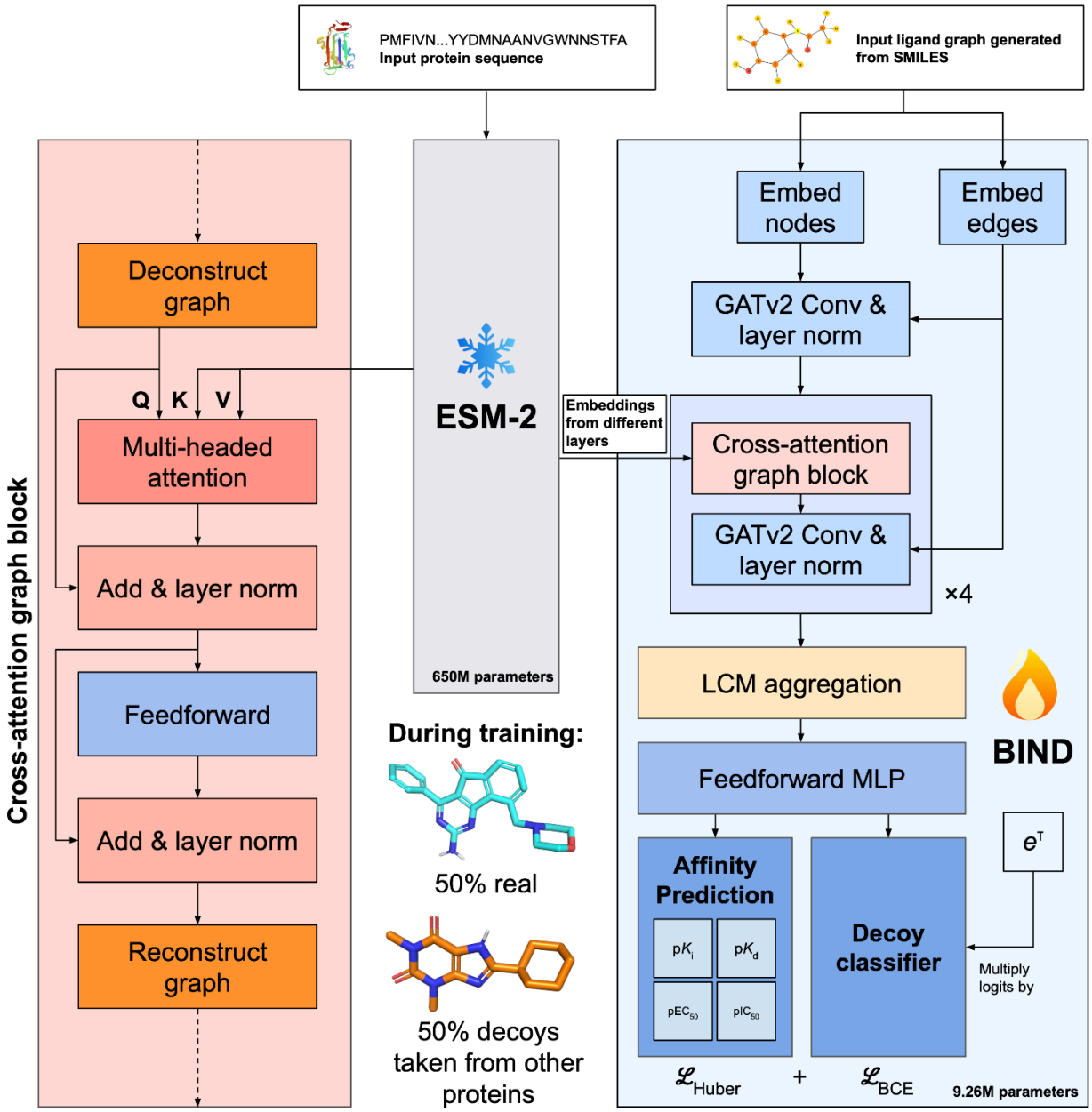



**Figure 2. BIND architecture incorporates a proposed cross-attention graph block, and is trained with both true ligands and decoys taken from other proteins in the same dataset.** The cross-attention graph block essentially deconstructs a graph and treats each node as a token for cross-attention - this allows the ligand to “query” the protein and its important parts. The loss is a summation of Huber loss for the affinity predictions, and binary cross entropy loss for the decoy classifier. During training, ESM-2’s weights are frozen such that only the 9.26M parameters from BIND are tuned. Q, K, and V in the diagram represent the query, key and value inputs characteristic of transformers.

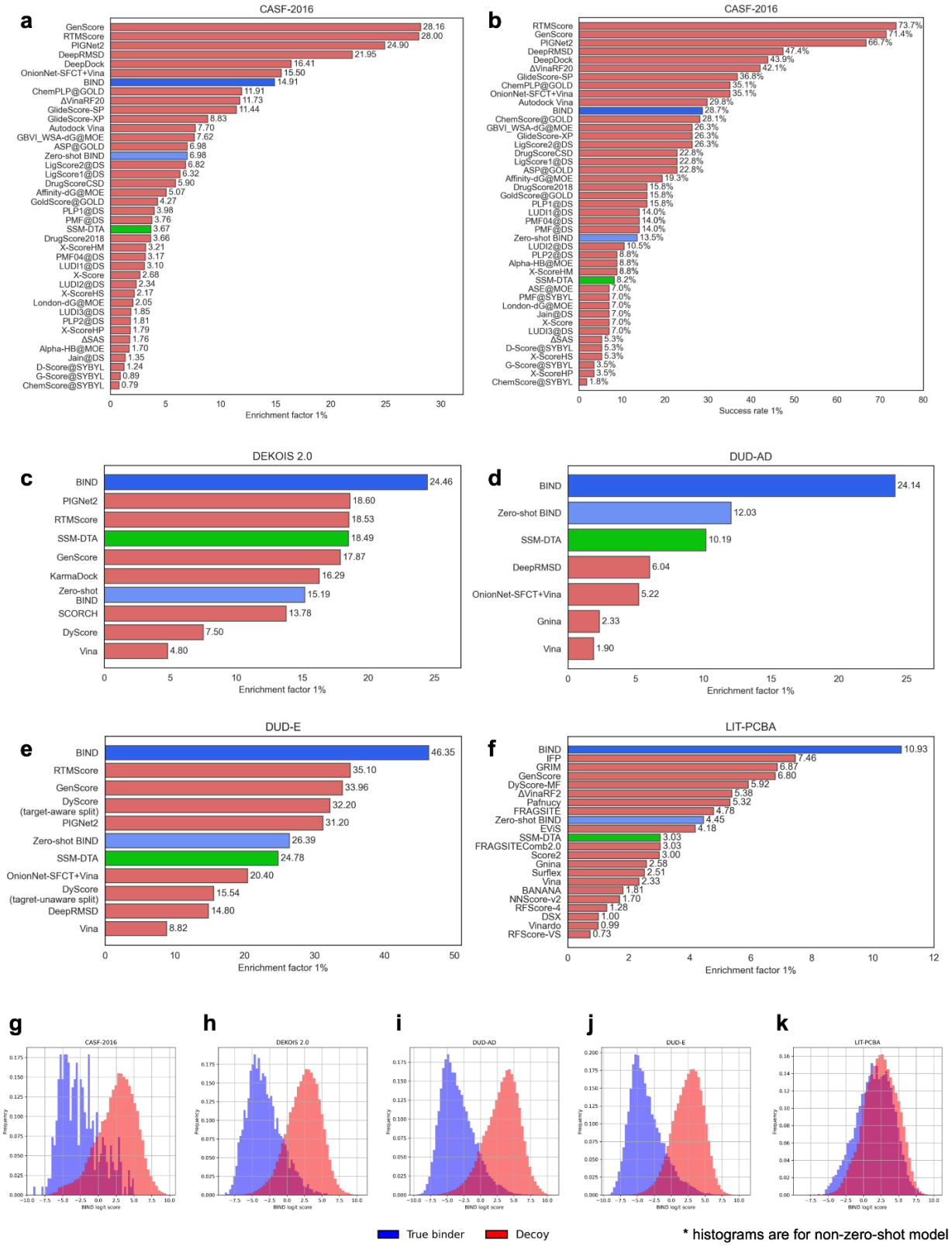

a
CASF-2016
GenScore 28.16
RTMScore 28.00
PIGNet2 24.90
DeepRMSD 21.95
DeepDock 16.41
OnionNet-SFCT+Vina 15.50
BIND 14.91
ChemPLP@GOLD 11.91
ΔVinaRF20 11.73
GlideScore-SP 11.44
GlideScore-XP 8.83
Autodock Vina 7.70
GBVI_WSA-dG@MOE 7.62
ASP@GOLD 6.98
Zero-shot BIND 6.98
LigScore2@DS 6.82
LigScore1@DS 6.32
DrugScoreCSD 5.90
Affinity-dG@MOE 5.07
GoldScore@GOLD 4.27
PLP1@DS 3.98
PMF@DS 3.76
SSM-DTA 3.67
DrugScore2018 3.66
X-ScoreHM 3.21
PMF04@DS 3.17
LUDI1@DS 3.10
X-Score 2.68
LUDI2@DS 2.34
X-ScoreHS 2.17
London-dG@MOE 2.05
LUDI3@DS 1.85
PLP2@DS 1.81
X-ScoreHP 1.79
ΔSAS 1.76
Alpha-HB@MOE 1.70
Jain@DS 1.35
D-Score@SYBYL 1.24
G-Score@SYBYL 0.89
ChemScore@SYBYL 0.79
Enrichment factor 1%
b
CASF-2016
RTMScore 73.7%
GenScore 71.4%
PIGNet2 66.7%
DeepRMSD 47.4%
DeepDock 43.9%
ΔVinaRF20 42.1%
GlideScore-SP 36.8%
ChemPLP@GOLD 35.1%
OnionNet-SFCT+Vina 35.1%
Autodock Vina 29.8%
BIND 28.7%
ChemScore@GOLD 28.1%
GBVI_WSA-dG@MOE 26.3%
GlideScore-XP 26.3%
LigScore2@DS 26.3%
DrugScoreCSD 22.8%
LigScore1@DS 22.8%
ASP@GOLD 22.8%
Affinity-dG@MOE 19.3%
DrugScore2018 15.8%
GoldScore@GOLD 15.8%
PLP1@DS 15.8%
LUDI1@DS 14.0%
PMF04@DS 14.0%
PMF@DS 14.0%
Zero-shot BIND 13.5%
LUDI2@DS 10.5%
PLP2@DS 8.8%
Alpha-HB@MOE 8.8%
X-ScoreHM 8.8%
SSM-DTA 8.2%
ASE@MOE 7.0%
PMF@SYBYL 7.0%
London-dG@MOE 7.0%
Jain@DS 7.0%
X-Score 7.0%
LUDI3@DS 7.0%
ΔSAS 5.3%
D-Score@SYBYL 5.3%
X-ScoreHS 5.3%
G-Score@SYBYL 3.5%
X-ScoreHP 3.5%
ChemScore@SYBYL 1.8%
Success rate 1%
c
DEKOIS 2.0
BIND 24.46
PIGNet2 18.60
RTMScore 18.53
SSM-DTA 18.49
GenScore 17.87
KarmaDock 16.29
Zero-shot BIND 15.19
SCORCH 13.78
DyScore 7.50
Vina 4.80
Enrichment factor 1%
d
DUD-AD
BIND 24.14
Zero-shot BIND 12.03
SSM-DTA 10.19
DeepRMSD 6.04
OnionNet-SFCT+Vina 5.22
Gnina 2.33
Vina 1.90
Enrichment factor 1%
e
DUD-E
BIND 46.35
RTMScore 35.10
GenScore 33.96
DyScore (target-aware split) 32.20
PIGNet2 31.20
Zero-shot BIND 26.39
SSM-DTA 24.78
OnionNet-SFCT+Vina 20.40
DyScore (tagret-unaware split) 15.54
DeepRMSD 14.80
Vina 8.82
Enrichment factor 1%
f
LIT-PCBA
BIND 10.93
IFP 7.46
GRIM 6.87
GenScore 6.80
DyScore-MF 5.92
ΔVinaRF2 5.38
Pafnucy 5.32
FRAGSITE 4.78
Zero-shot BIND 4.45
EVIS 4.18
SSM-DTA 3.03
FRAGSITEComb2.0 3.03
Score2 3.00
Gnina 2.58
Surflex 2.51
Vina 2.33
BANANA 1.81
NNScore-v2 1.70
RFScore-4 1.28
DSX 1.00
Vinardo 0.99
RFScore-VS 0.73
Enrichment factor 1%
g
CASF-2016
h
DEKOIS 2.0
i
DUD-AD
j
DUD-E
k
LIT-PCBA
Frequency
BIND logit score
True binder
Decoy
* histograms are for non-zero-shot model

**Figure 3. The model is performant on docking benchmarks and outperforms the state-of-the-art DTA only model in enrichment.** CASF-2016, DEKOIS 2.0, DUD-AD, DUD-E and LIT-PCBA were evaluated. (a-b) CASF-2016 1% enrichment factors and 1% success rate, (c-f) DEKOIS 2.0, DUD-AD, DUD-E and LIT-PCBA 1% enrichment factors, (g-k) probability-normalised histogram of the non-zero-shot BIND classifier logit output showing separation between true binders and decoys. All enrichment factors and success rates are averaged across the entire datasets evaluated, and histograms shown are distributions of logit scores for all proteins and ligands tested. The green bar indicates the enrichment factor of the published SSM-DTA model, which predicts the $pIC_{50}$ DTA. The zero-shot BIND model indicates the model in which proteins with > 90% homology comparative to the evaluation datasets are removed during training.

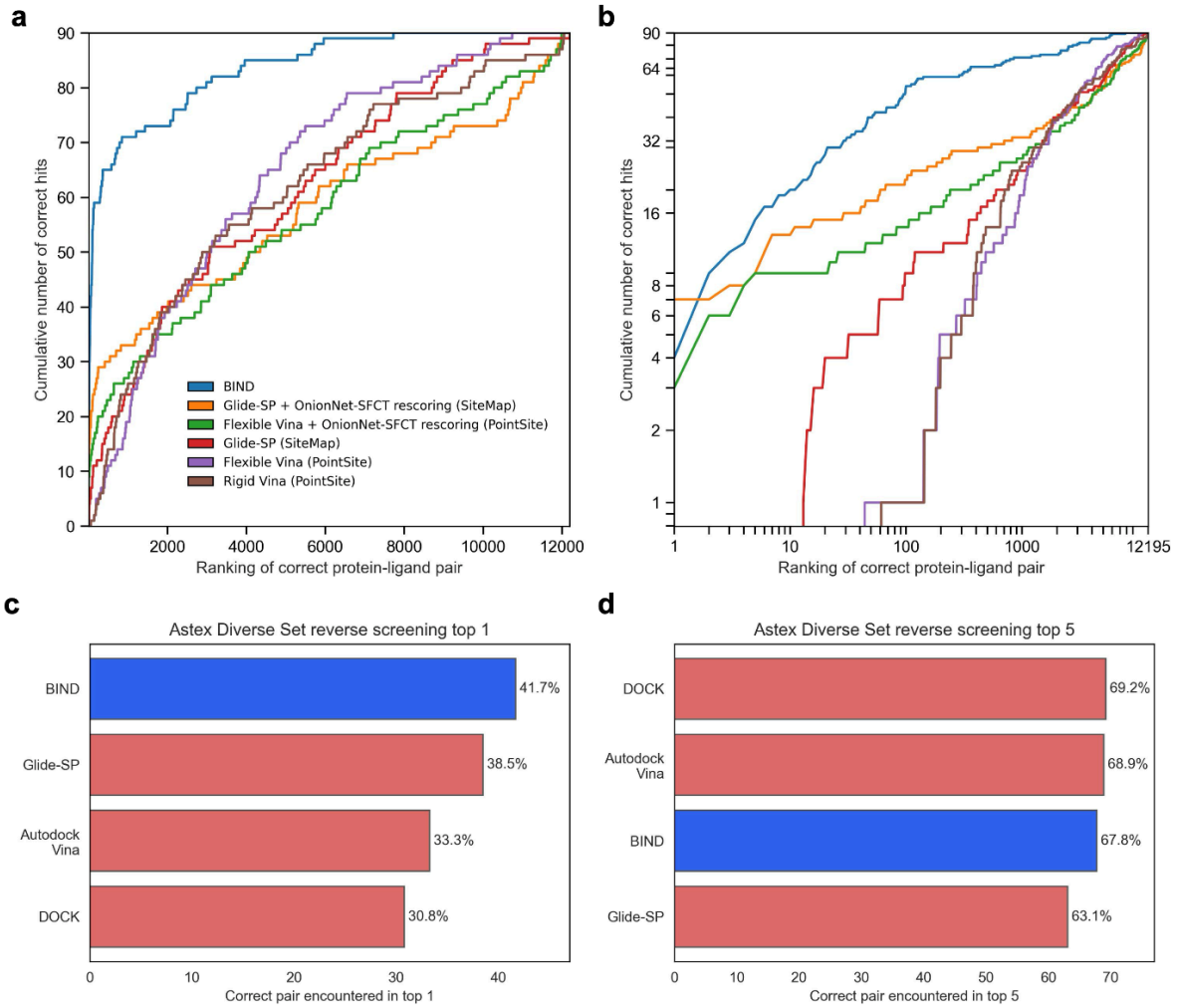



**Figure 4. Protein language models can outperform standard reverse docking pipelines using structures from AlphaFold2.** A total of 12,195 proteins and 90 ligands from Luo *et al*., 2023's AlphaFold2 reverse docking benchmark were individually pairwise scored using BIND and the ligands ranked by classification score for each protein. (a, b) The cumulative ranking of

BIND in standard and logarithmic plots respectively. The legend indicates the type of score function used, and the software used to determine the pocket locations in parentheses. (c, d) Ranking of 84 ligands against 85 proteins in the reverse docking benchmark on the Astex dataset, with other benchmark scores obtained from Luo *et al*., 2017.

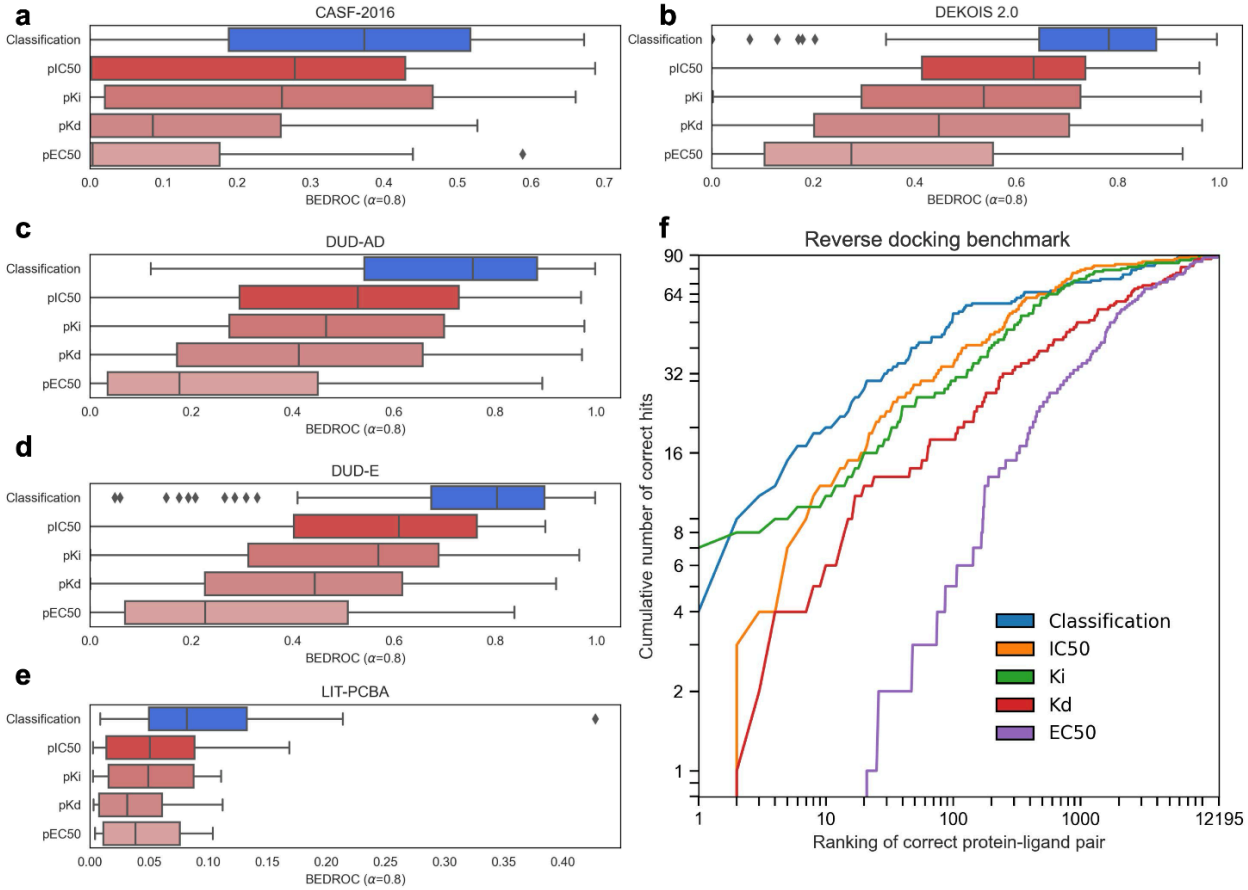



**Figure 5. Predicted binding affinity has lower enriching power compared to the classifier in both forward and reverse screening.** (a-e) Box-and-whisker plots of BEDROC values across the entire CASF-2016, DEKOIS 2.0, DUD-AD, DUD-E and LIT-PCBA datasets respectively, with boxes representing the interquartile range and the median demarcated in the box, whiskers showing the fence and diamonds showing outliers. (f) Cumulative logarithmic plot of ranking in reverse screening on Luo *et al*., 2023's AlphaFold2 reverse docking benchmark dataset.

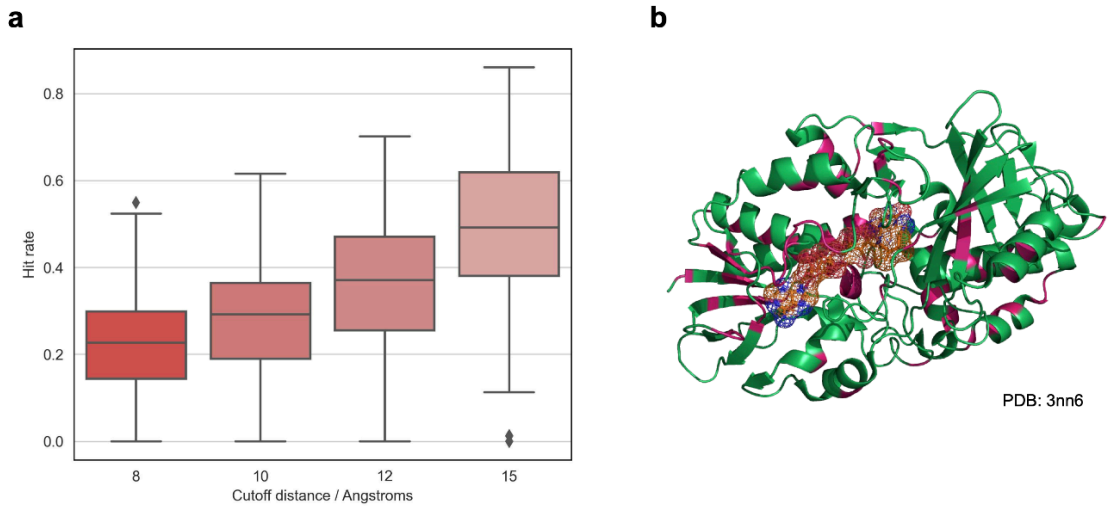
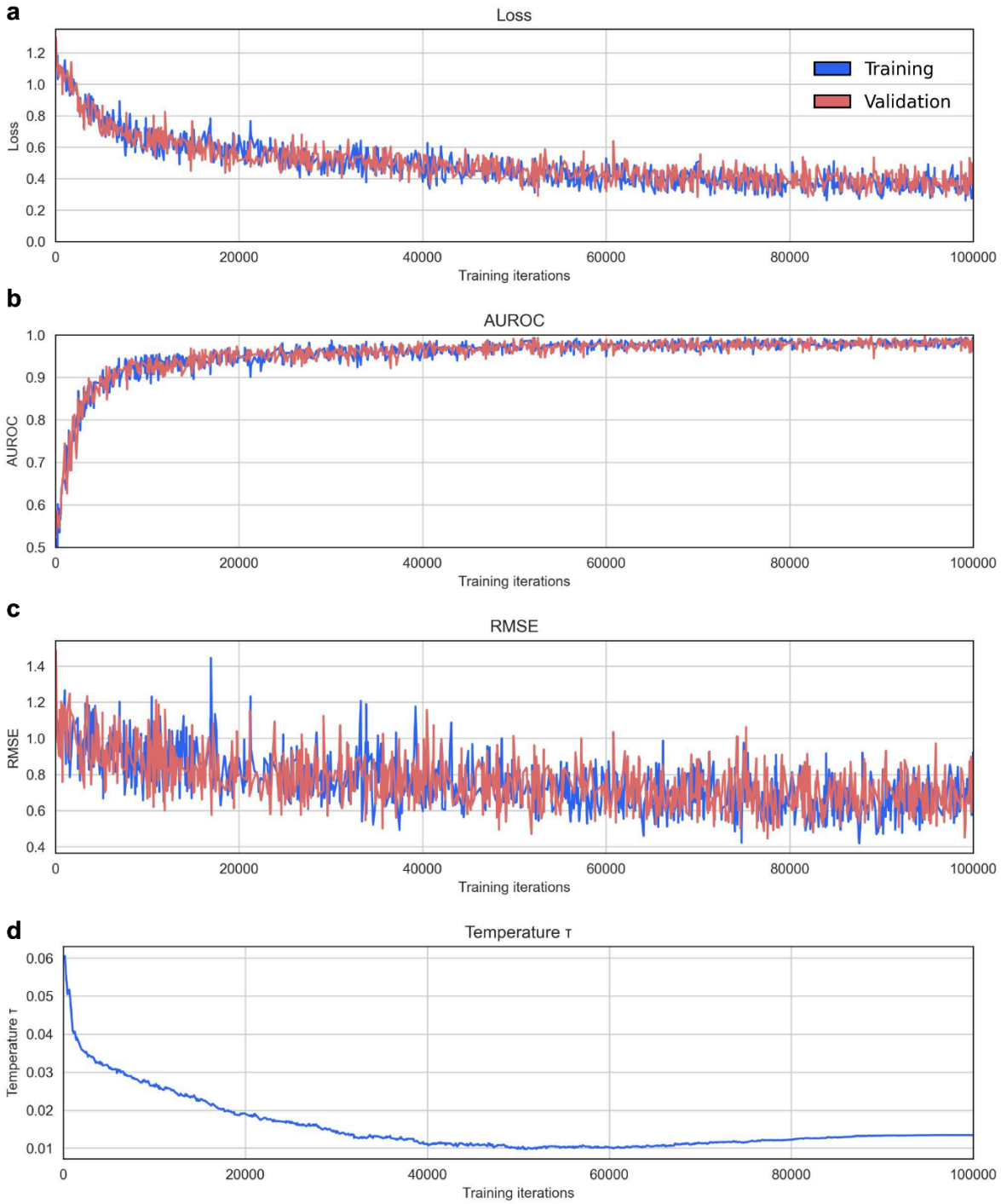
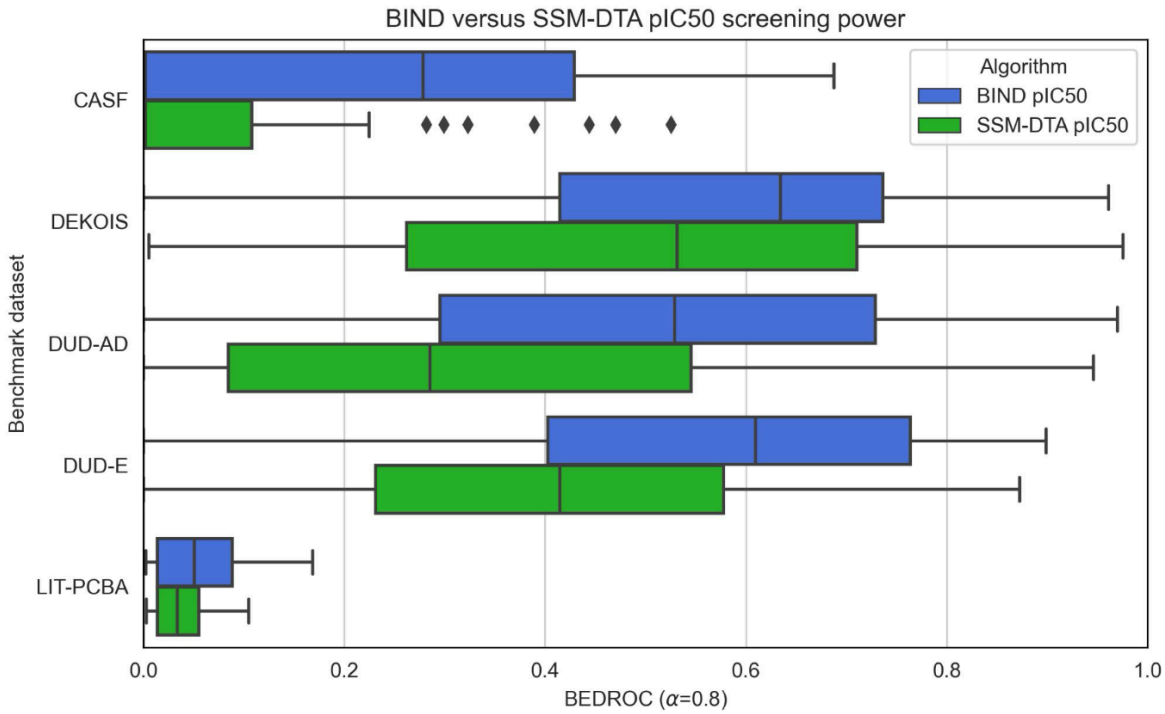



**Figure 6. Pocket residues can be elucidated from the cross-attention scores of the model despite the model not being trained to do so.** (a) 187 single-chained proteins from the CASF-2016 core set are inputted into BIND, with the attention scores averaged across all four cross-attention graph block layers. The proportion of binding pocket residues identified is shown at different cutoff distances. Boxes represent the interquartile range with the median demarcated in the box, whiskers showing the fence and diamonds showing outliers. (b) A visual example of the top 20% most attended to residues in 6-hydroxy-L-nicotine oxidase coloured in pink.

**S1 Model experiments.** In developing the model, different model configurations were trialled. The final, most successful version is described under the methods section of this work. Notably, flipping the order of ESM-2 latents input (i.e. cross-attending to downstream parts of the ESM-2 transformer before upstream) into the model resulted in worse results. Other experiments included having the ESM-2 embeddings cross-attend to the molecular graph instead, and then mean pooling over the tokens before a small feedforward layer. These resulted in similar RMSE predictions to having the molecular graph query the ESM-2 embeddings but drastically increased the computational overhead. When training with lower gradient accumulation steps of 128, certain instances of the model failed to converge. Lastly, an implicit hydrogen model was attempted, however these failed to perform as well as the explicit hydrogen models. The authors hypothesise that the addition of hydrogen atoms resulted in lesser compression of information in the form of latent space in the model, and hence culminated in better results overall.




**Supplementary Figure 1. Training and evaluation metrics for the model.** (a) Loss curve with a combined loss of Huber for the DTA regression and binary cross entropy for decoy/true binder classification; (b) AUROC train and validation curve; (c) DTA RMSE train and validation curve; (d) temperature parameter throughout training.

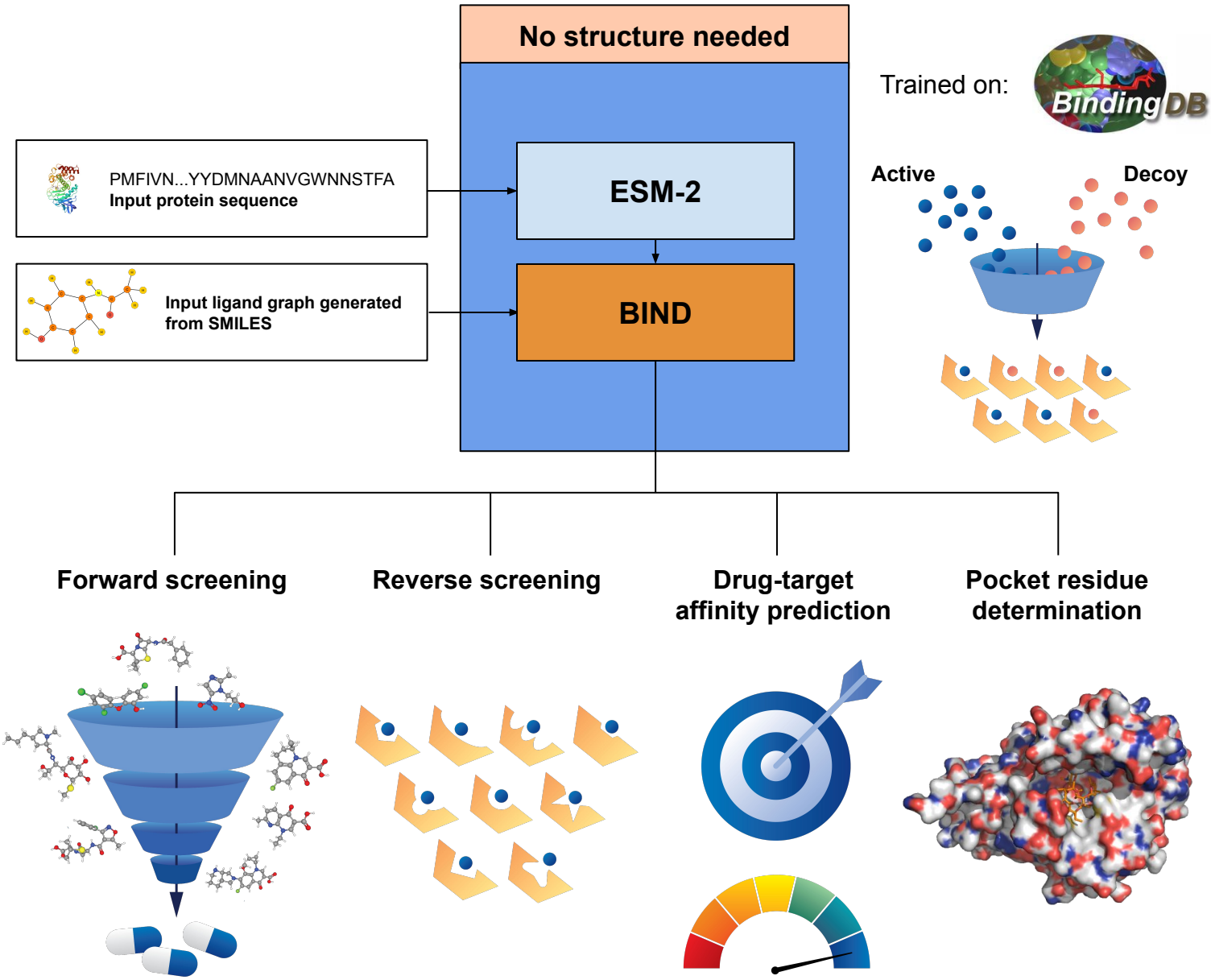


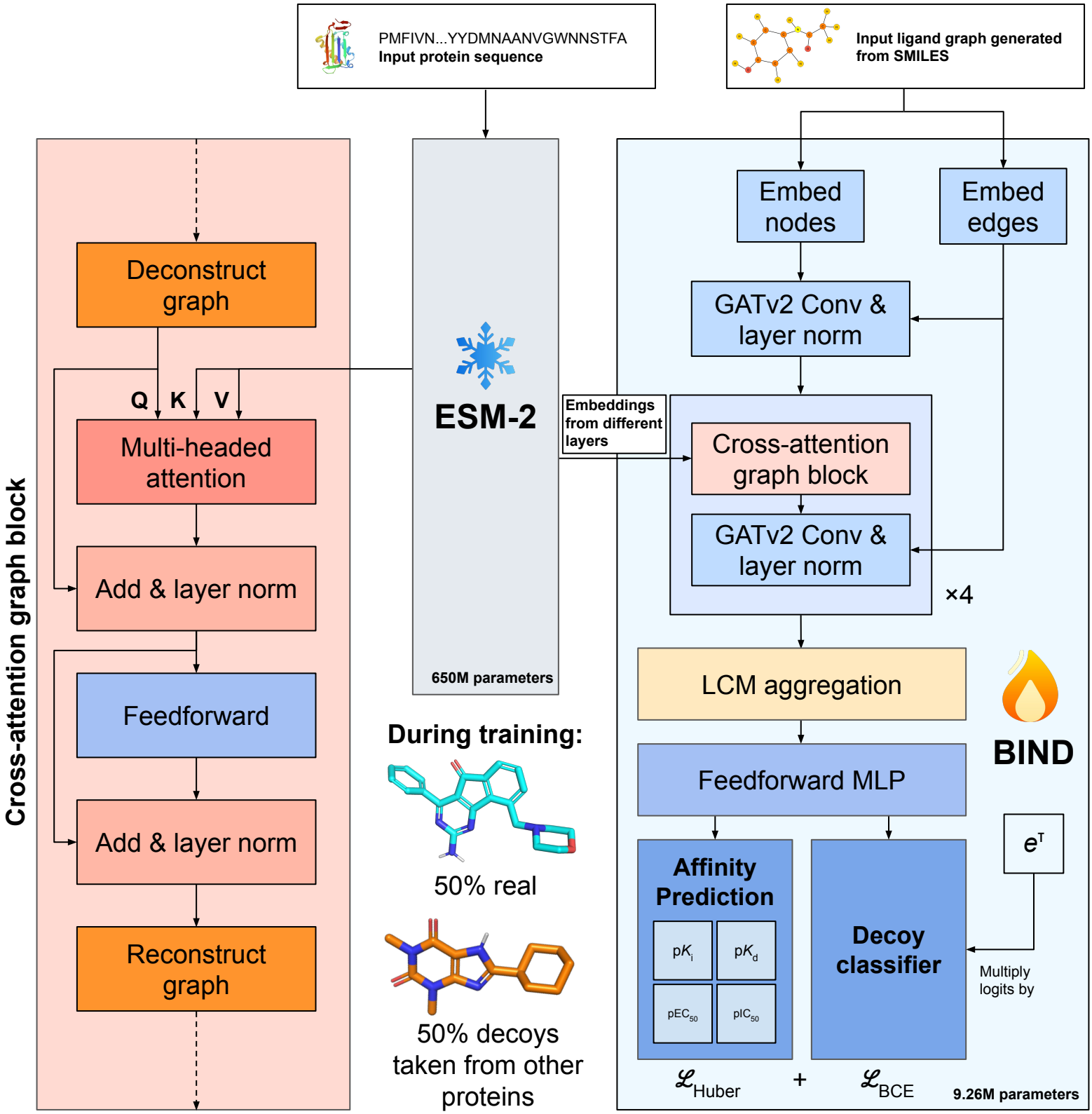

**Supplementary Figure 2. BIND still achieves higher screening power than SSM-DTA on its $pIC_{50}$ prediction despite SSM-DTA's being more accurate when measured via the RMSE metric.** The BEDROC values across the CASF-2016, DEKOIS, DUD-AD, DUD-E and LIT-PCBA datasets are compared and calculated by ranking on the $pIC_{50}$ predictions of both BIND and SSM-DTA. Boxes represent the interquartile range with the median demarcated in the box, whiskers showing the fence and diamonds showing outliers

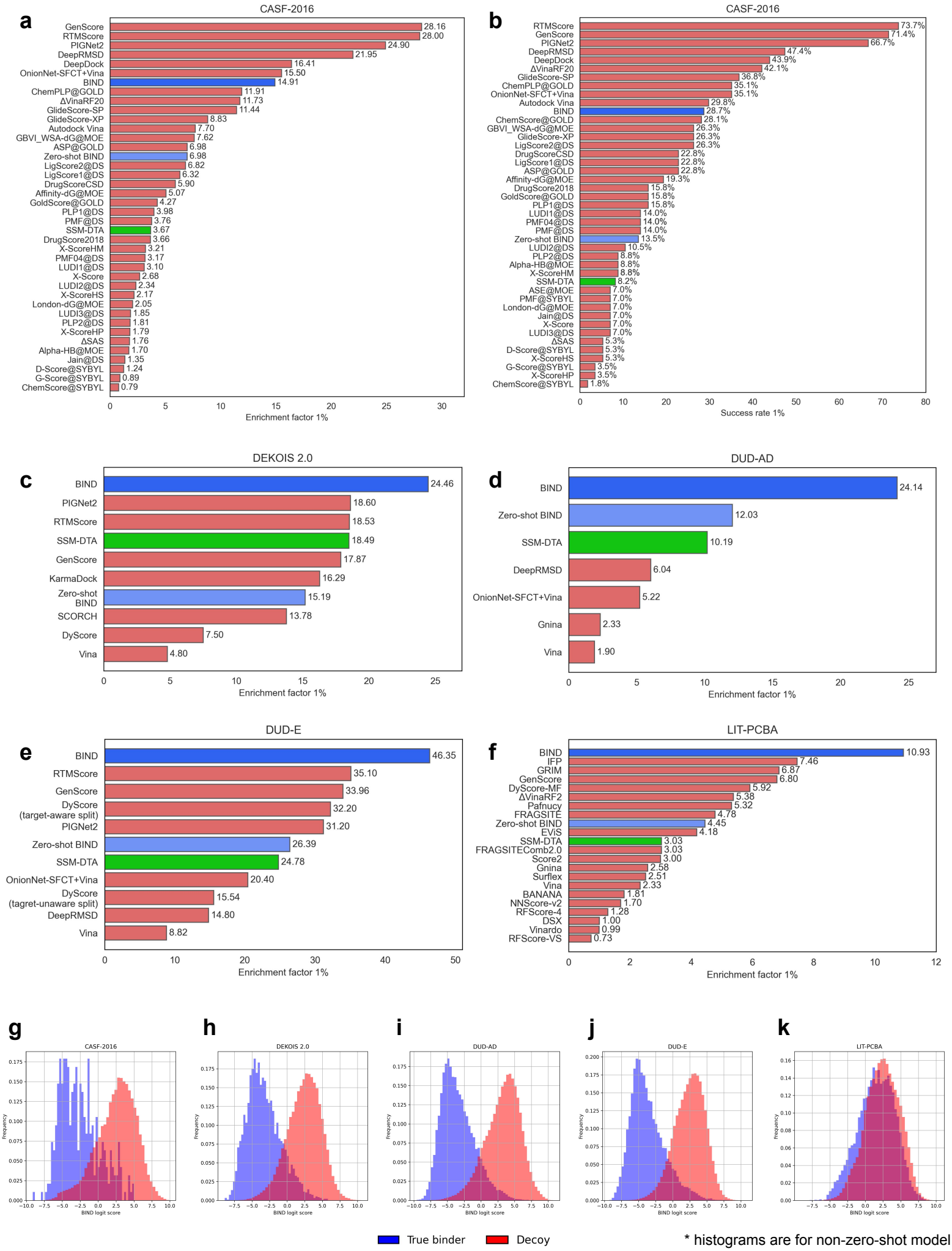
**Supplementary File 1. All evaluation metrics across individual targets in different evaluation datasets using BIND.**

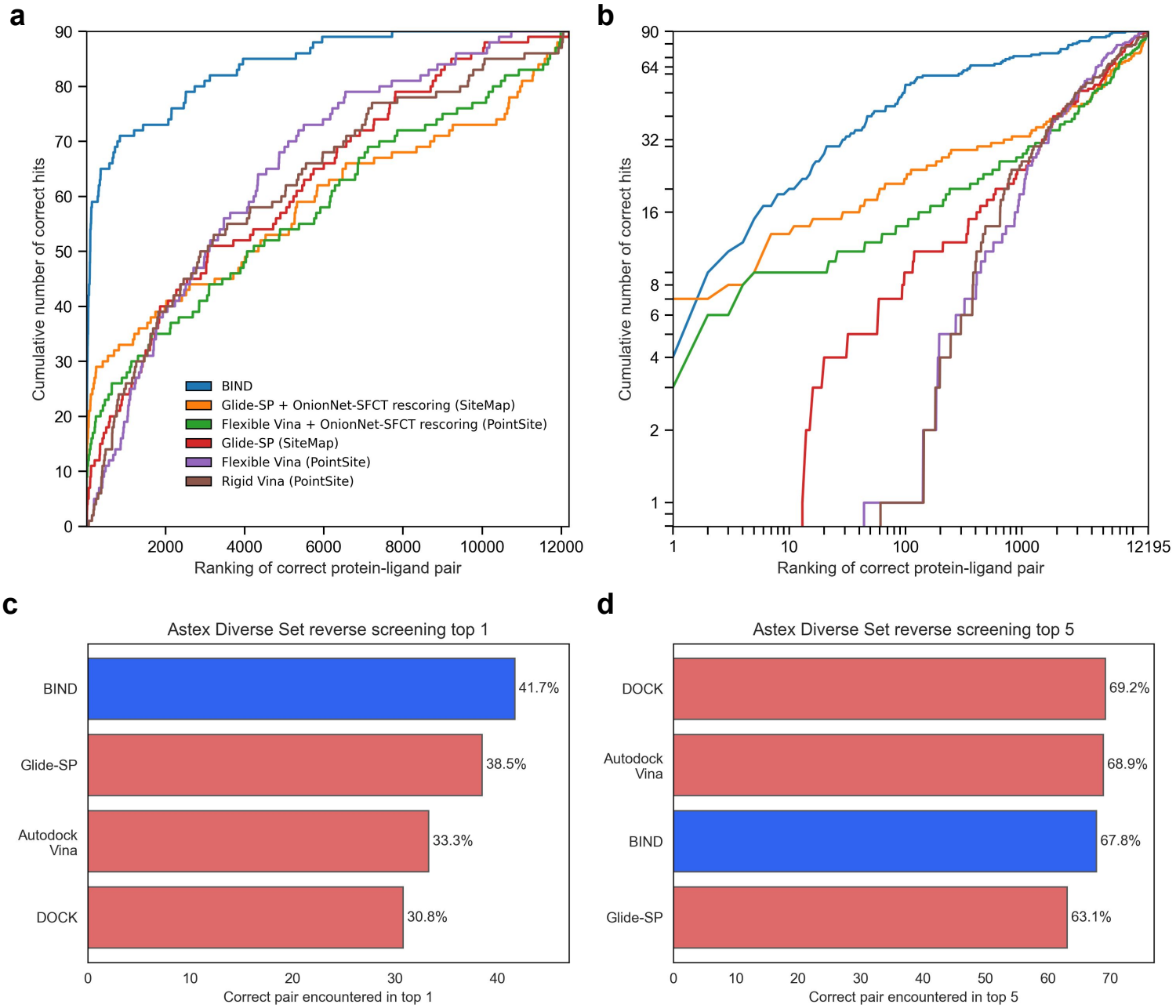
**Supplementary File 2. All evaluation metrics across individual targets in different evaluation datasets using zero-shot BIND.**

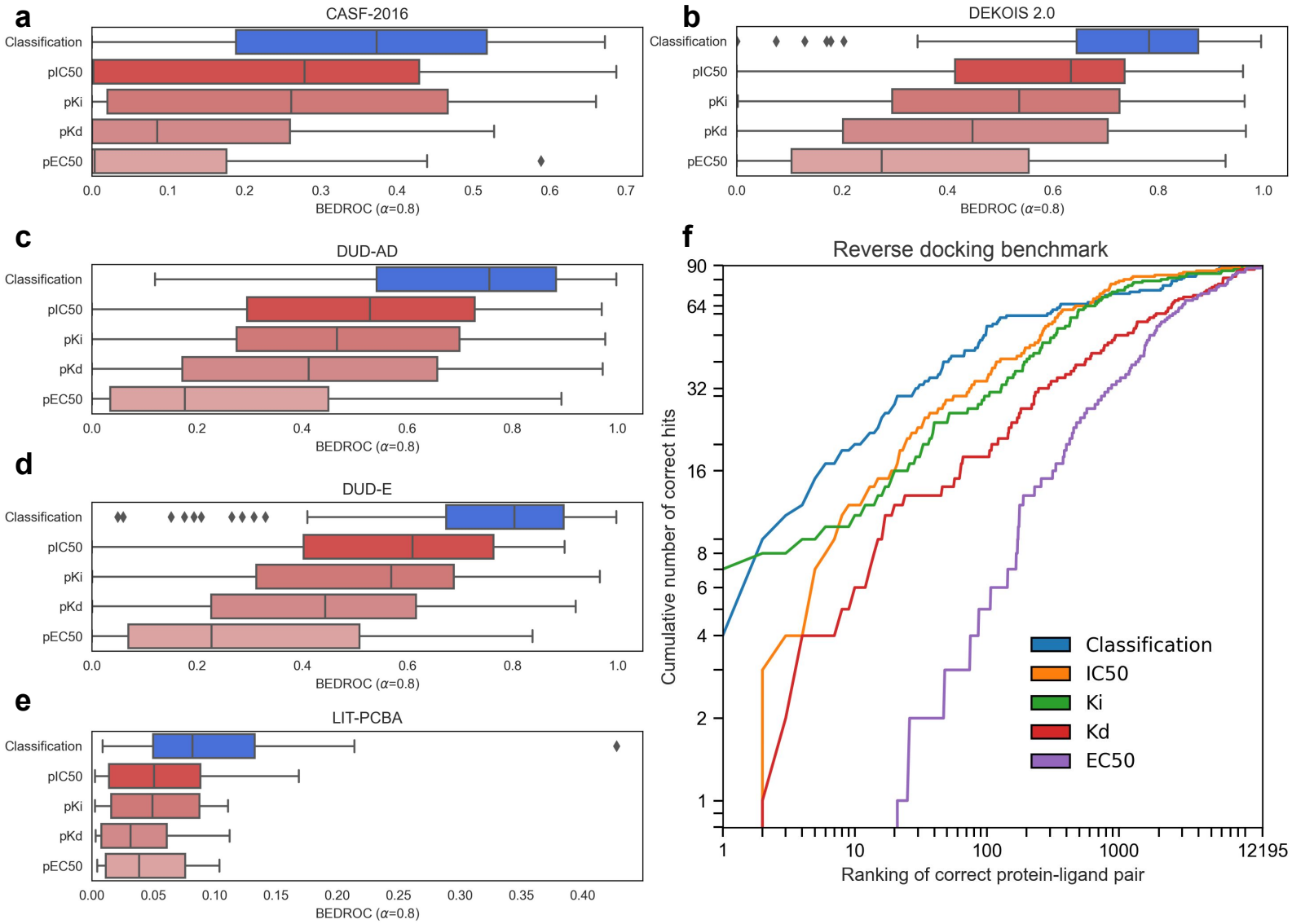
**Supplementary File 3. Protein sequences used in evaluation on CASF-2016, DEKOIS 2.0, DUD-AD, DUD-E, and LIT-PCBA.**

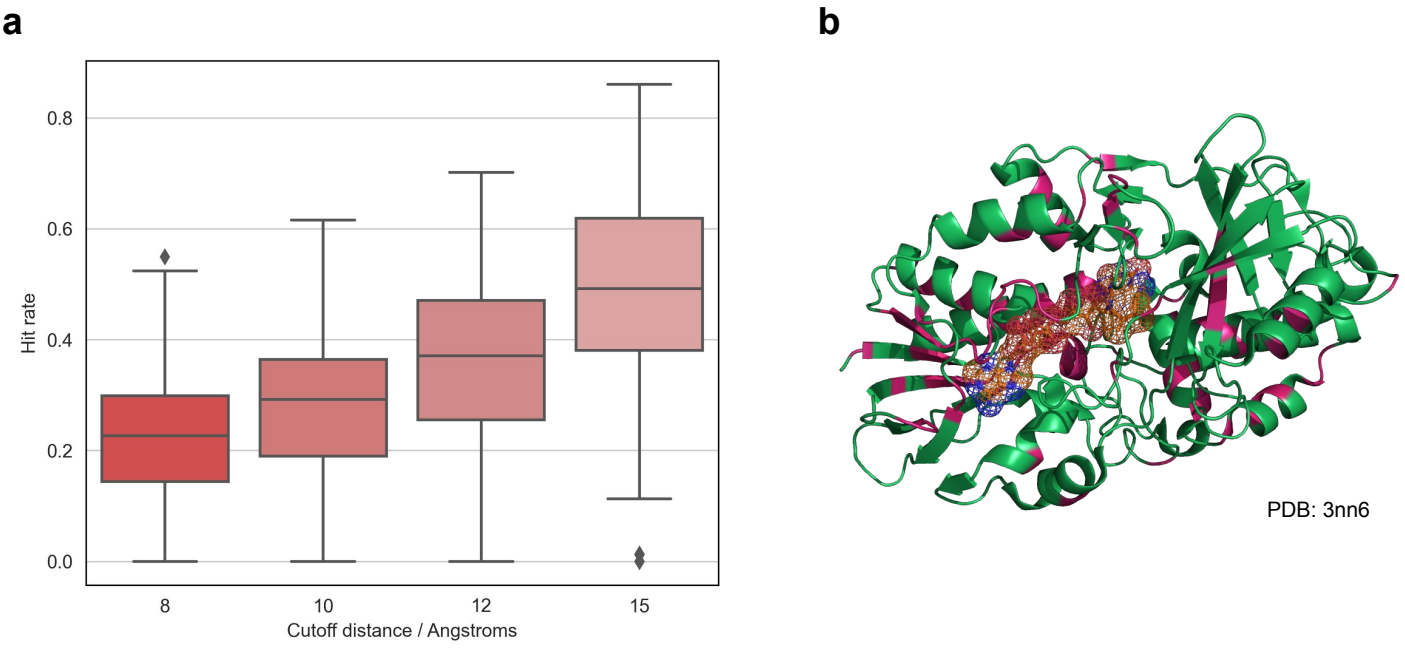

No structure needed
PMFIVN...YYDMNAANVGWNNSTFA
Input protein sequence
Input ligand graph generated from SMILES
ESM-2
BIND
Trained on:
BindingDB
Active
Decoy
Forward screening
Reverse screening
Drug-target affinity prediction
Pocket residue determination

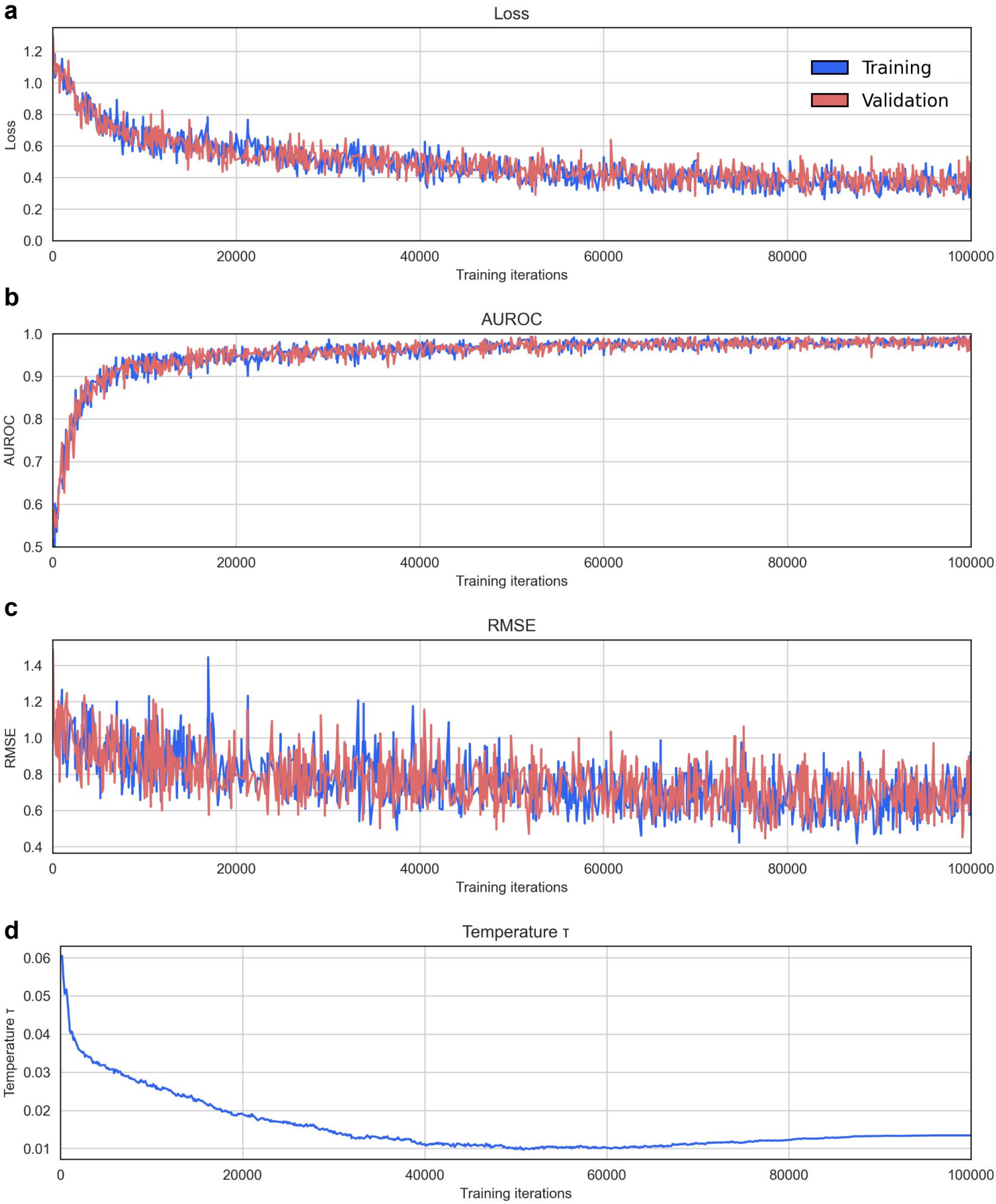

PMFIVN...YYDMNAANVGWNNSTFA
Input protein sequence
Input ligand graph generated from SMILES
Cross-attention graph block
Deconstruct graph
Q
K
V
Multi-headed attention
Add & layer norm
Feedforward
Add & layer norm
Reconstruct graph
ESM-2
650M parameters
Embeddings from different layers
Embed nodes
Embed edges
GATv2 Conv & layer norm
Cross-attention graph block
GATv2 Conv & layer norm
×4
LCM aggregation
Feedforward MLP
Affinity Prediction
pKi
pKd
pEC50
pIC50
Decoy classifier
eT
Multiply logits by
ℒHuber
+
ℒBCE
9.26M parameters
BIND
During training:
50% real
50% decoys taken from other proteins

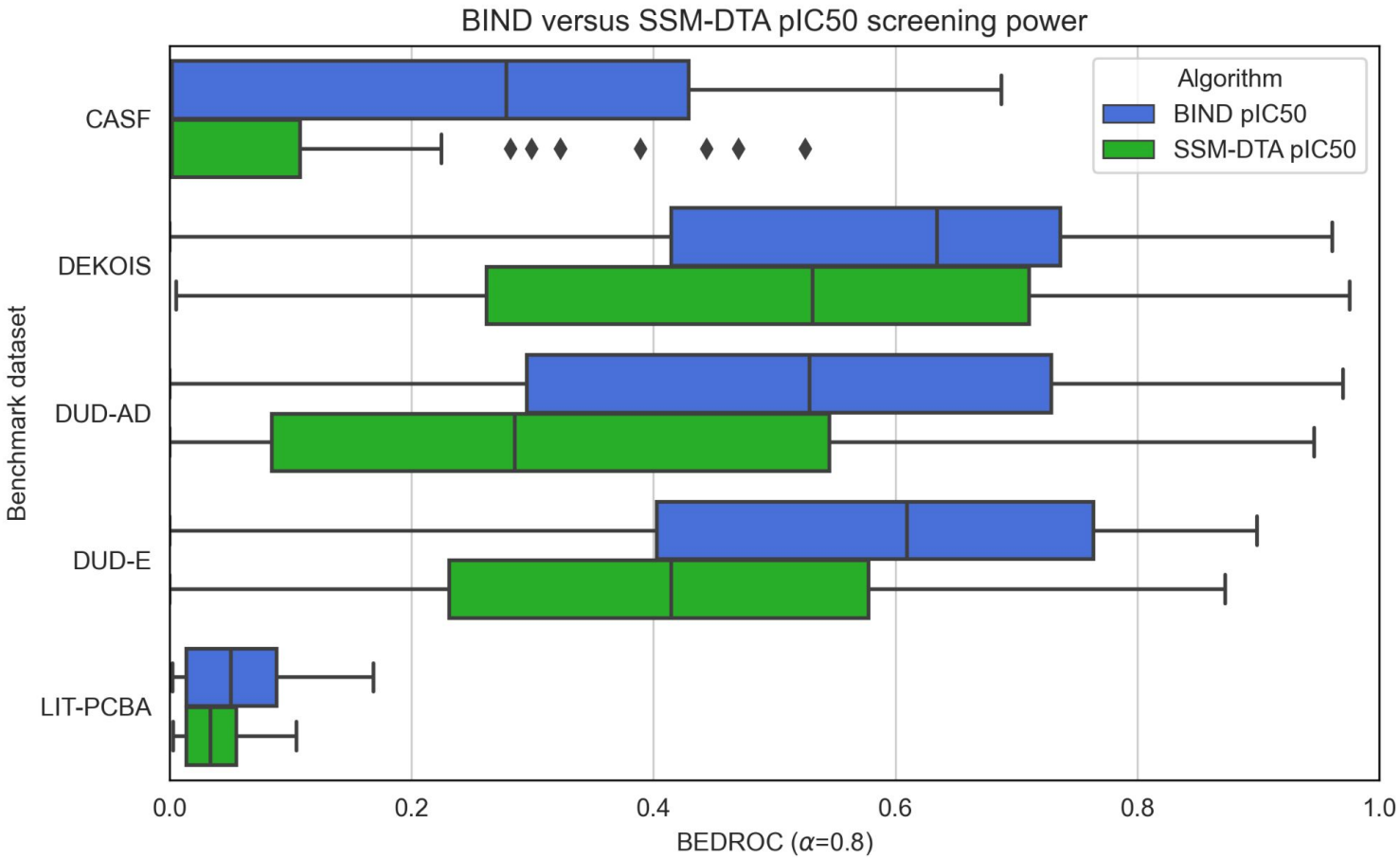

a
CASF-2016
GenScore 28.16
RTMScore 28.00
PIGNet2 24.90
DeepRMSD 21.95
DeepDock 16.41
OnionNet-SFCT+Vina 15.50
BIND 14.91
ChemPLP@GOLD 11.91
ΔVinaRF20 11.73
GlideScore-SP 11.44
GlideScore-XP 8.83
Autodock Vina 7.70
GBVI_WSA-dG@MOE 7.62
ASP@GOLD 6.98
Zero-shot BIND 6.98
LigScore2@DS 6.82
LigScore1@DS 6.32
DrugScoreCSD 5.90
Affinity-dG@MOE 5.07
GoldScore@GOLD 4.27
PLP1@DS 3.98
PMF@DS 3.76
SSM-DTA 3.67
DrugScore2018 3.66
X-ScoreHM 3.21
PMF04@DS 3.17
LUDI1@DS 3.10
X-Score 2.68
LUDI2@DS 2.34
X-ScoreHS 2.17
London-dG@MOE 2.05
LUDI3@DS 1.85
PLP2@DS 1.81
X-ScoreHP 1.79
ΔSAS 1.76
Alpha-HB@MOE 1.70
Jain@DS 1.35
D-Score@SYBYL 1.24
G-Score@SYBYL 0.89
ChemScore@SYBYL 0.79
0 5 10 15 20 25 30
Enrichment factor 1%
b
CASF-2016
RTMScore 73.7%
GenScore 71.4%
PIGNet2 66.7%
DeepRMSD 47.4%
DeepDock 43.9%
ΔVinaRF20 42.1%
GlideScore-SP 36.8%
ChemPLP@GOLD 35.1%
OnionNet-SFCT+Vina 35.1%
Autodock Vina 29.8%
BIND 28.7%
ChemScore@GOLD 28.1%
GBVI_WSA-dG@MOE 26.3%
GlideScore-XP 26.3%
LigScore2@DS 26.3%
DrugScoreCSD 22.8%
LigScore1@DS 22.8%
ASP@GOLD 22.8%
Affinity-dG@MOE 19.3%
DrugScore2018 15.8%
GoldScore@GOLD 15.8%
PLP1@DS 15.8%
LUDI1@DS 14.0%
PMF04@DS 14.0%
PMF@DS 14.0%
Zero-shot BIND 13.5%
LUDI2@DS 10.5%
PLP2@DS 8.8%
Alpha-HB@MOE 8.8%
X-ScoreHM 8.8%
SSM-DTA 8.2%
ASE@MOE 7.0%
PMF@SYBYL 7.0%
London-dG@MOE 7.0%
Jain@DS 7.0%
X-Score 7.0%
LUDI3@DS 7.0%
ΔSAS 5.3%
D-Score@SYBYL 5.3%
X-ScoreHS 5.3%
G-Score@SYBYL 3.5%
X-ScoreHP 3.5%
ChemScore@SYBYL 1.8%
0 10 20 30 40 50 60 70 80
Success rate 1%
c
DEKOIS 2.0
BIND 24.46
PIGNet2 18.60
RTMScore 18.53
SSM-DTA 18.49
GenScore 17.87
KarmaDock 16.29
Zero-shot BIND 15.19
SCORCH 13.78
DyScore 7.50
Vina 4.80
0 5 10 15 20 25
Enrichment factor 1%
d
DUD-AD
BIND 24.14
Zero-shot BIND 12.03
SSM-DTA 10.19
DeepRMSD 6.04
OnionNet-SFCT+Vina 5.22
Gnina 2.33
Vina 1.90
0 5 10 15 20 25
Enrichment factor 1%
e
DUD-E
BIND 46.35
RTMScore 35.10
GenScore 33.96
DyScore (target-aware split) 32.20
PIGNet2 31.20
Zero-shot BIND 26.39
SSM-DTA 24.78
OnionNet-SFCT+Vina 20.40
DyScore (tagret-unaware split) 15.54
DeepRMSD 14.80
Vina 8.82
0 10 20 30 40 50
Enrichment factor 1%
f
LIT-PCBA
BIND 10.93
IFP 7.46
GRIM 6.87
GenScore 6.80
DyScore-MF 5.92
ΔVinaRF2 5.38
Pafnucy 5.32
FRAGSITE 4.78
Zero-shot BIND 4.45
EViS 4.18
SSM-DTA 3.03
FRAGSITEComb2.0 3.03
Score2 3.00
Gnina 2.58
Surflex 2.51
Vina 2.33
BANANA 1.81
NNScore-v2 1.70
RFScore-4 1.28
DSX 1.00
Vinardo 0.99
RFScore-VS 0.73
0 2 4 6 8 10 12
Enrichment factor 1%
g
CASF-2016
h
DEKOIS 2.0
i
DUD-AD
j
DUD-E
k
LIT-PCBA
Frequency
BIND logit score
True binder
Decoy
* histograms are for non-zero-shot model


a
Cumulative number of correct hits
Ranking of correct protein-ligand pair
BIND
Glide-SP + OnionNet-SFCT rescoring (SiteMap)
Flexible Vina + OnionNet-SFCT rescoring (PointSite)
Glide-SP (SiteMap)
Flexible Vina (PointSite)
Rigid Vina (PointSite)
b
Cumulative number of correct hits
Ranking of correct protein-ligand pair
c
Astex Diverse Set reverse screening top 1
BIND
41.7%
Glide-SP
38.5%
Autodock Vina
33.3%
DOCK
30.8%
Correct pair encountered in top 1
d
Astex Diverse Set reverse screening top 5
DOCK
69.2%
Autodock Vina
68.9%
BIND
67.8%
Glide-SP
63.1%
Correct pair encountered in top 5


a
CASF-2016
Classification
pIC50
pKi
pKd
pEC50
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
BEDROC (α=0.8)
b
DEKOIS 2.0
Classification
pIC50
pKi
pKd
pEC50
0.0 0.2 0.4 0.6 0.8 1.0
BEDROC (α=0.8)
c
DUD-AD
Classification
pIC50
pKi
pKd
pEC50
0.0 0.2 0.4 0.6 0.8 1.0
BEDROC (α=0.8)
d
DUD-E
Classification
pIC50
pKi
pKd
pEC50
0.0 0.2 0.4 0.6 0.8 1.0
BEDROC (α=0.8)
e
LIT-PCBA
Classification
pIC50
pKi
pKd
pEC50
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40
BEDROC (α=0.8)
f
Reverse docking benchmark
Cumulative number of correct hits
90
64
32
16
8
6
4
2
1
1 10 100 1000 12195
Ranking of correct protein-ligand pair
Classification
IC50
Ki
Kd
EC50


a
b
Hit rate
0.0
0.2
0.4
0.6
0.8
8
10
12
15
Cutoff distance / Angstroms
PDB: 3nn6


a
Loss
Training
Validation
Loss
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
20000
40000
60000
80000
100000
Training iterations
b
AUROC
AUROC
1.0
0.9
0.8
0.7
0.6
0.5
0
20000
40000
60000
80000
100000
Training iterations
c
RMSE
RMSE
1.4
1.2
1.0
0.8
0.6
0.4
0
20000
40000
60000
80000
100000
Training iterations
d
Temperature τ
Temperature τ
0.06
0.05
0.04
0.03
0.02
0.01
0
20000
40000
60000
80000
100000
Training iterations


BIND versus SSM-DTA pIC50 screening power
Algorithm
BIND pIC50
SSM-DTA pIC50
CASF
DEKOIS
DUD-AD
DUD-E
LIT-PCBA
Benchmark dataset
0.0
0.2
0.4
0.6
0.8
1.0
BEDROC (α=0.8)